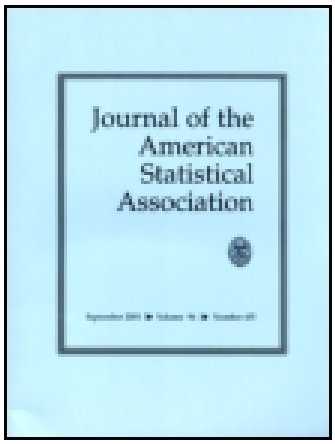

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Estimation for General Birth-Death Processes

Forrest W. Crawford, Vladimir N. Minin & Marc A. Suchard

Accepted author version posted online: 03 Dec 2013.Published online: 13 Jun 2014.

To cite this article: Forrest W. Crawford, Vladimir N. Minin & Marc A. Suchard (2014) Estimation for General Birth-Death Processes, Journal of the American Statistical Association, 109:506, 730-747, DOI: 10.1080/01621459.2013.866565

To link to this article: http://dx.doi.org/10.1080/01621459.2013.866565

Estimation for General Birth-Death Processes

Forrest W. CRAWFORD, Vladimir N. MININ, and Marc A. SUCHARD

Birth-death processes (BDPs) are continuous-time Markov chains that track the number of "particles" in a system over time. While widely used in population biology, genetics, and ecology, statistical inference of the instantaneous particle birth and death rates remains largely limited to restrictive linear BDPs in which per-particle birth and death rates are constant. Researchers often observe the number of particles at discrete times, necessitating data augmentation procedures such as expectation-maximization (EM) to find maximum likelihood estimates (MLEs). For BDPs on finite state-spaces, there are powerful matrix methods for computing the conditional expectations needed for the E-step of the EM algorithm. For BDPs on infinite state-spaces, closed-form solutions for the E-step are available for some linear models, but most previous work has resorted to time-consuming simulation. Remarkably, we show that the E-step conditional expectations can be expressed as convolutions of computable transition probabilities for any general BDP with arbitrary rates. This important observation, along with a convenient continued fraction representation of the Laplace transforms of the transition probabilities, allows for novel and efficient computation of the conditional expectations for all BDPs, eliminating the need for truncation of the state-space or costly simulation. We use this insight to derive EM algorithms that yield maximum likelihood estimation for general BDPs characterized by various rate models, including generalized linear models (GLM). We show that our Laplace convolution technique outperforms competing methods when they are available and demonstrate a technique to accelerate EM algorithm convergence. We validate our approach using synthetic data and then apply our methods to cancer cell growth and estimation of mutation parameters in microsatellite evolution.

KEY WORDS: Continuous-time Markov chain; EM algorithm; Maximum likelihood estimation; Microsatellite evolution; MM algorithm.

1. INTRODUCTION

A birth-death process (BDP) is a continuous-time Markov chain that models a nonnegative integer number of particles in a system (Feller 1971). The state of the system at a given time is the number of particles in existence. At any moment in time, one of the particles may "give birth" to a new particle, increasing the count by one, or one particle may "die," decreasing the count by one. BDPs are popular modeling tools in a wide variety of quantitative disciplines, such as population biology, genetics, and ecology (Thorne, Kishino, and Felsenstein 1991; Krone and Neuhauser 1997; Novozhilov, Karev, and Koonin 2006; Renshaw 2011). For example, BDPs can characterize epidemic dynamics (Bailey 1964; Andersson and Britton 2000), speciation and extinction (Nee, May, and Harvey 1994; Nee 2006), evolution of gene families (Cotton and Page 2005; Demuth et al. 2006), and the insertion and deletion events for probabilistic alignment of DNA sequences (Thorne, Kishino, and Felsenstein 1991; Holmes and Bruno 2001).

Traditionally, most modeling applications have used the "simple linear" BDP with constant per-particle birth and death rates, which arises from an assumption of independence among particles and no background birth and death rates. When individual birth and death rates instead depend on the size of the population as a whole, the model is called a "general" BDP. Previous statistical estimation in BDPs has focused mainly on estimating the constant per-particle birth and death rates of the simple linear BDP based on observations of the number of particles over time. However, the simple linear BDP is often unrealistic, and nonlinear dependence of the birth and death rates on the current number of particles provides the means to model more sophisticated and realistic patterns of stochastic population dynamics in a wide variety of biological disciplines (Novozhilov, Karev, and Koonin 2006). For example, populations sometimes exhibit logistic-like growth as their number approaches the carrying capacity of their environment (Tan and Piantadosi 1991). In genetic models, the rate of new offspring carrying an allele often depends on the proportions of both individuals already carrying the allele and those who do not (Moran 1958). In coalescent theory, the rate of coalescence changes with the square of the number of lineages (Kingman 1982). In addition, researchers may wish to assess the influence of covariates on birth and death rates by fitting a regression model (Kalbfleisch and Lawless 1985; Liu, Beckett, and DeNardo 2007).

Analytic studies of general BDPs have provided insight into theoretical properties including stationary distributions, moments, transition probabilities, and other quantities of interest. Karlin and McGregor (1957a, b) introduce a representation of BDP transition probabilities using orthogonal polynomials and spectral measure, but these can be extremely difficult to derive for general BDPs (Novozhilov, Karev, and Koonin 2006; Renshaw 2011). Several authors have characterized BDP transition probabilities and passage times in terms of continued fraction expressions for the Laplace transform of these quantities (Murphy and O'Donohoe 1975; Jones and Magnus 1977; Bordes and Roehner 1983; Guillemin and Pinchon 1998, 1999; Flajolet and Guillemin 2000; Crawford and Suchard 2012). However, none of these authors address the task of statistical inference using data observed from a general BDP.

Forrest W. Crawford, Department of Biostatistics, Yale School of Public Health, 60 College Street, Box 208034, New Haven, CT 06510 (E-mail: *forrest.crawford@yale.edu*). Vladimir N. Minin, Department of Statistics, University of Washington, Padelford Hall C-315, Box 354322, Seattle, WA 98195-4322 (E-mail: *vminin@u.washington.edu*). Marc A. Suchard, Department of Biomathematics, Department of Biostatistics, and Department of Human Genetics, University of California Los Angeles, 6558 Gonda Building, Los Angeles, CA 90095-1766 (E-mail: *msuchard@ucla.edu*). We thank Kenneth Lange, Hua Zhou, Gabriela Cybis, and two anonymous reviewers for insightful comments and criticism. We are grateful to Laurel Beckett, Hao Liu, Gerald Denardo, and Evan Tobin for providing the lymphoma data. This work was supported by NIH grants R01 GM086887, R01 AI107034, R01 HG006139, T32 GM008185, and NSF grants DMS-0856099 and DMS 1264153.

Journal of the American Statistical Association
June 2014, Vol. 109, No. 506, Theory and Methods
DOI: 10.1080/01621459.2013.866565

Progress in parameter estimation for general BDPs has also typically been limited to continuous observation of the process (Moran 1951, 1953; Anscombe 1953; Darwin 1956; Wolff 1965; Reynolds 1973; Keiding 1975). However, in practice researchers may observe data from BDPs only at discrete times through longitudinal observations. Estimating transition rates in continuous-time Markov processes using discrete observations is difficult since the state path between observations is not observed. Furthermore, direct analytic maximization of the likelihood for general BDPs remains infeasible for partially observed samples since the likelihood usually cannot be written in closed-form. Despite these challenges, several researchers have made progress in estimating parameters of the simple linear BDP under discrete observation (Keiding 1974; Thorne, Kishino, and Felsenstein 1991; Holmes and Bruno 2001; Rosenberg, Tsolaki, and Tanaka 2003; Dauxois 2004). However, none of these developments provides a robust method to find exact maximum likelihood estimates (MLEs) of parameters in discretely observed general BDPs with arbitrary birth and death rates.

A major insight comes from the fact that the likelihood of the continuously observed process has a simple form which easily yields expressions for estimation of rate parameters. This fact is the basis for expectation-maximization (EM) algorithms for maximum likelihood estimation in missing data problems (Dempster, Laird, and Rubin 1977). In finite state-space Markov chains, the relevant conditional expectations (the E-step of the EM algorithm) can often be computed efficiently (Minin and Suchard 2008); Hobolth and Jensen (2011) discussed eigendecomposition, uniformization, and integration of matrix exponentials. Using these matrix-algebraic tools, several researchers have derived EM algorithms for estimating transition rates in this context (Lange 1995a; Holmes and Rubin 2002; Bladt and Sorensen 2005; Hobolth and Jensen 2005; Metzner et al. 2007; Hobolth and Jensen 2011). Unfortunately, finding these conditional expectations for general BDPs poses challenges since the joint distribution of the states and waiting times (or the corresponding generating function) is usually not available in closed form. Notably, Holmes and Bruno (2001), Holmes and Rubin (2002), and Doss et al. (2013) were able to find analytic expressions or numerical approximations for these expectations in EM algorithms for certain BDPs whose rates depend linearly on the current number of particles. While these developments are promising, there remains a great need for estimation techniques that can be applied to more sophisticated infinite state-space BDPs under a variety of sampling scenarios. Indeed, more complex and realistic models like those reviewed by Novozhilov, Karev, and Koonin (2006) may be of little use to applied researchers if no practical method exists to estimate their parameters.

Here, we seek to fill this apparent void by providing the first framework for deriving EM algorithms for estimating the parameters of a discretely sampled general BDP. We first formally define the general BDP and give an exact expression for the Laplace transform of the transition probabilities in the form of a continued fraction. We then give the likelihood for continuously observed BDPs and outline the EM algorithm. Next, we describe a novel method to efficiently compute the expectations of the E-step for BDPs with arbitrary rates. Since these expectations are convolutions of transition probabilities, we perform the convolution in the Laplace domain, and then invert the Laplace transformed expressions to obtain the desired conditional expectation. This technique obviates the costly numerical integration, matrix computations, or repeated simulation that have plagued previous approaches. We provide examples of the maximization step for several different classes of BDPs and demonstrate a technique for accelerating convergence of the EM algorithm. We show that our method for performing the E-step is faster than competing simulation methods and matrix methods that require truncation of the state-space. We validate our method using simulated data and conclude with two applications. First, we analyze lymphoma cell growth under different treatment conditions by parameterizing the birth and death rates as a generalized linear model (GLM). Next, we study the evolution of DNA microsatellites in humans and chimpanzees to address an open question in evolutionary genomics.

2. GENERAL BDPS AND THEIR EM ALGORITHMS

2.1 Formal Description and Transition Probabilities

Consider a general BDP $X(\tau)$ counting the number of particles k in existence at times $\tau \geq 0$. From state $X(\tau) = k$, transitions to state $k+1$ happen with instantaneous rate λ_k, and transitions to state $k-1$ happen with instantaneous rate μ_k. The transition rates λ_k and μ_k may depend on k but are time-homogeneous. In this article, we assume that $X(\tau)$ is not explosive, that is $X(\tau)$ does not "run away" to infinity in finite time. As we show below, it is often necessary to evaluate finite-time transition probabilities to derive efficient EM algorithms for estimation of arbitrary birth and death rates in general BDPs. This proves useful both in completing the E-step of the EM algorithm and in computing incomplete data likelihoods for validation of our EM estimates. For a starting state $i \geq 0$, the finite-time transition probabilities $P_{ij}(\tau) = \Pr(X(\tau) = j \mid X(0) = i)$ obey the system of ordinary differential equations

$$\frac{\mathrm{d}P_{i0}(\tau)}{\mathrm{d}\tau} = \mu_1 P_{i1}(\tau) - \lambda_0 P_{i0}(\tau),$$

and

$$\frac{\mathrm{d}P_{ij}(\tau)}{\mathrm{d}\tau} = \lambda_{j-1} P_{i,j-1}(\tau) + \mu_{j+1} P_{i,j+1}(\tau) - (\lambda_j + \mu_j) P_{ij}(\tau), \tag{1}$$

for $j \geq 1$ with $P_{ii}(0) = 1$ and $P_{ij}(0) = 0$ for $i \neq j$ (Feller 1971).

For some simple parameterizations of λ_k and μ_k, closed-form solutions exist for the transition probabilities $P_{ij}(\tau)$, but this is not possible for most models. Karlin and McGregor (1957b) showed that for any parameterization of λ_k and μ_k, it is possible to express the transition probabilities in terms of orthogonal polynomials. However, in practice, these special polynomials are difficult to find, and even when they are available, they rarely yield solutions in closed-form or expressions that are amenable to computation (Novozhilov, Karev, and Koonin 2006; Renshaw 2011). In contrast, the continued fraction method we outline below does not require additional model-specific insight beyond specification of λ_k and μ_k.

To solve for the transition probabilities, it is advantageous to work in the Laplace domain (Karlin and McGregor 1957b). This transformation also proves essential in maintaining

numerical stability of transition probabilities in general BDPs and in computing the conditional expectations necessary for the EM algorithm derived in a subsequent section. Let

$$f_{ij}(s) = \mathscr{L}[P_{ij}(\tau)](s) = \int_0^\infty e^{s\tau} P_{ij}(\tau)\, \mathrm{d}\tau \tag{2}$$

be the Laplace transform of $P_{ij}(\tau)$ and let $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Laplace transforming Equation (1) yields

$$\begin{aligned} sf_{i0}(s) - \delta_{i0} &= \mu_1 f_{i1}(s) - \lambda_0 f_{i0}(s), \\ sf_{ij}(s) - \delta_{ij} &= \lambda_{j-1} f_{i,j-1}(s) + \mu_{j+1} f_{i,j+1}(s) \\ &\quad - (\lambda_j + \mu_j) f_{ij}(s). \end{aligned} \tag{3}$$

Letting $i = 0$ and rearranging (3), we obtain the recurrence relations

$$f_{00}(s) = \frac{1}{s + \lambda_0 - \mu_1\left(\frac{f_{01}(s)}{f_{00}(s)}\right)},$$

and

$$\frac{f_{0j}(s)}{f_{0,j-1}(s)} = \frac{\lambda_{j-1}}{s + \mu_j + \lambda_j - \mu_{j+1}\left(\frac{f_{0,j+1}(s)}{f_{0j}(s)}\right)}. \tag{4}$$

We can inductively combine these expressions for $j = 1, 2, 3, \ldots$ to arrive at the well-known generalized continued fraction

$$f_{00}(s) = \cfrac{1}{s + \lambda_0 - \cfrac{\lambda_0\mu_1}{s+\lambda_1+\mu_1 - \cfrac{\lambda_1\mu_2}{s+\lambda_2+\mu_2 - \cdots}}}. \tag{5}$$

This is an exact expression for the Laplace transform of the transition probability $P_{00}(\tau)$. In (5), let $a_1 = 1$ and $a_j = -\lambda_{j-2}\mu_{j-1}$, and let $b_1 = s + \lambda_0$ and $b_j = s + \lambda_{j-1} + \mu_{j-1}$ for $j \geq 2$. Then, (5) becomes

$$f_{00}(s) = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdots}}}. \tag{6}$$

We can write this more compactly as

$$f_{00}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots. \tag{7}$$

The kth convergent of $f_{0,0}(s)$ is

$$f_{00}^{(k)}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\cdots\frac{a_k}{b_k} = \frac{A_k(s)}{B_k(s)}, \tag{8}$$

where $A_k(s)$ and $B_k(s)$ are the numerator and denominator of the rational function $f_{0,0}^{(k)}$. The transition probabilities $P_{ij}(\tau)$ for $i, j > 0$ can be derived in continued fraction form by combining (3) and (5) to obtain

$$f_{ij}(s) = \begin{cases} \left(\prod_{k=j+1}^{i} \mu_k\right) \dfrac{B_j(s)}{B_{i+1}(s)+}\dfrac{B_i(s)a_{i+2}}{b_{i+2}+}\dfrac{a_{i+3}}{b_{i+3}+}\cdots & \text{for } j \leq i, \\ \left(\prod_{k=i}^{j-1} \lambda_k\right) \dfrac{B_i(s)}{B_{j+1}(s)+}\dfrac{B_j(s)a_{j+2}}{b_{j+2}+}\dfrac{a_{j+3}}{b_{j+3}+}\cdots & \text{for } i \leq j, \end{cases} \tag{9}$$

(Murphy and O'Donohoe 1975; Crawford and Suchard 2012).

Although the Laplace transforms of the transition probabilities are generally still not available in closed-form, a continued fraction representation is desirable for several reasons: (1) continued fraction representations of functions often converge much faster than equivalent power series; (2) there are efficient algorithms for evaluating continued fractions to a finite depth; and (3) there exist methods for bounding the error of truncated continued fractions (Bankier and Leighton 1942; Wall 1948; Blanch 1964; Lorentzen and Waadeland 1992; Craviotto, Jones, and Thron 1993; Abate and Whitt 1999; Cuyt et al. 2008). For an arbitrary BDP, we recover the transition probabilities through numerical inversion of the Laplace-transformed expressions. We evaluate the continued fraction to a monitored depth that controls the overall error and generates stable approximations to the transition probabilities unattainable by previous methods (Murphy and O'Donohoe 1975; Parthasarathy and Sudhesh 2006; Crawford and Suchard 2012). We derive approximate error bounds for this computation in the Appendix.

The ability to compute transition probabilities for general BDPs with arbitrary rate parameterizations proves useful in two ways. First, if we interpret finite-time transition probabilities as functions of an unknown parameter vector $\boldsymbol{\theta}$, then $P_{ab}(t)$ given $\boldsymbol{\theta}$ returns the *likelihood* of a discrete observation from a BDP such that $X(0) = a$ and $X(t) = b$, where the trajectory in time t between states a and b is unobserved. Second, transition probabilities play an important role in computing conditional expectations of sufficient statistics, as we shall see below.

2.2 Likelihood Expressions and Surrogate Functions

With a formal description of a general BDP and the finite-time transition probabilities in hand, we now proceed with our task of estimating the parameters of a general BDP using discrete observations. Given one or more independent observations of the form $\mathbf{Y} = (X(0) = a, X(t) = b, t)$ from a general BDP, we wish to find MLEs of the rate parameters λ_k and μ_k for $k = 0, 1, 2, \ldots$. We will assume that the birth and death rates at state k depend on both k and a finite-dimensional parameter vector $\boldsymbol{\theta}$, so that the form of $\lambda_k(\boldsymbol{\theta})$ and $\mu_k(\boldsymbol{\theta})$ is known for all k.

For a single realization of the process starting at $X(0) = a$ and ending at $X(t) = b$, let T_k be the total time spent in state k. Let U_k be the number of "up" steps (births) from state k, and let D_k be the number of "down" steps (deaths) from state k. Let the total number of up and down steps in a realization of the process be denoted by $U = \sum_{k=0}^\infty U_k$ and $D = \sum_{k=0}^\infty D_k$ respectively. We also define the total particle time,

$$T_{\text{particle}} = \int_0^t X(\tau)\, \mathrm{d}\tau = \sum_{k=0}^\infty k T_k, \tag{10}$$

that counts the amount of time lived by each particle since time $\tau = 0$. The total elapsed time is $t = \sum_{k=0}^\infty T_k$. We demonstrate these concepts schematically in Figure 1.

The log-likelihood for a continuously observed process takes a simple form when we sum over all possible states k (Wolff 1965):

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{k=0}^\infty &[U_k \log[\lambda_k(\boldsymbol{\theta})] + D_k \log[\mu_k(\boldsymbol{\theta})] \\ &- [\lambda_k(\boldsymbol{\theta}) + \mu_k(\boldsymbol{\theta})]T_k]. \end{aligned} \tag{11}$$

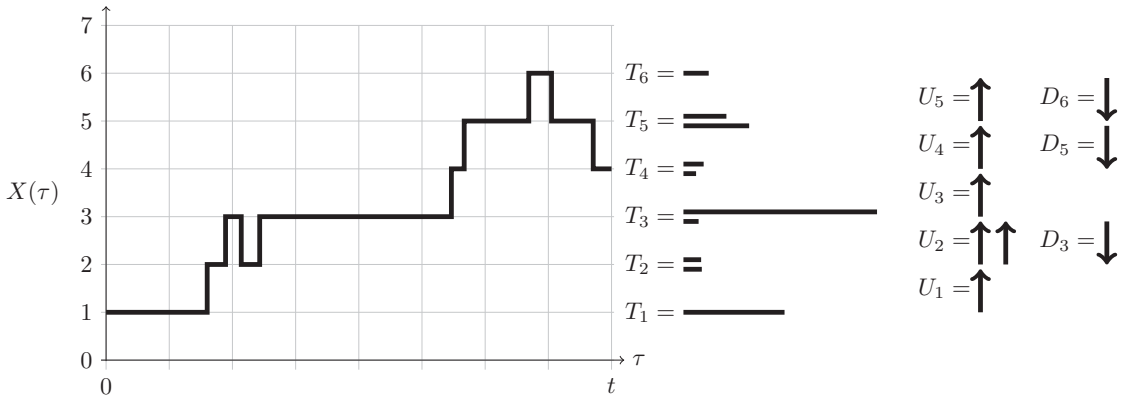

Figure 1. A sample path from a birth-death process (BDP) $X(\tau)$. The process starts at state $X(0) = 1$ and is at state $X(t) = 4$ at time t. At right are schematic representations of the time spent in each state T_k, the number of up steps U_k, and the number of down steps D_k. These quantities are the sufficient statistics for estimators of rate parameters in general birth-death processes.

However, when a BDP is sampled discretely such that only $X(0) = a$ and $X(t) = b$ are observed, the quantities U_k, D_k, and T_k are unknown for every state k, and we cannot maximize the log-likelihood (11) without them.

We therefore appeal to the EM algorithm for iterative maximum likelihood estimation with missing data (Dempster, Laird, and Rubin 1977). In the EM algorithm, we define a surrogate objective function Q by taking the expectation of the complete data log-likelihood (11), conditional on the observed data $\mathbf{Y}$ and the parameter values $\boldsymbol{\theta}^{(m)}$ from the previous iteration of the EM algorithm (the E-step). Then, we find the parameter values $\boldsymbol{\theta}^{(m+1)}$ that maximize this surrogate function (the M-step). This two-step process is repeated until convergence to the MLE of $\boldsymbol{\theta}$. Taking the expectation of (11) conditional on $\mathbf{Y}$ and $\boldsymbol{\theta}^{(m)}$, we form the surrogate function Q:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) &= \mathbb{E}[\ell(\boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}^{(m)}] \\ &= \sum_{k=0}^{\infty} [\mathbb{E}(U_k|\mathbf{Y}) \log[\lambda_k(\boldsymbol{\theta})] + \mathbb{E}(D_k|\mathbf{Y}) \log[\mu_k(\boldsymbol{\theta})] \\ &\quad - \mathbb{E}(T_k|\mathbf{Y})[\lambda_k(\boldsymbol{\theta}) + \mu_k(\boldsymbol{\theta})]], \end{aligned} \tag{12}$$

where for clarity we have omitted the dependence of the expectations on the parameter value $\boldsymbol{\theta}^{(m)}$ from the mth iterate. In general, we assume that the maximum likelihood estimator exists; see Bladt and Sorensen (2005) for a discussion of the issues of identifiability, existence, and uniqueness. In the following, we always assume that the BDP is nonexplosive (see Karlin and McGregor 1957a, b, for details) and that $\sum_k \lambda_k \mathbb{E}(U_k|\mathbf{Y})$, $\sum_k \mu_k \mathbb{E}(D_k|\mathbf{Y})$, and $\sum_k (\lambda_k + \mu_k)\mathbb{E}(T_k|\mathbf{Y})$ are finite.

2.3 Computing the Expectations of the *E*-Step

Computing the expectations of U_k, D_k, and T_k in the E-step is difficult in birth-death estimation since the unobserved state path and waiting times are not independent conditional on the observed data $\mathbf{Y}$. In addition, the state-space of a BDP is generally infinite, so the process may visit states $k \gg \max(a, b)$. It is tempting to approximate an infinite BDP as a similar process on the finite state-space $\{0, 1, \ldots, N\}$, where N is chosen so that the probability of the process visiting states greater than N is small. That is, we could choose N and ϵ so that

$$\Pr(X(s) > N \mid X(0) = a, X(t) = b, 0 < s < t) < \epsilon. \tag{13}$$

A priori truncation of the state space would allow one to take advantage of the methods for matrix-algebraic computation of conditional expectations such as eigendecomposition and uniformization, as developed in Hobolth and Jensen (2011). This turns out to be infeasible for two reasons. First, it is unclear how to evaluate (13) and whether knowledge of this probability can provide error bounds on expectations of BDP statistics; this makes the choice of N somewhat arbitrary. Second, as we demonstrate in Section 3.1 using numerical experiments, matrix methods for computation of expectations can suffer from catastrophic roundoff error.

Recently, some authors have made analytic progress for infinite state-space BDPs. Doss et al. (2013) adopted an approach for linear BDPs that combines analytic results with simulations. For some models, these authors were able to derive the generating function for the joint distribution of U, D, T_{particle}, and the state path conditional on $X(0) = a$ and can manipulated this generating function to complete the E-step. For a more complicated linear model, Doss et al. (2013) resorted to approximating the relevant conditional expectations by simulating sample paths, conditional on $\mathbf{Y}$ using the method introduced by Hobolth (2008).

Our solution is to recognize that we do not need to know very much about the missing data to find the conditional expectations used in the sufficient statistics above. In fact, the transition probabilities are all that we require. The following integral representations of the conditional expectations in the EM algorithm will prove useful:

$$\mathbb{E}(U_k|\mathbf{Y}) = \frac{\int_0^t P_{ak}(\tau)\lambda_k P_{k+1,b}(t-\tau)\,\mathrm{d}\tau}{P_{ab}(t)}, \tag{14a}$$

$$\mathbb{E}(D_k|\mathbf{Y}) = \frac{\int_0^t P_{ak}(\tau)\mu_k P_{k-1,b}(t-\tau)\,\mathrm{d}\tau}{P_{ab}(t)}, \tag{14b}$$

and

$$\mathbb{E}(T_k|\mathbf{Y}) = \frac{\int_0^t P_{ak}(\tau)P_{kb}(t-\tau)\,\mathrm{d}\tau}{P_{ab}(t)}. \tag{14c}$$

These formulas have appeared in many types of studies related to EM estimation for continuous-time Markov chains (Lange 1995a; Holmes and Rubin 2002; Bladt and Sorensen 2005; Hobolth and Jensen 2005; Metzner et al. 2007). For general

BDPs whose transition probabilities must be computed numerically, numerical integration over the product of the densities can be computationally prohibitive.

However, the numerators in (14) are convolutions of integrable time-domain functions. Since the Laplace transforms $f_{ab}(s)$ of these transition probabilities are available and easy to compute, we take advantage of the Laplace convolution property, arriving at the representations

$$\mathbb{E}(U_k|\mathbf{Y}) = \lambda_k \frac{\mathscr{L}^{-1}[f_{ak}(s)\ f_{k+1,b}(s)](t)}{P_{ab}(t)}, \tag{15a}$$

$$\mathbb{E}(D_k|\mathbf{Y}) = \mu_k \frac{\mathscr{L}^{-1}[f_{ak}(s)\ f_{k-1,b}(s)](t)}{P_{ab}(t)}, \tag{15b}$$

and

$$\mathbb{E}(T_k|\mathbf{Y}) = \frac{\mathscr{L}^{-1}[f_{ak}(s)\ f_{kb}(s)](t)}{P_{ab}(t)}. \tag{15c}$$

where $\mathscr{L}^{-1}$ denotes inverse Laplace transformation. Although these formulas are equivalent to (14), they offer substantial time savings over computing the integral directly, and render tractable the computation of expectations in the EM algorithm for arbitrary general BDPs. The Appendix shows how to calculate (15) numerically and control the total error using a discretized Laplace inversion method popularized by Abate and Whitt (1992b, 1995). This approach allows us to terminate the continued fraction evaluation dynamically at a depth that controls the error due to both truncation and discretization of the inversion integral. We emphasize that we do not need to choose a truncation index a priori as would be required in matrix truncation approaches.

2.4 Maximization Techniques for Various BDPs

In contrast to the generic technique outlined above for computing the expectations of the E-step, the M-step depends explicitly on the functional form of the birth and death rates $\lambda_k(\boldsymbol{\theta})$ and $\mu_k(\boldsymbol{\theta})$. Here, we give several representative examples of BDPs and techniques for completing the M-step of the EM algorithm, such as analytic maximization, minorize-maximize (MM), and Newton's method.

2.4.1 Simple Linear BDP. In the simple linear BDP, births and deaths happen at constant per-capita rates, so $\lambda_k = k\lambda$ and $\mu_k = k\mu$. The unknown parameter vector is $\boldsymbol{\theta} = (\lambda, \mu)$, and the surrogate function becomes

$$Q(\boldsymbol{\theta}) = \sum_{k=0}^{\infty}[\mathbb{E}(U_k|\mathbf{Y})\log[k\lambda] + \mathbb{E}(D_k|\mathbf{Y})\log[k\mu] - \mathbb{E}(T_k|\mathbf{Y})k(\lambda+\mu)]. \tag{16}$$

Taking the derivative of (16) with respect to the unknown parameters, setting the result to zero, and solving for λ and μ gives the M-step updates

$$\lambda^{(m+1)} = \frac{\mathbb{E}(U|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}, \tag{17a}$$

and

$$\mu^{(m+1)} = \frac{\mathbb{E}(D|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}. \tag{17b}$$

These updates correspond to the usual maximum likelihood estimators in the continuously observed process (Reynolds 1973). Note that the transition probabilities $P_{ab}(t)$ in the denominators of the expectations in (14) cancel out in (17a) and (17b). When this is the case, transition probabilities are not necessary to derive an EM algorithm.

2.4.2 Linear BDP With Immigration. Sometimes populations are not closed, and new individuals can enter; we call this action "immigration." Another interpretation arises in models of point mutations in DNA sequences. Suppose new mutations arise in a DNA sequence via two distinct processes: one inserts new mutants at a rate proportional to the number already present, and the other creates new mutations at a constant rate, regardless of how many already exist. To model this behavior, we augment the simple linear BDP above with a constant term ν representing immigration, so that $\lambda_k = k\lambda + \nu$ and $\mu_k = k\mu$. The log-likelihood becomes

$$\ell(\boldsymbol{\theta}) = \sum_{k=0}^{\infty}[U_k\log(k\lambda+\nu) + D_k\log(k\mu) - T_k[k(\lambda+\mu)+\nu]]. \tag{18}$$

Unfortunately, if we take the derivative of the log-likelihood with respect to λ or ν, the unknown appears in the denominator of the terms of the infinite sum. However, since each summand is a concave function of the unknown parameters, we can separate them in a minorizing function M such that for all $\boldsymbol{\theta}$, $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) \le \ell(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}) = \ell(\boldsymbol{\theta}^{(m)})$ as follows:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &\ge M\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\right)\\ &= \sum_{k=0}^{\infty}[U_k[p_k\log(p_k\lambda) + (1-p_k)\log((1-p_k)\nu)]\\ &\quad + D_k\log(\mu) - [k(\lambda+\mu)+\nu]T_k],\end{aligned} \tag{19}$$

where $p_k = k\lambda^{(m)}/(k\lambda^{(m)} + \nu^{(m)})$. Then, letting $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = \mathbb{E}(M(\boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}^{(m)})$ be the surrogate function, this minorization forms the basis for an EM algorithm in which a step of the minorize-maximize (MM) algorithm takes the place of the M-step, and the ascent property of the EM algorithm is preserved (Lange 2010). Maximizing Q with respect to λ and ν yields the updates

$$\lambda^{(m+1)} = \frac{\sum_{k=0}^{\infty} p_k\mathbb{E}(U_k|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}, \tag{20a}$$

and

$$\nu^{(m+1)} = \frac{1}{t}\sum_{k=0}^{\infty}(1-p_k)\mathbb{E}(U_k|\mathbf{Y}). \tag{20b}$$

Expression (20a) is similar to (17a), the update for λ in the simple BDP. The difference lies in that each $\mathbb{E}(U_k|\mathbf{Y})$ in this case is weighted by the proportion of additions at state k due to births, not immigrations. The update for μ is the same as (17b).

2.4.3 Logistic/Restricted Growth. To illustrate an EM algorithm for more complicated rate specifications in which no MM update is evident and the rates no longer depend on the current state k in a linear way, we examine a model for restricted population growth. Typical *deterministic* population models often

incorporate limitations on population size due to the carrying capacity K of the environment. One famous example is the logistic model of population growth (Murray 2002). Continuous-time stochastic analogs have previously required a finite cap on population size (Tan and Piantadosi 1991). These stochastic models roughly mimic the behavior of the deterministic model for population sizes below K, but are limited because they do not allow growth beyond K. Here, we present a model which supports transient growth beyond the carrying capacity, but where the population size tends to a balance between restricted growth and death.

Suppose births are cooperative, requiring two parents, but fecundity decays as the number of extant particles increases, and death remains an independent process such that $\lambda_k = \lambda k^2 e^{-\beta k}$ and $\mu_k = k\mu$. Here, we can interpret the carrying capacity roughly as the population size $k > 0$ at which $\lambda_k \approx \mu_k$. Ignoring irrelevant terms, the surrogate function becomes

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}\right) = \sum_{k=0}^{\infty} [\mathbb{E}(U_k|\mathbf{Y})[\log(\lambda) - \beta k] + \mathbb{E}(D_k|\mathbf{Y})\log(\mu) - \mathbb{E}(T_k|\mathbf{Y})[\lambda k^2 e^{-\beta k} + k\mu]]. \tag{21}$$

Since λ and β appear together, we opt for a numerical Newton step. Denoting the gradient of Q as $\mathbf{F}$ and the Hessian by $\mathbf{H}$, we update these parameters by

$$\begin{pmatrix} \lambda^{(m+1)} \\ \beta^{(m+1)} \end{pmatrix} = \begin{pmatrix} \lambda^{(m)} \\ \beta^{(m)} \end{pmatrix} - \mathbf{H}^{-1}\mathbf{F}. \tag{22}$$

The ascent property is preserved when a Newton step is used in place of an exact M-step (Lange 1995a). The update for μ is the same as (17b).

2.4.4 SIS Epidemic Models. Under a very common epidemic model, members of a finite population of size N are classified as either "susceptible" to a given disease or "infected" (Bailey 1964; Andersson and Britton 2000). Susceptibles become infected in proportion to the number of currently infected in the population, and infecteds revert to susceptible status with a certain rate independent of how many infecteds there are. This idealized susceptible-infectious-susceptible (SIS) infectious disease model specifies a general birth-death process in which we track the number of infecteds. Let $\lambda_k = \beta k(N-k)/N$ be the rate of new infections when there are already k infected in the population. Let $\mu_k = \gamma k/N$ be the rate of recovery of infecteds to susceptibles. Then if $\boldsymbol{\theta} = (\beta, \gamma)$, we have

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\right) = \sum_{k=0}^{N} [\mathbb{E}(U_k|\mathbf{Y})\log(\beta) + \mathbb{E}(D_k|\mathbf{Y})\log(\gamma) - \mathbb{E}(T_k|\mathbf{Y})(k(N-k)\beta + k\gamma)/N], \tag{23}$$

and the updates are

$$\beta^{(m+1)} = \frac{N\mathbb{E}(U|\mathbf{Y})}{\sum_{k=0}^{N}(N-k)k\mathbb{E}(T_k|\mathbf{Y})},$$

and

$$\gamma^{(m+1)} = \frac{N\mathbb{E}(D|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}. \tag{24}$$

2.4.5 Generalized Linear Models. Our general framework allows assessment of the influence of covariates on the rates of a general BDP in a novel way. Suppose we sample observations from independent processes $X_i(\tau)$, $i = 1, \ldots, N$ and observe $\mathbf{Y}_i = (X_i(0), X_i(t_i))$ associated with d covariates $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})'$. These processes may represent different subjects in a study. We model the birth and death rates λ_{ik} and μ_{ik} for each process/subject X_i as functions of $\mathbf{z}_i$ and unknown d-dimensional regression coefficients $\boldsymbol{\theta}_\lambda$ and $\boldsymbol{\theta}_\mu$ in a GLM framework. We link

$$\log(\lambda_{ik}) = g(k, \mathbf{z}_i'\boldsymbol{\theta}_\lambda) \quad \text{and} \quad \log(\mu_{ik}) = h(k, \mathbf{z}_i'\boldsymbol{\theta}_\mu), \tag{25}$$

where $g(\cdot)$ and $h(\cdot)$ are scalar-valued functions. We note the possibility that covariates may differ between $\boldsymbol{\theta}_\lambda$ and $\boldsymbol{\theta}_\mu$ through trivial modification; to ease notation, we do not explore this direction. Given N independent processes, we sum log-likelihoods to arrive at the multiple-subject surrogate function:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\right) = \sum_{i=1}^{N}\sum_{k=0}^{\infty}\left[\mathbb{E}(U_k|\mathbf{Y}_i)g(k, \mathbf{z}_i'\boldsymbol{\theta}_\lambda) + \mathbb{E}(D_k|\mathbf{Y}_i)h(k, \mathbf{z}_i'\boldsymbol{\theta}_\mu) - \mathbb{E}(T_k|\mathbf{Y}_i)\left(e^{g(k,\mathbf{z}_i'\boldsymbol{\theta}_\lambda)} + e^{h(k,\mathbf{z}_i'\boldsymbol{\theta}_\mu)}\right)\right]. \tag{26}$$

Although we cannot usually maximize this surrogate function for all elements of $(\boldsymbol{\theta}_\lambda, \boldsymbol{\theta}_\mu)$ simultaneously, a Newton step is often straightforward to derive.

As an example, consider a GLM extension of the simple linear BDP in which

$$\log(\lambda_{ik}) = \log(k) + \mathbf{z}_i'\boldsymbol{\theta}_\lambda, \quad \text{and} \quad \log(\mu_{ik}) = \log(k) + \mathbf{z}_i'\boldsymbol{\theta}_\mu. \tag{27}$$

Taking the gradient of the corresponding surrogate function Q with respect to the parameters $\boldsymbol{\theta}_\lambda$ yields

$$\nabla_{\boldsymbol{\theta}_\lambda} Q = \sum_{i=1}^{N} \mathbb{E}(U|\mathbf{Y}_i)\mathbf{z}_i - e^{\mathbf{z}_i'\boldsymbol{\theta}_\lambda}\mathbb{E}(T_{\text{particle}}|\mathbf{Y}_i)\mathbf{z}_i \tag{28}$$

and the second differential (Hessian) of Q is

$$\mathbf{d}^2_{\boldsymbol{\theta}_\lambda} Q = -\sum_{i=1}^{N} e^{\mathbf{z}_i'\boldsymbol{\theta}_\lambda}\mathbb{E}(T_{\text{particle}}|\mathbf{Y}_i)\mathbf{z}_i\mathbf{z}_i'. \tag{29}$$

Combining these, we arrive at the Newton step for the parameter vector $\boldsymbol{\theta}_\lambda$:

$$\boldsymbol{\theta}_\lambda^{(m+1)} = \boldsymbol{\theta}_\lambda^{(m)} - \left(\mathbf{d}^2_{\boldsymbol{\theta}_\lambda} Q\right)^{-1} \nabla_{\boldsymbol{\theta}_\lambda} Q. \tag{30}$$

A similar update can be found for $\boldsymbol{\theta}_\mu$. These updates are examples of the gradient EM algorithm for regression in Markov processes described by Wanek et al. (1993) and Lange (1995a). It is worth noting that the Hessian matrix $\mathbf{d}^2_{\boldsymbol{\theta}_\lambda} Q$ can become ill-conditioned, making it difficult to invert for the Newton step in (30) for some problems. Unfortunately there is no quasi-Newton option since in general $\mathbb{E}(T_{\text{particle}}|\mathbf{Y})e^{\mathbf{z}_i'\boldsymbol{\theta}_\lambda}$ is unbounded. An alternative to inversion of the Hessian matrix is cyclic coordinate descent in which a Newton step is performed for each coordinate $\boldsymbol{\theta}_j$ individually. This carries the advantage of avoiding matrix inversion, but convergence is slower and the ascent property must be checked at each Newton step.

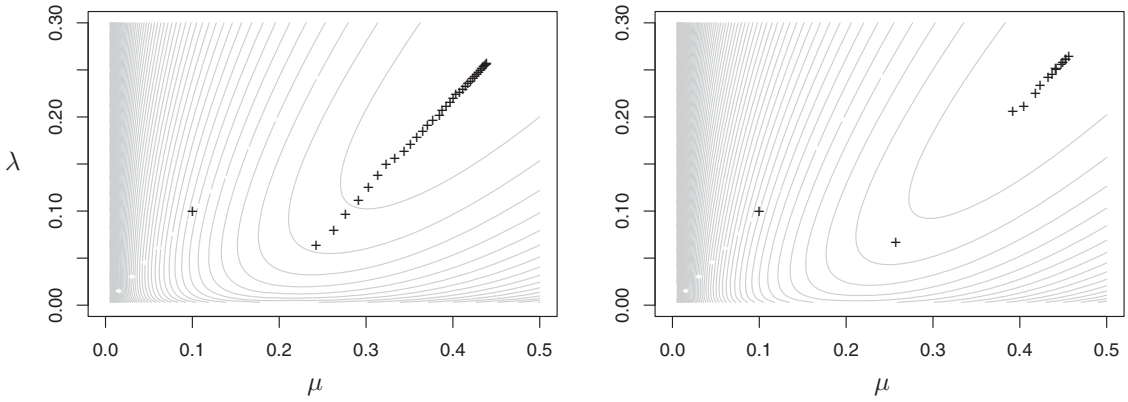

Figure 2. Effect of quasi-Newton acceleration on iterates of the EM algorithm for a simple linear BDP with birth rate λ and death rate μ. Contour lines sketch the log-likelihood from $N = 50$ discrete samples. Iterates are shown with the "+" symbol. On the left, ordinary EM iterates converge very slowly in the neighborhood of the maximum, for a total of 36 iterations. On the right, EM iterates using quasi-Newton acceleration make large jumps and converge rapidly in 15 iterations.

2.5 Implementation

2.5.1 E-Step Error and Acceleration. The E-step in these EM algorithms for BDP estimation usually involves infinite weighted sums of the conditional expectations $\mathbb{E}(U_k|\mathbf{Y})$, $\mathbb{E}(D_k|\mathbf{Y})$, and $\mathbb{E}(T_k|\mathbf{Y})$. For example, when estimating λ in the simple linear BDP, we must evaluate

$$\begin{aligned}\mathbb{E}(U|\mathbf{Y}) &= \sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y}) \\ &= \sum_{k=0}^{\infty} \frac{\lambda_k}{P_{ab}(t)} \mathscr{L}^{-1}[f_{ak}(s)\ f_{k+1,b}(s)](t). \qquad (31)\end{aligned}$$

We find an increase in computational efficiency by exchanging the order of Laplace inversion and summation. Then, (31) becomes

$$\mathbb{E}(U|\mathbf{Y}) = \frac{1}{P_{ab}(t)} \mathscr{L}^{-1}\left[\sum_{k=0}^{\infty} \lambda_k f_{ak}(s) f_{k+1,b}(s)\right](t), \qquad (32)$$

In practice, we can only evaluate a finite number of terms in the series, so we must truncate the infinite sum in (32). This truncation approach bears some similarity to matrix truncation methods, described further in Section 3.1 and Figure 2. The difference is that truncation of the infinite sum in (32) is *dynamic* and may depend on the magnitude of the summand at every step. In particular, we use a series acceleration method to compute (32) that provides ready approximation of the remaining tail sum at each step (Levin 1973; Press 2007). In contrast, the matrix approximation approach requires choosing the truncation index a priori, and does not allow for dynamic choice of the truncation index. The Appendix describes bounds for the numerical error in this computation in greater detail.

2.5.2 Acceleration of EM Iterates. EM algorithms are notorious for slow convergence, especially near optima. Although our purpose in the present article is limited to basic EM techniques for analyzing general BDPs, we exploit the quasi-Newton acceleration method introduced by Lange (1995b) in our implementations. Other acceleration methods exist, and may give better results, depending on the problem (Louis 1982; Meilijson 1989; Jamshidian and Jennrich 1993; Liu and Rubin 1994; Lange 1995a; Liu 1998; He and Liu 2012). Figure 3 shows

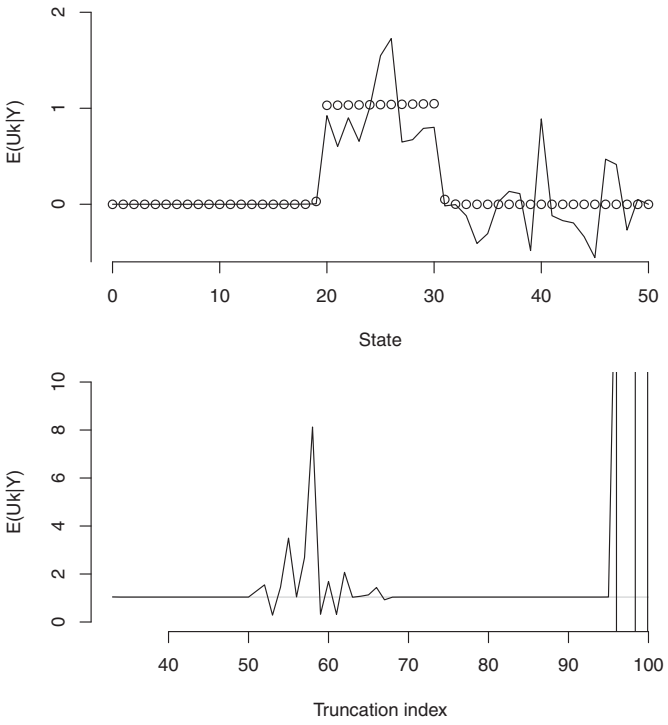

Figure 3. Ill-conditioned transition rate matrix causes the eigendecomposition method to fail. In the top panel, we calculate $\mathbb{E}(U_k|\mathbf{Y})$ for $\mathbf{Y} = (a = 20, b = 31, t = 1)$, $\beta = 1$, $\gamma = 1$, and $k = 0, 1, \ldots, 50$. The open circles denote values calculated by the Laplace method, and the line represents the values calculated using the EDecomp method. While not biologically unreasonable, these parameter values cause the transition rate matrix to become ill-conditioned, and eigendecomposition suffers from catastrophic numerical error. The Laplace method remains stable and is unaffected by matrix conditioning. In the bottom panel, we show $\mathbb{E}(U_{25}|\mathbf{Y})$ for the same $\mathbf{Y}$ as above, computed using the EDecomp method with different matrix truncation indices. The true value (approximately 1.038596) is shown in gray. The first group of inaccurate values is due to truncation of the matrix below states that are likely to be reached by the chain on its path from 20 to 31. The second group of inaccurate values (from about 95 to 100) is due to the numerical instability involved in inverting a large matrix.

the log-likelihood function and iterates for the basic EM and accelerated EM methods in the simple linear model. Since the quasi-Newton acceleration method does not guarantee that the likelihood increases at each step, "step-halving" is occasionally necessary to achieve ascent. Note that this requires likelihood evaluation at least once per iteration. Our approach is advantageous in that we can efficiently calculate this likelihood (product of transition probabilities) for any general BDP (Crawford and Suchard 2012).

2.5.3 Asymptotic Variance of EM Estimates. Finding the observed information matrix for an EM estimate can be challenging. Louis (1982) gives formulas for the observed information, which Doss et al. (2013) use to derive analytic expressions for the observed information for very simple BDPs. Direct calculation of the information matrix via the derivation given by Oakes (1999) is appealing, but the conditional expectations are usually only available to us numerically, and hence we cannot find the necessary analytic expressions. Analytic expressions for the asymptotic variance are generally hard to find for more complicated models. In some such cases, Liu (1998) suggested a normal approximation. We instead turn to the supplemented EM (SEM) algorithm of Meng and Rubin (1991), which computes the information matrix of the EM estimate of $\boldsymbol{\theta}$ after the MLE $\hat{\boldsymbol{\theta}}$ has been found. Although the SEM algorithm can be slow, it does not require that the expectations have analytic expressions. The observed information is $\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\mathrm{d}^2 Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})(\mathbf{I} - \mathrm{d}\mathbf{M}(\hat{\boldsymbol{\theta}}))$, where $\mathbf{M}(\boldsymbol{\theta})$ is the EM algorithm map such that $\boldsymbol{\theta}^{(m+1)} = \mathbf{M}(\boldsymbol{\theta}^{(m)})$. We numerically approximate the differential $\mathrm{d}\mathbf{M}$ at the termination of the EM algorithm.

We note also that since we are able to calculate transition probabilities directly, the observed data log-likelihood is easily computed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P_{a_i b_i}(t_i), \tag{33}$$

where $a_i = X_i(0)$ and $b_i = X_i(t_i)$. As an alternative to the approaches outlined above, we can calculate the Hessian using purely numerical techniques. If $\mathbf{H}(\hat{\boldsymbol{\theta}}) = \mathrm{d}^2\ell(\hat{\boldsymbol{\theta}})$ is the numerical Hessian evaluated at the estimated value $\hat{\boldsymbol{\theta}}$, then $\hat{\mathbf{I}} \approx -\mathbf{H}(\hat{\boldsymbol{\theta}})$.

3. RESULTS

3.1 Laplace Convolution E-Step Comparison

To illustrate the computational speedup that the Laplace convolution formulas (15) and their acceleration in section 2.5.1 achieve over existing methods, we calculate conditional expectations of the number of births using six different methods for various BDP models and report computing times in Table 1. The Reject method employs rejection sampling of trajectories where we condition on the starting state, and reject based on the ending state (Bladt and Sorensen 2005). The Endsim method uses an endpoint-conditioned simulation algorithm to sample trajectories on a truncated state space (Hobolth 2008; Hobolth and Stone 2009). In both the simulation methods, we repeated the simulation until the Monte Carlo error became small enough that we obtained the true value of the statistic to a certain accuracy with high probability. We terminated the simulation when a 95% confidence interval for the true value of the statistic has width less that 0.1. The TConv method involves naïve numerical time-domain convolution (Equation (14)) using the `integrate` function in R. The EDecomp method uses an eigendecomposition of the truncated rate matrix to compute the conditional expectation (Minin and Suchard 2008; Hobolth and Jensen 2011). The Unif method uses uniformization to compute the conditional expectations (Jensen 1953; Hobolth and Jensen 2011). Finally, the Laplace method uses our Laplace-domain convolution method outlined in Section 2.3.

To adapt finite state space methods to the problem of computing the number of births in a BDP, we choose a truncated rate matrix dimension of 100. We are aware that the size of the rate matrix affects the speed of the simulation routine, so

Table 1. Compute times (s) to perform the E-step for birth-death statistics using six different methods for four different BDP models

Model	Value $\mathbb{E}(U\|\mathbf{Y})$	Compute times (s)					
		Reject	EndSim	TConv	EDecomp	Unif	Laplace
Simple linear (2.4.1) $\lambda = 0.5, \mu = 0.3$ $\mathbf{Y} = (a = 19, b = 27, t = 1)$	13.51	81.95	851.22	9.28	0.23	5.04	0.18
Immigration (2.4.2) $\lambda = 0.5, \mu = 0.3, \alpha = 0.2$ $\mathbf{Y} = (a = 19, b = 27, t = 1)$	13.35	78.29	883.87	9.40	0.24	5.03	0.19
Logistic (2.4.3) $\lambda = 0.5, \mu = 0.3, \alpha = 0.2$ $\mathbf{Y} = (a = 10, b = 12, t = 1)$	5.67	12.21	571.28	2.85	*	2.29	0.09
SIS (2.4.4) $\beta = 0.5, \gamma = 0.3$ $\mathbf{Y} = (a = 10, b = 17, t = 1)$	7.05	0.17	11.66	3.47	*	2.23	0.06

NOTE: We report text section numbers in which the models are described in parentheses. The "Value" column reports the true value of the statistic $\mathbb{E}(U|\mathbf{Y}) = \sum_k \mathbb{E}(U_k|\mathbf{Y})$ for the simple linear, logistic, and SIS models, and $\sum_k p_k \mathbb{E}(U_k|Y)$ for the immigration model. Each E-step method obtained the same value up to at least two decimal places. Compute times are given for rejection simulation (Reject), endpoint-conditioned simulation (EndSim), numerical time-convolution (TConv), eigendecomposition (EDecomp), uniformization (Unif), and our Laplace convolution method (Laplace). In all cases, the Laplace method takes substantially less time. Eigendecomposition fails for the logistic and SIS models because the rate matrix becomes computationally singular.

we wish to keep the matrix as small as possible. On the other hand, the matrix must remain large enough to include states that may be visited with high probability in a path from a to b over time t. However, it is not practical to dynamically choose the dimension of the truncated rate matrix a priori. For example, we might choose the dimension d such that the process visits states greater than $d > b$ with low probability, so $P_{ak}(t) < \epsilon$ for $k > d$. However, rules for choosing the dimension of the truncated matrix that themselves depend on computation of transition probabilities can dramatically increase computational time.

In our implementation of all methods, we have made every effort to reuse as much shared R code as possible, with the aim of making the routines comparable in numerical accuracy and computational time. We consider four different BDPs: for the simple linear BDP and linear BDP with immigration, we use the discrete observation $\mathbf{Y} = (a = 19, b = 27, t = 1)$. Under the logistic and SIS models, the observation is $\mathbf{Y} = (a = 10, b = 12, t = 1)$ and $\mathbf{Y} = (a = 10, b = 17, t = 1)$. We list all model parameter values in Table 1.

In these examples, the Laplace convolution method generally outperforms other methods and remains stable even when the truncated rate matrix becomes ill-conditioned. The EDecomp and Uniformization methods perform reasonably well, but as we show in the next section, matrix decomposition methods can suffer from catastrophic numerical error when the rate matrix becomes large or nearly singular, as is often the case for the SIS model. While the simulation methods Reject and EndSim can provide quick estimates of the relevant expectations, the Monte Carlo error decays extremely slowly and therefore a large number of simulations are necessary to obtain convergence to a small interval about $\mathbb{E}(U|\mathbf{Y})$ with high probability. For example, in the simple linear model, the Reject method required more that 6000 successful simulants to achieve a Monte Carlo standard error small enough to terminate the simulation.

Finally, we give an example to illustrate one important benefit of our Laplace convolution method: since it does not depend explicitly on a decomposition of the transition rate matrix, the method is much more stable when this matrix is ill-conditioned. Consider the SIS model with $N = 50$ individuals and transition rates $\beta = 1$ and $\gamma = 1$. The top panel of Figure 2 shows the values of $\mathbb{E}(U_k|\mathbf{Y})$ for $\mathbf{Y} = (a = 20, b = 31, t = 1)$, with $k = 0, 1, \ldots, 50$, calculated using the Edecomp and Laplace methods. The eigendecomposition produces catastrophic numerical error for larger states because the transition rate matrix becomes ill-conditioned. The parameter values β and γ correspond to unit rates per unit time of susceptible individuals becoming ill, and ill individuals reverting to susceptible status. These values are not biologically unreasonable, and do not result in a process that is degenerate or ill-defined. The Laplace method handles rate specifications like these that result in ill-conditioned rate matrices without issue, and the continued fraction evaluation remains numerically stable. The bottom panel shows the computed value of $\mathbb{E}(U_{25}|\mathbf{Y})$ as above, but with different truncations of the rate matrix. It can be very difficult to determine a priori which truncations of the matrix will result in poor approximations. In contrast, the Laplace method allows evaluation of the terms of the sum (32) until the desired numerical accuracy has been reached.

Table 2. Point-estimates and their standard errors (SE) for simulated observations under various BDPs

Model	Parameter	True	Estimate	SE
Simple linear ($N = 500$)	λ	0.5	0.5039	0.0269
(2.4.1)	μ	0.2	0.1981	0.0254
Immigration ($N = 800$)	λ	0.2	0.2182	0.0129
(2.4.2)	ν	0.1	0.1016	0.0213
	μ	0.25	0.2488	0.0231
Logistic ($N = 1500$)	λ	0.3	0.2917	0.0035
(2.4.3)	α	0.5	0.4942	0.0397
	μ	0.05	0.0456	0.0633
SIS ($N = 1000$)	β	0.1	0.1025	0.0048
(2.4.4)	γ	2.0	2.1374	0.0367
GLM ($N = 1000$)	$\boldsymbol{\theta}_{\lambda,1}$	0.25	0.2585	0.0393
(2.4.5)	$\boldsymbol{\theta}_{\lambda,2}$	0.1	0.1143	0.0402
	$\boldsymbol{\theta}_{\mu,1}$	0.2	0.1973	0.0457
	$\boldsymbol{\theta}_{\mu,2}$	0.05	0.0877	0.0457

NOTE: We report the text section describing each of the models in parentheses. The method for generating the rates in the generalized linear model (GLM) BDP is described in the text.

3.2 Synthetic Examples

To evaluate the performance of our EM algorithms, we simulate discrete observations from several of the BDPs outlined above. For each sample, we draw starting points $X_i(0)$ uniformly from the integers 0 to 20, and times t_i uniformly from 0.1 to 3. We then simulate a trajectory of the BDP and record the state $X_i(t_i)$. For the GLM, we employ the simple linear parameterization with a log link with $d = 2$ covariates. We specify the covariates $\mathbf{z}_i = (z_{i,1}, z_{i,2})$ as follows: $z_{i,1} \sim N(1, \sigma^2)$, $z_{i,2} \sim N(2, \sigma^2)$ for $i = 1, \ldots, N/2$, $z_{i,1} \sim N(2, \sigma^2)$ and $z_{i,2} \sim N(1, \sigma^2)$ for $i = N/2 + 1, \ldots, N$, where $\sigma^2 = 0.1$.

Table 2 reports the number of simulated observations, true parameter values, point-estimates, asymptotic standard error estimates for all model parameters. It is important to note that the MLEs can differ substantially from the parameter values used to perform the simulation, regardless of the algorithm used to find the MLEs. This is due to several factors, including: (1) missing state paths; (2) stochasticity of the BDP generating the state paths; (3) arbitrary choice of starting states $X_i(0)$; and (4) finite sample sizes. Despite these limitations inherent in learning from partially observed stochastic processes, the point-estimates match the true parameter values rather well.

3.3 Application to Lymphoma Cell Growth

Cancer researchers often use in vitro experiments to evaluate the efficacy of novel therapies. They subject cultured cancer cells to treatment and count the number of cells that survive. Liu, Beckett, and DeNardo (2007) studied the effect of a mixture of two monoclonal antibodies, chLym-1 and rituximab, on proliferation of human lymphoma cells. These antibodies exhibit strong antitumor cell effects (Liu et al. 2004; DeNardo 2005). The data of Liu, Beckett, and DeNardo (2007) consist of repeated experiments in which the outcome is the number of viable tumor cells in a test tube. They counted the number of lymphoma cells at antibody concentrations 0, 0.025, 0.25, 2.5, and 10 μml and studied the effects over incubation times of 1, 2, and

3 days. The data consist of observations $\mathbf{Y}_i = (X_i(0), X_i(t_i), t_i)$, where $X_i(0)$ is the number of viable cells at the beginning of the ith experiment, and $X_i(t_i)$ is the number of viable cells at time $t_i \in \{1, 2, 3\}$ days. Liu, Beckett, and DeNardo (2007) fit a model for the mean behavior of a simple linear BDP using the antibody concentration as a covariate through a quasi-likelihood approach that models conditional expected count at time t_i:

$$\begin{aligned}\mathbb{E}(X_i(t_i) \mid X_i(0) = x) &= x \exp[(\lambda - \mu)t_i] \\ &= x \exp[\mathbf{z}_i'(\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}_\mu)t_i]. \qquad (34)\end{aligned}$$

Modeling the mean behavior of the BDP allows only the difference of $\boldsymbol{\theta}_\lambda$ and $\boldsymbol{\theta}_\mu$ to be estimated. The resulting log-linear model for the deterministic mean behavior of the BDP is limiting because it is essentially equivalent to Poisson regression, and does not capture the stochastic branching structure of the underlying BDP.

We now extend the work of Liu, Beckett, and DeNardo (2007) by using a full stochastic BDP model instead of fitting the deterministic mean behavior. Although in the experiments described by Liu, Beckett, and DeNardo, $X_i(0)$ is unknown, they estimated its mean as 23 under one model. To avoid conditioning on a random variable, we also follow Liu, Beckett, and DeNardo (2007) and treat $X_i(0) = 23$ as fixed in our analysis. Since the concentration covariates vary nonlinearly, we transform them as $\log(1 + c_i)$, where c_i is the concentration in the ith observation. Consider a GLM for the rates of a simple linear BDP, as described in section 2.4.5. The rates for the ith observation are $\log \lambda_{in} = \log n + \mathbf{z}_i'\boldsymbol{\theta}_\lambda$ and $\log \mu_{in} = \log n + \mathbf{z}_i'\boldsymbol{\theta}_\mu$. Here, the covariate vector is $\mathbf{z}_i = (1, \log(1 + c_i))'$, consisting of an intercept and the log-transformed antibody concentration. The surrogate function is given by (26). Table 3 shows the results of fitting this model, and Table 4 shows the estimated birth and death rates for cells grown under each each antibody concentration.

We draw several tentative conclusions from our stochastic BDP analysis of these data. First, both baseline birth and death rates, in the absence of antibody, are large. Birth rate is dramatically decreased by higher concentrations of antibody. Interestingly, death rate also decreases slightly with antibody concentration. At the highest experimental antibody concentration, the death rate becomes larger than the birth rate, resulting in dramatic reduction in cell counts. These observations agree with the known properties of cancer cells in general—that they reproduce very rapidly when uninhibited by therapeutic agents. However, since we are able to estimate the effect of antibody concentration on both birth and death rates separately, our method provides additional insight into the branching nature of the underlying process.

Table 3. Parameter estimates and asymptotic standard errors for the cancer cell model

	$\boldsymbol{\theta}_\lambda$		$\boldsymbol{\theta}_\mu$	
Parameter	Estimate	SE	Estimate	SE
Intercept	1.4719	0.1614	1.2038	0.2134
$\log(1 + c_i)$	−0.1190	0.0753	−0.0018	0.0927

NOTE: We fit the simple linear BDP to the data of Liu, Beckett, and DeNardo (2007) using the regression framework outlined in Section 2.4.5. The covariate c_i is the concentration of antibody added to the lymphoma cell culture. Standard errors were obtained using the numerical Hessian of the log-likelihood.

Table 4. Birth and death rates stratified by different concentrations of antibody, where $\lambda = e^{\mathbf{z}_i'\boldsymbol{\theta}_\lambda}$ and $\mu = e^{\mathbf{z}_i'\boldsymbol{\theta}_\mu}$

	λ		μ	
Antibody concentration	Estimate	SE	Estimate	SE
0	4.357	0.703	3.333	0.711
0.025	4.345	0.696	3.333	0.705
0.25	4.243	0.643	3.331	0.66
2.5	3.754	0.478	3.325	0.488
10	3.276	0.494	3.319	0.505

NOTE: We obtained asymptotic standard errors by applying the delta method to the asymptotic variance matrix of our estimate of $\boldsymbol{\theta}$. Note that the birth rate λ decreases much more rapidly than death rate μ as antibody concentration increases.

3.4 Application to Microsatellite Evolution

Microsatellites are short tandem repeats of characters in a DNA sequence (Schlötterer 2000; Ellegren 2004; Richard, Kerrest, and Dujon 2008). The number of repeated "motifs" in a microsatellite often changes over evolutionary timescales. The molecular mechanism responsible for changes in repeat numbers is known as "polymerase slippage" (Schlötterer 2000). Several researchers have proposed linear BDPs for use in analyzing evolution of microsatellite repeat numbers (Calabrese and Durrett 2003; Whittaker et al. 2003; Sainudiin et al. 2004). However, many investigations demonstrate that microsatellite mutability depends on the number of repeats already present, motif size, and motif nucleotide composition (Chakraborty et al. 1997; Kelkar et al. 2008; Eckert and Hile 2009; Amos 2010). Exactly how these factors affect addition and deletion rates remains an open question (Bhargava and Fuentes 2010). To our knowledge, no previous study formulates or fits a general BDP in which motif size and composition are treated as a covariates in a generalized regression framework, despite the scientific interest in examining such effects on microsatellite evolution.

Webster, Smith, and Ellegren (2002) studied the evolution of 2467 microsatellites common (orthologous) to both humans and chimpanzees, providing an ideal dataset for studying the influence of repeat number and motif size on addition and deletion rates. For each of these observed microsatellites, Webster, Smith, and Ellegren (2002) recorded the motif nucleotide pattern and the number of repeats of this motif found in chimpanzees and humans, and estimate a mutability parameter that controls the rate of addition and deletion. We now apply our BDP inference technique to chimpanzee-human microsatellite evolution, drawing on the data in Table S6 of the supplementary information in Webster, Smith, and Ellegren (2002). We introduce several novel modeling and inferential techniques relevant to the study of microsatellites, and deduce the effect of motif size and composition on microsatellite addition and deletion rates. While the likelihood takes a slightly more complicated form, our BDP regression technique is straightforward to implement, yielding insight into the complicated process of microsatellite evolution.

To analyze the data as realizations from a BDP, we must acknowledge the evolutionary relationship between chimpanzees

and humans. Suppose the most recent common ancestor of chimpanzees and humans lived at time t in the past, so that an evolutionary time of $2t$ separates contemporary humans and chimpanzees. We note that under mild conditions, general BDPs are reversible Markov chains (Renshaw 2011). Therefore, assuming stationarity of the chimpanzee microsatellite length distributions, we stand justified in reversing the evolutionary process from the ancestor to chimpanzee, so that for estimation purposes we may regard humans as direct descendants of modern chimpanzees (or vice-versa) over an evolutionary time of $2t$. If c is the number of repeats in a chimpanzee microsatellite and h is the number of repeats in the corresponding human microsatellite, then the likelihood of the observation $\mathbf{Y} = (c, h, t)$ is

$$\begin{aligned}\Pr(\mathbf{Y}) &= \sum_{k=0}^{\infty} \pi_k P_{kc}(t)\ P_{kh}(t) = \pi_c \sum_{k=0}^{\infty} P_{ck}(t)\ P_{kh}(t) \\ &= \pi_c P_{ch}(2t), \end{aligned} \tag{35}$$

where π_k is the equilibrium probability of the microsatellite having k repeats. The second equality follows by reversibility and the third by the Chapman–Kolmogorov equality. Therefore, the log-likelihood of the observation $\mathbf{Y}$ is now $\log \pi_c(\boldsymbol{\theta}) + \ell(\boldsymbol{\theta}; \mathbf{Y})$. Figure 4 shows a schematic representation of this reversibility argument.

The observed data for microsatellite i are $\mathbf{Y}_i = (X_i(0), X_i(1), 1)$, where $X_i(0)$ is the number of repeats observed in chimpanzees, $X_i(1)$ is the number of repeats observed in humans, and the evolutionary time separating humans and chimpanzees is scaled to unity. In addition to the evolutionary relationship explained above, there are other complications: in the Webster, Smith, and Ellegren (2002) dataset, it is evident that microsatellites with small numbers of repeats are not detected. Rose and Falush (1998) argued that there is a minimum number of repeats necessary for microsatellite mutation via polymerase slippage. Sainudiin et al. (2004) interpreted this finding as justification for truncating the state-space of BDP at $x_{\min}$, so that $X(\tau) \geq x_{\min}$. To avoid questions of ascertainment bias (see, e.g., Vowles and Amos 2006), and to make our results comparable to those of past researchers, we *define* a microsatellite to be a collection of more than $x_{\min}$ repeated motifs, where $x_{\min}$ is 9 for repeats of size 1, 5 for repeats of size 3 and 4, and 2 for repeats of size 5. Researchers have also observed that microsatellites do not tend to grow indefinitely (Kruglyak et al. 1998). The maximum number of repeats in the Webster, Smith, and Ellegren (2002) dataset is 47. This suggests a finite nonzero equilibrium distribution of microsatellite lengths. To achieve such an equilibrium distribution, we preliminarily view the evolution as a linear BDP with immigration on a state-space that is truncated below $x_{\min}$. It is reasonable to assume that rates of addition and deletion depend linearly on how many repeats are already present. Then for a microsatellite that currently has k repeats, the birth and death rates are

$$\lambda_k = \begin{cases} k\lambda + \lambda & k \geq x_{\min} \\ 0 & k < x_{\min} \end{cases} \quad \text{and} \quad \mu_k = \begin{cases} k\mu & k > x_{\min} \\ 0 & k \leq x_{\min}. \end{cases} \tag{36}$$

This gives a geometric equilibrium distribution for the number of repeats:

$$\pi_k = \begin{cases} \left(1 - \dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^{k - x_{\min} - 1} & k \geq x_{\min} \\ 0 & k < x_{\min}, \end{cases} \tag{37}$$

when $\lambda < \mu$ (Renshaw 2011). We choose this simple model so that the BDP has a simple closed-form nonzero equilibrium solution that is easy to incorporate into the log-likelihood. Note that the constraint $\lambda < \mu$ does not mean that the rate of microsatellite repeat addition is always less than the rate of deletion, since it is possible that $\lambda_k > \mu_k$ for small k. Additionally, $\lambda < \mu$ does not mean that the number of repeats in a microsatellite tends to zero over long evolutionary times—the equilibrium distribution (37) assigns positive probability to all repeat numbers greater than or equal to $x_{\min}$. Now we augment the log-likelihood with the log-equilibrium probability of observing $X_i(0)$ chimpanzee repeats

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log[\pi_{X_i(0)}(\boldsymbol{\theta})] + \ell(\boldsymbol{\theta}; \mathbf{Y}_i), \tag{38}$$

where $\ell(\boldsymbol{\theta}; \mathbf{Y}_i)$ is equivalent to (11). Including the influence of the equilibrium distribution is similar to imposing a prior distribution on λ and μ.

To incorporate and evaluate the influence of motif size and composition heterogeneity, we now treat λ and μ in the ith observation as functions of the covariate vector $\mathbf{z}_i$ in a general BDP. Suppose microsatellite i has motif size r_i. We code the vectors $\mathbf{z}_i$ as follows:

$$\mathbf{z}_i = \begin{cases} (0, 0, p_a, p_c, p_t)' & r_i = 1 \\ (1, 0, p_a, p_c, p_t)' & r_i = 2 \\ (1, 1, p_a, p_c, p_t)' & r_i \geq 3, \end{cases} \tag{39}$$

where p_x is the proportion of x nucleotides per repeat. We define a single parameter α that controls the difference between λ and μ. Then in the ith microsatellite, the complete model

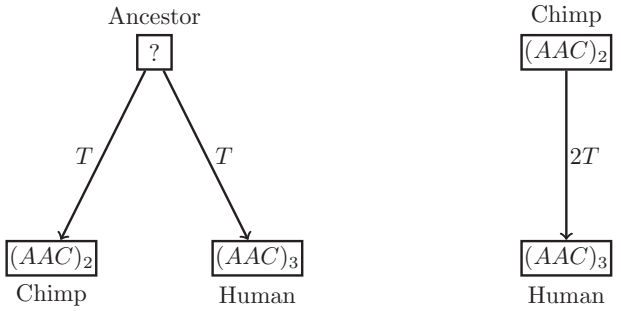

Figure 4. Reversibility of the BDP implies that the evolutionary relationship between contemporary chimpanzees and the most recent common ancestor can be inverted. On the left, the most recent common ancestor of chimpanzees and humans lived at time T in the past. At a certain locus, chimpanzees have a microsatellite consisting of two repeats of the motif AAC, and at an orthologous locus, humans have theee repeats of the motif. The number of repeats in the ancestor is unknown. On the right, using a probabilistic justification explained in the text, we may interpret the evolutionary relationship between chimpanzees and humans as unidirectional, while "integrating out" the number of repeats at the ancestral locus.

Table 5. Maximum likelihood estimates of parameters in the microsatellite evolution model and their asymptotic standard errors

Parameter	Covariate	Estimate	SE
α	birth	−0.132	0.007
$\boldsymbol{\theta}_1$	$r_i = 2$	0.063	0.107
$\boldsymbol{\theta}_2$	$r_i \geq 3$	−1.390	0.127
$\boldsymbol{\theta}_3$	p_a	0.224	0.261
$\boldsymbol{\theta}_4$	p_c	−1.510	0.370
$\boldsymbol{\theta}_5$	p_t	−0.355	0.054

NOTE: The first two elements of $\boldsymbol{\theta}$ correspond to the motif size r_i, and the last three correspond to the motif nucleotide composition. The parameter α controls the difference between the birth and death rates. The ith microsatellite birth rate is then $\lambda = \exp(\alpha + \mathbf{z}_i'\boldsymbol{\theta})$ and the death rate is $\mu = \exp(\mathbf{z}_i'\boldsymbol{\theta})$. Estimated birth and death rates are higher for dinucleotide repeats than for mononucleotide repeats or microsatellites whose motifs have 3, 4, or 5 nucleotides. Microsatellites whose motif consists, for example, of A nucleotides have higher birth and death rates compared to C and T nucleotides.

becomes

$$\log(\lambda_{k,i}) = \log(k+1) + \alpha + \mathbf{z}_i'\boldsymbol{\theta} \quad \text{and}$$
$$\log(\mu_{k,i}) = \log(k) + \mathbf{z}_i'\boldsymbol{\theta}. \tag{40}$$

Therefore $(\alpha, \boldsymbol{\theta})'$ is the 6×1 vector of unknown parameters. Putting all this together, the surrogate function becomes

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\right) \propto \left(\sum_{i=1}^{N}(X_i(0) - x_{\min,i} - 1)\alpha + \log\left(1 - e^{\alpha}\right) + \left[\sum_{k=0}^{\infty}\mathbb{E}(U_k|\mathbf{Y}_i)(\alpha + \mathbf{z}_i'\boldsymbol{\theta}) + \mathbb{E}(D_k|\mathbf{Y}_i)\mathbf{z}_i'\boldsymbol{\theta} - \mathbb{E}(T_k|\mathbf{Y}_i)\left((k+1)e^{\alpha+\mathbf{z}_i'\boldsymbol{\theta}} + ke^{\mathbf{z}_i'\boldsymbol{\theta}}\right)\right]\right) \tag{41}$$

where $\alpha < 0$ since $\lambda < \mu$. We use a gradient EM algorithm to find the MLE of $(\alpha, \boldsymbol{\theta})$.

Table 5 reports the parameter estimates, along with asymptotic standard errors. From these results, we infer that motifs of different sizes and composition have different characteristics under our evolutionary model. As an example, a microsatellite consisting of AAC repeats has $z = (1, 1, 0.667, 0.333, 0)$, $\lambda = 0.163$ (0.082) and $\mu = 0.186$ (0.087) where the standard errors obtained by the delta method are given in parenthesis. Specifically, λ and μ are greatest for dinucleotide repeats, as compared to motifs with one or at least three repeats. Motifs consisting mostly of A nucleotides also give rise to higher λ and μ than those consisting of C and T nucleotides. These conclusions are largely consistent with the descriptive results obtained by Webster, Smith, and Ellegren (2002). Our analysis also provides a natural probabilistic justification for the existence of a finite nonzero equilibrium distribution of microsatellite repeat numbers and a formal statistical framework for deducing the effect of motif size and repeat number on mutation rates.

4. DISCUSSION

Application of stochastic models in statistics requires a flexible and general approach to parameter estimation, without which even the most realistic model becomes unappealing to researchers who wish to learn from the data they have collected. Estimation for continuously observed BDPs is straightforward and well-established. For partially observed BDPs, our approach is unique because it requires only two simple ingredients: the functional form of the birth and death rates $\lambda_k(\boldsymbol{\theta})$ and $\mu_k(\boldsymbol{\theta})$ for all k, and an exact or approximate M-step. A third ingredient is optional: the Hessian of the surrogate function is useful when asymptotic standard errors are desired. However, this matrix can often be approximated numerically upon convergence of the EM algorithm, since the observed-data likelihood is available numerically via (33). With these ingredients in hand, even elusive general BDPs become tractable.

In previous work on estimation for BDPs, completion of the E-step typically relies on rate matrix truncation, time-domain numerical integration or simulation of BDP trajectories. As we show in Table 1, both rejection sampling and endpoint-conditioned simulation can occasionally perform satisfactorily, especially in comparison to time-domain convolution. However, endpoint-conditioning is designed for finite state-space Markov chains, and it relies on a rate matrix eigendecomposition to calculate transition probabilities. In the logistic and SIS models, this matrix can become nearly singular, causing the both simulation and matrix exponentiation methods to fail, even when we choose parameter values that are not biologically unreasonable. Matrix-algebraic approaches provide powerful methods for computing conditional expectations in finite state-space processes, but have serious drawbacks when used to approximate processes on infinite state-spaces. The Laplace convolution in the E-step of our algorithm for BDPs is more generic and flexible than matrix methods, and provides equivalent or better performance. This is partly due to the dynamic nature of the continued fraction evaluation—we can descend in the fraction to a depth that achieves acceptable truncation error without needing to specify this depth a priori. In addition, continued fraction evaluation is numerically more stable than eigendecomposition or uniformization operations on truncated rate matrices. For this reason, a variation on our Laplace convolution method for computing the E-step may offer further use in estimation for non-BDP finite Markov chains as well, such as nucleotide or codon substitution models. For some linear BDPs, the availability of a generating function furnishes analytic E- and M-steps yielding very fast parameter updates in closed-form (Doss et al. 2013). For some models, these tools provide the asymptotic variance of the MLE in closed-form. However, for the majority of BDPs, we must return to the Laplace convolution method outlined in this article.

If one cannot find analytic parameter updates in the M-step, several options remain available. With a minorizing function as in Section 2.4.2, an EM-MM algorithm is viable. Further, one or more numerical Newton steps offers an alternative, as in Sections 2.4.3 and 2.4.5. One may employ other gradient-based methods as well. Although the MM update derived for the BDP with immigration (Section 2.4.2) is appealing in its simplicity, multiple minorizations of the likelihood can result in very slow convergence, since the surrogate function lies far from the true likelihood for most values of $\boldsymbol{\theta}$. In addition, Newton steps that require matrix inversion may suffer since the Hessian of the surrogate can become ill-conditioned.

Even with the substantial speedup offered by our Laplace convolution method for performing the E-step and quasi-Newton acceleration of the EM iterates, our algorithms can move slowly toward the MLE. Here, naïve numerical optimization of the

incomplete data likelihood can sometimes run computationally faster. However, such techniques perform very poorly when the number of parameters increases and they often require specification of tuning constants to reach the global optimum. For BDP estimation problems, EM algorithms offer several other advantages over naïve numerical optimization, and these benefits are especially stark when the M-step is available in closed-form. First, when the log-likelihood is locally convex, the EM algorithm is robust with respect to the initial parameter values near the maximum, and EM algorithms generally do not need tuning parameters. Further, the ascent property ensures the iterates will approach a maximum. Perhaps the most important reason to consider EM algorithms is that they can accommodate higher-dimensional parameter spaces without substantially increasing the computational complexity of the algorithm. This is especially useful in models with many unknown parameters when performing regression with covariates (Section 2.4.5), or our microsatellite example. Maximum likelihood approaches to regression can suffer in high-dimensional regression problems, and the resulting estimates of the information matrix can be inaccurate (DasGupta and Lahiri 2012). One way to alleviate to this problem is regularization of parameter estimates via L_1 or L_2 penalties, which may be interpreted equivalently as maximum a priori estimation in a Bayesian context. Penalties or prior distributions can easily be accommodated in our BDP regression formulation via modification of the M-step (Green 1990). We also note the potential for substantial computational speedup by parallelizing the E-step. When discrete observations from a BDP are independent, the E-step may be performed in parallel for every observation. For example, $\mathbb{E}(U|\mathbf{Y}_i)$ can be computed simultaneously for $i = 1, \ldots, N$. When speed is an issue, graphics processing units may prove useful in reducing the computational cost of EM algorithms (Zhou, Lange, and Suchard 2010).

With regard to our examples, we present a novel way of studying the dynamics of count data in laboratory experiments and the evolution of microsatellite repeats using a GLM. Previous efforts often ignore the branching nature of the underlying process, use incomplete or equilibrium models of counts, or fit separate models for experiments or observations of different types. In our lymphoma analysis, we use a realistic simple linear birth-death model with covariates to discern the relationship of antibody concentration to per-cell birth and death rates. This results in parameter estimates that have a natural biological interpretation. In our microsatellite application, we treat motif size as a categorical variable and incorporate motif nucleotide composition, allowing us to fit a single regression model to all the microsatellite observations simultaneously. Although our microsatellite example is limited in scope, it is easy to imagine a more comprehensive study. For example, incorporating more sophisticated motif nucleotide composition covariates and location of the microsatellite on the chromosome might provide additional insight into the evolutionary process. Our EM framework is nearly ideal for these types of studies, since the number of unknown parameters does not substantially increase the computational burden of the M-step, and the E-step is completely unaffected by the number of parameters.

Interestingly, we attempted to use the generic nonlinear regression R function `nlm` to validate the MLEs obtained by our EM algorithm for the microsatellite evolution problem, starting at a variety of initial values, including the MLE found by our EM algorithm. This naïve optimizer failed to converge in every case. We speculate that this is because the small numerical errors in the likelihood evaluation have similar order of magnitude as the curvature of the likelihood function near the maximum. Our EM algorithms take advantage of analytic derivatives of the surrogate function instead of the likelihood, and hence are less susceptible to small errors in the numerical gradient.

5. SOFTWARE

The R package `birth.death`, available at *http://crawford.research.yale.edu/software* provides functions to calculate transition probabilities and conditional expectations used in this article.

APPENDIX A: CONTROL OF NUMERICAL ERROR IN THE E-STEP

Completing the E-step in our EM algorithms requires several levels of numerical approximation. Fortunately, the error in these computations can be controlled dynamically to achieve the necessary numerical accuracy under fairly general conditions. In this Appendix, we derive error bounds for transition probabilities $P_{ab}(t)$ and conditional expectations $\mathbb{E}(U_k|\mathbf{Y})$ and $\mathbb{E}(U|\mathbf{Y}) = \sum_k \mathbb{E}(U_k|\mathbf{Y})$ where $\mathbf{Y} = (a, b, t)$, and show how the overall error can be controlled to provide accurate calculations for use in the EM algorithms outlined in this article. The bounds for conditional expectations of T_k and D_k are essentially the same. Throughout this Appendix, we assume that the BDP is nonexplosive and that the expectations of the sufficient statistics are finite. We begin by stating several results that will be useful in deriving overall error bounds.

In practice, infinite continued fractions such as (5) can only be evaluated computationally to finite depth M. For such continued fractions, we have the following truncation bound.

Lemma 1. Without loss of generality, suppose $a = 0$ and $b = 0$, and $f_{00}^{(M)}(s) = A_M(s)/B_M(s)$ converges to $f_{00}(s)$ as $M \to \infty$. Then

$$\left|f_{00}(s) - f_{00}^{(M)}(s)\right| \le \frac{\left|\frac{B_M(s)}{B_{M-1}(s)}\right|}{\left|\mathrm{Im}\left(\frac{B_M(s)}{B_{M-1}(s)}\right)\right|}\left|f_{00}^{(M)}(s) - f_{00}^{(M-1)}(s)\right|, \quad \text{(A.1)}$$

when the denominator is nonzero (Craviotto, Jones, and Thron 1993).

Suppose we have a continuous real-valued function $g(t)$ with Laplace transform $G(s)$ and let

$$s_0 = \frac{A}{2t} \quad \text{and} \quad s_j = \frac{A + 2j\pi i}{2t} \quad \text{(A.2)}$$

for $j \ge 1$, where $i = \sqrt{-1}$. The general inversion formula we will use is

$$\tilde{g}(t) = \frac{e^{A/2}}{2t}\mathrm{Re}(G(s_0)) + \frac{e^{A/2}}{t}\sum_{j=1}^{\infty}(-1)^j \mathrm{Re}(G(s_j)), \quad \text{(A.3)}$$

where $\mathrm{Re}(s)$ is the real part of the complex variable s and A is a positive tuning constant that we will set to control the error. When there is no error in the evaluation of the Laplace transform $G(s)$, we have the following bound for the discretization error.

Lemma 2. The discretization error in (A.3) is

$$|g(t) - \tilde{g}(t)| \le \sum_{j=1}^{\infty} e^{-jA} g((2j+1)t) \le \sum_{j=1}^{\infty} e^{-jA} = \frac{e^{-A}}{1-e^{-A}} \tag{A.4}$$

since $g(t) \le 1$ (Abate and Whitt 1992a).

When computing infinite sums, acceleration methods can be useful if the series is slow to converge. When the terms in the summand of (A.3) are alternating in sign and rapidly decreasing in magnitude, a reasonable estimate of the remainder ω_J after J terms in (A.3) is the first term in the tail sum (Levin 1973; Weiger 1989; Abate and Whitt 1992a; Press 2007). We assume that there exists a J large enough that

$$|\omega_J| = \left| \sum_{j=J+1}^{\infty} (-1)^j \mathrm{Re}(g(s_j)) \right| \le |\mathrm{Re}(g(s_J))|. \tag{A.5}$$

We now analyze the error that arises when evaluating BDP likelihoods and expectations in the EM algorithms developed in this article.

A.1 Transition Probability Error

To find the transition probability $P_{ab}(t)$, we set $G(s) = f_{ab}(s)$ in (A.3). We can only evaluate the infinite continued fraction $f_{ab}(s)$ to a finite depth, so we approximate $f_{ab}(s_j)$ by the M_jth convergent $f_{ab}^{(M_j)}(s_j)$. Here, M_j is a positive integer chosen dynamically so that the error due to truncation is $|f_{ab}(s_j) - f_{ab}^{(M_j)}(s_j)| \le \epsilon$ using Lemma 1, where we have selected $\epsilon > 0$ in advance. Let

$$\widetilde{P}_{ab}(t) = \frac{e^{A/2}}{2t} \mathrm{Re}\,(f_{ab}(s_0)) + \frac{e^{A/2}}{t} \sum_{j=1}^{\infty} (-1)^j \mathrm{Re}(f_{ab}(s_j)) \tag{A.6}$$

be the discretized Laplace inversion (A.3) computed using the *infinite* continued fraction $f_{ab}(s)$. The discretization error is

$$|P_{ab}(t) - \widetilde{P}_{ab}(t)| \le \sum_{j=1}^{\infty} e^{-jA} P_{ab}((2j+1)t) \le \frac{e^{-A}}{1-e^{-A}} \tag{A.7}$$

by Lemma 2. Let

$$\widehat{P}_{ab}(t) = \frac{e^{A/2}}{2t} \mathrm{Re}\big(f_{ab}^{(M_0)}(s_0)\big) + \frac{e^{A/2}}{t} \sum_{j=1}^{J} (-1)^j \mathrm{Re}\big(f_{ab}^{(M_j)}(s_j)\big) \tag{A.8}$$

be the inversion sum computed using the continued fraction truncated at depth M_j in the jth term in the sum. The infinite sum in (A.6) has been replaced by a J-term sum in (A.8), where the maximum summation index J is also chosen dynamically based on an estimate of the tail sum (A.5). Now we consider the error due to truncation of the infinite continued fraction and termination of the infinite sum after J terms, where J is chosen so that the remainder $|\omega_J| \le \delta$.

$$\begin{aligned} |\widetilde{P}_{ab}(t) - \widehat{P}_{ab}(t)| &\le \frac{e^{A/2}}{t} \left[\left| \frac{1}{2} \mathrm{Re}\big(f_{ab}(s_0) - f_{ab}^{(M_0)}(s_0)\big) \right| \right. \\ &\quad + \left| \sum_{j=1}^{J} (-1)^j \mathrm{Re}\big(f_{ab}(s_j) - f_{ab}^{(M_j)}(s_j)\big) \right| \\ &\quad \left. + \left| \sum_{j=J+1}^{\infty} (-1)^j \mathrm{Re}(f_{ab}(s_j)) \right| \right] \\ &\le \frac{e^{A/2}}{t} \left[\frac{1}{2}\epsilon + J\epsilon + \omega_J \right] \le \frac{e^{A/2}}{t} \left[\left(\frac{1}{2} + J \right) \epsilon + \delta \right] \end{aligned} \tag{A.9}$$

by Lemma 2 and (A.5). Both types of truncation occur dynamically: the continued fraction in the jth term in the sum is terminated at depth M_j when the error given by Lemma 1 is less than ϵ; likewise, truncation of the infinite sum happens when the estimated tail sum remainder ω_J is smaller than δ. Putting these bounds together, we find that the overall error is, by the triangle inequality,

$$\begin{aligned} |P_{ab}(t) - \widehat{P}_{ab}(t)| &\le |P_{ab}(t) - \widetilde{P}_{ab}(t)| + |\widetilde{P}_{ab}(t) - \widehat{P_{abab}}(t)| \\ &\le \frac{e^{-A}}{1-e^{-A}} + \frac{e^{A/2}}{t} \left[\left(\frac{1}{2} + J \right) \epsilon + \delta \right]. \end{aligned} \tag{A.10}$$

Following Abate and Whitt (1995), a simple way to choose the constant A is to approximate $e^{-A}/(1-e^{-A}) \approx e^{-A}$ and put $\epsilon = \delta = e^{-3A/2}$, resulting in

$$e^{-A} \left[1 + \frac{1}{t} \left(\frac{3}{2} + J \right) \right]. \tag{A.11}$$

Then to achieve an error at most $10^{-\gamma}$, set $A = \log[10^{\gamma}(1 + (3/2 + J)/t]$. To provide a rough bound, set $t = 1$ and $\gamma = 8$. Since the truncation index is determined dynamically and J is not usually known in advance, we specify $J = 100$, giving $A = 23$ as a conservative choice of the error tuning constant.

A.2 Error in Computation of $\mathbb{E}(U_k|\mathbf{Y})$

Recall from (14a) that the numerator of $\mathbb{E}(U_k|\mathbf{Y})$ is a convolution of transition probabilities. Let

$$g_k(t) = \int_0^t P_{ak}(u)\, P_{k+1,b}(t-u)\, \mathrm{d}u \tag{A.12}$$

be the time-domain convolution integral and let $G_k(s) = f_{ak}(s) f_{k+1,b}(s)$ be its Laplace transform. Let $\tilde{g}_k(t)$ be given by (A.3) and let $\hat{g}_k(t)$ be the same quantity but with truncation of the infinite sum (A.3) at the Jth term. Fix a small error tolerance $\epsilon > 0$ and suppose that for each j, we evaluate the continued fractions $f_{ak}(s_j)$ and $f_{k+1,b}(s_j)$ to depths M_j and N_j, respectively, so that the truncation error in the difference of these convergent products is less than ϵ,

$$\left| f_{ak}(s_j) f_{k+1,b}(s_j) - f_{ak}^{(M_j)}(s_j) f_{k+1,b}^{(N_j)}(s_j) \right| \le \epsilon \tag{A.13}$$

using Lemma 1. Using Lemma 2, the discretization error in (A.3) is approximated by

$$\begin{aligned} |g_k(t) - \tilde{g}_k(t)| &= \sum_{j=1}^{\infty} e^{-jA} \int_0^{(2j+1)t} P_{ak}(u)\, P_{k+1,b}((2j+1)t - u)\, \mathrm{d}u \\ &\le \sum_{j=1}^{\infty} e^{-jA} (2j+1)t \\ &= t \frac{3e^A - 1}{(e^A - 1)^2} \end{aligned} \tag{A.14}$$

since the integrand is less than one. The error due to truncation of the continued fractions and infinite sum, analogous to (A.9), becomes

$$\begin{aligned}
&|\widetilde{g}_k(t) - \hat{g}_k(t)| \\
&= \frac{e^{A/2}}{t}\Bigg[\left|\frac{1}{2}\mathrm{Re}\big(f_{ak}(s_0)f_{k+1,b}(s_0) - f_{ak}^{(M_0)}(s_0)f_{k+1,b}^{(N_0)}(s_0)\big)\right| \\
&\quad + \left|\sum_{j=1}^{J}(-1)^j\mathrm{Re}\big(f_{ak}(s_j)f_{k+1,b}(s_j) - f_{ab}^{(M_j)}(s_j)f_{k+1,b}^{(N_j)}(s_j)\big)\right| \\
&\quad + \left|\sum_{j=J+1}^{\infty}(-1)^j\mathrm{Re}\big(f_{ak}(s_j)f_{k+1,b}(s_j)\big)\right|\Bigg] \\
&\le \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right], \qquad \text{(A.15)}
\end{aligned}$$

where J is chosen so that the remainder $|\omega_J| \le \delta$. Putting these together, we have the overall numerator error

$$\begin{aligned}
|g_k(t) - \hat{g}_k(t)| &\le |g_k(t) - \tilde{g}_k(t)| + |\tilde{g}_k(t) - \hat{g}_k(t)| \\
&\le t\frac{3e^A-1}{(e^A-1)^2} + \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right]. \qquad \text{(A.16)}
\end{aligned}$$

Now, recall that the quantity we wish to compute is $\mathbb{E}(U_k|\mathbf{Y}) = \lambda_k g_k(t)/P_{ab}(t)$, where the numerator and denominator are evaluated separately. First we seek a lower bound for $P_{ab}(t)$ in terms of $\widehat{P}_{ab}(t)$ and A. Let

$$\chi = \frac{e^{-A}}{1-e^{-A}} + \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right] \qquad \text{(A.17)}$$

be the error bound for $|P_{ab}(t) - \widehat{P}_{ab}(t)|$ from (A.10). We assume that $\widehat{P}_{ab}(t) - \chi > 0$ so that $\widehat{P}_{ab}(t) - \chi$ is a lower bound for $P_{ab}(t)$, that is

$$P_{ab}(t) \ge \widehat{P}_{ab} - \chi > 0. \qquad \text{(A.18)}$$

Then the error in the ratio is given by

$$\begin{aligned}
&\left|\frac{g_k(t)}{P_{ab}(t)} - \frac{\hat{g}_k(t)}{\widehat{P}_{ab}(t)}\right| \\
&= \left|\frac{g_k(t)\widehat{P}_{ab}(t) - \hat{g}_k(t)P_{ab}(t)}{P_{ab}(t)\widehat{P}_{ab}(t)}\right| \\
&= \left|\frac{g_k(t)\widehat{P}_{ab}(t) - g_k(t)P_{ab}(t) + g_k(t)P_{ab}(t) - \hat{g}_k(t)P_{ab}(t)}{P_{ab}(t)\widehat{P}_{ab}(t)}\right| \\
&\le \frac{g_k(t)}{P_{ab}(t)\widehat{P}_{ab}(t)}|\widehat{P}_{ab}(t) - P_{ab}(t)| + \frac{1}{\widehat{P}_{ab}(t)}|g_k(t) - \hat{g}_k(t)| \\
&\le \frac{1}{\widehat{P}_{ab}(t)}\left(t\frac{|\widehat{P}_{ab}(t) - P_{ab}(t)|}{P_{ab}(t)} + |g_k(t) - \hat{g}_k(t)|\right) \\
&\le \frac{1}{\widehat{P}_{ab}(t)}\left(\frac{t\chi}{\widehat{P}_{ab}(t) - \chi} + |g_k(t) - \hat{g}_k(t)|\right). \qquad \text{(A.19)}
\end{aligned}$$

Now using the transition probability error (A.10), the lower bound for $P_{ab}(t)$ (A.18), and the error for $g_k(t)$ (A.16), we find that

$$\begin{aligned}
\left|\frac{g_k(t)}{P_{ab}(t)} - \frac{\hat{g}_k(t)}{\widehat{P}_{ab}(t)}\right| \le \frac{1}{\widehat{P}_{ab}(t)}\Bigg(&\frac{t\left[\frac{e^{-A}}{1-e^{-A}} + \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right]\right]}{\widehat{P}_{ab}(t) - \frac{e^{-A}}{1-e^{-A}} - \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right]} \\
&+ t\frac{3e^A-1}{(e^A-1)^2} + \frac{e^{A/2}}{t}\left[\left(\frac{1}{2}+J\right)\epsilon+\delta\right]\Bigg). \qquad \text{(A.20)}
\end{aligned}$$

To find the tuning constant A that keeps the error less than 10^{-8}, we again approximate $e^{-A} \approx e^{-A}/(1-e^{-A})$ and put $\delta = \epsilon = e^{-3A/2}$, $J = 100$, and $\widehat{P}_{ab}(t) = 1/2$ to obtain $A = 25$.

A.3 Error in Computation of $\mathbb{E}(U|\mathbf{Y})$

We use a slightly different approach for $\mathbb{E}(U|\mathbf{Y}) = \sum_{k=0}^{\infty}\mathbb{E}(U_k|\mathbf{Y})$ because the expectation itself incorporates an infinite sum that can only be evaluated to finitely many terms in practice. Let

$$g(t) = \int_0^t \sum_{k=0}^{\infty}\lambda_k P_{ak}(u)\, P_{k+1,b}(t-u)\,\mathrm{d}u \qquad \text{(A.21)}$$

be the infinite sum of time-domain convolutions from (32) and let

$$G(s) = \sum_{k=0}^{\infty}\lambda_k f_{ak}(s)f_{k+1,b}(s) \qquad \text{(A.22)}$$

be its Laplace transform. Recall from (A.13) that we can choose continued fraction truncation depths M_{jk} and N_{jk} so that

$$\left|f_{ak}(s_j)f_{k+1,b}(s_j) - f_{ak}^{(M_{jk})}(s_j)f_{k+1,b}^{(N_{jk})}(s_j)\right| \le \epsilon \qquad \text{(A.23)}$$

for every j and k. First, note that the product $P_{ak}(u)P_{k+1,b}(t-u) \le 1$ for all k. To proceed, we must make two weak assumptions about the decay of the transition probabilities and the growth of the birth rates λ_k. Fix a number $\eta > 0$ and $C > \max(a, b)$ such that

$$P_{ak}(u)P_{k+1,b}(t-u) \le e^{-\eta k}$$

and

$$\widetilde{P}_{ak}(u)\widetilde{P}_{k+1,b}(t-u) \le e^{-\eta k} \qquad \text{(A.24)}$$

for $k > C$. Assume also that there exists $\Lambda > 0$ such that $\lambda_k \le \Lambda k^2$ for $k \ge C$, meaning that the birth rates do not grow too rapidly. These assumptions are intuitively reasonable and agree with our empirical observations of the decay of (A.24) with k. Indeed, a process whose birth rates increase faster than $O(k^2)$ for large k is likely to be explosive. These assumptions imply a bound for $g(t)$:

$$\begin{aligned}
\sum_{k=0}^{\infty}\lambda_k g_k(t) &\le \left(\sum_{k=1}^{C-1}\lambda_k\right) + \sum_{k=C}^{\infty}\lambda_k g_k(t) \\
&\le \left(\sum_{k=0}^{C-1}\lambda_k\right) + \sum_{k=0}^{\infty}\lambda_{k+C}e^{-\eta(k+C)} \\
&\le \left(\sum_{k=0}^{C-1}\lambda_k\right) + \Lambda e^{-\eta C}\sum_{k=0}^{\infty}(k+C)^2e^{-\eta k} \\
&\le \left(\sum_{k=0}^{C-1}\lambda_k\right) + \Lambda e^{-\eta C}\xi_C, \qquad \text{(A.25)}
\end{aligned}$$

where we have used (A.24), $\lambda_k \le \Lambda k^2$ for $k \ge C$, and

$$\begin{aligned}
\xi_C &= \sum_{k=0}^{\infty}(C+k)^2e^{-\eta k} \\
&= \frac{-2C^2e^{\eta} + C^2e^{2\eta} + 2Ce^{\eta} + e^{\eta} + C^2 - 2C + 1}{(e^{\eta}-1)^3}. \qquad \text{(A.26)}
\end{aligned}$$

Note that (A.25) holds for $\sum_k \lambda_k \tilde{g}(t)$ as well. Since the birth rates λ_k are known in advance, Λ can be readily determined or a tighter bound than Λk^2 can be found.

The discretization error for the numerator is given by

$$\begin{aligned}
&|g(t) - \tilde{g}(t)| \\
&\le \sum_{j=1}^{\infty}e^{-jA}\int_0^{(2j+1)t}\sum_{k=0}^{\infty}\lambda_k P_{ak}(u)P_{k+1,b}((2j+1)t-u)\,\mathrm{d}u \\
&\le \sum_{j=1}^{\infty}e^{-jA}\int_0^{(2j+1)t}\left[\left(\sum_{k=0}^{C-1}\lambda_k\right) + \sum_{k=C}^{\infty}\lambda_k e^{-\eta k}\right]\mathrm{d}u \\
&\le \left[\left(\sum_{k=0}^{C-1}\lambda_k\right) + e^{-\eta C}\Lambda\xi_C\right]t\frac{3e^A-1}{(e^A-1)^2}. \qquad \text{(A.27)}
\end{aligned}$$

The truncation error for the numerator is

$$
\begin{aligned}
&|\tilde{g}(t) - \hat{g}(t)| \\
&= \frac{e^{A/2}}{t}\left|\frac{1}{2}\mathrm{Re}\left[\sum_{k=0}^{\infty}\lambda_k f_{ak}(s_0)f_{k+1,b}(s_0) - \sum_{k=0}^{D}\lambda_k f_{ak}^{(M_{0k})}(s_0)f_{k+1,b}^{(N_{0k})}(s_0)\right]\right. \\
&\quad + \sum_{j=1}^{J}(-1)^j\mathrm{Re}\left[\sum_{k=0}^{\infty}\lambda_k f_{ak}(s_j)f_{k+1,b}(s_j)\right. \\
&\quad \left. - \sum_{k=0}^{D}\lambda_k f_{ak}^{(M_{jk})}(s_j)f_{k+1,b}^{(N_{jk})}(s_j)\right] \\
&\quad \left. + \sum_{j=J+1}^{\infty}(-1)^j\mathrm{Re}\left[\sum_{k=0}^{\infty}\lambda_k f_{ak}(s_j)f_{k+1,b}(s_j)\right]\right| \\
&= \frac{e^{A/2}}{t}\left[(J+1/2)\left(\sum_{k=0}^{D}\lambda_k\right)\epsilon + \sum_{k=D+1}^{\infty}\lambda_k\tilde{g}_k(t)\right] \\
&\le \frac{e^{A/2}}{t}\left[(J+1/2)\left(\sum_{k=0}^{D}\lambda_k\right)\epsilon + \Lambda e^{-\eta D}\xi_D\right], \qquad \text{(A.28)}
\end{aligned}
$$

where we have truncated the inversion sum at J and the innermost sum at $k = D > C$ so that the remainder estimate $e^{-\eta D}\xi_D$ is small. Putting these together, we find that

$$
\begin{aligned}
|g(t) - \hat{g}(t)| &\le |g(t) - \tilde{g}(t)| + |\tilde{g}(t) - \hat{g}(t)| \\
&\le \left[\left(\sum_{k=0}^{C-1}\lambda_k\right) + e^{-\eta C}\Lambda\xi_C\right] t\frac{3e^A - 1}{(e^A-1)^2} + \frac{e^{A/2}}{t} \\
&\quad \times \left[(J+1/2)\left(\sum_{k=0}^{D}\lambda_k\right)\epsilon + \Lambda e^{-\eta D}\xi_D\right]. \qquad \text{(A.29)}
\end{aligned}
$$

Similar to (A.20), the overall error for the numerator obeys the following inequality:

$$
\begin{aligned}
&\left|\frac{g(t)}{P_{ab}(t)} - \frac{\hat{g}(t)}{\widehat{P}_{ab}(t)}\right| \\
&\le \frac{g(t)}{P_{ab}(t)\widehat{P}_{ab}(t)}|\widehat{P}_{ab}(t) - P_{ab}(t)| + \frac{1}{\widehat{P}_{ab}(t)}|g(t) - \hat{g}(t)| \\
&\le \frac{1}{\widehat{P}_{ab}(t)}\left(\left(\sum_{k=0}^{C-1}\lambda_k + \Lambda e^{-\eta C}\xi_C\right)\frac{|\widehat{P}_{ab}(t) - P_{ab}(t)|}{P_{ab}(t)} + |g(t) - \hat{g}(t)|\right) \\
&\le \frac{1}{\widehat{P}_{ab}(t)}\left(\frac{\left(\sum_{k=0}^{C-1}\lambda_k + \Lambda e^{-\eta C}\xi_C\right)\chi}{\widehat{P}_{ab}(t) - \chi} + |g(t) - \hat{g}(t)|\right), \qquad \text{(A.30)}
\end{aligned}
$$

where χ is given by (A.17) and $|g(t) - \hat{g}(t)|$ is given by (A.29). Then to find the constant A such that we achieve a total error less than 10^{-8}, we again approximate $e^{-A} \approx e^{-A}/(1 - e^{-A})$. As an example, suppose $\lambda_k = 2k$, so $\Lambda = 2$. Setting $C = D = J = 100$, $\widehat{P}_{ab}(t) = 1/2$, and $e^{-\eta C}\xi_C < 1$ we find that a very generous value of A (resulting in a loose bound and more than enough numerical precision) is $A = 34$.

[Received July 2012. Revised October 2013.]

REFERENCES

Abate, J., and Whitt, W. (1992a), "The Fourier-Series Method for Inverting Transforms of Probability Distributions," *Queueing Systems*, 10, 5–87. [743]

——— (1992b), "Numerical Inversion of Probability Generating Functions," *Operations Research Letters*, 12, 245–251. [734]

——— (1995), "Numerical Inversion of Laplace Transforms of Probability Distributions," *ORS A Journal on Computing*, 7, 36–43. [734,743]

——— (1999), "Computing Laplace Transforms for Numerical Inversion via Continued Fractions," *INFORMS Journal on Computing*, 11, 394–405. [732]

Amos, W. (2010), "Mutation Biases and Mutation Rate Variation Around Very Short Human Microsatellites Revealed by Human-Chimpanzee-Orangutan Genomic Sequence Alignments," *Journal of Molecular Evolution*, 71, 192–201. [739]

Andersson, H., and Britton, T. (2000), *Stochastic Epidemic Models and their Statistical Analysis (Lecture Notes in Statistics)*, New York: Springer. [730,735]

Anscombe, F. J. (1953), "Sequential Estimation," *Journal of the Royal Statistical Society*, Series B, 15, 1–29. [731]

Bailey, N. T. J. (1964), *The Elements of Stochastic Processes With Applications to the Natural Sciences*, New York: Wiley. [730,735]

Bankier, J. D., and Leighton, W. (1942), "Numerical Continued Fractions," *American Journal of Mathematics*, 64, 653–668. [732]

Bhargava, A., and Fuentes, F. (2010), "Mutational Dynamics of Microsatellites," *Molecular Biotechnology*, 44, 250–266. [739]

Bladt, M., and Sorensen, M. (2005), "Statistical Inference for Discretely Observed Markov Jump Processes," *Journal of The Royal Statistical Society*, Series B, 67, 395–410. [731,733,737]

Blanch, G. (1964), "Numerical Evaluation of Continued Fractions," *SIAM Review*, 6, 383–421. [732]

Bordes, G., and Roehner, B. (1983), "Application of Stieltjes Theory for S-Fractions to Birth and Death Processes," *Advances in Applied Probability*, 15, 507–530. [730]

Calabrese, P., and Durrett, R. (2003), "Dinucleotide Repeats in the Drosophila and Human Genomes Have Complex, Length-Dependent Mutation Processes," *Molecular Biology and Evolution*, 20, 715–725. [739]

Chakraborty, R., Kimmel, M., Stivers, D., Davison, L., and Deka, R. (1997), "Relative Mutation Rates at di-, tri-, and Tetranucleotide Microsatellite Loci," *Proceedings of the National Academy of Sciences of the United States of America*, 94, 1041–1046. [739]

Cotton, J. A., and Page, R. D. M. (2005), "Rates and Patterns of Gene Duplication and Loss in the Human Genome," *Proceedings of the Royal Society B*, 272, 277–283. [730]

Craviotto, C., Jones, W. B., and Thron, W. J. (1993), "A Survey of Truncation Error Analysis for Padé and Continued Fraction Approximants," *Acta Applicandae Mathematicae*, 33, 211–272. [732,742]

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. [730,732,737]

Cuyt, A., Petersen, V., Verdonk, B., Waadeland, H., and Jones, W. (2008), *Handbook of Continued Fractions for Special Functions*, Berlin/Heidelberg: Springer. [732]

Darwin, J. H. (1956), "The Behaviour of an Estimator for a Simple Birth and Death Process," *Biometrika*, 43, 23–31. [731]

DasGupta, A., and Lahiri, S. (2012), "Density Estimation in High and Ultra High Dimensions, Regularization, and the l1 Asymptotics," in *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman* (vol. 8), D. Fourdrinier, É. Marchand, and A. L. Rukhin, eds., Beachwood, Ohio: Institute of Mathematical Statistics, pp. 1–23. [742]

Dauxois, J. (2004), "Bayesian Inference for Linear Growth Birth and Death Processes," *Journal of Statistical Planning and Inference*, 121, 1–19. [731]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [731,733]

Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N., and Hahn, M. W. (2006), "The Evolution of Mammalian Gene Families," *PLoS ONE*, 1, 1–16. [730]

DeNardo, G. L. (2005), "Concepts in Radioimmunotherapy and Immunotherapy: Radioimmunotherapy From a lym-1 Perspective," *Seminars in Oncology*, 32, 27–35. [738]

Doss, C. R., Suchard, M. A., Holmes, I., Kato-Maeda, M., and Minin, V. N. (2013), "Fitting Birth-Death Processes to Panel Data with Applications to Bacterial DNA Fingerprinting," *The Annals of Applied Statistics*, 4, 2315–2335. [731,733,737,741]

Eckert, K. A., and Hile, S. E. (2009), "Every Microsatellite is Different: Intrinsic DNA Features Dictate Mutagenesis of Common Microsatellites Present in the Human Genome," *Molecular Carcinogenesis*, 48, 379–388. [739]

Ellegren, H. (2004), "Microsatellites: Simple Sequences With Complex Evolution," *Nature Reviews Genetics*, 5, 435–445. [739]

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, New York: Wiley. [730,731]

Flajolet, P., and Guillemin, F. (2000), "The Formal Theory of Birth-and-Death Processes, Lattice Path Combinatorics and Continued Fractions," *Advances in Applied Probability*, 32, 750–778. [730]

Green, P. J. (1990), "On Use of the EM for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society*, Series B, 52, 443–452. [742]

Guillemin, F., and Pinchon, D. (1998), "Continued Fraction Analysis of the Duration of an Excursion in an $M/M/\infty$ System," *Journal of Applied Probability*, 35, 165–183. [730]

——— (1999), "Excursions of Birth and Death Processes, Orthogonal Polynomials, and Continued Fractions," *Journal of Applied Probability*, 36, 752–770. [730]

He, Y., and Liu, C. (2012), "The Dynamic 'Expectation–Conditional Maximization Either' Algorithm," *Journal of the Royal Statistical Society*, Series B, 74, 313–336. [736]

Hobolth, A. (2008), "A Markov Chain Monte Carlo Expectation Maximization Algorithm for Statistical Analysis of DNA Sequence Evolution With Neighbor-Depending Substitution Rates," *Journal of Computational and Graphical Statistics*, 17, 1–25. [733,737]

Hobolth, A., and Jensen, J. (2011), "Summary Statistics for Endpoint-Conditioned Continuous-Time Markov Chains," *Journal of Applied Probability*, 48, 911–924. [731,733,737]

——— (2005), "Statistical Inference in Evolutionary Models of DNA Sequences via the EM Algorithm," *Statistical Applications in Genetics and Molecular*, 4, 1–19. [731,733]

Hobolth, A., and Stone, E. A. (2009), "Simulation From Endpoint-Conditioned, Continuous-Time Markov Chains on a Finite State Space, With Applications to Molecular Evolution," *Annals of Applied Statistics*, 3, 1024–1231. [737]

Holmes, I., and Bruno, W. J. (2001), "Evolutionary HMMs: A Bayesian Approach to Multiple Alignment," *Bioinformatics*, 17, 803–820. [730,731]

Holmes, I., and Rubin, G. (2002), "An Expectation Maximization Algorithm for Training Hidden Substitution Models," *Journal of Molecular Biology*, 317, 753–764. [731,733]

Jamshidian, M., and Jennrich, R. I. (1993), "Conjugate Gradient Acceleration of the EM Algorithm," *Journal of the American Statistical Association*, 88, 221–228. [736]

Jensen, A. (1953), "Markoff Chains as an Aid in the Study of Markoff Processes," *Scandinavian Actuarial Journal*, 1953, 87–91. [737]

Jones, W. B., and Magnus, A. (1977), "Application of Stieltjes Fractions to Birth-Death Processes," in *Padé and Rational Approximation*, eds. E. B. Saff, and R. S. Varga, New York: Academic Press, pp. 173–179. [730]

Kalbfleisch, J. D., and Lawless, J. F. (1985), "The Analysis of Panel Data Under a Markov Assumption," *Journal of the American Statistical Association*, 80, 863–871. [730]

Karlin, S., and McGregor, J. (1957a), "The Classification of Birth and Death Processes," *Transactions of the American Mathematical Society*, 86, 366–400. [730,733]

——— (1957b), "The Differential Equations of Birth-and-Death Processes, and the Stieltjes Moment Problem," *Transactions of the American Mathematical Society*, 85, 589–646. [730,731,733]

Keiding, N. (1974), "Estimation in the Birth Process," *Biometrika*, 61, 71–80. [731]

——— (1975), "Maximum Likelihood Estimation in the Birth-and-Death Process," *The Annals of Statistics*, 3, 363–372. [731]

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., and Makova, K. D. (2008), "The Genome-Wide Determinants of Human and Chimpanzee Microsatellite Evolution," *Genome Research*, 18, 30–38. [739]

Kingman, J. F. C. (1982), "On the Genealogy of Large Populations," *Journal of Applied Probability*, 19, 27–43. [730]

Krone, S. M., and Neuhauser, C. (1997), "Ancestral Processes With Selection," *Theoretical Population Biology*, 51, 210–237. [730]

Kruglyak, S., Durrett, R. T., Schug, M. D., and Aquadro, C. F. (1998), "Equilibrium Distributions of Microsatellite Repeat Length Resulting From a Balance Between Slippage Events and Point Mutations," *Proceedings of the National Academy of Sciences of the United States of America*, 95, 10, 774–10, 778. [740]

Lange, K. (1995a), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of The Royal Statistical Society*, Series B, 57, 425–437. [731,733,735,736]

——— (1995b), "A Quasi-Newton Acceleration of the EM Algorithm," *Statistica Sinica*, 5, 1–18. [736]

——— (2010), *Numerical Analysis for Statisticians (Statistics and Computing)* (2nd ed.), New York: Springer. [734]

Levin, D. (1973), "Development of Non-Linear Transformations for Improving Convergence Of Sequences," *International Journal of Computer Mathematics*, 3, 371–388. [736,743]

Liu, C. (1998), "Information Matrix Computation From Conditional Information via Normal Approximation," *Biometrika*, 85, 973–979. [736,737]

Liu, C., DeNardo, G., Tobin, E., and DeNardo, S. (2004), "Antilymphoma Effects of Anti-HLA-DR and CD20 Monoclonal Antibodies (Lym-1 and Rituximab) on Human Lymphoma Cells," *Cancer Biotherapy and Radiopharmaceuticals*, 19, 545–561. [738]

Liu, C., and Rubin, D. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Faster Monotone Convergence," *Biometrika*, 81, 633–648. [736]

Liu, H., Beckett, L. A., and DeNardo, G. L. (2007), "On the Analysis of Count Data of Birth-and-Death Process Type: With Application to Molecularly Targeted Cancer Therapy," *Statistics in Medicine*, 26, 1114–1135. [730,738,739]

Lorentzen, L., and Waadeland, H. (1992), *Continued Fractions With Applications*, Amsterdam: North-Holland. [732]

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of The Royal Statistical Society*, Series B, 44, 226–233. [736,737]

Meilijson, I. (1989), "A Fast Improvement to the EM Algorithm on Its Own Terms," *Journal of The Royal Statistical Society*, Series B, 51, 127–138. [736]

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909. [737]

Metzner, P., Dittmer, E., Jahnke, T., and Schütte, C. (2007), "Generator Estimation of Markov Jump Processes," *Journal of Computational Physics*, 227, 353–375. [731,733]

Minin, V., and Suchard, M. (2008), "Counting Labeled Transitions in Continuous-Time Markov Models of Evolution," *Journal of Mathematical Biology*, 56, 391–412. [731,737]

Moran, P. A. P. (1951), "Estimation Methods for Evolutive Processes," *Journal of The Royal Statistical Society*, Series B, 13, 141–146. [731]

——— (1953), "The Estimation of the Parameters of a Birth and Death Process," *Journal of The Royal Statistical Society*, Series B, 15, 241–245. [731]

——— (1958), "Random Processes in Genetics," *Mathematical Proceedings of the Cambridge*, 54, 60–71. [730]

Murphy, J. A., and O'Donohoe, M. R. (1975), "Some Properties of Continued Fractions With Applications in Markov Processes," *IMA Journal of Applied Mathematics*, 16, 57–71. [730,732]

Murray, J. (2002), *Mathematical Biology: An Introduction (Interdisciplinary Applied Mathematics) (Vol 1.)*, New York: Springer. [735]

Nee, S. (2006), "Birth-Death Models in Macroevolution," *Annual Review of Ecology, Evolution*, S 37, 1–17. [730]

Nee, S., May, R. M., and Harvey, P. H. (1994), "The Reconstructed Evolutionary Process," *Philosophical Transactions of the Royal Society*, Series B, 344, 305–311. [730]

Novozhilov, A. S., Karev, G. P., and Koonin, E. V. (2006), "Biological Applications of the Theory of Birth-and-Death Processes," *Brief Bioinformation*, 7, 70–85. [730,731]

Oakes, D. (1999), "Direct Calculation of the Information Matrix via the EM," *Journal of the Royal Statistical Society*, Series B, 61, 479–482. [737]

Parthasarathy, P. R., and Sudhesh, R. (2006), "Exact Transient Solution of a State-Dependent Birth-Death Process," *Journal of Applied Mathematics and Stochastic Analysis*, 82, 1–16. [732]

Press, W. H. (2007), *Numerical Recipes: The Art of Scientific Computing*, New York: Cambridge University Press. [736,743]

Renshaw, E. (2011), *Stochastic Population Processes: Analysis, Approximations, Simulations*, New York: Oxord University Press. [730,731,740]

Reynolds, J. F. (1973), "On Estimating the Parameters of a Birth-Death Process," *Australian & New Zealand Journal of Statistics*, 15, 35–43. [731,734]

Richard, G. F., Kerrest, A., and Dujon, B. (2008), "Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes," *Microbiology and Molecular Biology Reviews*, 72, 686–727. [739]

Rose, O., and Falush, D. (1998), "A Threshold Size for Microsatellite Expansion," *Molecular Biology and Evolution*, 15, 613–615. [740]

Rosenberg, N. A., Tsolaki, A. G., and Tanaka, M. M. (2003), "Estimating Change Rates of Genetic Markers Using Serial Samples: Applications to the Transposon IS6110 in Mycobacterium Tuberculosis," *Theoretical Population Biology*, 63, 347–363. [731]

Sainudiin, R., Durrett, R. T., Aquadro, C. F., and Nielsen, R. (2004), "Microsatellite Mutation Models," *Genetics*, 168, 383–395. [739,740]

Schlötterer, C. (2000), "Evolutionary Dynamics of Microsatellite DNA," *Chromosoma*, 109, 365–371. [739]

Tan, W. Y., and Piantadosi, S. (1991), "On Stochastic Growth Processes With Application to Stochastic Logistic Growth," *Statistica Sinica*, 1, 527–540. [730,735]

Thorne, J., Kishino, H., and Felsenstein, J. (1991), "An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences," *Journal of Molecular Evolution*, 33, 114–124. [730]

Vowles, E. J., and Amos, W. (2006), "Quantifying Ascertainment Bias and Species-Specific Length Differences in Human and Chimpanzee Microsatellites Using Genome Sequences," *Molecular Biology and Evolution*, 23, 598–607. [740]

Wall, H. S. (1948), *Analytic Theory of Continued Fractions*, New York: University Series in Higher Mathematics, D. Van Nostrand Company, Inc. [732]

Wanek, L. A., Goradia, T. M., Elashoff, R. M., and Morton, D. L. (1993), "Multi-Stage Markov Analysis of Progressive Disease Applied to Melanoma," *Biometrical Journal*, 35, 967–983. [735]

Webster, M. T., Smith, N. G. C., and Ellegren, H. (2002), "Microsatellite Evolution Inferred From Human and Chimpanzee Genomic Sequence Alignments," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 8748–8753. [739,740,741]

Weiger, E. (1989), "Nonlinear Sequence Transformations for the Acceleration of Convergence and the Summation of Divergent Series," *Computer Physics Reports*, 10, 189–371. [743]

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003), "Likelihood-Based Estimation of Microsatellite Mutation Rates," *Genetics*, 164, 781–787. [739]

Wolff, R. W. (1965), "Problems of Statistical Inference for Birth and Death Queuing Models," *Operations Research*, 13, 343–357. [731,732]

Zhou, H., Lange, K., and Suchard, M. (2010), "Graphics Processing Units and High-Dimensional Optimization," *Statistical Science*, 25, 311–324. [742]