Applied Probability Trust (17 September 2018)

THE LINEAR BIRTH-DEATH PROCESS: AN INFERENTIAL RETROSPECTIVE

SIMON TAVARÉ,* *Cambridge*

Abstract

This article provides an introduction to statistical inference for the classical linear birth-death process, focusing on computational aspects of the problem in the setting of discretely observed processes. The basic probabilistic properties are given in Section 2, focusing on computation of the transition functions. This is followed by a brief discussion of simulation methods in Section 3, and of frequentist methods in Section 4. Section 5 is devoted to Bayesian methods, from rejection sampling to Markov chain Monte Carlo and Approximate Bayesian Computation. Section 6 considers the time-inhomogeneous case, and the article ends with a brief discussion in Section 7.

Keywords: Approximate Bayesian Computation; MCMC; estimation

2010 Mathematics Subject Classification: Primary 62M05

Secondary 62F15;62P10

* Postal address: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, England

1. Introduction

It is a pleasure to contribute to a volume that celebrates Peter Jagers' considerable contributions to the mathematics of population dynamics, and to branching processes in particular. I have chosen as my theme a brief overview of what is known about the inferential aspects of the linear birth-death process, a simple example of a Markov branching process, in part because it reflects Jagers' interest in all things branching, and in part because 2018 is the 75th anniversary of Palm's derivation of the transition function of the process.

The linear birth-death process, a model for population growth and extinction, appears in many elementary probability textbooks as an example of a Markov process for which many explicit results can be found, beginning from Feller's [9] calculation of the expected number of individuals alive at a given time. Statistical inference for this model when observed at discrete equidistant time points is closely related to parallel work on inference for branching processes (cf. Harris [11]). The seminal paper of Keiding [15] derives many asymptotic results for estimation in this process, as well as providing more historical perspective.

Immel [12] noted that the complexity of the transition functions made inference from discrete observations infeasible, a situation that has been resolved only in part to the present day. Indeed, recently there has been a flurry of interest in methods for robust computation of the transition functions of general birth-death processes, and applications to inference of parameters. See, for example, [5, 6, 7, 29].

In this chapter I will review some of the basic results for the process, and discuss computational aspects of classical and Bayesian inference for

the process observed at discrete, but not necessarily equidistantly spaced, time points. I hope the article will provide a useful introduction to computational methods for inference for stochastic processes, in a setting in which the approaches may be compared. R code for performing the computations is available from the author on request.

2. Linear birth and death processes

The constant-rate linear birth-death process $\{Z(t), t \geq 0\}$, with state space $S = \{0, 1, 2, \ldots\}$ is a classical example of a continuous-time Markov process, in which $Z(t)$ gives the number of individuals in the population at time t. Individuals give birth to a new individual at rate $\lambda > 0$ and die at rate $\mu > 0$, independently for all individuals. In terms of transition rates, we have

$$m \to m + 1 \text{ at rate } m\lambda$$

and

$$m \to m - 1 \text{ at rate } m\mu.$$

Feller [9] showed that if the population starts from a single individual, the mean population size at time t is

$$\mathbb{E}Z(t) = e^{(\lambda-\mu)t} \tag{1}$$

and

$$\mathbb{V}\text{ar}\, Z(t) = \begin{cases} \frac{(\lambda+\mu)}{\lambda-\mu}\, e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t} - 1), & \lambda \neq \mu, \\ 2\lambda t, & \lambda = \mu. \end{cases} \tag{2}$$

It is known that the process eventually dies out or tends to ∞, and

$$\mathbb{P}(\text{population dies out}) = \min\left(1, \frac{\mu}{\lambda}\right) \tag{3}$$

The distribution of $Z(t)$ is due to Palm (1943), in an unpublished article. Kendall [16] studied the non-homogeneous process in which both λ and μ could be functions of t; this is discussed further in Section 6. Specialised to the constant-value case, he showed that the distribution of the number of progeny of a single individual is modified geometric. Writing

$$p_{1m}(t) := \mathbb{P}(Z(t) = m | Z(0) = 1), \quad m = 0, 1, \ldots,$$

we have

$$\begin{aligned} p_{1m}(t) &= (1 - \alpha(t))\,(1 - \beta(t))\,\beta(t)^{m-1}, \quad m = 1, 2, \ldots \qquad (4) \\ p_{10}(t) &= \alpha(t), \end{aligned}$$

where, for $\lambda \neq \mu$,

$$\alpha(t) = \frac{\mu(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}, \qquad \beta(t) = \frac{\lambda}{\mu}\,\alpha(t),$$

while for $\lambda = \mu$,

$$\alpha(t) = \frac{\lambda t}{1 + \lambda t} = \beta(t).$$

From (4), for $m = 1, 2, \ldots$,

$$p^*_{1m}(t) := \mathbb{P}(Z(t) = m | Z(t) > 0, Z(0) = 1) = (1 - \beta(t))\,\beta(t))^{m-1}. \qquad (5)$$

The distribution in (4) is usually found by deriving and solving a pde for the probability generating function of $Z(t)$, obtaining ([16], eq. (9))

$$\mathbb{E}(s^{Z(t)} \mid Z(0) = 1) = \frac{\alpha(t) + (1 - \alpha(t) - \beta(t))s}{1 - \beta(t)s}. \qquad (6)$$

Writing this in the form

$$\mathbb{E}(s^{Z(t)} \mid Z(0) = 1) = \alpha(t) + (1 - \alpha(t))\,\frac{(1 - \beta(t))s}{1 - \beta(t)s} \qquad (7)$$

leads directly to (4) and (5). For an ingenious graphical construction using binary planted plane trees, see Branson [3].

It follows immediately from (7) that when $\lambda > \mu$ the rescaled population size $Y(t) = e^{-\rho t}Z(t)$ has asymptotically the distribution of a point mass at 0 with size μ/λ (corresponding to extinction), and conditional on non-extinction, has an exponential distribution with mean $\lambda/(\lambda - \mu)$. On the set of non-extinction, this convergence is almost sure.

Remark 1. The process is sometimes parametrised in terms of the Malthusian parameter

$$\rho = \lambda - \mu, \tag{8}$$

and the ratio

$$\tau = \lambda/\mu; \tag{9}$$

then

$$\mu = \frac{\rho}{\tau - 1}, \quad \lambda = \frac{\tau\rho}{\tau - 1}.$$

2.1. The process starting from $Z(0) = n$

The birth-death process is an example of a Markov branching process in which individuals behave independently; the process starting from n individuals behaves as the sum of n independent copies of the process starting from a single individual. This shows that the extinction probability starting from n individuals is the nth power of that for the process starting with a single individual, and the expected value and variance of the population size are given by (1) and (2) multiplied by n, respectively.

The distribution $\{p_{nm}(t), m = 0, 1, \ldots\}$ is the n-fold convolution of that in (4). There is an explicit formula for this; for example, Bailey [2] showed

that, for $m = 0, 1, 2, \ldots$,

$$p_{nm}(t) = \sum_{j=0}^{\min(m,n)} \binom{n}{j} \binom{n+m-j-1}{n-1} \alpha(t)^{n-j}\, \beta(t)^{m-j}\, (1-\alpha(t)-\beta(t))^j. \tag{10}$$

It is useful to note (Keiding [15]) that the family initiated by each of the n individuals at time 0 either survives to time t (probability $1-\alpha(t)$), or dies out before time t (probability $\alpha(t)$). Since the families evolve independently, the number of families $F(t)$ that survive to time t has the Binomial distribution with parameters n and $1-\alpha(t)$, so that for $j = 0, 1, \ldots, n$

$$\mathbb{P}(F(t) = j \mid Z(0) = n) = \mathrm{Bin}(n, 1-\alpha(t))\{j\} := \binom{n}{j} (1-\alpha(t))^j \alpha(t)^{n-j}, \tag{11}$$

and each surviving family has the geometric distribution given in (5). Hence we can write

$$Z(t) = \sum_{j=1}^{F(t)} Z_i^*(t),$$

where $F(t)$ and the $Z_i^*(t)$ are independent rvs, and the $Z_i^*(t)$ are distributed as (5). It follows immediately that, for $m = 0, 1, 2, \ldots$,

$$p_{nm}(t) = \sum_{j=0}^{\min(m,n)} \binom{n}{j} (1-\alpha(t))^j \alpha(t)^{n-j}\, p_{jm}^*(t), \tag{12}$$

where, for $j = 1, 2, \ldots$, and $m = j, j+1, \ldots$

$$p_{jm}^*(t) = \mathrm{NB}(j, \beta(t))\{m\} := \binom{m-1}{m-j} (1-\beta(t))^j \beta(t)^{m-j}, \tag{13}$$

is the Negative Binomial distribution, the j-fold convolution of the distribution in (5), and $p_{00}^*(t) = 1$; cf. Guttorp [10], p47.

2.2. Computing $p_{nm}(t)$

The form of the classical result in (10) hides its real origin, and may contain large binomial coefficients and alternating terms that make calculation difficult for some parameter values. On the other hand, the expression in (12) is a sum of positive terms, an advantage that can be exploited in several ways.

To illustrate the point, values for both expressions are given in Table 1 for a variety of values of n, m for $\lambda, \mu \in \{1, 2\}$. The middle column on the right of Table 1 shows the potential danger of formula (10), where large binomial coefficients and oscillating terms are causing the trouble; here $1 - \alpha - \beta = 1 - 1.16190 = -0.16190$.

n	m	λ	μ	value from (12)	value from (10)	value from (18)
4	10	1	2	0.00085036	0.00085036	0.00085036
4	10	2	1	0.054423	0.054423	0.054423
40	60	1	2	1.4864e-09	1.4849e-09	1.4869e-09
40	60	2	1	0.0015586	0.0015570	0.0015586
40	100	1	2	1.4509e-20	6.9489e-21	5.3020e-13
40	100	2	1	0.016728	0.0080116	0.016728
100	100	1	2	1.1507e-09	6.6500e+01	1.1501e-09
100	100	2	1	1.1507e-09	6.6500e+01	1.1515e-09

TABLE 1: Values of $p_{nm}(t)$ for $t = 1$ for various values of n, m, λ, and μ, evaluated in R using (12), (10), and (18) with $\eta = 10$.

Remark 2. Note that, for any $t \geq 0$,

$$\alpha(t) := \alpha(t; \lambda, \mu) = \beta(t; \mu, \lambda), \quad \beta(t) := \beta(t; \lambda, \mu) = \alpha(t; \mu, \lambda). \tag{14}$$

Denoting a typical summand in (12) by $s(\lambda, \mu)$, elementary algebra using (14) shows that

$$s(\mu, \lambda) = \left(\frac{\lambda}{\mu}\right)^{n-m} s(\lambda, \mu), \tag{15}$$

so that, in particular, if the birth and death rates are swapped,

$$p_{nm}(t; \mu, \lambda) = \left(\frac{\mu}{\lambda}\right)^{m-n} p_{nm}(t; \lambda, \mu). \tag{16}$$

The values in the table reflect this fact.

The identity in (16) was found by a probabilistic argument by Waugh [28], where he showed *inter alia* that a supercritical process ($\lambda > \mu$) conditioned on ultimate extinction behaves just like the process with λ and μ swapped.

Remark 3. The identity in (16) also provides an alternative method for computing the transition functions when $n - m$ is large in absolute value. For example, if $\mu > \lambda$ and $m >> n$ then the transition probability is likely to be very small, so is best calculated by swapping the roles of μ and λ and using the identity in (16) to extract the leading exponential coefficient.

2.3. Inverting probability generating functions

From (6), the probability generating function of $Z(t)$, given $Z(0) = n$, is given by

$$G(s) := \mathbb{E}(s^{Z(t)} \mid Z(0) = n) = \left(\frac{\alpha(t) + (1 - \alpha(t) - \beta(t))s}{1 - \beta(t)s}\right)^n. \tag{17}$$

One approach to calculate $p_{nm}(t)$ is therefore to invert the transform in (17) numerically. A convenient way to do this is described in [1] as the Lattice-Poisson method, which provides bounds on the error in the

calculation. Define, for $m \geq 1$ and $0 < r < 1$,

$$\tilde{p}_{nm}(t) = \frac{1}{2kr^k}\left(G(r) + (-1)^m G(-r) + 2\sum_{j=1}^{m-1}(-1)^j \mathrm{Re}(G(re^{\pi ij/m}))\right), \tag{18}$$

where Re() denotes real part. It is shown in [1] that

$$|p_{nm}(t) - \tilde{p}_{nm}(t)| \leq \frac{r^{2m}}{1-r^{2m}},$$

which suggests a method for choosing r: for accuracy to $10^{-\eta}$, choose $r = 10^{-\eta/2m}$. Values returned by the method, as implemented in R, are shown in Table 1 for comparison.

There are several observations to make here. First, the anomalous result in the fifth row in Table 1 is due to roundoff error. Indeed, Abate and Whitt [1] show that accuracy to $10^{-\eta}$ requires about $3\eta/2$ digit precision; for $\eta = 10$, this is pushing the limits of precision in the standard implementation of complex arithmetic in R. The last two rows of the table illustrate the need for higher precision too; both the rightmost terms should be equal (as was the case with the estimates from the two explicit formulae in the first two columns.

These numerical issues are readily resolved by using high precision computation. This can be achieved for real calculations in R using the multiple precision floating point package `Rmfpr`; this is not currently implemented for complex arithmetic. An alternative is to use an F95 implementation of Smith's FM multiple-precision software package (cf. [26]). Using 40 significant digit computations within FM and $\eta = 20$ produced lattice-Poisson estimates that agree with the results in the first column of the table. Higher precision arithmetic also can be used to resolve the numerical instability displayed in the second column of the

table. In the remainder of this paper, we use standard R computation without further mention.

2.4. How many families?

As an example of historical inference in the process, we can find the distribution of the number of families at t, given the total number of individuals at time t. For $j = 1, 2, \ldots, \min(m, n)$, we have

$$\begin{aligned}
&\mathbb{P}(F(t) = j \mid Z(t) = m, Z(0) = n) \\
&\quad = \frac{\mathbb{P}(Z(t) = m \mid F(t) = j)\, \mathbb{P}(F(t) = j | Z(0) = n)}{\mathbb{P}(Z(t) = m | Z(0) = n)} \\
&\quad = \frac{\mathrm{NB}(j, \beta(t))\{m\}\, \mathrm{Bin}(n, 1 - \alpha(t))\{j\}}{\mathbb{P}(Z(t) = m | Z(0) = n)}
\end{aligned} \tag{19}$$

Unconditionally, we have

$$\mathbb{E}(F(t)|Z(0) = n) = n(1 - \alpha(t)), \tag{20}$$

and

$$\mathbb{V}\mathrm{ar}(F(t)|Z(0) = n) = n\alpha(t)(1 - \alpha(t)). \tag{21}$$

The values in (20) and the mean and variance of the distribution in (19) for the case $n = 10, m = 100$ are given in Table 2. The symmetry reflected in (15) shows that the conditional distribution of $F(t)$ given $Z(t)$ is invariant under interchange of λ and μ, on the face of it a surprising result.

3. Simulating the process

The earlier results can be exploited for simulating the process at discrete time points

$$0 = t_0 < t_1 < \cdots < t_m = t. \tag{22}$$

t	$\mathbb{E}_{100}F(t)$	$\mathbb{V}\mathrm{ar}_{100}\,F(t)$	$\mathbb{E}F(t)$	$\mathbb{V}\mathrm{ar}\,F(t)$
0.0625	9.99	0.008	8.86	1.01
0.125	9.97	0.031	7.90	1.66
0.25	9.88	0.118	6.38	2.31
0.50	9.54	0.416	4.35	2.46
1.00	8.48	1.103	2.25	1.75
2.50	5.12	1.595	0.43	0.41
5.00	1.99	0.697	0.03	0.03

TABLE 2: Mean and variance of conditional (columns 2 and 3, from (19)) and unconditional (columns 4 and 5, from (20) and (21)) number of surviving families (or lineages), for $\lambda = 1, \mu = 2$ with $Z(0) = 10, Z(t) = 100$ for various values of t.

We write

$$Z_i = Z(t_i), i = 1, 2, \ldots, m, \tag{23}$$

and

$$\Delta_i = t_i - t_{i-1}, i = 1, 2, \ldots, m, \tag{24}$$

and use the construction described in (11) and (13). Starting from $Z(0) = Z_0 = n$, simulate, for $l = 1, 2, \ldots, m$,

$$F_l \sim \mathrm{Bin}(Z_{l-1}, 1 - \alpha(\Delta_l)), \quad Z_l \sim \mathrm{NB}(F_l, \beta(\Delta_l)). \tag{25}$$

This generates observations on a richer process, namely $\{(F_l, Z_l), l = 1, 2, \ldots, m\}$ by keeping track of the number of surviving families in each interval. Note that the process is elementary to simulate in R.

This approach is making use of the explicit nature of the transition functions of the process, an option not typically available for other birth-

Observation l	1	2	3	4	5	6	7	8	9	10
time t_l	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36	0.40
F_l	8	11	11	15	14	17	19	20	21	27
Z_l	11	13	16	17	17	21	23	23	28	32
Observation l	11	12	13	14	15	16	17	18	19	20
time t_l	0.44	0.48	0.52	0.56	0.60	0.64	0.68	0.72	0.76	0.80
F_l	30	35	42	54	67	72	93	111	119	146
Z_l	36	48	61	68	80	96	123	135	161	173
Observation l	21	22	23	24	25					
time t_l	0.84	0.88	0.92	0.96	1.00					
F_l	157	163	173	202	230					
Z_l	87	197	221	253	289					

TABLE 3: Simulation of a process with Z(0) = 10, $\lambda = 6, \mu = 3$

death processes. Instead, we can make use of the usual "wait-jump-wait-$\cdots$" construction, which simulates exponential waiting times between changes of state, and chooses whether to have a birth event or a death event at the jump. If the process is now at n, the waiting time to the next event is exponential with rate $n(\lambda + \mu)$, and the next state results in a birth to a randomly chosen individual with probability $\lambda/(\lambda + \mu)$ or the death of that individual with probability $\mu/(\lambda + \mu)$. This approach will be exploited when we discuss Approximate Bayesian Computation in Section 5.3.

4. Frequentist methods

There is an extensive literature concerning estimation of λ and μ (or, equivalently, ρ and τ defined in (8) and (9) respectively). Parameter estimation depends, of course, on the sampling scheme used to study the population. For continuous observation over $[0, t]$, Keiding [15] showed that the maximum likelihood estimators of λ and μ are

$$\hat{\lambda} = \frac{B_t}{S_t}, \quad \hat{\mu} = \frac{D_t}{S_t},$$

where $S_t = \int_0^t Z(u)du$ is the total time lived in the population in $[0, t]$, B_t is the number of births and D_t the number of deaths in $[0, t]$. Asymptotics for large values of $Z(0)$, and for large values of t, conditional on non-extinction and on ultimate extinction, are described there.

4.1. Moment estimators

We will focus instead on discretely observed samples $Z_0, Z_1, \ldots, Z_m$, as described in Section 3. It is simple to construct moment estimators of the Malthusian parameter ρ. Since $\mathbb{E}(Z_i|Z_{i-1}) = Z_{i-1}e^{\rho\Delta_i}$, it is natural to compare the values of Z_i with those of $Z_{i-1}e^{\rho\Delta_i}$. For example, we might minimise the quantity

$$D(\rho) = \sum_{i=1}^{m} \left(Z_i - Z_{i-1}e^{\rho\Delta_i}\right)^2,$$

to find that the estimator $\tilde{\rho}$ satisfies

$$\sum_{i=1}^{m} Z_i Z_{i-1}\Delta_i e^{\tilde{\rho}\Delta_i} = \sum_{i=1}^{m} Z_i^2 \Delta_i e^{2\tilde{\rho}\Delta_i},$$

which can be solved numerically using R.

When the $\{t_i\}$ are equally spaced, so that $\Delta_i \equiv \Delta$, an explicit solution is available, namely

$$\tilde{\rho} = \frac{1}{\Delta} \log \left(\sum_{i=1}^{m} Z_i Z_{i-1} / \sum_{i=0}^{m-1} Z_i^2 \right). \tag{26}$$

It is straightforward to show that on the set of non-extinction, $\tilde{\rho}$ is a consistent estimator of ρ.

4.2. Likelihood-based methods

The Markov property shows that the likelihood of the data $Z_0, Z_1, \ldots, Z_m$ is given by

$$L(\lambda, \mu) = \prod_{l=1}^{m} p_{Z_{l-1} Z_l}(\Delta_l), \tag{27}$$

the transition function $p_{nm}(t)$ being defined in (12). When the Δ_l are equal, Keiding (1975) showed, using a clever argument based on knowing the values of the F_l, that the MLE of the Malthusian parameter is

$$\hat{\rho} = \frac{1}{\Delta} \log \left(\sum_{l=1}^{m} Z_l / \sum_{l=0}^{m-1} Z_l \right), \tag{28}$$

which should be compared to that in (26). For the example in Table 3, we have $\hat{\rho} = 3.19$, while $\tilde{\rho} = 2.97$; the true value is $\rho = 3$. A small simulation study based on 100,000 runs gave an estimated mean and MSE of 2.85 and 0.35 for $\hat{\rho}$ and 2.82 and 0.45 for $\tilde{\rho}$, suggesting the superiority of the MLE in this example.

In principle, numerical minimization of the likelihood function presents no difficulties, as long as the transition functions can be calculated accurately. R code gave $\hat{\lambda} = 5.23, \hat{\mu} = 2.04$, and $\hat{\rho} = 3.19$, in agreement with the value from (28).

If the F_l are not observed, as is assumed here, evaluation of the transition functions $p_{nm}(t)$ becomes more crucial, and if the population sizes are large, this is likely to be difficult. Alternative methods to evaluate $p_{nm(t)}$ for this model and related birth-death processes appear in [5] and [7]. Hautphenne at al. [7] use saddlepoint approximations for the linear case considered here.

5. Bayesian methods

We now move on to the Bayesian setting. We use $\pi(\cdot)$ to denote the prior for the parameter $\theta = (\lambda, \mu)$, and discuss methods for computing, or simulating observations from, the posterior $f(\theta|\mathcal{D})$ of θ given observations such as $\mathcal{D} = (Z_1, Z_2, \ldots, Z_m)$ or $\mathcal{D} = (F_1, \ldots, F_m, Z_1, \ldots, Z_m)$.

Since the likelihood can, in principle, be computed, the normalising constant

$$\mathbb{P}(\mathcal{D}) = \int_\theta \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta \tag{29}$$

can be evaluated by numerical integration, whence

$$f(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)/\mathbb{P}(\mathcal{D}) \tag{30}$$

can be evaluated over a grid of θ-values. Rather than continue this theme, we resort instead to methods for simulating samples from $f(\theta|\mathcal{D})$.

5.1. Rejection

The simplest of these is the rejection method. The idea is to simulate an observation θ from $\pi(\cdot)$, and accept θ as a draw from $f(\theta|\mathcal{D})$ with probability proportional to $\mathbb{P}(\mathcal{D}|\theta)$.

The constant of proportionality can make a big difference to the effectiveness of the method. Suppose for example, that $\mathcal{D} = (F_1, \ldots, F_m, Z_1, \ldots, Z_m)$. From (25) we know that the likelihood is

$$\prod_{l=1}^{m} \mathrm{Bin}(Z_{l-1}, 1-\alpha(\Delta_l))\{F_l\}\,\mathrm{NB}(F_l, \beta(\Delta_l))\{Z_l\},$$

which can be reordered to give

$$\left[\prod_{l=1}^{m} \mathrm{Bin}(Z_{l-1}, 1-\alpha(\Delta_l))\{F_l\}\right]\,\left[\prod_{l=1}^{m} \mathrm{NB}(F_l, \beta(\Delta_l))\{Z_l\}\right]. \tag{31}$$

The lth term in the left-hand product can be bounded above by

$$\mathrm{Bin}(Z_{l-1}, F_l/Z_{l-1})\{F_l\}, \tag{32}$$

while the lth term in the right-hand product can be bounded above by

$$\mathrm{NB}(F_l, F_{l-1}/Z_l)\{Z_l\}, \tag{33}$$

leading to an upper bound for the likelihood term in (31) and therefore for that in (27).

To give a feel for the effect of the bound, we note that for the example in Table 3, in which just the Z_l are observed and for which $\lambda = 6, \mu = 3$, the likelihood is 4.63×10^{-32}, while the upper bound is 1.97×10^{-6}. Thus using the bound saves a factor of about 500,000 over the naive version; nonetheless, the method is not going to be useful, as the success rate is still 1 in 4.27×10^{25} simulations.

5.2. Markov chain Monte Carlo

The next approach is to use Markov chain Monte Carlo to generate observations that have, approximately, the required distribution $f(\theta|\mathcal{D})$ for a given prior $\pi(\theta)$. MCMC generates a Markov chain of observations on the parameter θ (here, $\theta = (\lambda, \mu)$)) as follows.

1. If now at θ, propose a move to θ' according to a transition kernel $q(\theta \to \theta')$.

2. Calculate the Hastings Ratio

$$h = \min\left(1, \frac{\mathbb{P}(\mathcal{D}|\theta')\,\pi(\theta')\,q(\theta' \to \theta)}{\mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta)\,q(\theta \to \theta')}\right)$$

3. Move to θ' with probability h, else stay at θ. Go to step 1.

Under suitable regularity conditions, f is the stationary (and limiting) distribution of the chain. The practical difficulties of implementing MCMC methods are well known [4], and I will not rehearse them here.

For illustration, and for comparison with other methods, I will take $\mathcal{D} = (Z_0, Z_1, \ldots, Z_m)$ as in Table 3, and assume independent, uniform priors for λ and μ. The implementation takes $\mathrm{U}(a, b)$ for λ, and $\mathrm{U}(c, d)$ for μ. The update mechanism for λ chooses a value that is uniformly distributed with centre the current parameter value, and width $w = (b-a)/g$, suitably modified to keep the proposed value in (a, b); a similar proposal is used for μ. The parameter g tunes the method, and can be used to get any required acceptance rate. In the implementation reported below, both priors were $\mathrm{U}(0, 20)$, and $g = 10$ gave an acceptance rate of about 20%. The chain was run for 100,000 steps, and the first 2,500 were used for burn in. Values taken every 50 steps were used in the subsequent analysis.

Figures 1 and 2 show histograms of the 1950 observations. Figure 3 provides a contour plot of the joint posterior of (λ, μ).

Note that if we assume a wide, flat prior then the mode of the posterior will correspond to the maximum likelihood estimator. In this example, the contour plot reveals a maximum at $\lambda = 5.26, \mu = 2.06$, so that the

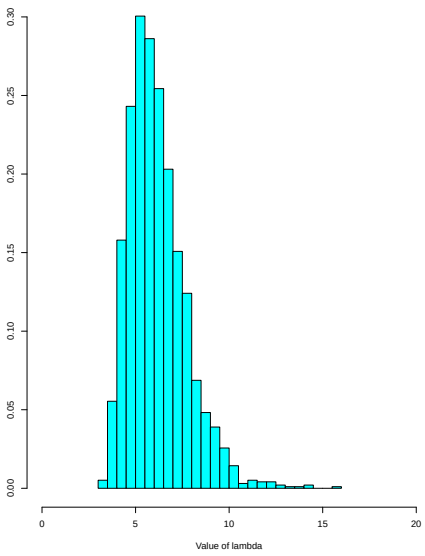

FIGURE 1: Posterior for λ, MCMC.

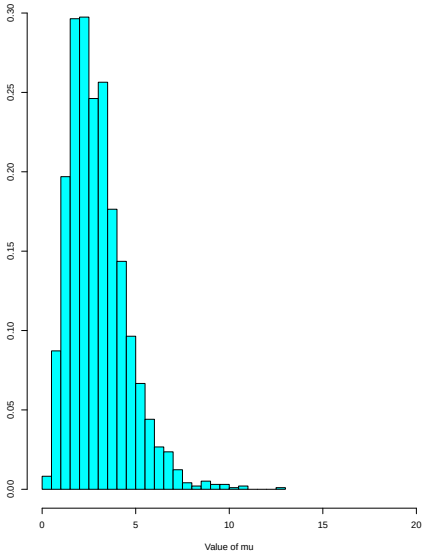

FIGURE 2: Posterior for μ, MCMC.

estimated value of ρ is 3.20. These estimates agree well with those given in Section 4.2.

5.3. Approximate Bayesian Computation

The earlier inference methods depend on being able to calculate the likelihood accurately. However, it is often the case that likelihoods are impossible to compute, either accurately or, indeed, at all. This problem arises in many fields of science, particularly when trying to fit complex mechanistic stochastic models to data.

One approach to inference in such settings is provided by what is now called ABC – Approximate Bayesian Computation. This relies on our ability to simulate observations from the stochastic model $\mathcal{M}$ of interest. The likelihood-based version goes as follows:

1. Generate an observation θ from the prior $\pi()$.

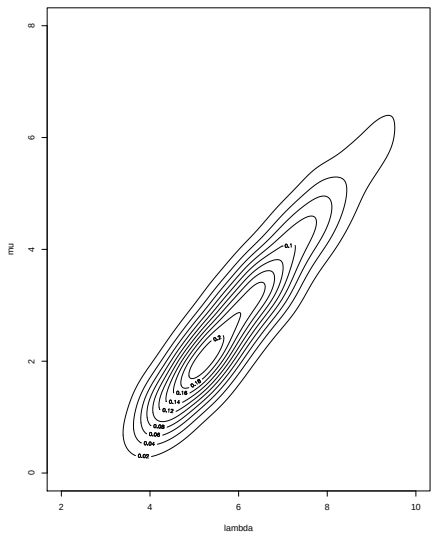

FIGURE 3: Contour plot of posterior density for (λ, μ), MCMC.

2. Accept θ as an observation from the posterior with probability proportional to $\mathbb{P}(\mathcal{D}|\theta)$.

Accepted values clearly have the required distribution, $f(\theta|\mathcal{D})$.

The basis of the ABC method replaces step 2 with:

2′. Simulate an observation $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ

3′. Accept θ as an observation from the posterior if $\mathcal{D}' = \mathcal{D}$

It is clear that observations accepted in this procedure also have the required distribution, but the method is freed from the tyranny of calculating likelihoods. This approach arose in Rubin [23], although not in the present setting. Of course, the simulation version only works if the chance of hitting the target in step 3 is sufficiently large. This led Pritchard et al. [22] to propose a version of the following ABC method. Start with a measure of similarity d on the space of data sets (so that small positive

values of d correspond to more similar data sets), and a tolerance $\epsilon > 0$. Then,

1. Generate an observation θ from the prior $\pi()$.
2. Simulate an observation $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ
3. Accept θ as an approximate observation from the posterior if $d(\mathcal{D}', \mathcal{D}) < \epsilon$.

In many applications the metric is used to compare summary statistics of the data, rather than the data themselves; this has led to a large literature on choosing summary statistics; see [25]. In our birth-death example, we are assuming that all that can be measured are the population sizes $Z_i = Z(t_i), i = 0, 1, \ldots, m$. Our example is therefore simpler, a case of likelihood-free inference, for which we do not need to summarise the data.

Fan and Sisson [8] provide a series of more sophisticated algorithms for implementing ABC, but for our illustration we use the simplest method. It also has the advantage of being an example of embarrassingly parallel computation, which can be spread across multiple cores easily. We need to choose a metric for comparing two sequences of population sizes, and we chose

$$d((Z'_1, \ldots, Z'_m), (Z_1, \ldots, Z_m)) = \left(\sum_{l=1}^{m} (Z'_l - Z_l)^2 \right)^{1/2}. \tag{34}$$

The wait-jump-wait approach was used to simulate observations from the process, obviating the need for knowing the transition functions of the process. Because this approach produces observations at increasing timepoints, and the function in (34) increases with m, it is possible to

	25%	median	mean	75%
MCMC				
λ	5.07	5.93	6.16	6.97
μ	1.86	2.73	2.98	3.78
ABC				
λ	4.51	5.94	6.14	7.63
μ	1.32	2.71	2.97	4.46

TABLE 4: Comparison of percentiles and mean of posterior values for MCMC and ABC methods.

abort runs that result in large values of the metric without simulating the entire trajectory, thereby saving computation time. 500,000 runs were obtained with cutoff chosen as $d = 500$. Figures 4 and 5 show histograms of the values of λ and μ from the 1948 observations with the lowest d-values (corresponding to values of $d \leq 53$) with both priors uniform on $(0, 10)$. Figure 3 provides a contour plot of the joint posterior of (λ, μ). As can be seen by comparing the contour plots in Figures 3 and 8, the values are over-dispersed relative to the MCMC estimates, a fact that is further verified from the data in Table 4.

Remark 4. If we could observe more details of the process than merely the population sizes $\mathcal{D} = (Z_1, \ldots, Z_m)$, for example if $\mathcal{D}$ included the family sizes $F_1, \ldots, F_m$ then ABC can be used to find approximate posterior distribution of θ by choosing a metric such as the sum or maximum of two metrics of the form in (34), one comparing simulated Z-values with the observed target, and one the simulated F values with their observed target. We leave the reader to explore this and other similar scenarios.

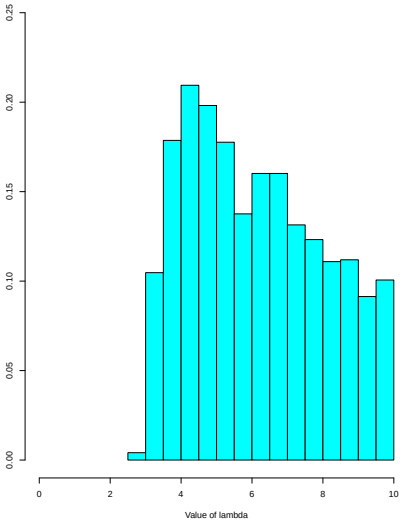

FIGURE 4: Posterior for λ, ABC.

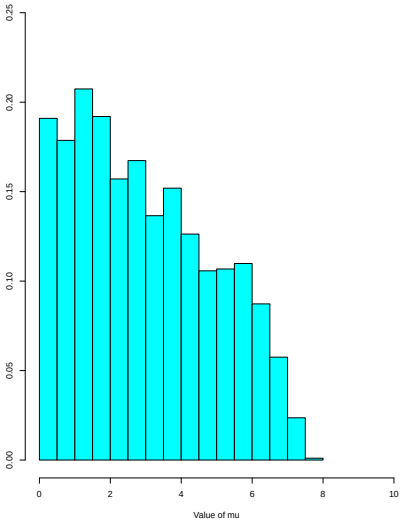

FIGURE 5: Posterior for μ, ABC.

6. Variable rates

Kendall's paper (1948) generalised earlier work by making the birth and death rates depend on time. Here we summarise some of the results. Define

$$\rho(t) = \int_0^t \{\mu(u) - \lambda(u)\} du,$$

and

$$W(t) = 1 + e^{-\rho(t)} \int_0^t e^{\rho(u)} \lambda(u) du \tag{35}$$

$$= e^{-\rho(t)} \left\{ 1 + \int_0^t e^{\rho(u)} \mu(u) du \right\} \tag{36}$$

and set

$$\alpha(t) = 1 - \frac{e^{-\rho(t)}}{W(t)} = \frac{\int_0^t e^{\rho(u)} \mu(u) du}{1 + \int_0^t e^{\rho(u)} \mu(u) du} \tag{37}$$

and

$$\beta(t) = 1 - \frac{1}{W(t)} = \frac{e^{-\rho(t)} \int_0^t e^{\rho(u)} \lambda(u) du}{1 + e^{-\rho(t)} \int_0^t e^{\rho(u)} \lambda(u) du}. \tag{38}$$

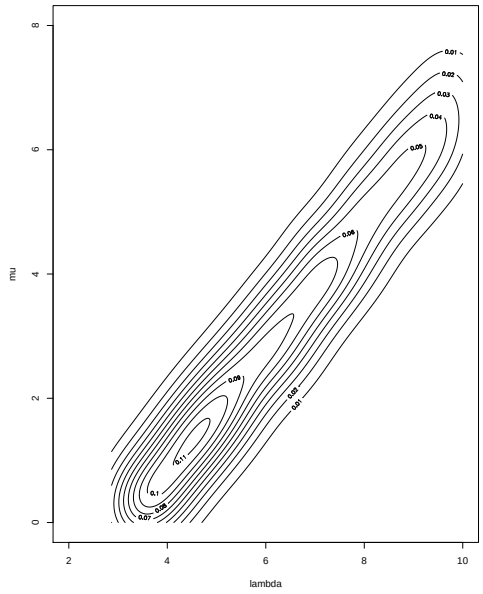

FIGURE 6: Contour plot of posterior density for (λ, μ), ABC.

If the process starts from a single individual, the mean population size at time t is

$$\mathbb{E}Z(t) = e^{\int_0^t \{\lambda(u)-\mu(u)\}du}$$

and the variance is

$$\mathbb{V}\text{ar}\, Z(t) = e^{-2\rho(t)} \int_0^t e^{\rho(u)\{\lambda(u)+\mu(u)\}du};$$

these values are to be multiplied by n if $Z(0) = n$. The probability that the process eventually dies out, starting from a single individual, is given by

$$\mathbb{P}(\text{population dies out}) = \lim_{t\to\infty} \alpha(t) = \frac{\int_0^\infty e^{\rho(u)}\mu(u)du}{1 + \int_0^\infty e^{\rho(u)}\mu(u)du} \qquad (39)$$

and starting from $Z(0) = n$, this should be raised to the power n. Finally, we note that the distribution $p_{1m}(t), m \geq 1$ of $Z(t)$ is given by (4), and, when $Z(0) = n$, $p_{nm}(t)$ is given by (12) and (13).

6.1. Simulation

The wait-jump-wait simulation method works as follows. Suppose the process at time s has value r. The distribution function of the time to the next jump is then

$$F(t|Z(s)=r) = 1 - e^{-\int_s^{s+t} r(\lambda(u)+\mu(u))du}, t \geq 0.$$

Thus to simulate the next waiting time T we let $U \sim \mathrm{U}(0,1)$, and solve for T the equation

$$-\log(U) = \int_s^{s+T} r(\lambda(u)+\mu(u))du. \tag{40}$$

6.2. Inference

Once again we focus on inference for the process observed at time points $0 = t_0 < t_1 < \cdots < t_m = t$, and for $i = 1, 2, \ldots, m$ we set $Z_i = Z(t_i)$. The likelihood of the observations $Z_0 = n, Z_1, \ldots, Z_m$ is

$$L = \prod_{i=1}^{m} p_{Z_{i-1},Z_i}(t_{i-1}, t_i), \tag{41}$$

so we are left with the problem of identifying the transition functions

$$p_{nm}(s,t) = \mathbb{P}(X(t) = m \mid X(s) = n),\ t \geq s,$$

of the process; fortunately, this is easy. They can be read off from the result for $p_{nm}(0,t) \equiv p_{nm}(t)$ by letting

$$\lambda(u) = \mu(u) = 0,\ 0 < u < s;$$

this traps the process at n until time s. Specifically, for $0 < s < t$, define

$$\rho(s,t) = \int_s^t \{\mu(u) - \lambda(u)\}du := \rho(t) - \rho(s),$$

and set

$$\begin{aligned} W(s,t) &= 1 + e^{-\rho(s,t)} \int_s^t e^{\rho(s,u)} \lambda(u) du \\ &= e^{-\rho(s,t)} \left\{ 1 + \int_s^t e^{\rho(s,u)} \mu(u) du \right\} \\ \alpha(s,t) &= 1 - e^{-\rho(s,t)} / W(s,t) \\ \beta(s,t) &= 1 - 1/W(s,t). \end{aligned}$$

The values of $\alpha(s,t)$ and $\beta(s,t)$ can be inserted into the formulae (12) and (13) to get the transition function starting from time s.

6.3. Example

For illustration, consider the case in which

$$\mu(t) \equiv \mu, \quad \lambda(t) = \lambda + \frac{\gamma}{t+1}, \ t \geq 0, \tag{42}$$

and $\lambda + \gamma > 0$. In this case,

$$\rho(t) = (\mu - \lambda)t - \gamma \log(t+1),$$

and

$$\mathbb{E}Z(t) = e^{(\lambda-\mu)t} (t+1)^{\gamma},$$

and the probability of eventual extinction is 1 if $\lambda \leq \mu$, and $I/(1+I)$, where

$$I = e^{\lambda-\mu} \mu (\lambda - \mu)^{\gamma-1} \int_{\lambda-\mu}^{\infty} v^{-\gamma} e^{-v} dv, \tag{43}$$

if $\lambda > \mu$. Values of the extinction probability as a function of γ for $\lambda = 6, \mu = 3$ are given in Fig. 7.

To implement the simulation method in (40), note that if $Z(s) = r \geq 1$ then the next waiting time T is found as the solution of the equation

$$\log(U)/r - \gamma \log(s+1) + (\lambda + \mu)T + \gamma \log(s+1+T) = 0$$

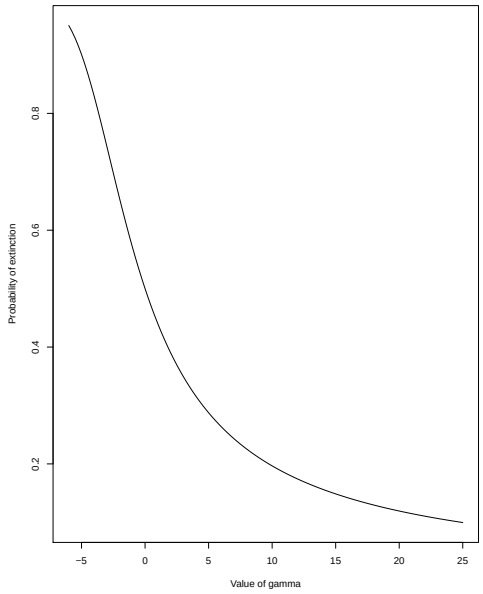

FIGURE 7: Probability of extinction as a function of γ for $\lambda = 6, \mu = 3$ computed from (43).

and, if $T = t$, the next state is $r + 1$ with probability

$$\frac{\lambda + \gamma/(s + t + 1)}{\lambda + \gamma/(s + t + 1) + \mu},$$

and is otherwise $r - 1$.

Table 5 shows the results of a simulation of the process that will be used to illustrate the inference problem. Computation of the MLEs of λ, μ, γ via (41) offer no new problems. A grid search to determine good starting values resulted in estimates of $\hat{\lambda} = 6.18, \hat{\mu} = 1.74$, and $\hat{\gamma} = 2.49$. The ABC approach is also illustrated here, based on independent priors uniform on (0,10) for each parameter. The metric in (34) is used once more, and 10,000 simulations were generated using the rates method, rejecting any runs with $d > 500$. The 1951 values with $d < 239$ were chosen to represent the approximate joint posterior of the parameters, and resulted in median

Observation l	1	2	3	4	5	6	7	8	9	10
time t_l	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36	0.40
F_l	8	12	14	22	24	28	38	52	68	84
Z_l	12	16	24	28	35	43	55	76	92	115
Observation l	11	12	13	14	15	16	17	18	19	20
time t_l	0.44	0.48	0.52	0.56	0.60	0.64	0.68	0.72	0.76	0.80
F_l	102	115	162	209	264	341	431	564	707	891
Z_l	126	178	227	299	380	472	628	778	993	1267
Observation l	21	22	23	24	25					
time t_l	0.84	0.88	0.92	0.96	1.00					
F_l	1147	1477	1814	2310	2882					
Z_l	1623	2019	2546	3209	3995					

TABLE 5: Simulation of the variable-rate process with Z(0) = 10, $\lambda = 6, \mu = 3, \gamma = 5$.

values of 7.24, 3.13 and 2.74 for λ, μ and γ respectively. The histograms in Figs. 8, 9 and 10 illustrate the posteriors. There is, indeed, information in the data about the parameters, the locations of the marginals being broadly consistent with the MLEs.

7. Discussion

There are other models for which explicit results are available, including the linear birth-and-death processes with immigration [14]. Explicit availability of transition functions makes these models a useful calibration for other computational approaches.

Computation of the transition function of continuous-time Markov chains

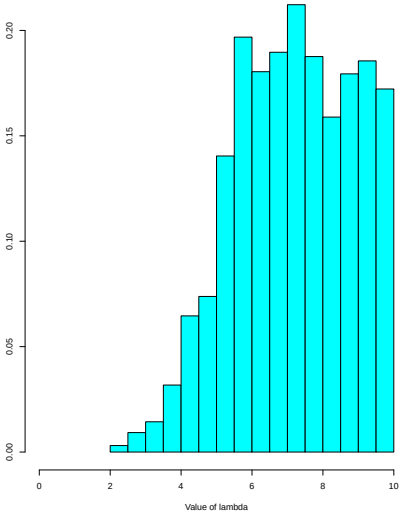

FIGURE 8: Posterior for λ, ABC.

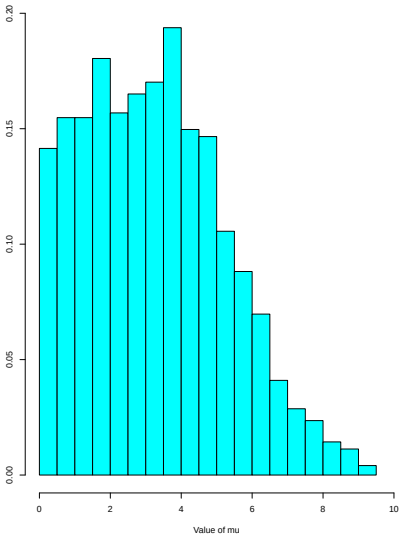

FIGURE 9: Posterior for μ, ABC.

can be a challenge. For a finite process with generator matrix $Q = (q_{ij})$, this is equivalent to computing e^{Qt}. Moler and Van Loan [20, 21] discuss a number of numerical methods for this for arbitrary matrices Q, and point to the scaling and squaring method as a good general choice. Melloy and Bennett [19] specialise to the stochastic case. Their method is connected to the so-called Poissonization, or uniformization, representation [13]. If we define

$$\rho = \max_i |q_{ii}|$$

and

$$P = \rho^{-1} Q + I$$

then P is a stochastic matrix, and

$$e^{Qt} = \sum_{n \geq 0} \frac{e^{-\rho t} (\rho t)^n}{n!} P^n. \tag{44}$$

Thus the process jumps at the points of a Poisson process, making moves

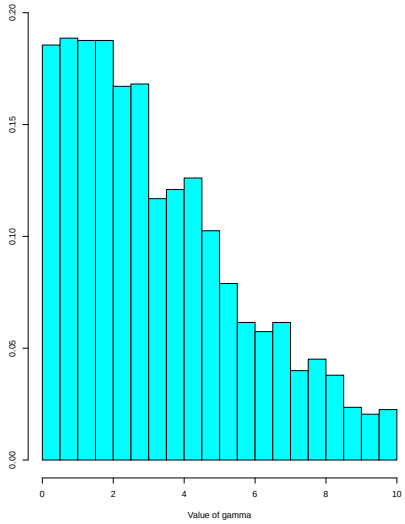

FIGURE 10: Posterior for γ, ABC.

according to the transition matrix P, which might not result in changing state. This can be used to compute e^{Qt} as a sum of non-negative terms. Uniformization for inhomogeneous chains was described in [27]; see also [17]. Uniformization can also be used to simulate the process.

For infinite stochastic matrices with unbounded entries, things are more complicated. Northwest corner truncation approaches are available; see for example [24], Chapter 7 and [18]. Because of likelihood approaches to inference of model parameters, there has recently been a resurgence of interest in calculating the transition probabilities of general birth-and-death processes, those in which the birth and death rates may depend on the state in more complicated ways than the linear case. See, for example, the continued fraction method of [6, 5], the compressed sensing approach of [29], and the saddlepoint approach in [7] for the linear case.

Acknowledgement

I thank Dr. Sophie Hautphenne for alerting me to her work in [7].

References

[1] Abate, J. and Whitt, W. (1992). Numerical inversion of probability generating functions. *Operations Research Letters*, **12**, 245–251.

[2] Bailey, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley, New York-London.

[3] Branson, D. (1991). Inhomogeneous birth-death and birth-death-immigration processes and the logarithmic series distribution. *Stochastic Processes and their Applications*, **39**, 131–137.

[4] Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (eds.) (2011). *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, Chapman & Hall/CRC Press.

[5] Crawford, F. W., Minin, V. N. and Suchard, M. A. (2014). Estimation for general birth-death processes. *Journal of the American Statistical Association*, **109,** 730–747.

[6] Crawford, F. W. and Suchard, M. A. (2012). Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *Journal of Mathematical Biology*, **65**, 553–580.

[7] Davison, A. C., Hautphenne, S. and Kraus, A. (2018). Parameter estimation for discretely-observed linear birth-and-death processes. arXiv:1802.05015v1.

[8] Fan, Y. and Sisson, S. A. (2018). ABC samplers. arXiv:1802.09650v1.

[9] Feller, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, **5**, 11–40.

[10] Guttorp, P. (1991). *Statistical Inference for Branching Processes.* Wiley, New York.

[11] Harris, T. (1948). Branching processes. *Annals of Mathematical Statistics*, **19**, 474–494.

[12] Immel, E. R. (1951). Problems of estimation and of hypothesis testing connected with birth-and-death Markov processes. Thesis, University of California, Los Angeles.

[13] Jensen, A. (1953). Markoff chains as an aid in the study of Markoff processes. *Skandinavisk Akrtuarietidskrift*, **36**, 87–91.

[14] Karlin, S. and McGregor, J. (1967). The number of mutant forms maintained in a population. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. LeCam L & Neyman J pp. 415–438. University of California Press, Berkeley.

[15] Keiding, N. (1975). Maximum likelihood estimation in the birth-and-death process. The Annals of Statistics, **3**, 363–372.

[16] Kendall, D. G. (1948). On the generalized "birth-and-death" process. *Annals of Mathematical Statistics*, **19**, 1–15.

[17] Li, Y. F., Zio, E. and Lin Y. H. (2014). Methods of solutions of inhomogeneous continuous time Markov chains for degradation process modeling. Chapter 1, *Applied Reliability Engineering and Risk Analysis: Probabilistic Models and Statistical Inference*, eds. Ilia B. Frenkel, I. B., Karagrigoriou, A. Lisnianski, A. and Kleyner, A. John Wiley & Sons, Ltd.

[18] Masuyama, H. (2016). Limit formulas for the normalized fundamental matrix of the northwest-corner truncation of Markov chains: Matrix-infinite-product-form solutions of block-Hessenberg Markov chains. arXiv:1603:07877v5.

[19] Melloy, B. J. and Bennett, G. K. (1993). Computing the exponential of an intensity matrix. *Journal of Computational and Applied mathematics*, **46**, 405–413.

[20] Moler, C. and Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, **20**, 801–836.

[21] Moler, C. and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, **45**, 3–49.

[22] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

[23] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics,* **12**, 1151–1172.

[24] Seneta, E. (2006). *Non-negative Matrices and Markov Chains,* 2nd edn. Springer Series in Statistics, Springer.

[25] Sisson, S. A., Fan, Y. and Beaumont, M. A. (eds.) (2018). *Handbook of Approximate Bayesian Computation*, Chapman & Hall/CRC Press.

[26] Smith, D. M. Multiple precision complex arithmetic and functions. `https://dmsmith.lmu.build/toms1998.pdf`

[27] van Dijk, N. M. (1992). Uniformization for nonhomogeneous Markov chains. *Operations Research Letters*, **12**, 283–291.

[28] Waugh, W. A. O'N. (1958). Conditioned Markov processes. *Biometrika*, **45**, 241–249.

[29] Xu, J. and Minin, V. N. (2015). Efficient transition probability computation for continuous-time branching processes via compressed sensing. `auai.org/uai2015/proceedings/papers/239.pdf`