Title
General birth-death processes: probabilities, inference, and applications

Author
Crawford, Forrest Wrenn

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

General birth-death processes: probabilities, inference, and applications

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomathematics

by

Forrest Wrenn Crawford

2012

Abstract of the Dissertation

General birth-death processes: probabilities, inference, and applications

by

Forrest Wrenn Crawford

Doctor of Philosophy in Biomathematics

University of California, Los Angeles, 2012

Professor Marc A. Suchard, Chair

A birth-death process is a continuous-time Markov chain that counts the number of particles in a system over time. Each particle can give birth to another particle or die, and the rate of births and deaths at any given time depends on how many extant particles there are. Birth-death processes are popular modeling tools in evolution, population biology, genetics, epidemiology, and ecology. Despite the widespread interest in birth-death models, no efficient method exists to evaluate the finite-time transition probabilities in a process with arbitrary birth and death rates. Statistical inference of the instantaneous particle birth and death rates also remains largely limited to continuously-observed processes in which per-particle birth and death rates are constant. The lack of theoretical progress in developing statistical tools for dealing with data from birth-death processes has hindered their adoption by applied researchers, and represents a major research frontier in statistical inference for stochastic processes. In this dissertation, I seek to fill this apparent void in three ways. First, I develop mathematical theory and computational tools for computing transition probabilities for general birth-death processes. Second, I develop algorithms for maximum likelihood estimation of rate parameters in discretely observed processes. Third, I derive probability distributions for characteristics of certain birth-death models that are fundamental in macroevolutionary studies. In each case, I give practical applications of the methodology, and show how unsolved problems can be attacked using these techniques.

The dissertation of Forrest Wrenn Crawford is approved.

John Novembre

Kenneth Lange

Janet Sinsheimer

Marc A. Suchard, Committee Chair

University of California, Los Angeles

2012

For Theodore

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I thank my advisor, Marc Suchard, for his wisdom, patience, and uncompromising approach to research. I could not have asked for a better guide than Marc through graduate school and the beginnings of my scientific career. I also want to acknowledge Marc's skill as a harsh but principled editor of manuscripts. Thank you for teaching me how to do science. I also thank the other members of my dissertation committee, Janet Sinsheimer, Kenneth Lange, and John Novembre, for their many valuable comments, questions, and great ideas for future work.

I am indebted to my wife Vanessa for moving to Los Angeles with me and supporting me during these crazy four years. Thank you. I thank my son Theodore for always being happy to see me at the end of a long day. I thank my parents – Jennifer Wrenn & Andrew Kayner, Mark Crawford & Lee Crawford – for their love and support.

There are many others whose advice, criticism, and companionship has helped along the way: the Biomathematics department faculty, including Elliot Landaw, for giving me the opportunity to study here; Van Savage, for providing helpful advice about the academic job search and negotiations; and Tom Chou, for discussions about whether it is better to be a physicist or statistician. I also thank Charles Taylor, for introducing me to forensic genetics; Robert Weiss, for teaching me the importance of statistical practice; Mike Alfaro and Graham Slater, for introducing me to interesting problems in macroevolution; Vladimir Minin, for valuable advice on statistics and jobs; Hua Zhou, for many helpful comments on manuscripts and job advice; Eric Chi, for new ideas in optimization; Alexander Alekseyenko, for helpful advice about school and the academic job market; David Alexander, for many fruitful discussions about mathematics, computing, and DNA sequencing; Gabriela Cybis, for indulging my frequent requests for manuscript critiques; Jennifer Tom, for encouragement and job advice; Joshua Chang, for always being willing to bounce around ideas; Mitchell Johnson, for many thoughtful discussions about computers; and other graduate students and postdocs, including Mandev Gill, Sibon Li, Lewis Lee, and Darren Kessner. Thanks also to David Tomita and Martha Reimer for their outstanding work in support of the Department

and for helping to guide me through the UCLA academic and financial bureaucracies.

Some parts of this dissertation have appeared elsewhere or are under review. A version of Chapter 1 will appear in Crawford and Suchard (2012). It is joint work with Marc A. Suchard. Chapter 2 is joint work with Marc A. Suchard, and appears as Crawford and Suchard (2011) in *The Journal of Mathematical Biology*. Chapter 3 is joint work with Vladimir N. Minin and Marc A. Suchard, and is currently under review (Crawford et al, 2012a). Chapter 4 is joint work with Marc A. Suchard; the project arose from stimulating discussions with Michael Alfaro and Graham Slater. Chapter 5 is joint work with Robert E. Weiss and Marc A. Suchard, and appears as Crawford et al (2012b).

Throughout my graduate career at UCLA, my studies and research work have been funded by three sources. First, the Department of Biomathematics provided academic fees and a stipend during my first year. The Systems and Integrative Biology Training Grant (NIH T32GM008185) paid fees and a generous stipend during years 2-4. In addition, Marc Suchard supplemented my income with graduate student researcher funds.

Vita

2009	MS, Biomathematics, UCLA
2002	BA Neuroscience, Minor in Computer Science
2009-2012	Graduate student researcher, UCLA
2002-2006	Research Associate, Department of Radiology, UCSF
2001	Research Assistant, Department of Psychology, Oberlin College
2000	Summer Intern, Keck Center for Integrative Neuroscience, UCSF

Publications and Presentations

Crawford FW, Suchard MA. Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *J Math Biol* (In press)

Crawford FW, Khayal IS, McGue C, Saraswathy S, Pirzkall A, Cha S, Lamborn KR, Chang S, Berger MS, Nelson SJ. Relationship of pre-surgery metabolic and physiological MR imaging parameters to survival for patients with untreated GBM. *J Neuro-Oncology* 2008; 91(3):337-351.

Saraswathy S, Crawford FW, Lamborn KL, Pirzkall A, Chang S, Cha S, Nelson SJ. Evaluation of MR markers that predict survival in patients with newly diagnosed GBM prior to adjuvant therapy. *J Neuro-Oncology* 2008; 91(1):69-81.

Khayal IS, Crawford FW, Saraswathy S, Lamborn KR, Chang SM, Cha S, McKnight TR, Nelson SJ. Relationship between choline and apparent diffusion coefficient in patients with

gliomas. *J Magn Reson Imaging* 2008; 27(4):718-725.

Crane JC, Crawford FW, Nelson SJ. Grid enabled magnetic resonance scanners for near real-time medical image processing. *J Parallel Distrib Comput* 2006; 66:1524-1533.

Li X, Vigneron DB, Cha S, Graves EE, Crawford FW, Chang SM, Nelson SJ. Relationship of MR-derived lactate, mobile lipids, and relative blood volume for gliomas in vivo *Am J Neuroradiol.* 2005; 26(4):760-769.

Cha S, Tihan T, Crawford FW, Fischbein NJ, Chang S, Bollen A, Nelson SJ, Prados M, Berger MS, Dillon WP. Differentiation of low-grade oligodendrogliomas from low-grade astrocytomas by using quantitative blood-volume measurements derived from dynamic susceptibility contrast-enhanced MR imaging. *Am J Neuroradiol* 2005;26(2):266-273.

Seminar presentation, Carnegie Mellon Statistics; Pittsburgh, PA 2012

Seminar presentation, Yale Biostatistics; New Haven, CT 2012

Young researcher presentation, Case Studies in Bayesian Statistics and Machine Learning; Pittsburgh, PA 2011

Presentation, Systems and Integrative Biology Retreat; Los Angeles, CA 2010-2012

Presentation, International Society for Magnetic Resonance in Medicine Scientific Meeting; Seattle, WA 2006

Presentation, International Society for Magnetic Resonance in Medicine Scientific Meeting; Kyoto, Japan 2004

Presentation, American Society for Neuroradiology Scientific Meeting; Seattle, WA 2004

Presentation, Advances in Imaging Research Symposium; San Francisco, CA 2004

CHAPTER 1

Introduction to birth-death processes in evolution and ecology

1.1 Motivation

Stochastic processes play a vital role in modern evolutionary theory and inference. Models of evolutionary change that incorporate some kind of randomness are appealing for two reasons: first, macroevolutionary change arises from such complex mechanisms that it can reasonably be regarded as fundamentally stochastic; second, and more practically, stochastic models give researchers a means to form hypotheses about the nature of evolutionary change in a rigorous statistical framework that allows model selection and evolutionary hypothesis testing.

One fundamental property of evolutionary change is its branching structure. The simplest branching models for species evolution assume that one species splits into two (a speciation event) and bequeaths its characteristics to both offspring, which evolve independently and simultaneously. This *branching process* continues in both new species over long evolutionary timescales. The evolutionary relationship for a collection of species can be expressed as a "family tree", called a *phylogeny*, and researchers often attempt to reconstruct phylogenies for observed species. For example, modelers often treat species evolution as a stochastic branching process that generates a phylogenetic tree. Then, on the branches of this tree, another stochastic process gives rise to the observed data: often DNA sequence evolution is treated as a continuous-time Markov chain on the states $\{A, G, C, T\}$; quantitative trait changes are modeled by Brownian motion. Using simple stochastic models for speciation and DNA sequence/trait evolution, researchers have developed sophisticated methods for

statistical inference of important evolutionary quantities of interest, including time to most recent common ancestor for a group of species, rate of speciation, and rates of nucleotide mutation, insertion, or deletion.

When modern-day scientists measure the characteristics of a collection of extant species, they glimpse only a contemporary snapshot of this process – the number of species and their characteristics, which may include DNA sequences. But the evolutionary genealogy that produced the species remains unobserved. This raises the question of how researchers can infer the parameters underlying the evolution of those species. Despite the widespread usefulness of probabilistic ideas in explanatory models of evolution, performing statistical inference using evolutionary data can be dauntingly complex. One reason for this is that even simple models of evolutionary change are often mathematically intractible. Part of the difficulty arises from the branching structure of evolutionary change, which confers statistical dependencies on observed data. This dependency, or covariance structure, is largely unobserved, since one generally does not know the phylogeny *a priori*. The difficulty of handling corellated data from species whose evolutionary relationship is unknown can make evolutionary inference extremely challenging. Indeed, in order for a probabilistic model to be useful for researchers interested in learning about evolution, they must be able to compute the probability, or *likelihood*, of the data they observe, given some evolutionary parameters. Stochastic data whose dependencies arise from another stochastic process can thwart principled statistical analysis, and many researchers are actively working on finding tractable models and advanced inference schemes for dealing with such datasets.

One intuitively resonable assumption that can dramatically simplify the evolutionary inference task is the principle of *conditional independence*. This assumption has two consequences relevant to the statistical inference problem for evolutionary models. First, when a species splits into two, the offspring evolve independently and the statistical distribution of their traits is identical, conditional on the trait of their common ancestor. Second, the further waiting time until an evolutionary event is "memoryless" – it does not depend on how much time has already elapsed. This assumption is also known as the *Markov property*, and leads to waiting times between speciation events whose statistical distribution is expo-

nential. Together, these consequences of the conditional independence assumption render tractable many useful evolutionary models.

Counts are one of the simplest types of evolutionary data. For example, researchers may count inserted or deleted nucleotides in DNA sequences, genes, chromosomes, species, or even individual organisms in a population. In this introductory chapter, we introduce a simple but analyticially troublesome class of stochastic counting models called *birth-death processes* (BDPs). BDPs have gained wide use in both evolutionary theory and applications, including population genetics, ecology, and phylogenetic reconstruction. We shall see that the general BDP is simple to describe analytically, but likelihoods for discrete observations from the process can be perniciously difficult to compute. Our treatment of BDPs serves as a case study for work at the mathematical/statistical frontiers of evolutionary inference, and illustrates the successful marriage of probabilistic insight and classical statistical methodology in an important applied setting.

In the remainder of this chapter, we describe in detail some theoretical background on BDPs, a numerical method for computing transition probabilities, and techniques for likelihood-based inference. Our treatment is largely theoretical, since the methodology we develop is new. Indeed, so much mathematical work remains to be done in evolutionary modeling and inference, we do not feel it is inappropriate to devote our entire chapter to new methods for calculating likelihoods and doing statistical inference for BDPs. We conclude the chapter with a brief numerical example of evolutionary inference for microsatellite repeat number evolution in humans and chimpanzees.

1.2 Background and mathematical description

BDPs are a flexible class of continuous-time Markov chains that model the number of "particles" in a system, where each particle can "give birth" to another particle or "die" (Feller, 1971; Karlin and Taylor, 1975). BDPs are a type of branching process in which we do not keep track of the ancestry of each particle, only the total number in a system or population (Kimmel and Axelrod, 2002). Figure 1.2 shows the relationship between a branching

process and the corresponding birth-death process. The rate of births and deaths at any given time depends on how many extant particles there are. When there are k particles, a birth occurs with instantaneous rate λ_k and a death with instantaneous rate μ_k. In the classical "simple linear" BDP, $\lambda_k = k\lambda$ and $\mu_k = k\mu$ so that per-particle birth and death rates remain constant. In a "general" BDP, λ_k and μ_k can be any function of k but remain time-homogeneous (Kendall, 1948, 1949).

The usefulness of BDPs lies in the fact that "particle" can refer to a member of any discrete potentially interacting system in which one only keeps track of the number of objects in existence. BDPs are popular modeling tools in evolution, population biology, genetics, and ecology (Novozhilov et al, 2006). For example, if we interpret the particles as species in a macroevolutionary setting, BDPs can be used to study speciation and extinction over evolutionary timescales (Nee et al, 1994; Nee, 2006). BDPs can also be used to study infectious disease dynamics in a finite population, where the number of individuals infected is the quantity of interest (Bailey, 1964; Andersson and Britton, 2000). In molecular evolution, BDPs can model inserted and deleted nucleotides in a DNA or RNA sequence as part of a probabilistic alignment method (Thorne et al, 1991; Holmes and Bruno, 2001), mobile/transposable genetic elements (Rosenberg et al, 2003), gene families (Demuth et al, 2006), or even whole chromosomes (Mayrose et al, 2010). BDPs can model populations of organisms in a resource-limited environment (Tan and Piantadosi, 1991; Renshaw, 1993, 2011). In finite populations, BDPs are commonly used to model quantities of interest in an evolutionary setting, such as allele frequencies, selection, or coalescence (Moran, 1958; Krone and Neuhauser, 1997; Kingman, 1982b).

There is a rich history of theoretical research into the properties of BDPs. Kendall (1948, 1949) introduces the process with constant per-particle birth and death rates and finds the transition probabilities by a generating function argument. In their groundbreaking series of papers, Karlin and McGregor analyze properties of BDPs, including stationary distributions, moments, transition probabilities, recurrence and passage times, and other quantities of interest (Karlin and McGregor, 1957b,a). They also explore in depth applications of this theory to BDPs whose rates depend linearly on k (Karlin and McGregor, 1958a), and queuing

processes (Karlin and McGregor, 1958b).

Beyond the pioneering work of Karlin and McGregor, many authors have discovered extensions and deeper interpretations for the theoretical properties of BDPs. For example, the theory of BDPs is intimately related to properties of continued fractions (Guillemin and Pinchon, 1999). Flajolet and Guillemin (2000) elucidate the relationship between sample trajectories (or state paths) of a BDP and lattice path combinatorics via continued fractions and develop expressions for a variety of recurrence and passage time variables in terms of continued fractions. Lenin and Parthasarathy (2000) and Parthasarathy et al (1998) discuss further some well-known continued fractions whose deep connection to BDPs previously went unappreciated.

To make our discussion more formal, consider a continuous-time Markov chain $X(t)$ representing the number of particles in a system at time t, taking values on the non-negative integers. To construct a general BDP in a formal way, we first define the rules according to which the number of particles evolves. We do this by specifying the behavior of the process for a very short time Δt, when there are k particles in the system. Intuitively, if Δt is very small, the probability of an event during $(t, t+\Delta t)$ that occurs with rate r is approximately $r\Delta t$. Therefore, the probability of a birth in the interval $(t, t+\Delta t)$, given $X(t) = k$, is

$$\Pr\left(X(t+\Delta t) - X(t) = 1 \mid X(t) = k\right) = \lambda_k \Delta t + o(\Delta t). \tag{1.1}$$

By $o(\Delta t)$ we mean terms that have smaller order than Δt, so that

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0. \tag{1.2}$$

Intuitively, this means that the probability of more than one birth in a small time Δt is negligibly small. The probability of a death in $(t, t+\Delta t)$ is likewise

$$\Pr\left(X(t+\Delta t) - X(t) = -1 \mid X(t) = k\right) = \mu_k \Delta t + o(\Delta t). \tag{1.3}$$

The chance of more than one event of either kind is

$$\Pr\left(|X(t+\Delta t) - X(t)| > 1 \mid X(t) = k\right) = o(\Delta t). \tag{1.4}$$

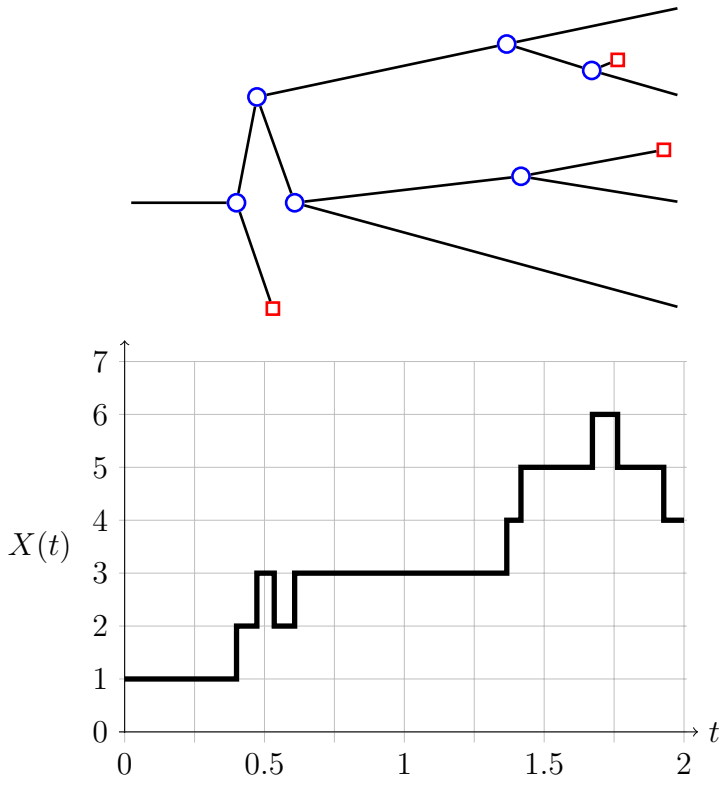


Figure 1.1: Equivalence of branching and counting process interpretations for birth-death processes. The branching process diagram at top shows the genealogy of all individuals during $t \in (0, 2)$, where circles represent births and squares represent deaths. The lower diagram shows the equivalent BDP $X(t)$ counting the number of extant individuals during the same time interval, where the ancestry implicit in the branching diagram is ignored.

Together, these assumptions imply that the probability of no births or deaths occurring during $(t, t + \Delta t)$ is

$$\Pr\left(X(t + \Delta t) - X(t) = 0 \mid X(t) = k\right) = 1 - (\lambda_k + \mu_k)\Delta t + o(\Delta t). \tag{1.5}$$

1.2.1 Transition probabilities

Let $P_{ab}(t) = \Pr(X(t) = b \mid X(0) = a)$ be the transition probability from state $X(0) = a$ to $X(t) = b$. We can use the above equations to form a differential equation describing the change in transition probabilities over time as the process evolves. Suppose that $X(0) = a$. At the current time t, we want to know the probability that in the next Δt units of time, the process will reach state b. We look into the future by writing the probabilities of three types of events that can take the process to state b: birth from $b - 1$, death from $b + 1$, or no change from b:

$$P_{ab}(t + \Delta t) = \lambda_{b-1}P_{a,b-1}(t)\Delta t + \mu_{b+1}P_{a,b+1}(t)\Delta t + (1 - \lambda_b - \mu_b)P_{ab}(t)\Delta t + o(\Delta t). \tag{1.6}$$

The probability of moving from any other state to b is negligibly small. Subtracting $P_{ab}(t)$ from both sides, dividing by Δt, and sending Δt to zero, we obtain the Kolmogorov forward equations:

$$\frac{\mathrm{d}P_{ab}(t)}{\mathrm{d}t} = \lambda_{b-1}P_{a,b-1}(t) + \mu_{b+1}P_{a,b+1}(t) - (\lambda_b + \mu_b)P_{ab}(t), \tag{1.7}$$

where $P_{ab}(0) = 1$ if $a = b$ and zero otherwise. We always assume $\mu_0 = \lambda_{-1} = 0$. In matrix form, (1.7) becomes

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{A}\mathbf{P}(t), \tag{1.8}$$

where $\mathbf{A}$ is the generator matrix with entries $\mathbf{A} = \{a_{ij}\}$, $a_{i,i-1} = \mu_i$, $a_{ii} = -(\lambda_i + \mu_i)$, and $a_{n,n+1} = \lambda_i$. In the matrix case, the initial condition becomes $\mathbf{P}(0) = \mathbf{I}$. This infinite sequence of coupled ordinary differential equations (1.7) or (1.8) can be difficult to solve for many general BDPs (Novozhilov et al, 2006; Renshaw, 2011).

Fortunately, in the simple linear BDP where $\lambda_k = k\lambda$ and $\mu_k = k\mu$, it is possible to solve for these transition probabilities explicitly by finding a generating function solution to the forward equations (Bailey, 1964; Lange, 2010a). To illustrate, let $G_a(s, t) = \sum_{k=0}^{\infty} s^k P_{ak}(t)$.

Let $b = k$ in (1.7), multiply both sides by s^k, and sum on k to obtain

$$\begin{aligned}\frac{\partial G_a(s,t)}{\partial t} &= \sum_{k=0}^{\infty} s^k \frac{\mathrm{d}P_{ak}(t)}{\mathrm{d}t} \\ &= \lambda s^2 \sum_{k=1}^{\infty}(k-1)s^{k-2}P_{a,k-1}(t) + \mu \sum_{k=0}^{\infty}(k+1)s^k P_{a,k+1}(t) \\ &\quad - (\lambda+\mu)s\sum_{k=0}^{\infty} k s^{k-1}P_{ak}(t) \\ &= (\lambda s - \mu)(s-1)\frac{\partial G_a(s,t)}{\partial s},\end{aligned} \tag{1.9}$$

with the initial condition $G_a(s,0) = s^a$. Using Lagrange's method, we suppose s is a function of t. Differentiating $G(s,t) = s_0$, we have

$$\frac{\partial G_a}{\partial t} = \frac{\partial G_a}{\partial s}\frac{\partial s}{\partial t} = 0. \tag{1.10}$$

Comparing this to (1.9), the problem reduces to solving the auxilliary ordinary differential equation

$$\frac{\mathrm{d}s}{\mathrm{d}t} = (\lambda s - \mu)(1-s). \tag{1.11}$$

Equation (1.11), and the initial condition $s(0) = s_0$, lead to the solution

$$s_0 = \frac{\mu(s-1) + (\lambda s - \mu)e^{-(\lambda-\mu)t}}{\lambda(s-1) + (\lambda s - \mu)e^{-(\lambda-\mu)t}}. \tag{1.12}$$

Now $G_a(s,0) = s^a$, so in general

$$G_a(s,t) = \left(\frac{\mu(s-1) + (\lambda s - \mu)e^{-(\lambda-\mu)t}}{\lambda(s-1) + (\lambda s - \mu)e^{-(\lambda-\mu)t}}\right)^a. \tag{1.13}$$

Inverting and finding the bth coefficient of the power series $G_a(s,t)$ we find the transition probabilities

$$P_{ab}(t) = \sum_{j=0}^{\min(a,b)} \binom{a}{j}\binom{a+b-j-1}{a-1}\alpha^{a-j}\beta^{b-j}(1-\alpha-\beta)^j, \tag{1.14}$$

where

$$\alpha(t) = \frac{\mu(e^{(\lambda-\mu)t}-1)}{\lambda(e^{(\lambda-\mu)t}-\mu)} \quad \text{and} \quad \beta(t) = \frac{\lambda(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu}. \tag{1.15}$$

The problem becomes much more complicated for general BDPs. Karlin and McGregor (1957b) present the definitive treatment of the existence of transition probabilities and other properties of BDPs. They obtain the following integral form for the transition probabilities:

$$P_{ab}(t) = \omega_b \int_0^\infty e^{-xt} Q_a(x) Q_b(x) \, \mathrm{d}\psi(x), \tag{1.16}$$

where $\omega_0 = 1$ and $\omega_k = (\lambda_0 \cdots \lambda_{k-1})/(\mu_1 \cdots \mu_k)$ for $k \geq 1$. To see intuitively why this is so, consider a sequence of polynomials $\{Q_k(x)\}$ satisfying the three-term recurrence relation

$$\begin{aligned} -xQ_0(x) &= -(\lambda_0 + \mu_0)Q_0(x) + \lambda_0 Q_1(x), \quad \text{and} \\ -xQ_k(x) &= \mu_k Q_{k-1}(x) - (\lambda_k + \mu_k)Q_k(x) + \lambda_k Q_{k+1}(x) \quad \text{for } k \geq 1. \end{aligned} \tag{1.17}$$

Here, ψ is the spectral measure of the process with respect to which the polynomials $\{Q_k(x)\}$ are orthogonal via the inner product:

$$\langle Q_a, Q_b \rangle_\psi = \int_0^\infty Q_a(x) Q_b(x) \, \mathrm{d}\psi(x) = \begin{cases} 0 & \text{when } a \neq b \\ 1/\omega_b & \text{when } a = b, \end{cases} \tag{1.18}$$

where $Q_0(x) = 1$. In vector-matrix form, the relationship (1.17) becomes

$$-x\mathbf{Q}(x) = \mathbf{A}\mathbf{Q}(x). \tag{1.19}$$

That is, $\mathbf{Q}(x)$ is a right-eigenvector of $\mathbf{A}$ with corresponding eigenvalue $-x$. To proceed, we study the sequence of functions

$$f_a(x,t) = \sum_{k=0}^\infty P_{ak}(t) Q_k(x), \quad a = 0, 1, 2, \ldots. \tag{1.20}$$

In matrix-vector notation, the sequence (1.20) becomes

$$\mathbf{f}(x,t) = \mathbf{P}(t)\mathbf{Q}(x). \tag{1.21}$$

We ignore the details of defining infinite-dimensional vector-matrix multiplication. Now differentiating (1.21), we obtain

$$\frac{\partial \mathbf{f}(x,t)}{\partial t} = -x\mathbf{f}(t), \tag{1.22}$$

with the initial condition $\mathbf{f}(x,0) = \mathbf{P}(0)\mathbf{Q}(x) = \mathbf{Q}(x)$ since $\mathbf{P}(0) = \mathbf{I}$. Solving the differential equation (1.22), we have

$$\mathbf{f}(x,t) = e^{-xt}\mathbf{Q}(x). \tag{1.23}$$

The ath element of $\mathbf{f}(x,t)$ is

$$f_a(x,t) = \sum_{k=0}^{\infty} P_{ak}(t)Q_k(x) = e^{-xt}Q_a(x). \tag{1.24}$$

Now multiplying both sides of (1.24) by $Q_b(x)$ and integrating with respect to the measure $\psi(x)$, we return to

$$\sum_{k=0}^{\infty} P_{a,k}(t) \int_0^{\infty} Q_k(x)Q_b(x)\ \mathrm{d}\psi(x) = \int_0^{\infty} e^{-xt}Q_a(x)Q_b(x)\ \mathrm{d}\psi(x). \tag{1.25}$$

The integrand on the left side of (1.25) is zero when $b \neq k$, so by (1.18), we are left with

$$\frac{P_{ab}(t)}{\omega_b} = \int_0^{\infty} e^{-xt}Q_a(x)Q_b(x)\ \mathrm{d}\psi(x). \tag{1.26}$$

Multiplying by ω_b, we obtain (1.16), completing our informal derivation.

It is interesting to note that we can regard $P_{ab}(t)$ is the bth (generalized) Fourier coefficient of $f_a(x,t)$. This integral representation of the transition probabilities for a BDP is intuitively satisfying because the time-dependency of $P_{ab}(t)$ is contained entirely in the exponential term, and $P_{ab}(t)$ depends on $Q_a(x)$ and $Q_b(x)$ in a simple way. In addition, we have the obvious corollary that

$$\frac{P_{ab}(t)}{P_{ba}(t)} = \frac{\omega_b}{\omega_a}. \tag{1.27}$$

Beyond these simple results related to the interpretation of (1.26), the formalism developed by Karlin and McGregor (1957b) makes possible deep analytic insight into the behavior of general BDPs, including recurrence times and first passage times. One result of special interest to us gives the conditions under which a BDP with a given transition rate matrix $\mathbf{A}$ is unique: Karlin and McGregor show that there is only one transition probability matrix $\mathbf{P}(t)$ that satisfies (1.8) if and only if

$$\sum_{k=0}^{\infty} \left(\omega_k + \frac{1}{\lambda_k \omega_k} \right) = \infty. \tag{1.28}$$

This property assumes that probability is conserved on the non-negative integers and hence $\mu_0 = 0$. We will always assume this is the case in what follows.

Equilibrium solutions are straightforward to obtain (Renshaw, 2011). Setting the left-hand side of the Kolmogorov forward equations (1.7) to zero and replacing the finite-time

transition probabilities $P_{ab}(t)$ with the equilibrium probabilities π_b, we find that

$$\mu_{b+1}\pi_{b+1} - \lambda_b\pi_b = \mu_b\pi_b - \lambda_{b-1}\pi_{b-1}. \tag{1.29}$$

Since this is the case for every b, it is true for $b = 0$ in particular, and $\mu_0 = \lambda_{-1} = 0$, so both sides of (1.29) are zero for every b by induction. This gives the detailed balance condition for continuous-time Markov chains,

$$\mu_k\pi_k = \lambda_{k-1}\pi_{k-1} \quad \text{for } k = 1, 2, \ldots. \tag{1.30}$$

Therefore every general BDP is a reversible Markov chain. Iterating the recurrence (1.30), we find that

$$\pi_k = \frac{\lambda_0\lambda_1\cdots\lambda_{k-1}}{\mu_1\mu_2\cdots\mu_k}\pi_0, \tag{1.31}$$

where we have choosen π_0 so that $\sum_k \pi_k = 1$. Note that $\pi_k \propto \omega_k$ for every k.

Despite the elegant representation (1.16) for the transition probabilities, it can be very difficult to find the polynomials $\{Q_k(x)\}$ (Renshaw, 2011; Novozhilov et al, 2006). In addition, the task of finding these polynomials and measure ψ is a fundamentally analytical task, and is generally not amenable to computational solution. In other words, one cannot simply compute $P_{ab}(t)$ using a computer for an arbitrary set of birth and death rates $\{\lambda_k\}$ and $\{\mu_k\}$ using the formula (1.16) alone. For this reason, nearly all modeling applications use the simple linear BDP since it is analytically tractable. Renshaw (2011) writes of the need for an alternative approach to solving the forward system in order to find transition probabilities for general BDPs:

> "A worthwhile and potentially rewarding challenge would be to develop a simplified and user-friendly version of this technique which would work over a wide range of stochastic processes."

The next section is devoted to this task.

1.3 Numerical transition probabilities

We now outline a method, first presented in Crawford and Suchard (2011), for numerically computing the transition probabilities for a general BDP with arbitrary birth and death rates. To proceed, denote the Laplace transform of $P_{ab}(t)$ as

$$f_{ab}(s) = \mathscr{L}\left[P_{ab}(t)\right](s) = \int_0^\infty e^{-st} P_{ab}(t)\ \mathrm{d}t. \tag{1.32}$$

Now, applying the Laplace transform to (1.7) with $a = 0$, we have

$$\begin{aligned} sf_{00}(s) - P_{00}(0) &= \mu_1 f_{01}(s) - \lambda_0 f_{00}(s), \text{ and} \\ sf_{0b}(s) - P_{0b}(0) &= \lambda_{n-1} f_{0,b-1}(s) + \mu_{b+1} f_{0,b+1}(s) - (\lambda_b + \mu_b) f_{0b}(s) \end{aligned} \tag{1.33}$$

for $b \geq 1$. Recalling that $P_{00}(0) = 1$, and $P_{0b}(0) = 0$ for $b \geq 1$, we rearrange (1.33) to find

$$\begin{aligned} f_{00}(s) &= \frac{1}{s + \lambda_0 - \mu_1 \left(\frac{f_{01}(s)}{f_{00}(s)}\right)}, \text{ and} \\ \frac{f_{0b}(s)}{f_{0,b-1}(s)} &= \frac{\lambda_{b-1}}{s + \mu_b + \lambda_b - \mu_{b+1}\left(\frac{f_{0,b+1}(s)}{f_{0b}(s)}\right)}. \end{aligned} \tag{1.34}$$

By combining these recurrence relations, we obtain the generalized continued fraction

$$f_{00}(s) = \cfrac{1}{s + \lambda_0 - \cfrac{\lambda_0\mu_1}{s + \lambda_1 + \mu_1 - \cfrac{\lambda_1\mu_2}{s + \lambda_2 + \mu_2 - \cdots}}}, \tag{1.35}$$

which is an exact expression for the Laplace transform of the transition probability $P_{0,0}(t)$ (Karlin and McGregor, 1957b; Bordes and Roehner, 1983; Guillemin and Pinchon, 1999; Flajolet and Guillemin, 2000). Now define $a_1 = 1$ and $a_n = -\lambda_{n-2}\mu_{n-1}$, and $b_1 = s + \lambda_0$ and $b_n = s + \lambda_{n-1} + \mu_{n-1}$ for $n \geq 2$. Then (1.35) becomes

$$f_{00}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots. \tag{1.36}$$

We denote the kth convergent of the Laplace transform $f_{00}(s)$ by

$$f_{00}^{(k)}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\cdots\frac{a_k}{b_k} = \frac{A_k(s)}{B_k(s)}. \tag{1.37}$$

The main result of Crawford and Suchard (2011) is the following theorem giving continued fraction expressions for the Laplace transform of the transition probability in a general birth-death process.

Theorem 1. *The Laplace transform of the transition probability $P_{ab}(t)$ is given by*

$$f_{ab}(s) = \begin{cases} \left(\prod_{j=b+1}^{a} \mu_j \right) \dfrac{B_b(s)}{B_{a+1}(s)+} \dfrac{B_a(s)a_{a+2}}{b_{a+2}+} \dfrac{a_{a+3}}{b_{a+3}+} \cdots & \text{for } b \leq a, \\ \left(\prod_{j=a}^{b-1} \lambda_j \right) \dfrac{B_a(s)}{B_{b+1}(s)+} \dfrac{B_b(s)a_{b+2}}{b_{b+2}+} \dfrac{a_{b+3}}{b_{b+3}+} \cdots & \text{for } a \leq b, \end{cases} \tag{1.38}$$

where a_n, b_n, and B_n are as defined above.

The proof of this fact relies on elementary manipulation of the continued fraction recurrences (1.34). Crawford and Suchard (2011) obtain time-domain transition probabilities $P_{ab}(t)$ from (1.38) by numerically inverting the Laplace transforms. We refer the reader to that publication for the computational details. The method returns transition probabilities for many general BDPs that have eluded previous analytical and numerical methods.

1.4 Inference

Another factor hindering more widespread adoption of BDPs by applied researchers is the difficulty in performing statistical estimation of the unknown parameters in a BDP using real-world data (Holmes and Bruno, 2001; Doss et al, 2010). Typically efforts in estimation for BPDs have been limited to continuous observation of the process (Moran, 1951, 1953; Anscombe, 1953; Darwin, 1956; Wolff, 1965; Reynolds, 1973). In addition, most work to date has focused on the simple linear BDP because it is analytically tractable (Keiding, 1975; Thorne et al, 1991; Dauxois, 2004; Rosenberg et al, 2003). However, in practice researchers often observe data from BDPs only at discrete times through longitudinal sampling. In addition, the simple linear BDP may be unappealing because it fails to capture more complicated dynamics of population growth and decay that arise when particles do not

behave independently. To learn from discretely observed general BDPs, we will need more advanced statistical tools.

1.4.1 Likelihood for the continuously-observed process

In a discretely observed general BDP, the likelihood cannot be written in closed form, making analytic maximum likelihood estimation impossible. However, the likelihood of a continuously-observed BDP is straightforward to express (Reynolds, 1973; Keiding, 1975). To develop the likelihood for continuously observed data from a general BDP, we note the following important fact: the exponentially distributed waiting time of a continuous-time Markov process in a certain state is independent of the destination of the next jump (Lange, 2010a). Recall that the waiting time W for the first event to occur from state k is exponentially distributed with rate $\lambda_k + \mu_k$. If the waiting time in the current state k is $W = \tau$, and the next change is a birth,

$$\begin{aligned}
\Pr(W = \tau \mid \text{birth}, X(0) = k) &= \Pr(W = \tau \mid X(0) = k) \Pr(\text{birth} \mid X(0) = k) \\
&= (\lambda_k + \mu_k) e^{-(\lambda_k + \mu_k)\tau} \left(\frac{\lambda_k}{\lambda_k + \mu_k} \right) \\
&= \lambda_k e^{-(\lambda_k + \mu_k)\tau}.
\end{aligned} \tag{1.39}$$

Likewise, the probability of a waiting time $W = \tau$ followed by a death is

$$\Pr(W = \tau \mid \text{death}, X(0) = k) = \mu_k e^{-(\lambda_k + \mu_k)\tau}. \tag{1.40}$$

Since we can only observe the process for a finite time t, the last observation will be the waiting time in some state k from the time of the jump to k to the end of observation. Using the same reasoning,

$$\Pr(W \geq \tau \mid \text{no births or deaths}, X(0) = k) = e^{-(\lambda_k + \mu_k)\tau}. \tag{1.41}$$

To write the likelihood of a continuously-observed BDP from time 0 to t, we introduce some notation to ease our presentation. Suppose we observe $i = 1, \ldots, n$ jumps. Let W_i be the waiting time in the current state just before the ith jump. Define the indicator $B_i = 1$ if the ith jump is a birth, and $B_i = 0$ if the ith jump is a death. Let $t_1, \ldots, t_n$ be

the times of the n jumps, with $t_0 = 0$. Then the likelihood of a sequence of observations $\mathbf{Y} = \{X(\tau), 0 < \tau < t\}$ is

$$\begin{aligned} L = &\prod_{i=1}^{n} \Pr\left(W_i = t_i - t_{i-1} \mid X(t_{i-1})\right) \\ &\times \Pr\left(\text{birth} \mid X(t_{i-1})\right)^{B_i} \Pr(\text{death} \mid X(t_{i-1}))^{1-B_i} \\ &\times \Pr\left(W_{n+1} = t - t_n \mid \text{no births or deaths}, X(t_n)\right) \end{aligned}$$

$$\begin{aligned} = &\prod_{i=1}^{n} (\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})}) \exp\left[-(\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})})(t_i - t_{i-1})\right] \\ &\times \left(\frac{\lambda_{X(t_{i-1})}^{B_i} \mu_{X(t_{i-1})}^{1-B_i}}{\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})}}\right) \times \exp\left[-(\lambda_{X(t_n)} + \mu_{X(t_n)})(t - t_n)\right] \end{aligned} \tag{1.42}$$

$$\begin{aligned} = &\prod_{i=1}^{n} \lambda_{X(t_{i-1})}^{B_i} \mu_{X(t_{i-1})}^{1-B_i} \exp\left[-(\lambda_{X(t_{i-1})} + \mu_{X(t_{i-1})})(t_i - t_{i-1})\right] \\ &\times \exp\left[-(\lambda_{X(t_n)} + \mu_{X(t_n)})(t - t_n)\right], \end{aligned}$$

where $X(t_{i-1})$ is the state just before the ith jump. This cumbersome notation can be eliminated if we instead keep track of the total waiting time in each state and the number of births and deaths from each state. Define $\mathbf{1}\{E\}$ to be the indicator of an event E, and let

$$T_k = \sum_{i=1}^{n} (t_i - t_{i-1}) \mathbf{1}\{X(t_{i-1}) = k\} \tag{1.43}$$

be the total time spent in state k over all visits to k. Then let

$$U_k = \sum_{i=1}^{n} \mathbf{1}\{X(t_{i-1}) = k, B_i = 1\} \tag{1.44}$$

be the number of up steps (births) from state k, and let

$$D_k = \sum_{i=1}^{n} \mathbf{1}\{X(t_{i-1}) = k, B_i = 0\} \tag{1.45}$$

be the number of down steps (deaths) from state k. Then we can re-write the likelihood (1.42) in much simpler and more transparent form as

$$L = \prod_{k=0}^{\infty} \lambda_k^{U_k} \mu_k^{D_k} \exp[-(\lambda_k + \mu_k) T_k]. \tag{1.46}$$

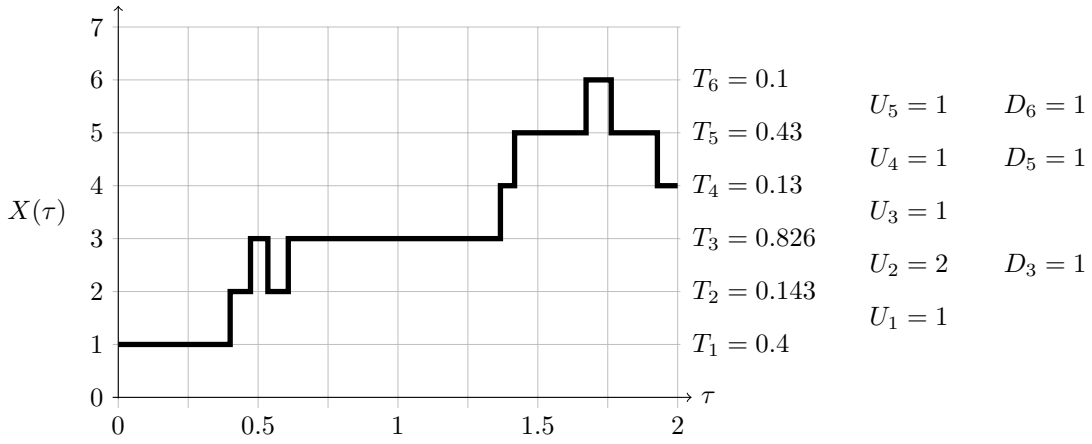


Figure 1.2: Sample trajectory of a birth-death counting process starting at $X(0) = 1$ showing how to calculate the statistics U_k, D_k, and T_k from the continuously observed trajectory.

Of course, in a BDP observed continuously for a finite time (for which (1.28) holds), there are only finitely many jumps observed, so the product above is not really infinite in practice. Figure 3.1 shows an example of how to calculate the statistics U_k, D_k, and T_k for a given continuously observed trajectory from a BDP. With the continuously-observed data likelihood in hand, we seek a means of obtaining maximum likelihood estimates from discretely observed data.

Maximum likelihood estimation for continuously-observed BDPs is often straightforward. For example, consider the simple linear BDP with birth rate $\lambda_k = k\lambda$ and death rate $\mu_k = k\mu$. The likelihood (1.46) of a single observation becomes, up to a normalizing constant,

$$L \propto \lambda^U \mu^D \exp\left[-(\lambda + \mu) \int_0^t X(\tau)\, \mathrm{d}\tau\right], \tag{1.47}$$

where $U = \sum_k U_k$ is the total number of up steps (births), $U = \sum_k D_k$ is the total number of down steps (deaths) during the interval $(0, t)$, and

$$\int_0^t X(\tau)\, \mathrm{d}\tau = \sum_{k=0}^{\infty} k T_k. \tag{1.48}$$

Maximizing (1.47) with respect to the unknown parameters λ and μ, we obtain the maximum

likelihood estimators

$$\hat{\lambda} = \frac{U}{\int_0^t X(\tau)\ \mathrm{d}\tau} \quad \text{and} \quad \hat{\mu} = \frac{D}{\int_0^t X(\tau)\ \mathrm{d}\tau}. \tag{1.49}$$

Although the estimators provided by (1.49) involve an integral over the state path of the process, $X(t)$ is simply a step function that is fully observed over $(0, t)$.

1.4.2 Likelihood for the discretely observed process

Suppose now that the process $X(\tau)$ is observed only discretely, once at time 0 and again at time t. Let us label the state of the BDP at these times as $X(0) = a$ and $X(t) = b$. Then given that $X(0) = a$, the probability that $X(t) = b$ is the transition probability $P_{ab}(t)$. In section 1.3 we outlined a method for numerically computing this probability for any general BDP. If we regard the transition probability $P_{ab}(t)$ as a function of some unknown parameters θ which control the birth and death rates, writing $P_{ab}(t|\theta)$, then we have the *likelihood* of our observation. In principle, we could numerically maximize the likelihood for discrete observations to find an estimate of θ. However, as the number of parameters increases, naïve numerical optimization often suffers from poor convergence.

1.5 Estimation via the EM algorithm

In this section, we review the estimation machinery developed by Crawford et al (2012a) for maximum likelihood or maximum a posteriori estimation in BDPs. When a BDP is discretely sampled, U_k, D_k, and T_k are unobserved for every k; we cannot maximize the likelihood without knowing these statistics. We therefore appeal to the expectation-maximization (EM) algorithm for iterative maximum likelihood estimation with missing data (Dempster et al, 1977). When the incomplete data likelihood is intractable but the complete data likelihood has a simple form, the EM algorithm operates by replacing the each missing datum by a conditional expectation as follows. If X is the complete (unobserved data), Y represents the incomplete (observed) data, and $\ell(\theta|X)$ is the complete data log-likelihood, we form a

surrogate function Q as the expectation of the complete data likelihood, conditional on the observed data Y and the current (mth) parameter iterate:

$$Q\left(\theta \mid \theta^{(m)}\right) = \mathbb{E}\left(\ell(\theta|X) \mid Y = y, \theta^{(m)}\right). \tag{1.50}$$

This is the E-step of the EM algorithm, and it accomplishes a minorization of $\ell(\theta)$ at $\theta^{(m)}$. The M-step maximizes (or takes a step toward the maximum of) Q. By alternating these steps — minorizing ℓ by Q, then finding the θ that maximizes Q — the EM algorithm drives succeeding iterates toward the MLE.

Taking the expectation of the logarithm of (1.46), conditional on the observed data $Y = (X(0) = a, X(t) = b, t)$ and the current parameter estimate $\theta^{(m)}$, we write the surrogate function for the BDP as follows:

$$\begin{aligned} Q\big(\theta \mid \theta^{(m)}\big) &= \mathbb{E}\big[\ell(\theta) \mid Y, \theta^{(m)}\big] \\ &= \sum_{k=0}^{\infty} \mathbb{E}(U_k|Y) \log\big[\lambda_k(\theta)\big] + \mathbb{E}(D_k|Y) \log\big[\mu_k(\theta)\big] - \mathbb{E}(T_k|Y)\big[\lambda_k(\theta) + \mu_k(\theta)\big], \end{aligned} \tag{1.51}$$

In the above equation and many that follow, we omit the dependence of the conditional expectations on $\theta^{(m)}$ from the mth iterate for visual clarity.

To calculate the conditional expectations necessary for the E-step of the EM algorithm, we appeal to the following integral expressions

$$\mathbb{E}(U_k|Y) = \frac{\displaystyle\int_0^t P_{ak}(\tau)\lambda_k P_{k+1,b}(t-\tau)\, \mathrm{d}\tau}{P_{ab}(t)}, \tag{1.52a}$$

$$\mathbb{E}(D_k|Y) = \frac{\displaystyle\int_0^t P_{ak}(\tau)\mu_k P_{k-1,b}(t-\tau)\, \mathrm{d}\tau}{P_{ab}(t)}, \quad \text{and} \tag{1.52b}$$

$$\mathbb{E}(T_k|Y) = \frac{\displaystyle\int_0^t P_{ak}(\tau)P_{kb}(t-\tau)\, \mathrm{d}\tau}{P_{ab}(t)} \tag{1.52c}$$

(Lange, 1995a; Holmes and Rubin, 2002; Bladt and Sorensen, 2005; Hobolth and Jensen, 2005; Metzner et al, 2007). Since the Laplace transforms $f_{a,b}(s)$ of these transition probabilities can be computed via Theorem 1, we appeal to the Laplace convolution property to

obtain

$$\mathbb{E}(U_k|Y) = \lambda_k \frac{\mathscr{L}^{-1}\Big[f_{ak}(s)\ f_{k+1,b}(s)\Big](t)}{P_{ab}(t)}, \tag{1.53a}$$

$$\mathbb{E}(D_k|Y) = \mu_k \frac{\mathscr{L}^{-1}\Big[f_{ak}(s)\ f_{k-1,b}(s)\Big](t)}{P_{ab}(t)}, \quad \text{and} \tag{1.53b}$$

$$\mathbb{E}(T_k|Y) = \frac{\mathscr{L}^{-1}\Big[f_{ak}(s)\ f_{kb}(s)\Big](t)}{P_{ab}(t)}, \tag{1.53c}$$

where $\mathcal{L}^{-1}$ denotes inverse Laplace transformation. These expressions are formally equivalent to (3.12), but they offer substantial computational time savings over numerical integration of (3.12), and make possible efficient computation of conditional expectations for EM algorithms for any BDP.

1.6 EM algorithms in evolutionary inference

1.6.1 Simple linear BDP

In the simple linear BDP (also known as the "Kendall process"), births and deaths happen at constant per-particle rates, so $\lambda_k = k\lambda$ and $\mu_k = k\mu$. The unknown is $\theta = (\lambda, \mu)$. The surrogate function Q becomes

$$Q(\theta) = \sum_{k=0}^{\infty} \mathbb{E}(U_k|Y)\log[k\lambda] + \mathbb{E}(D_k|Y)\log[k\mu] - \mathbb{E}(T_k|Y)k(\lambda+\mu). \tag{1.54}$$

Maximizing (3.14) with respect to the θ yields the updates:

$$\lambda^{(m+1)} = \frac{\mathbb{E}(U|Y)}{\mathbb{E}(T_{\text{particle}}|Y)}, \text{ and} \tag{1.55a}$$

$$\mu^{(m+1)} = \frac{\mathbb{E}(D|Y)}{\mathbb{E}(T_{\text{particle}}|Y)}. \tag{1.55b}$$

1.6.2 Linear BDP with immigration

The linear BDP with immigration is similar to the simple linear BDP, but there is a source of new arrivals whose rate is not proportional to the number of individual particles already in existence. The rate of new immigrants is constant, making the immigration component of

this BDP a Poisson process. This yields the birth and death rates $\lambda_k = k\lambda + \nu$ and $\mu_k = k\mu$. The log-likelihood becomes

$$\ell(\theta) = \sum_{k=0}^{\infty} U_k \log(k\lambda + \nu) + D_k \log(\mu) - T_k[k(\lambda + \mu) + \nu]. \tag{1.56}$$

Unfortunately, it is difficult to maximize this surrogate function analytically. Since each term in the sum is a concave function of the unknown parameters, we can separate them in a second minorizing function H such that for all θ, $H\big(\theta|\theta^{(m)}\big) \leq \ell(\theta)$ and $H\big(\theta^{(m)}|\theta^{(m)}\big) = \ell\big(\theta^{(m)}\big)$. To accomplish the minorization, note that

$$\log(k\lambda+\nu) \geq \frac{k\lambda^{(m)}}{k\lambda^{(m)} + \nu^{(m)}} \log\left[\frac{k\lambda^{(m)}}{k\lambda^{(m)} + \nu^{(m)}}\lambda\right] + \frac{\nu^{(m)}}{k\lambda^{(m)} + \nu^{(m)}} \log\left[\frac{\nu^{(m)}}{k\lambda^{(m)} + \nu^{(m)}}\nu\right]. \tag{1.57}$$

We form a minorizing log-likelihood function H as follows:

$$\begin{aligned}\ell(\theta) &\geq H\big(\theta|\theta^{(m)}\big) \\ &= \sum_{k=0}^{\infty} U_k\big[p_k \log\big(p_k\lambda\big) + (1-p_k)\log\big((1-p_k)\nu\big)\big] + D_k\log(\mu) - \big[k(\lambda+\mu)+\nu\big]T_k,\end{aligned} \tag{1.58}$$

where

$$p_k = \frac{k\lambda^{(m)}}{k\lambda^{(m)} + \nu^{(m)}}. \tag{1.59}$$

Then letting $Q\big(\theta \mid \theta^{(m)}\big) = \mathbb{E}\left(H(\theta) \mid Y, \theta^{(m)}\right)$ be the surrogate function and maximizing with respect to the unknowns gives the updates

$$\lambda^{(m+1)} = \frac{\displaystyle\sum_{k=0}^{\infty} p_k \mathbb{E}(U_k|Y)}{\mathbb{E}(T_{\text{particle}}|Y)}, \text{ and} \tag{1.60a}$$

$$\nu^{(m+1)} = \frac{\displaystyle\sum_{k=0}^{\infty} (1-p_k)\mathbb{E}(U_k|Y)}{t}. \tag{1.60b}$$

The update for μ is the same as (3.15b).

1.6.3 Moran model

The Moran model is a popular model for genetic drift in continuous time. In its simplest incarnation, the process keeps track of the number of alleles of a certain type at a biallelic

locus in a haploid population of constant size $N < \infty$. Call the two alleles A and B, and suppose we wish to keep track of the number of A carriers in the population. In the Moran model with selection, carriers of A have fitness α, and carriers of B have fitness β. For the sake of identifiability in a statistical setting, we specify $\beta = 1$ and let α denote the relative fitness of A carriers over B carriers. In the moran model with mutation, A mutates to B in one generation with probability u, and B mutates to A with probability v. Selection enters the process when an existing individual dies. A replacement allele is drawn from the population (including the individual that dies). Once this replacement allele is chosen, it is subjected to mutation. If we let $X(t)$ be the number of A alleles in the population at time t, the rate of additions of new A carriers is

$$\lambda_n = \frac{N-n}{N}\left[\alpha\frac{n}{N}(1-u) + \frac{N-n}{N}v\right], \tag{1.61}$$

for $n = 0, \ldots, N$ with $\lambda_n = 0$ when $n > N$. The rate of removals of A carriers is

$$\mu_n = \frac{n}{N}\left[\frac{N-n}{N}(1-v) + \alpha\frac{n}{N}u\right], \tag{1.62}$$

Then $X(t)$ is a general BDP with these birth and death rates.

Maximizing the log-likelihood with respect to the unknowns α, u, and v is difficult. However, following the example of the previous section, we can construct a minorizing function to separate the parameters in the logarithm terms. Note that we can minorize the birth rate as

$$\begin{aligned}
\log(\lambda_n) &\propto \log\left[n\alpha(1-u) + (N-n)v\right] \\
&\geq p_n^{(m)} \log\left(p_n^{(m)} n\alpha(1-u)\right) + (1-p_n^{(m)}) \log\left((1-p_n^{(m)})(N-n)v\right) \\
&\propto p_n^{(m)}\big(\log(\alpha) + \log(1-u)\big) + (1-p_n^{(m)})\log(v),
\end{aligned} \tag{1.63}$$

where

$$p_n^{(m)} = \frac{n\alpha^{(m)}(1-u^{(m)})}{n\alpha^{(m)}(1-u^{(m)}) + (N-n)v^{(m)}}. \tag{1.64}$$

Likewise, we minorize the death rate as

$$\begin{aligned}
\log(\mu_n) &\propto \log\left[(N-n)(1-v) + n\alpha u\right] \\
&\geq q_n^{(m)} \log\left(q_n^{(m)}(N-n)(1-v)\right) + (1-q_n^{(m)}) \log\left(q_n^{(m)} n\alpha u\right) \\
&\propto q_n^{(m)} \log(1-v) + (1-q_n^{(m)})\big(\log(\alpha) + \log(u)\big),
\end{aligned} \tag{1.65}$$

where

$$q_n^{(m)} = \frac{(N-n)(1-v^{(m)})}{(N-n)(1-v^{(m)}) + n\alpha^{(m)}u^{(m)}}. \tag{1.66}$$

Now, we form the minorizing function H as

$$\begin{aligned}
\ell(\theta) &\geq H(\theta) \\
&= \sum_{k=0}^{N} B_k \Big[p_k^{(m)} \big(\log(\alpha) + \log(1-u) \big) + (1 - p_k^{(m)}) \log(v) \Big] \\
&\quad + D_k \Big[q_k^{(m)} \log(1-v) + (1 - q_k^{(m)}) \big(\log(\alpha) + \log(u) \big) \Big] \\
&\quad - \frac{T_k}{N^2} \Big[(N-k)k\alpha(1-u) + (N-k)^2 v + (N-k)k(1-v) + k^2\alpha u \Big].
\end{aligned} \tag{1.67}$$

One simple way to proceed is to find updates for each of the unknowns, conditional on the previous (mth) estimate of the others. The update for α is

$$\alpha^{(m+1)} = \frac{\displaystyle\sum_{k=0}^{N} p_k^{(m)} B_k + (1 - q_k^{(m)}) D_k}{\displaystyle\frac{1}{N^2} \sum_{k=0}^{N} T_k \big[(N-k)k(1-u^{(m)}) + k^2 u^{(m)} \big]}. \tag{1.68}$$

For the sake of brevity, we give the update for u as the solution of the quadratic equation

$$\sum_{k=0}^{N} -uB_k p_k^{(m)} + (1-u) D_k (1 - q_k^{(m)}) - u(1-u) \frac{T_k}{N^2} \big[k^2 \alpha^{(m)} - (N-k)k\alpha^{(m)} \big] = 0. \tag{1.69}$$

The update for v is similar.

1.6.4 Maximum *a posteriori* estimation

In a Bayesian setting, a prior distribution $f(\theta)$ on the unknown parameters θ is given, and we seek to maximize the log-posterior distribution of the parameters, given the data, $\Pr(\theta \mid Y) \propto \Pr(Y \mid \theta) f(\theta)$ to obtain the maximum *a posteriori* (MAP) estimate of θ. Then the surrogate function becomes $Q\big(\theta|\theta^{(m)}\big) = \mathbb{E}\left(\ell(\theta)|Y, \theta^{(m)}\right) + \log\left[f\left(\theta\right)\right]$.

To illustrate, suppose that independent observations Y_i from a BDP follow the simple linear model, and we believe that λ and μ are *a priori* independent and are Gamma-distributed:

$$\lambda \sim \text{Gamma}(k_\lambda, \beta_\lambda) \qquad \text{and} \qquad \mu \sim \text{Gamma}(k_\mu, \beta_\mu). \tag{1.70}$$

Then the unknowns are $\theta = (\lambda, \mu)$ and the log-prior for θ is

$$\log f(\theta) \propto (k_\lambda - 1)\log(\lambda) + (k_\mu - 1)\log(\mu) - \frac{\lambda}{\beta_\lambda} - \frac{\mu}{\beta_\mu}. \tag{1.71}$$

Ignoring irrelevant terms, the surrogate function becomes

$$\begin{aligned} Q\big(\theta|\theta^{(m)}\big) = \mathbb{E}(U|Y)\log(\lambda) + \mathbb{E}(D|Y)\log(\mu) - \mathbb{E}(T_{\text{particle}}|Y)(\lambda+\mu) \\ + (k_\lambda - 1)\log(\lambda) + (k_\mu - 1)\log(\mu) - \frac{\lambda}{\beta_\lambda} - \frac{\mu}{\beta_\mu} \end{aligned} \tag{1.72}$$

The updates are

$$\lambda^{(m+1)} = \frac{\mathbb{E}(U|Y) + k_\lambda - 1}{\mathbb{E}(T_{\text{particle}}|Y) + \frac{1}{\beta_\lambda}}, \quad \text{and} \tag{1.73a}$$

$$\mu^{(m+1)} = \frac{\mathbb{E}(D|Y) + k_\mu - 1}{\mathbb{E}(T_{\text{particle}}|Y) + \frac{1}{\beta_\mu}}. \tag{1.73b}$$

1.7 Numerical example

Microsatellites are short repeated motifs of bases in DNA sequences (Schlötterer, 2000; Ellegren, 2004; Richard et al, 2008). The number of repeated motifs in a microsatellite can change during the DNA replication phase of meiosis. One widely accepted theory regarding the origin of microsatellite repeat number changes posits that the molecular machinery that places new nucleotide bases on the template DNA strand can slip and misalign the new bases so that they are staggered with respect to the template strand; this is known as "polymerase slippage" (Schlötterer, 2000). Intuitively, a microsatellite with k repeats offers k opportunities for strand mismatch due to polymerase slippage during each meiosis. Generalizing to continuous evolutionary time, it is reasonable to model the process of repeat additions and deletions as a Markov chain in which only jumps to adjacent positive integers are possible. Indeed, many researchers have proposed linear BDPs to study repeat number evolution in microsatellites (Whittaker et al, 2003; Calabrese and Durrett, 2003; Sainudiin et al, 2004).

In this section, we apply our methods to the problem of chimpanzee-human microsatellite evolution, drawing on the data in Table 6 of the supplementary information in Webster et al (2002). Figure 3.3 shows the repeat numbers for the orthologous microsatellites in

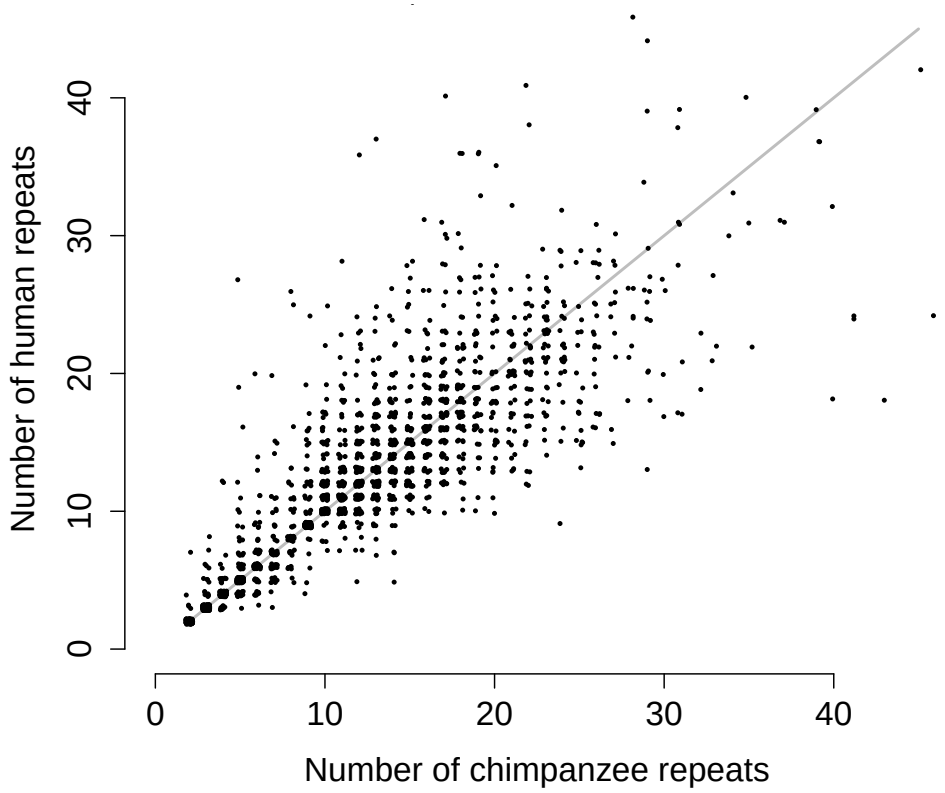


Figure 1.3: Number of repeats in chimpanzees and humans for 2467 orthologous microsatellite loci. The counts are jittered to show the density, and the gray line denotes equal number of repeats between each species.

chimpanzees plotted against the corresponding number of repeats in humans. The data are jittered slightly to show the density of observations for each set of counts. To simplify the inference task, we regard chimpanzees as the ancient ancestors of modern-day humans and model the evolution of microsatellite repeat numbers as a simple linear BDP over a fixed evolutionary time scaled to unity. Since we do not observe the evolutionary trajectory of microsatellite repeat numbers from chimpanzees to humans, we regard the counts as discrete observations from a BDP. For microsatellite i, we let $X_i(0)$ be the number of repeats in chimpanzees, and $X_i(1)$ be the number in humans. We further assume that the birth and death rates are the same for every microsatellite in the dataset, and that microsatellites are ascertained randomly and without regard to repeat number. Crawford et al (2012a) explore these issues in far greater detail. Figure 3.2 shows the iterates of the EM algorithm to fit the simple linear BDP to the microsatellite repeat number data, superimposed on the contours of the likelihood function.

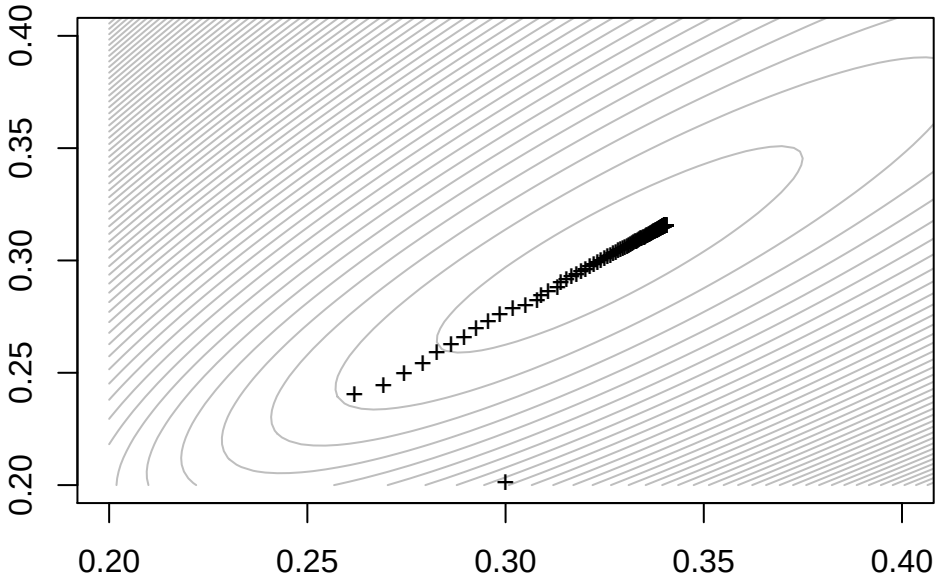


Figure 1.4: Iterates of the EM algorithm for the microsatellite evolution dataset converging to the MLE ($\lambda = 0.3405,\ \mu = 0.2147$).

CHAPTER 2

Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution

A birth-death process is a continuous-time Markov chain that counts the number of particles in a system over time. In the general process with n current particles, a new particle is born with instantaneous rate λ_n and a particle dies with instantaneous rate μ_n. Currently no robust and efficient method exists to evaluate the finite-time transition probabilities in a general birth-death process with arbitrary birth and death rates. In this paper, we first revisit the theory of continued fractions to obtain expressions for the Laplace transforms of these transition probabilities and make explicit an important derivation connecting transition probabilities and continued fractions. We then develop an efficient algorithm for computing these probabilities that analyzes the error associated with approximations in the method. We demonstrate that this error-controlled method agrees with known solutions and outperforms previous approaches to computing these probabilities. Finally, we apply our novel method to several important problems in ecology, evolution, and genetics.

2.1 Introduction

Birth-death processes (BDPs) have a rich history in probabilistic modeling, including applications in ecology, genetics, and evolution (Thorne et al, 1991; Krone and Neuhauser, 1997; Novozhilov et al, 2006). Traditionally, BDPs have been used to model the number of organisms or particles in a system, each of which reproduce and die in continuous time. A general

BDP is a continuous-time Markov chain on the non-negative integers in which instantaneous transitions from state $n \geq 0$ to either $n+1$ or $n-1$ are possible. These transitions are called "births" and "deaths". Starting at state n, jumps to $n+1$ occur with instantaneous rate λ_n and jumps to $n-1$ with instantaneous rate μ_n. The simplest BDP has linear rates $\lambda_n = n\lambda$ and $\mu_n = n\mu$ with no state-independent terms (Kendall, 1948; Feller, 1971). This model is the most widely-used BDP since there exist closed-form expressions for its transition probabilities (Bailey, 1964; Novozhilov et al, 2006). Many applications of BDPs require convenient methods for computing the probability $P_{m,n}(t)$ that the system moves from state m to state n in finite time $t \geq 0$. These probabilities exhibit their usefulness in many modeling applications since the probabilities do not depend on the possibly unobserved path taken by the process from m to n and hence make possible analyses of discretely sampled or partially observed processes. Despite the relative simplicity of specifying the rates of a general BDP, it can be remarkably difficult to find closed-form solutions for the transition probabilities even for simple models (Renshaw, 1993; Mederer, 2003; Novozhilov et al, 2006).

In a pioneering series of papers, Karlin and McGregor develop a formal theory of general BDPs that expresses their transition probabilities in terms of a sequence of orthogonal polynomials and a spectral measure (Karlin and McGregor, 1957a,b, 1958b). While the work of Karlin and McGregor yields valuable theoretical insights regarding the existence of unique solutions and properties of recurrence and transience for a given process, there remains no clear recipe for determining the orthogonal polynomials and measure corresponding to an arbitrary set of birth and death rates. Additionally, even when the polynomials and measure are known, the transition probabilities may not have an analytic representation or a convenient computational form.

Possibly due to the difficulty of finding computationally useful formulas for transition probabilities in general BDPs, many applied researchers resort to easier analyses using moments, first passage times, equilibrium probabilities, and other tractable quantities of interest. Referring to the system of Kolmogorov forward differential equations for transition probabilities that we give below, Novozhilov et al (2006, page 73) write,

> "The problem with exact solutions of system (1) is that, in many cases, the expressions for the state probabilities, although explicit, are intractable for analysis and include special polynomials. In such cases, it may be sensible to solve more modest problems concerning the birth-and-death process under consideration, without the knowledge of the time-dependent behavior of state probabilities $p_n(t)$."

Indeed, closed-form analytic expressions for transition probabilities of general BDPs are only known for a few types of processes. Some examples include constant birth and death rates (Bailey, 1964), zero birth or death rates (pure-death and pure-birth) (Yule, 1925; Taylor and Karlin, 1998), and certain linear rates (Karlin and McGregor, 1958a). As a seemingly straightforward example, in the BDP with linear birth and death rates $\lambda_n = n\lambda + \nu$ and $\mu_n = n\mu + \gamma$ including state-independent terms, Ismail et al (1988) offer the orthogonal polynomials and associated measure, but still no closed form is available for the transition probabilities.

Despite the difficulty in obtaining analytic expressions, several authors have made progress in approximate numerical methods for solution of transition probabilities in general BDPs. Murphy and O'Donohoe (1975) develop an appealing numerical method for the transition probabilities based on a continued fraction representation of Laplace-transformed transition probabilities. They invert these transformed probabilities by first truncating the continued fraction. Several other authors give similar expressions derived from truncation of the state space (Grassmann, 1977b,a; Rosenlund, 1978; Sharma and Dass, 1988; Mohanty et al, 1993). However, Klar et al (2010) find that methods based on continued fraction truncation and then subsequent analytical transformation can suffer from instability. As an alternative, Parthasarathy and Sudhesh (2006a) express the infinite continued fraction representation given by Murphy and O'Donohoe as a power series. Unfortunately, the small radius of convergence of this series makes it less useful for numerical computation.

We also note that for general BDPs that take values on a finite state space (usually $n \in \{0, 1, \ldots, N\}$), it is possible to write a finite-dimensional stochastic transition rate ma-

trix and solve for the matrix of transition probabilities. If the rate matrix is diagonalizable, computation of transition probabilities in this manner can be computationally straightforward. To illustrate, let Q be a finite-dimensional stochastic rate matrix with $Q = U\Lambda U^{-1}$ where U is an orthogonal matrix and Λ is diagonal. The matrix of transition probabilities P satisfies the matrix differential equation $P' = PQ$ with initial condition $P(0) = I$. The solution is $P(t) = \exp[Qt] = U \text{ diag}(e^{z_1 t}, e^{z_2 t}, \ldots, e^{z_N t})\ U^{-1}$, where $z_1, \ldots, z_N$ are the eigenvalues of Q. However, it is possible to specify reasonable rate parameters in a general BDP that satisfy requirements for the existence of a unique solution, but do not result in a diagonalizable rate matrix. Also, if the state space over which the BDP takes values is large, numerical eigendecomposition of Q may be computationally expensive and could introduce serious roundoff errors.

To our knowledge, no robust computational method currently exists for finding the finite-time transition probabilities of general BDPs with arbitrary rates. Such a technique would allow rapid development of rich and sophisticated ecological, genetic, and evolutionary models. Additionally, in statistical applications, transition probabilities can serve as observed data likelihoods, and are thus often useful in estimating transition rate parameters from partially observed BDPs. We believe more sophisticated BDPs can be very useful for applied researchers. In spite of the numerical difficulties presented by approximant methods, we are surprised that continued fraction methods like that of Murphy and O'Donohoe (1975) are not more widely explored. This may be due to omission of important details in their derivation of continued fraction expressions for the Laplace transform of the transition probabilities.

In this paper, we build on continued fraction expressions for the Laplace transforms of the transition probabilities of a general BDP using techniques similar to those introduced by Murphy and O'Donohoe, and we fill in the missing details in the proof of this representation. We then apply the Laplace inversion formulae of Abate and Whitt (1992a,b) to obtain an efficient and robust method for computation of transition probabilities in general BDPs. Our method relies on three observations: 1) it is possible to find exact expressions for Laplace transforms of the transition probabilities of a general BDP using continued fractions (Murphy and O'Donohoe, 1975); 2) evaluation of continued fractions is typically very

fast, requires far fewer evaluations than equivalent power series, and there exist robust algorithms for evaluating them efficiently (Bankier and Leighton, 1942; Wall, 1948; Blanch, 1964; Lorentzen and Waadeland, 1992; Craviotto et al, 1993; Abate and Whitt, 1999; Cuyt et al, 2008); and 3) recovery of probability distributions by Laplace inversion using a Riemann sum approximation is often more computationally stable than analytical methods of inversion (Abate and Whitt, 1992a,b, 1995). Finally, we demonstrate the advantages of our error-controlled method through its application to several birth-death models in ecology, genetics, and evolution whose solution remains unavailable by other means.

2.2 Transition probabilities

2.2.1 Background

A general birth-death process is a continuous-time Markov process $\mathcal{X} = \{X(t), t \geq 0\}$ counting the number of arbitrarily defined "particles" in existence at time $t \geq 0$, with $X(0) = m \geq 0$. To characterize the process, we define non-negative instantaneous birth rates λ_n and death rates μ_n for $n \geq 0$, with $\mu_0 = 0$ and transition probabilities $P_{m,n}(t) = \Pr(X(t) = n \mid X(0) = m)$. While λ_n and μ_n are time-homogeneous constants, they may depend on n. We refer to the classical linear BDP in which $\lambda_n = n\lambda$ and $\mu_n = n\mu$ as the "simple birth-death process" (Kendall, 1948; Feller, 1971). The general BDP transition probabilities satisfy the infinite system of ordinary differential equations

$$\begin{aligned} \frac{\mathrm{d}P_{m,0}(t)}{\mathrm{d}t} &= \mu_1 P_{m,1}(t) - \lambda_0 P_{m,0}(t), \text{ and} \\ \frac{\mathrm{d}P_{m,n}(t)}{\mathrm{d}t} &= \lambda_{n-1} P_{m,n-1}(t) + \mu_{n+1} P_{m,n+1}(t) - (\lambda_n + \mu_n) P_{m,n}(t) \text{ for } n \geq 1, \end{aligned} \tag{2.1}$$

with boundary conditions $P_{m,m}(0) = 1$ and $P_{m,n}(0) = 0$ for $n \neq m$ (Feller, 1971).

Karlin and McGregor (1957b) show that for arbitrary starting state m, transition probabilities can be represented in the form

$$P_{m,n}(t) = \pi_n \int_0^\infty e^{-xt} Q_m(x) Q_n(x) \psi(\mathrm{d}x), \tag{2.2}$$

where $\pi_0 = 1$ and $\pi_n = (\lambda_0 \cdots \lambda_{n-1})/(\mu_1 \cdots \mu_n)$ for $n \geq 1$. Here, $\{Q_n(x)\}$ is a sequence of polynomials satisfying the three-term recurrence relation

$$\begin{aligned} \lambda_0 Q_1(x) &= \lambda_0 + \mu_0 - x, \text{ and} \\ \lambda_n Q_{n+1}(x) &= (\lambda_n + \mu_n - x)Q_n(x) - \mu_n Q_{n-1}(x), \end{aligned} \tag{2.3}$$

and ψ is the spectral measure of $\mathcal{X}$ with respect to which the polynomials $\{Q_n(x)\}$ are orthogonal. The system (5.2) has a unique solution if and only if

$$\sum_{k=0}^{\infty} \left(\pi_k + \frac{1}{\lambda_k \pi_k} \right) = \infty. \tag{2.4}$$

In what follows, we assume that the rate parameters $\{\lambda_n\}$ and $\{\mu_n\}$ satisfy (2.4). Closed-form solutions to (5.2) are available for a surprisingly small number of choices of $\{\lambda_n\}$ and $\{\mu_n\}$. We therefore need another approach to find useful formulae for computation of the transition probabilities.

2.2.2 Continued fraction representation of Laplace transform

To find an expression that is useful for computing $P_{m,n}(t)$ for an arbitrary general BDP, a fruitful approach is often to Laplace transform each equation of the system (5.2) and form a recurrence relationship relating back to the Laplace transform of $P_{m,n}(t)$. We base our presentation on that of Murphy and O'Donohoe (1975). Denote the Laplace transform of $P_{n,m}(t)$ as

$$f_{m,n}(s) = \mathcal{L}\left[P_{m,n}(t)\right](s) = \int_0^{\infty} e^{-st} P_{m,n}(t) \, \mathrm{d}t. \tag{2.5}$$

Applying the Laplace transform to (5.2), with the starting state $m = 0$, we arrive at

$$\begin{aligned} s f_{0,0}(s) - P_{0,0}(0) &= \mu_1 f_{0,1}(s) - \lambda_0 f_{0,0}(s), \text{ and} \\ s f_{0,n}(s) - P_{0,n}(0) &= \lambda_{n-1} f_{0,n-1}(s) + \mu_{n+1} f_{0,n+1}(s) - (\lambda_n + \mu_n) f_{0,n}(s) \end{aligned} \tag{2.6}$$

for $n \geq 1$. Rearranging and recalling that $P_{0,0}(0) = 1$ and $P_{0,n}(0) = 0$ for $n \geq 1$, we simplify (5.4) to

$$\begin{aligned} f_{0,1}(s) &= \frac{1}{\mu_1}\big[(s + \lambda_0) f_{0,0}(s) - 1\big], \text{and} \\ f_{0,n}(s) &= \frac{1}{\mu_n}\Big[(s + \lambda_{n-1} + \mu_{n-1}) f_{0,n-1}(s) - \lambda_{n-2} f_{0,n-2}(s)\Big] \text{for } n \geq 2. \end{aligned} \tag{2.7}$$

Some rearranging of (2.7) yields the forward system of recurrence relations

$$\begin{aligned} f_{0,0}(s) &= \frac{1}{s + \lambda_0 - \mu_1 \left(\frac{f_{0,1}(s)}{f_{0,0}(s)} \right)}, \text{ and} \\ \frac{f_{0,n}(s)}{f_{0,n-1}(s)} &= \frac{\lambda_{n-1}}{s + \mu_n + \lambda_n - \mu_{n+1} \left(\frac{f_{0,n+1}(s)}{f_{0,n}(s)} \right)}. \end{aligned} \tag{2.8}$$

Then combining these expressions, we arrive at the generalized continued fraction

$$f_{0,0}(s) = \cfrac{1}{s + \lambda_0 - \cfrac{\lambda_0 \mu_1}{s + \lambda_1 + \mu_1 - \cfrac{\lambda_1 \mu_2}{s + \lambda_2 + \mu_2 - \cdots}}}. \tag{2.9}$$

This is an exact expression for the Laplace transform of the transition probability $P_{0,0}(t)$. Let the partial numerators in (5.5) be $a_1 = 1$ and $a_n = -\lambda_{n-2}\mu_{n-1}$, and the partial denominators $b_1 = s + \lambda_0$ and $b_n = s + \lambda_{n-1} + \mu_{n-1}$ for $n \geq 2$. Then (5.5) becomes

$$f_{0,0}(s) = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdots}}}. \tag{2.10}$$

To express (3.6) in more typographically economical notation, we write

$$f_{0,0}(s) = \frac{a_1}{b_1+} \frac{a_2}{b_2+} \frac{a_3}{b_3+} \cdots . \tag{2.11}$$

We denote the kth convergent (approximant) of $f_{0,0}(s)$ as

$$f_{0,0}^{(k)}(s) = \frac{a_1}{b_1+} \frac{a_2}{b_2+} \cdots \frac{a_k}{b_k} = \frac{A_k(s)}{B_k(s)}. \tag{2.12}$$

There are deep connections between the orthogonal polynomial representation (2.3), Laplace transforms (2.7), and continued fractions of the form (5.5) that are beyond the scope of this paper (Karlin and McGregor, 1957b; Bordes and Roehner, 1983; Guillemin and Pinchon, 1999). Interestingly, Flajolet and Guillemin (2000) demonstrate a close relationship between the Laplace transforms of transition probabilities and state paths of the underlying Markov chain.

Before stating a theorem supporting this representation, we give two lemmas that will be useful in what follows.

Lemma 1. *Both the numerator A_k and denominator B_k of* (3.7) *satisfy the same recurrence, due to Wallis (1695):*

$$\begin{aligned} A_k &= b_k A_{k-1} + a_k A_{k-2}, \text{ and} \\ B_k &= b_k B_{k-1} + a_k B_{k-2}, \end{aligned} \tag{2.13}$$

with $A_0 = 0$, $A_1 = a_1$, $B_0 = 1$, and $B_1 = b_1$.

Lemma 2. *By repeated application of Lemma 1, we arrive at the determinant formula*

$$\begin{aligned} A_k B_{k-1} - A_{k-1} B_k &= (b_k A_{k-1} + a_k A_{k-2}) B_{k-1} - A_{k-1}(b_k B_{k-1} + a_k B_{k-2}) \\ &= -a_k (A_{k-1} B_{k-2} - A_{k-2} B_{k-1}) \\ &= (-1)^{k-1} \prod_{i=1}^{k} a_i. \end{aligned} \tag{2.14}$$

Now we state and prove a theorem giving expressions for the Laplace transform of $P_{m,n}(t)$. Although Murphy and O'Donohoe (1975) first report this result, they do not provide a detailed derivation in their paper.

Theorem 1. *The Laplace transform of the transition probability $P_{m,n}(t)$ is given by*

$$f_{m,n}(s) = \begin{cases} \left(\prod_{j=n+1}^{m} \mu_j \right) \dfrac{B_n(s)}{B_{m+1}(s)+} \dfrac{B_m(s) a_{m+2}}{b_{m+2}+} \dfrac{a_{m+3}}{b_{m+3}+} \cdots & \text{for } n \le m, \\ \left(\prod_{j=m}^{n-1} \lambda_j \right) \dfrac{B_m(s)}{B_{n+1}(s)+} \dfrac{B_n(s) a_{n+2}}{b_{n+2}+} \dfrac{a_{n+3}}{b_{n+3}+} \cdots & \text{for } m \le n, \end{cases} \tag{2.15}$$

where a_n, b_n, and B_n are as defined above.

Proof. To simplify notation, we sometimes omit the dependence of f_k, A_k, and B_k on the Laplace variable s. Suppose the process starts at $X(0) = m$. We can re-write the Laplace-transformed equations (5.4) with $P_{m,m}(0) = 1$ and $P_{m,n}(0) = 0$ for all $n \neq m$ as

$$s f_{m,0}(s) - \delta_{m0} = \mu_1 f_{m,1}(s) - \lambda_0 f_{m,0}(s), \tag{2.16a}$$

$$s f_{m,n}(s) - \delta_{mn} = \lambda_{n-1} f_{m,n-1}(s) + \mu_{n+1} f_{m,n+1}(s) - (\lambda_n + \mu_n) f_{m,n}(s), \tag{2.16b}$$

where $\delta_{mn} = 1$ if $m = n$ and zero otherwise. We first derive the expression for $n \leq m$. If $m = 0$, $f_{0,0}(s)$ is given by (2.11), so we assume in what follows when $n \leq m$, that $m \geq 1$. Rearranging (2.16a), we see that since $B_0 = 1$ and $s + \lambda_0 = b_1 = B_1$,

$$f_{m,0} = \frac{B_0}{B_1}\mu_1 f_{m,1}. \tag{2.17}$$

Now, to show the general case by induction, assume that for $n \leq m$,

$$f_{m,n-1} = \frac{B_{n-1}}{B_n}\mu_n f_{m,n}. \tag{2.18}$$

Substituting (2.18) into (4.58) when $n < m$, we have

$$b_{n+1}f_{m,n} = \lambda_{n-1}\frac{B_{n-1}}{B_n}\mu_n f_{m,n} + \mu_{n+1}f_{m,n+1} \tag{2.19}$$

$$\left(b_{n+1} + a_{n+1}\frac{B_{n-1}}{B_n}\right) f_{m,n} = \mu_{n+1}f_{m,n+1} \tag{2.20}$$

$$f_{m,n} = \frac{B_n}{B_{n+1}}\mu_{n+1}f_{m,n+1} \tag{2.21}$$

and so (2.18) is true for any $n < m$. Letting $n = m$, we have by (2.18) and (4.58),

$$b_{m+1}f_{m,m} = 1 + \lambda_{m-1}\left(\frac{B_{m-1}}{B_m}\mu_m f_{m,m}\right) + \mu_{m+1}f_{m,m+1}. \tag{2.22}$$

Recalling that $s + \lambda_m + \mu_m = b_{m+1}$ and using Lemma 1,

$$\mu_{m+1}f_{m,m+1} = 1 - \frac{B_{m+1}}{B_m}f_{m,m}. \tag{2.23}$$

Rearranging the previous equation, we find that

$$f_{m,m} = \frac{1}{\frac{B_{m+1}}{B_m} + \mu_{m+1}\frac{f_{m,m+1}}{f_{m,m}}}. \tag{2.24}$$

Likewise, we can write (4.58) as a continued fraction recurrence:

$$\frac{f_{m,n}}{f_{m,n-1}} = \frac{\lambda_{n-1}}{s + \mu_n + \lambda_n + \mu_{n+1}\frac{f_{m,n+1}}{f_{m,n}}}. \tag{2.25}$$

Then plugging (2.25) into (2.24) and iterating, we obtain the continued fraction for $f_{m,m}$:

$$\begin{aligned} f_{m,m} &= \frac{1}{\frac{B_{m+1}}{B_m}+}\frac{a_{m+2}}{b_{m+2}+}\frac{a_{m+3}}{b_{m+3}+}\cdots \\ &= \frac{B_m}{B_{m+1}+}\frac{B_m a_{m+2}}{b_{m+2}+}\frac{a_{m+3}}{b_{m+3}+}\cdots. \end{aligned} \tag{2.26}$$

This is an exact formula for the Laplace transform of $P_{m,m}(t)$, and proves the case $m = n$. For $n \leq m$, we iterate (2.18) to get

$$\begin{aligned} f_{m,n} &= \frac{B_n}{B_{n+1}} \mu_{n+1} f_{m,n+1} \\ &= \frac{B_n}{B_{n+1}} \frac{B_{n+1}}{B_{n+2}} \mu_{n+1}\mu_{n+2} f_{m,n+2} \\ &= \frac{B_n}{B_{n+1}} \frac{B_{n+1}}{B_{n+2}} \cdots \frac{B_{m-1}}{B_m} \mu_{n+1}\mu_{n+2}\cdots\mu_m f_{m,m} \\ &= \left(\prod_{j=n+1}^{m} \mu_j \right) \frac{B_n}{B_m} f_{m,m}. \end{aligned} \tag{2.27}$$

Substituting (2.26) for $f_{m,m}$ completes the proof for $n \leq m$.

To find the formula for $f_{m,n}$ when $n > m$, we adopt a similar approach. From (2.24) we arrive at

$$B_{m+1} f_{m,m} = B_m - B_m \mu_{m+1} f_{m,m+1}. \tag{2.28}$$

We proceed inductively. Assume that for $n > m$,

$$B_{n+1} f_{m,n} = \left(\prod_{j=m}^{n-1} \lambda_j \right) B_m + \mu_{n+1} B_n f_{m,n+1}. \tag{2.29}$$

From (4.58), we have

$$b_{n+2} f_{m,n+1} = \lambda_n f_{m,n} + \mu_{n+2} f_{m,n+2}. \tag{2.30}$$

Solving for $f_{m,n}$ in (2.29) and plugging this into the above equation, we have

$$b_{n+2} f_{m,n+1} = \lambda_n \left(\prod_{j=m}^{n-1} \lambda_j \right) \frac{B_m}{B_{n+1}} + \lambda_n \mu_{n+1} \frac{B_n}{B_{n+1}} f_{m,n+1} + \mu_{n+2} f_{m,n+2}. \tag{2.31}$$

Recalling that $-\lambda_n \mu_{n+1} = a_{n+2}$,

$$\left(b_{n+2} B_{n+1} + a_{n+2} B_m\right) f_{m,n+1} = \left(\prod_{j=m}^{n} \lambda_j \right) B_n + \mu_{n+2} B_{n+1} f_{m,n+2}, \tag{2.32}$$

and by Lemma 1,

$$B_{n+2} f_{m,n+1} = \left(\prod_{j=m}^{n} \lambda_j \right) B_m + \mu_{m+2} B_{n+1} f_{m,m+2}. \tag{2.33}$$

This establishes the recurrence (2.29). Then for any $n \geq m$, we can rearrange (2.29) to obtain

$$f_{m,n} = \left(\prod_{j=m}^{n-1} \lambda_j \right) \frac{B_m}{B_{n+1} - B_n \mu_{n+1} \frac{f_{m,n+1}}{f_{m,n}}}. \tag{2.34}$$

This completes the proof. □ □

2.2.3 Obtaining transition probabilities

Murphy and O'Donohoe (1975) find transition probabilities by truncating (2.15) at a pre-specified depth, forming a partial fractions sum, and inverse transforming. Parthasarathy and Sudhesh (2006a) give a series solution for transition probabilities based on an equivalence between continued fractions like (2.15) and power series. However, both of these approaches suffer from serious drawbacks, as we explore in detail in the Appendix.

We instead seek an efficient and robust numerical method for evaluating and inverting (2.15). We first note that continued fractions typically converge rapidly, and in our experience, evaluation of (2.15) is very fast and stable using the Lentz algorithm and its subsequent improvements (Lentz, 1976; Thompson and Barnett, 1986; Press, 2007). We therefore invert (2.15) numerically by a summation formula.

To do this, we treat the continued fraction representation (2.15) of the Laplace transform of $P_{m,n}(t)$ as an unknown but computable function of the complex Laplace variable s. We base our presentation on that of Abate and Whitt (1992a). If ϵ is a positive real number such that all singularities of $f_{m,n}(s)$ lie to the left of ϵ in the complex plane, the inverse Laplace transform of $f_{m,n}(s)$ is given by the Bromwich integral

$$P_{m,n}(t) = \mathcal{L}^{-1}\left(f_{m,n}(s)\right) = \frac{1}{2\pi i} \int_{\epsilon - i\infty}^{\epsilon + i\infty} e^{st} f_{m,n}(s) \, \mathrm{d}s. \tag{2.35}$$

Letting $s = \epsilon + iu$,

$$\begin{aligned} P_{m,n}(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{(\epsilon+iu)t} f_{m,n}(\epsilon+iu)\ \mathrm{d}u \\ &= \frac{e^{\epsilon t}}{2\pi}\int_{-\infty}^{\infty} \big[\cos(ut) + i\sin(ut)\big] f_{m,n}(\epsilon+iu)\ \mathrm{d}u \\ &= \frac{e^{\epsilon t}}{2\pi}\Bigg[\int_{-\infty}^{\infty} \Big[\mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut) - \mathrm{Im}\big(f_{m,n}(\epsilon+iu)\big)\sin(ut)\Big]\ \mathrm{d}u \\ &\quad + i\int_{-\infty}^{\infty} \Big[\mathrm{Im}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut) + \mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\sin(ut)\Big]\ \mathrm{d}u\Bigg], \end{aligned} \tag{2.36}$$

but $P_{m,n}(t)$ is real-valued, so the imaginary part of the last equality in (2.36) is zero. Then

$$P_{m,n}(t) = \frac{e^{\epsilon t}}{2\pi}\int_{-\infty}^{\infty} \Big[\mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut) - \mathrm{Im}\big(f_{m,n}(\epsilon+iu)\big)\sin(ut)\Big]\ \mathrm{d}u. \tag{2.37}$$

But since $P_{m,n}(t) = 0$ for $t < 0$, we also have that

$$\int_{-\infty}^{\infty} \Big[\mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut) + \mathrm{Im}\big(f_{m,n}(\epsilon+iu)\big)\sin(ut)\Big]\ \mathrm{d}u = 0. \tag{2.38}$$

Then applying (2.38) to (2.37), we obtain

$$P_{m,n}(t) = \frac{e^{\epsilon t}}{\pi}\int_{-\infty}^{\infty} \mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut)\ \mathrm{d}u. \tag{2.39}$$

Finally, we note that since

$$\mathrm{Re}\big(f(\epsilon-iu)\big) = \int_0^{\infty} e^{-\epsilon t}\cos(ut)P_{m,n}(t)\ \mathrm{d}t = \mathrm{Re}\big(f(\epsilon+iu)\big), \tag{2.40}$$

it must be the case that $\mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)$ is even in u for every ϵ. Therefore,

$$P_{m,n}(t) = \frac{2e^{\epsilon t}}{\pi}\int_0^{\infty} \mathrm{Re}\big(f_{m,n}(\epsilon+iu)\big)\cos(ut)\ \mathrm{d}u. \tag{2.41}$$

Following Abate and Whitt (1992a), we approximate the integral above by a discrete Riemann sum via the trapezoidal rule with step size h:

$$\begin{aligned} P_{m,n}(t) &\approx \frac{he^{\epsilon t}}{\pi}\mathrm{Re}\left(f_{m,n}(\epsilon)\right) + \frac{2he^{\epsilon t}}{\pi}\sum_{k=1}^{\infty}\mathrm{Re}\left(f_{m,n}(\epsilon+ikh)\right)\cos(kht) \\ &= \frac{e^{A/2}}{2t}\mathrm{Re}\left(f_{m,n}\left(\frac{A}{2t}\right)\right) + \frac{e^{A/2}}{t}\sum_{k=1}^{\infty}(-1)^k\mathrm{Re}\left(f_{m,n}\left(\frac{A+2k\pi i}{2t}\right)\right), \end{aligned} \tag{2.42}$$

where the second line is obtained by setting $h = \pi/(2t)$ and $\epsilon = A/(2t)$; this change of variables eliminates the cosine term.

2.2.4 Numerical considerations

While (4.66) presents a method for numerical solution of the transition probabilities $P_{m,n}(t)$ for a BDP with arbitrary birth and death rates, it is not yet an algorithm for reliable evaluation of these probabilities. In order to develop a reliable numerical method, we must: 1) characterize the error introduced by discretization of the integral in (2.41); 2) determine a suitable method to evaluate this nearly alternating sum while controlling the error; and 3) accurately and rapidly evaluate the infinite continued fraction in (2.15).

Abate and Whitt show that the discretization error that arises in (4.66) is

$$e_d = \sum_{k=1}^{\infty} e^{-kA} P_{m,n}\big((2k+1)t\big), \tag{2.43}$$

and when $P_{m,n}(t) \leq 1$,

$$e_d \leq \sum_{k=1}^{\infty} e^{-kA} = \frac{e^{-A}}{1-e^{-A}} \approx e^{-A}, \tag{2.44}$$

when e^{-A} is small. Then to obtain $e_d \leq 10^{-\gamma}$, we set $A = \gamma \log(10)$. As Abate and Whitt point out, the terms of the series (4.66) alternate in sign when

$$\mathrm{Re}\left(f_{m,n}\left(\frac{A+2k\pi i}{2t}\right)\right) \tag{2.45}$$

has constant sign. This suggests that a series acceleration method may be helpful in keeping the terms of the sum manageable and avoiding roundoff error due to summands of alternating sign. We opt to use the Levin transform for this purpose (Levin, 1973; Press, 2007; Numerical Recipes Software, 2007).

Evaluation of rational approximations to continued fractions by repeated application of Lemma 1 is appealing, but suffers from roundoff error when denominators are small (Press, 2007). To evaluate the infinite continued fraction in the summand of (4.66), we use the modified Lentz method (Lentz, 1976; Thompson and Barnett, 1986; Press, 2007). To demonstrate, suppose we wish to approximate the value of $f_{0,0}(s)$, given by (5.5) by truncating at depth k. Then

$$f_{0,0}^{(k)}(s) = \frac{A_k(s)}{B_k(s)} \tag{2.46}$$

is the kth rational approximant to the infinite continued fraction $f_{0,0}(s)$. In the modified Lentz method, we stabilize the computation by finding the ratios

$$C_k = \frac{A_k}{A_{k-1}} \quad \text{and} \quad D_k = \frac{B_{k-1}}{B_k} \tag{2.47}$$

so that $f_{0,0}^{(k)}$ can be found iteratively by

$$f_{0,0}^{(k)} = f_{0,0}^{(k-1)} C_k D_k. \tag{2.48}$$

Using Lemma 1, we can iteratively compute C_k and D_k via the updates

$$C_k = b_k + \frac{a_k}{C_{k-1}} \quad \text{and} \quad D_k = \frac{1}{b_k + a_k D_{k-1}}. \tag{2.49}$$

In practice, we must evaluate the continued fraction to only a finite depth, but we must evaluate to a depth sufficient to control the error. Suppose we wish to evaluate the infinite continued fraction $f_{0,0}(s)$ given by (5.5) at some complex number s. Intuitively, we wish to terminate the Lentz algorithm when the difference between successive convergents is small. However, it is not immediately clear how the difference between convergents $f_{0,0}^{(k)}(s) - f_{0,0}^{(k-1)}(s)$ is related to the absolute error $f_{0,0}(s) - f_{0,0}^{(k)}$. Craviotto et al (1993) make this relationship clear by furnishing an *a posteriori* truncation error bound for Jacobi fractions of the same form as (5.5) in this paper. Assuming that $f_{0,0}^{(k)}(s) = A_k(s)/B_k(s)$ converges to $f_{0,0}(s)$ as $k \to \infty$, Craviotto et al (1993) give the bound

$$\left| f_{0,0}(s) - f_{0,0}^{(k)}(s) \right| \leq \frac{\left| \frac{B_k(s)}{B_{k-1}(s)} \right|}{\left| \operatorname{Im}\left(\frac{B_k(s)}{B_{k-1}(s)} \right) \right|} \left| f_{0,0}^{(k)}(s) - f_{0,0}^{(k-1)}(s) \right|, \tag{2.50}$$

that is valid when $\operatorname{Im}(s)$ is nonzero. Note that $B_k(s)/B_{k-1}(s) = 1/D_k(s)$, so (2.50) is easy to evaluate during iteration under the Lentz algorithm. Therefore, we stop at depth k in the Lentz algorithm when

$$\frac{|1/D_k(s)|}{|\operatorname{Im}(1/D_k(s))|} \left| f_{0,0}^{(k)}(s) - f_{0,0}^{(k-1)}(s) \right| \tag{2.51}$$

is small.

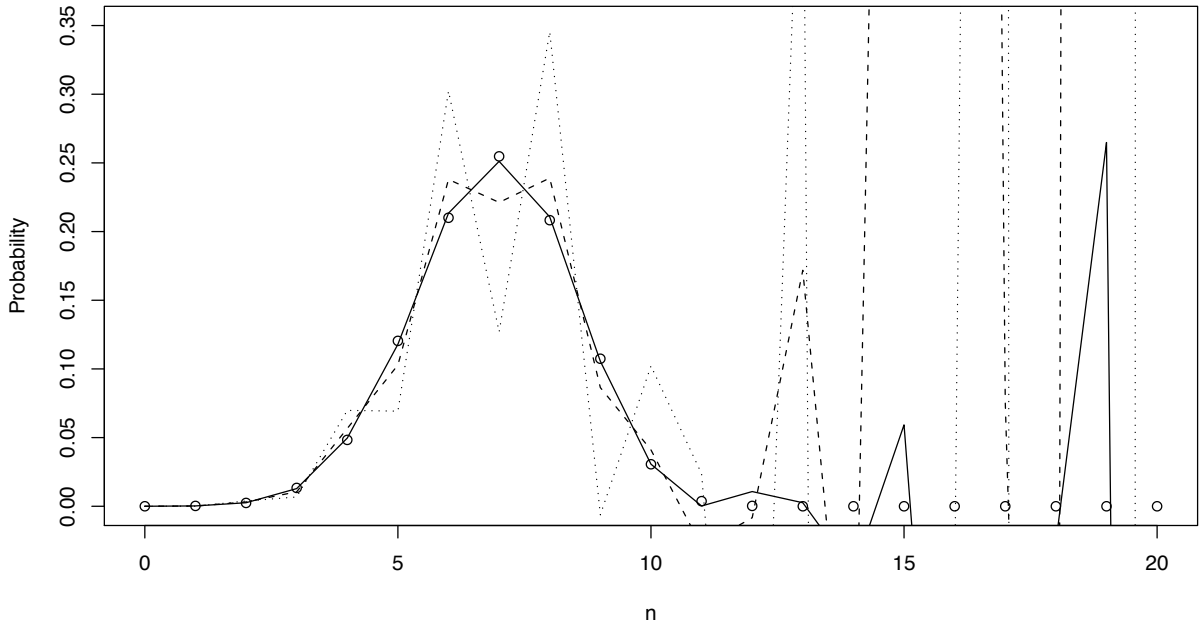


Figure 2.1: Comparison of transition probabilities $P_{10,n}(t = 1)$ computed by our error-controlled method and that of Murphy and O'Donohoe (1975) for the immigration-death model with $\lambda_n = 0.2$ and $\mu_n = 0.4n$. The open circles are the values given by our method. The solid line corresponds with the approximant method of Murphy and O'Donohoe with $k = 2$ (solid line), $k = 3$ (dashed line), and $k = 4$ (dotted line). In our experience, the approximant method fails whenever $n + m + k$ is greater than approximately 20. It is interesting to note that increasing the depth of truncation k in the approximant method actually worsens the approximation.

2.2.5 Numerical results

Although our error-controlled method is designed to be used when an analytic solution cannot be found, we seek to validate our numerical results by comparison to available analytic and numerical solutions. For the simple BDP with $\lambda_n = n\lambda$ and $\mu_n = n\mu$, our numerical results agree with the values from the well-known closed-form solution given explicitly in Bailey (1964) as

$$\begin{aligned} P_{m,n}(t) &= \sum_{j=0}^{\min(m,n)} \binom{m}{j} \binom{m+n-j-1}{m-1} \alpha^{m-j} \beta^{n-j} (1-\alpha-\beta)^j \\ P_{m,0}(t) &= \alpha^m \end{aligned} \tag{2.52}$$

where

$$\alpha = \frac{\mu\left(e^{(\lambda-\mu)t}-1\right)}{\lambda e^{(\lambda-\mu)t}-\mu} \quad \text{and} \quad \beta = \frac{\lambda\left(e^{(\lambda-\mu)t}-1\right)}{\lambda e^{(\lambda-\mu)t}-\mu}. \tag{2.53}$$

Murphy and O'Donohoe (1975) give numerical probabilities for four general birth-death models: a) immigration-death with $\lambda_n = 0.2$ and $\mu_n = 0.4n$; b) immigration-emigration with $\lambda_n = 0.3$, $\mu_0 = 0$, and $\mu_n = 0.1$; c) queue with $\lambda_n = 0.6$, $\mu_0 = 0$, $\mu_1 = \mu_2 = 0.2$, $\mu_3 = \mu_4 = 0.4$, and $\mu_n = 0.6$ for $n \geq 5$; and d) $\lambda_n = 0.4$, $\mu_n = 0.1\sqrt{n}$. Our results agree with those computed by Murphy and O'Donohoe for each of the four models given in Tables 2 through 7 in their paper (Murphy and O'Donohoe, 1975). We note that Murphy and O'Donohoe did not report probabilities for $m > 2$ or $n > 5$ in any of their four models. In our experience, their method performs poorly when $n+m+k$ is greater than approximately 20.

As a demonstration of the instability of the approximant method, we contrast the numerical results given by our error-controlled method with those obtained using the approximant method, that we implemented as described in Murphy and O'Donohoe (1975), except for some rescaling of intermediate quantities to avoid obvious sources of roundoff error. Figure 2.1 shows this comparison, using model (a) above, for three values of the truncation index k. Note that increasing the truncation depth k in the approximant method does not improve the error.

2.3 Applications

Drawing on the robustness and generality of our error-controlled method, we conclude with four models in ecology, genetics, and evolution whose analytic solutions remain elusive and where past numerical approaches have fallen short. Using our approach, computation of transition probabilities is straightforward, and the techniques outlined above may be used without modification. Some of the examples are well-known models, and others are novel. In some cases, the orthogonal polynomials satisfying (2.3) are known, and hence a solution could be numerically computed using (2.2), provided there are good ways of evaluating the polynomials. Often, a severe drawback of using known orthogonal polynomials to compute a solution based on (2.2) is that the polynomials are model-specific. This makes experimentation and model selection difficult, since computation of transition probabilities depends on a priori analytic information about the polynomials and measure associated with the BDP. Our method does not rely on a priori information about the process, other than the birth and death rates for each state.

2.3.1 Immigration and emigration

Consider a population model for the number of organisms in an area, and suppose new immigrants arrive at rate ν, and emigrants leave at rate γ. Organisms living in the area reproduce with per-capita birth rate λ and die with rate μ. Define the linear rates

$$\lambda_n = n\lambda + \nu \quad \text{and} \quad \mu_n = n\mu + \gamma. \tag{2.54}$$

For the case $\gamma = 0$, an analytic expression for the orthogonal polynomials is known (Karlin and McGregor, 1958a). For nonzero γ, orthogonal polynomials are available from which a solution of the form (2.2) may be computed (Karlin and McGregor, 1958a; Ismail et al, 1988). However, using our error-controlled method, we can easily find the transition probabilities without additional analytic information. Figure 2.2 shows an example of the time-evolution of $P_{10,n}(t)$ for various times t and states n, with the parameters $\lambda = 0.5$, $\nu = 0.2$, $\mu = 0.3$, and $\gamma = 0.1$. The approximant method method of Murphy and O'Donohoe fails to produce

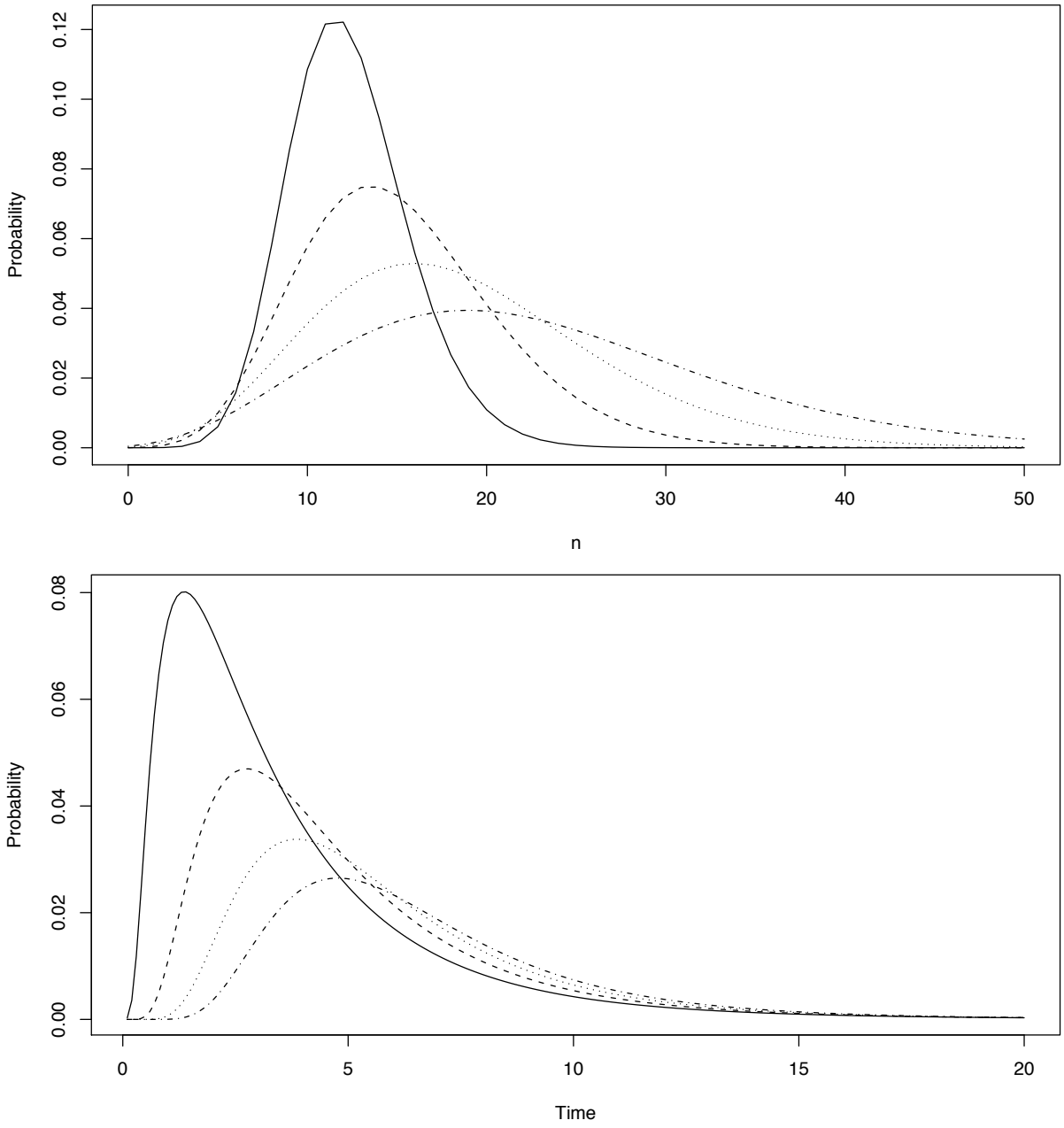


Figure 2.2: Transition probabilities for the immigration/emigration model with $\lambda = 0.5$, $\nu = 0.2$, $\mu = 0.3$, and $\gamma = 0.1$. The top panel shows $P_{10,n}(t)$ with $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), and $t = 4$ (dash-dotted line) for $n = 0, \ldots, 50$. The bottom panel shows $P_{10,n}(t)$ with $n = 15$ (solid line), $n = 20$ (dashed line), $n = 25$ (dotted line), and $n = 30$ (dash-dotted line) for $t \in (0, 20)$.

useful probabilities for $n > 10$ (not shown).

2.3.2 Logistic growth with Allee effects

Populations of organisms that occupy a finite space may be subject to various constraints on their growth. The per-capita birth rate may decline when there are more organisms than the ecosystem can sustain (Tan and Piantadosi, 1991). This can happen when there are too many organisms competing for the same food supply. The decay of population size above some carrying capacity is usually called logistic growth by ecologists. Another density-dependent constraint is known as the Allee effect, in which per-capita birth rate increases superlinearly with n once a small population has been established, due to favorable consequences of density, such as cooperation and mutual protection from predators (Allee et al, 1949). As a realistic example of a general BDP that has no obvious solution by orthogonal polynomials, we seek a model that both transiently supports growth above the carrying capacity, and reflects these two density-dependent constraints, similar in spirit to models described by Tan and Piantadosi (1991) and Dennis (2002).

Qualitatively, if the per-capita birth rate with no density effects is λ, then the total birth rate should rise faster than $n\lambda$ when n is small, slower than $n\lambda$ for intermediate n near the carrying capacity, and should decay toward zero for n greater than the carrying capacity. Tan and Piantadosi introduce a logistic birth rate $\lambda_n = n\lambda\left(1 - \frac{n}{N}\right)$ for a finite state space model that takes values $\{0, 1, \ldots, N\}$. However, to allow for temporary growth beyond the carrying capacity, we choose $\lambda_n \propto \lambda n^2 e^{-\alpha n}$ for intermediate and large n. To achieve attenuated growth for small n as well, we scale this rate by a logistic function, yielding

$$\lambda_n = \frac{\lambda n^2 e^{-\alpha n}}{1 + e^{\beta(n-M)}} \quad \text{and} \quad \mu_n = n\mu, \tag{2.55}$$

where M is the population size with highest birth rate, and the death rate is assumed to be proportional only to the number of existing individuals. Figure 2.3 shows the resulting rates for various states n, with the different phases of population change shaded. To illustrate that the model produces the desired behavior, several realizations of the process are given in the lower panel for various starting values. The shaded regions correspond with the three

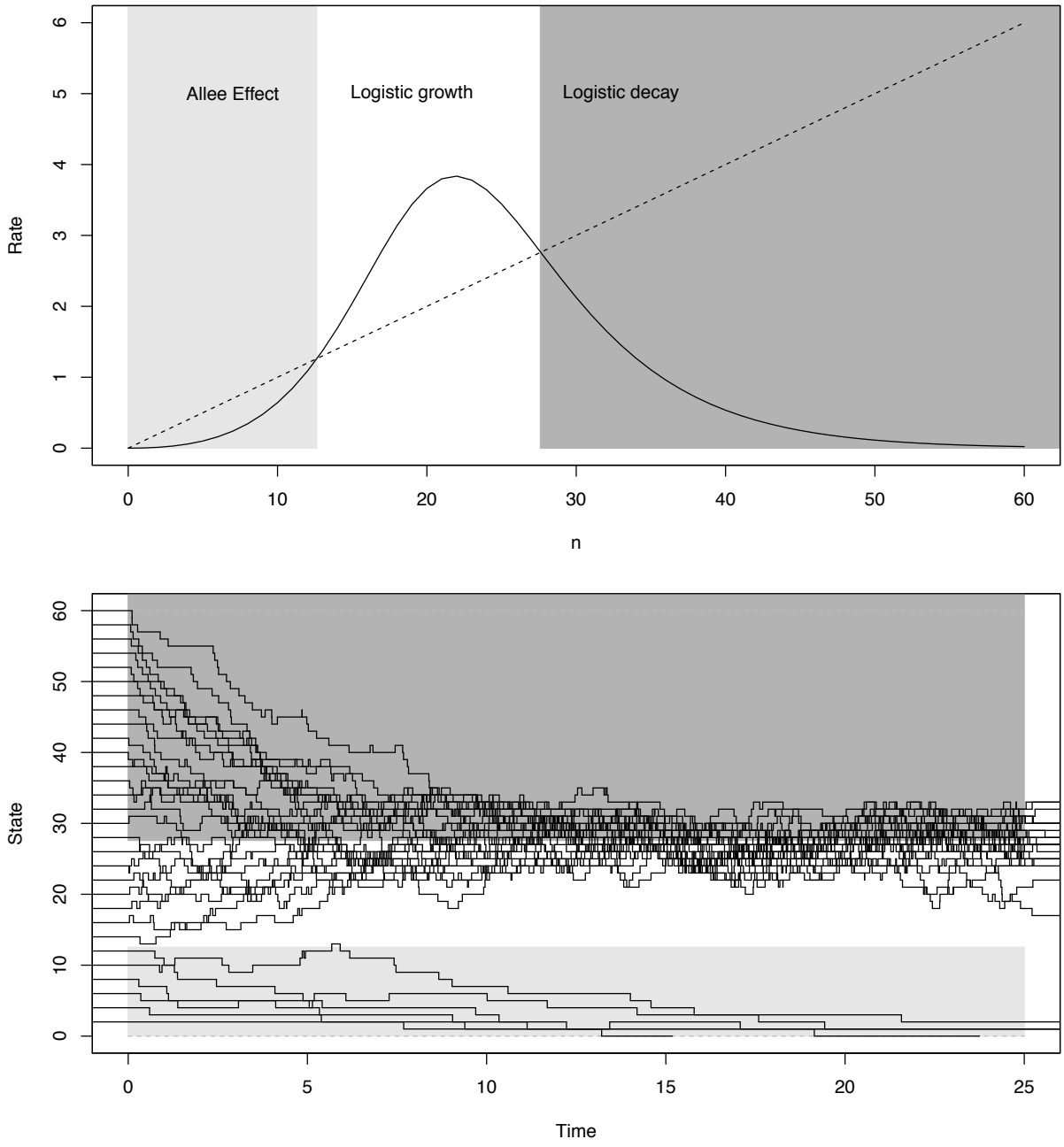


Figure 2.3: Behavior of logistic/Allee model. The upper panel shows a plot of birth (solid line) and death (dashed line) rates for states $n = 0, \ldots, 60$, and parameters $\lambda = 1$, $\mu = 0.1$, $M = 20$, $\alpha = 0.2$, and $\beta = 0.3$. The different phases of growth are labeled in the shaded regions. The lower panel shows stochastic realizations of the logistic/Allee model for various starting values. The shaded regions correspond with the shaded phases of growth in the upper panel.

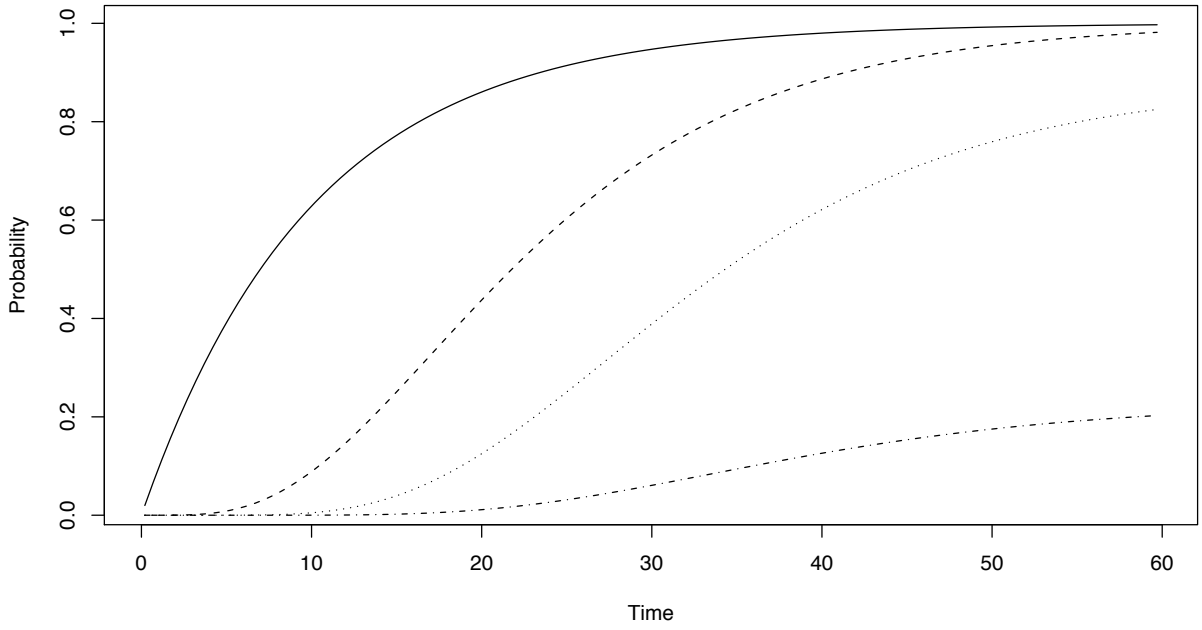


Figure 2.4: Logistic/Allee model probabilities of extinction $P_{m,0}(t)$ for initial population sizes $m = 1$ (solid line), $m = 5$ (dashed line), $m = 10$ (dotted line), and $m = 15$ (dash-dotted line). The full model parametrization is found in the text.

phases of growth. Note that most paths in the lower panel of Figure 2.3 center near $n = 27$, where the birth rate and death rate are equal. The lower panel corresponds with Figure 1 in Dennis (2002). Figure 2.4 demonstrates the success of the error-controlled method in computing time-dependent extinction probabilities $P_{m,0}$ for various starting values with $\lambda = 1$, $\alpha = 0.2$, $\beta = 0.3$, $M = 20$, and $\mu = 0.1$.

2.3.3 Moran models with mutation and selection

The probability of fixation or extinction of an allele in finite populations is frequently of interest to researchers in genetics. However, publications often rely on the probability of eventual extinction $P_{m,0}(t \to \infty)$, or the probability of fixation of a novel mutation in a population of constant size N, $P_{1,N}(t \to \infty)$. While these asymptotic probabilities do reveal important properties of the underlying models, the information they provide about the distribution of time to fixation/extinction is incomplete. In practice, researchers may

observe that m organisms in a sample exhibit a certain trait at a certain time. Then $P_{m,0}(t)$, the probability of extinction of that trait at *finite* times t in the future should presumably be of great interest, since researchers cannot reliably observe the process for infinitely long times. Additionally, the finite-time probability of fixation/extinction may exhibit threshold effects or unexpected dynamics that are not revealed by the asymptotic probability of such an event.

Moran (1958) introduces a model for the time-evolution of a biallelic locus when the population size is constant through time. A biallelic locus is a location in an organism's genome in which two different genetic variants or alleles exist in a population. We are interested in how the number of individuals carrying each allele changes from generation to generation. Krone and Neuhauser (1997) exploit the Moran model to derive a BDP counting the number of individuals with a certain allele in the context of ancestral genealogy reconstruction in which one allele offers a selective advantage to individuals that carry it. Selection greatly complicates the problem and remains an active area of research. In a limiting case, this process corresponds to Kingman's coalescent process when there is no mutation or selection (Kingman, 1982a,b).

To construct the Moran process with mutation and selection, suppose a finite population of N haploid organisms has 2 alleles at a certain locus: A_1 and A_2. Individuals that carry A_1 reproduce at rate α and A_2 individuals reproduce at rate β. Suppose further that individuals carrying the A_1 allele have a selective advantage over individuals carrying A_2, so $\alpha > \beta$. When an individual dies, it is replaced by the offspring of a random parent chosen from all N individuals, including the one that dies. This parent contributes a gamete carrying its allele that is also subject to mutation. Mutation from A_1 to A_2 happens with probability u and in reverse with rate v. The new offspring receives the possibly mutated haplotype and the process continues.

Let $X(t)$ be a BDP counting the number of A_1 individuals on the state space $n \in \{0, \ldots, N\}$. To construct the transition rates of the process, suppose there are currently n individuals of type A_1. We first consider the addition of a new individual of type A_1, so that $n \to n+1$. For this to happen, the individual that dies must be of type A_2. If the

parent of the replacement is one of the n of type A_1, the parent contributes its allele without mutation, and this happens with probability $1-u$. If the parent of the replacement is one of the $N-n$ of type A_2, the parent contributes its allele, which then mutates with probability v. Therefore, the total rate of addition is

$$\lambda_n = \frac{N-n}{N}\left[\alpha\frac{n}{N}(1-u) + \beta\frac{N-n}{N}v\right], \tag{2.56}$$

for $n = 0, \ldots, N$ with $\lambda_n = 0$ when $n > N$. Likewise, the removal of an individual of type A_1 can happen when one of the n individuals of type A_1 is chosen for replacement. If the parent of the replacement is one of the $N-n$ of type A_2, the parent contributes A_2 without mutation, with probability $1-v$. If the parent is one of the n of type A_1, the allele must mutate to A_2 with probability u. The total rate of removals becomes

$$\mu_n = \frac{n}{N}\left[\beta\frac{N-n}{N}(1-v) + \alpha\frac{n}{N}u\right], \tag{2.57}$$

for $n = 1, \ldots, N$ with $\mu_0 = \lambda_N = 0$ and $\mu_n = 0$ when $n > N$. Note that if $v > 0$, then $\lambda_0 > 0$ so the A_1 allele cannot go extinct. Also, if $u > 0$, then $\mu_N > 0$, so the A_1 allele cannot be fixed in the population.

Karlin and McGregor (1962) derive the relevant polynomials and measure for the Moran process described above, but without selection, so that $\alpha = \beta$. Donnelly (1984) gives expressions for the transition probabilities in the case where $\alpha = \beta = 1$, noting that when selection is introduced (via differing α and β), his approach is no longer fruitful. Using our technique, computation of the transition probabilities under selection is straightforward. The upper panel of Figure 2.5 shows the probability of fixation by time t. The lower panel shows the finite-time fixation probability of A_1, $P_{m,100}(t)$, with $u = 0$ so the state $n = 100$ is absorbing.

Since the state space in the Moran model is finite, it is natural to consider the matrix exponentiation method discussed in the Introduction. We write the stochastic transition

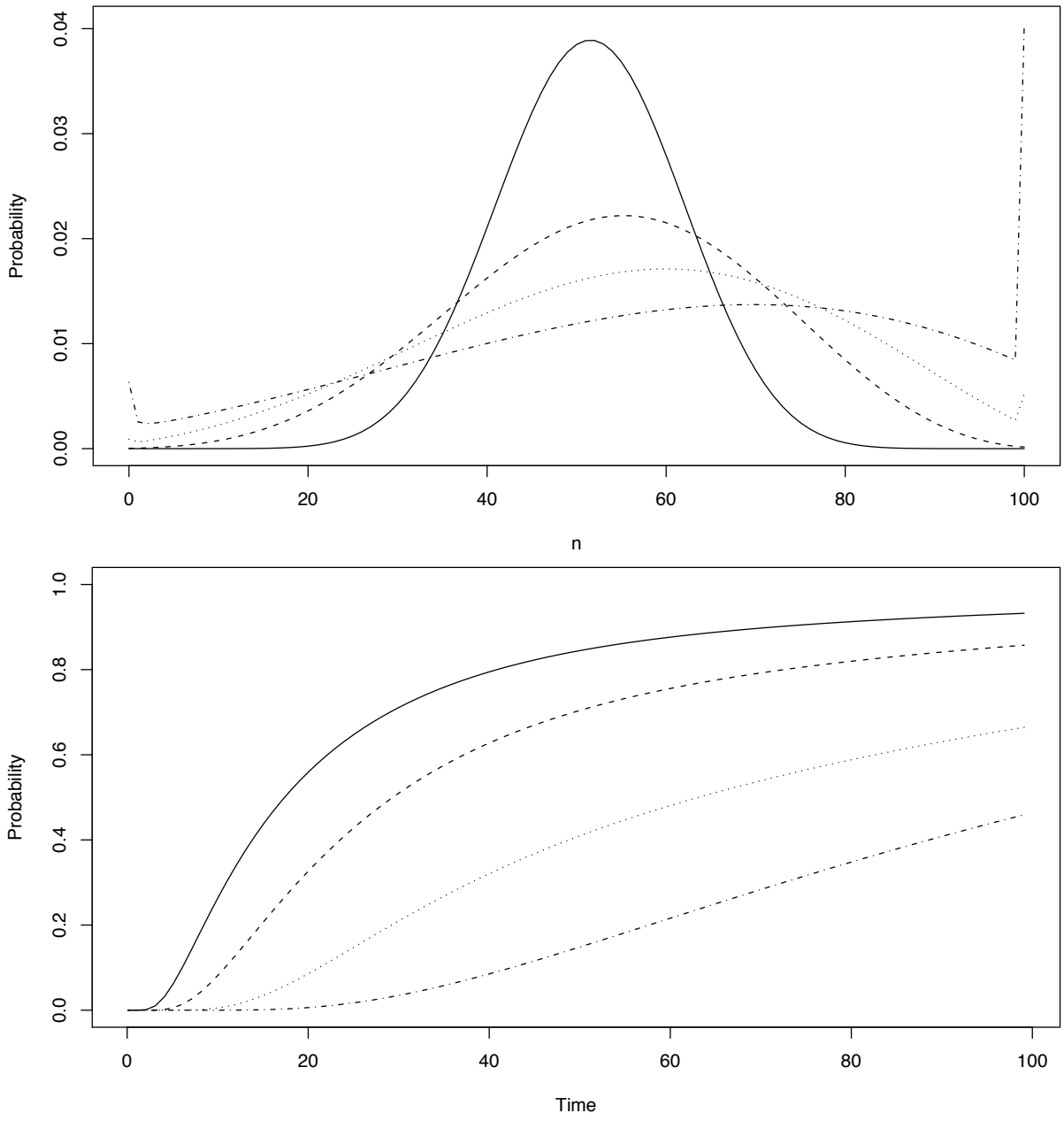


Figure 2.5: Transition probabilities for the Moran model with selection. The upper panel shows the probability of n individuals having allele A_1 at time t, $P_{50,n}(t)$ for the Moran model with $N = 100$, starting from $m = 50$ with $u = 0.02$, $v = 0.01$, $\alpha = 60$, and $\beta = 10$. We show the probabilities for $t = 1$ (solid line), $t = 3$ (dashed line), $t = 5$ (dotted line), $t = 8$ (dash-dotted line). Note that although the states 0 and 100 are not absorbing, the mutation rates u and v are small enough that probability accumulates significantly in these end states. Note also the asymmetry in the distribution at longer times. The lower panel reports the probability of fixation by time t, $P_{50,100}(t)$, for the same model, but with $u = 0$ so the state $n = 100$ is absorbing. The probabilities shown are for $m = 70$ (solid line), $m = 50$ (dashed line), $m = 20$ (dotted line), and $m = 1$ (dash-dotted line). Note the starkly different

matrix as

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_N & -(\lambda_N+\mu_N) & \lambda_N \end{pmatrix} \tag{2.58}$$

where λ_n and μ_n are defined by (2.56) and (2.57), respectively. In our experience, the matrix exponentiation method often works well, and its computational cost is similar to that of our error-controlled method. However, it is highly sensitive to rate matrix conditioning. For example, Figure 2.6 shows a comparison of transition probabilities from the error-controlled method and the matrix exponentiation method for the Moran model with $N = 100$, $\alpha = 210$, $\beta = 20$, $u = 0.002$, and $v = 0$. In evolutionary terms, this means that mutation from A_1 to A_2 is impossible, and the A_2 haplotype suffers from low fitness. Computationally, this has the effect of making μ_n small for most n, and hence the rate matrix grows ill-conditioned.

Although the rate matrix in this example is nearly defective, this choice of parameter values is not unreasonably extreme. For example, researchers in population genetics often wish to test the hypothesis that selection occurs in a dataset. They fit parameters for models with selection (full model) and without selection (restricted model) and perform a likelihood ratio test of this hypothesis. If the estimates of β and u in the full model are small, they may be unable to reliably compute the probability (likelihood) of the data, given the estimated parameter values under the full model.

2.3.4 A frameshift-aware indel model

Thorne et al (1991) introduce a BDP modeling insertion and deletion of nucleotides in DNA for applications in molecular evolution. The authors model the process of sequence length evolution by assuming that a new nucleotide can be inserted adjacent to every existing nucleotide, and every existing nucleotide is subject to deletion, at a constant per-nucleotide rate. This corresponds to the simple BDP with $\lambda_n = n\lambda$ and $\mu_n = n\mu$. If a sequence has

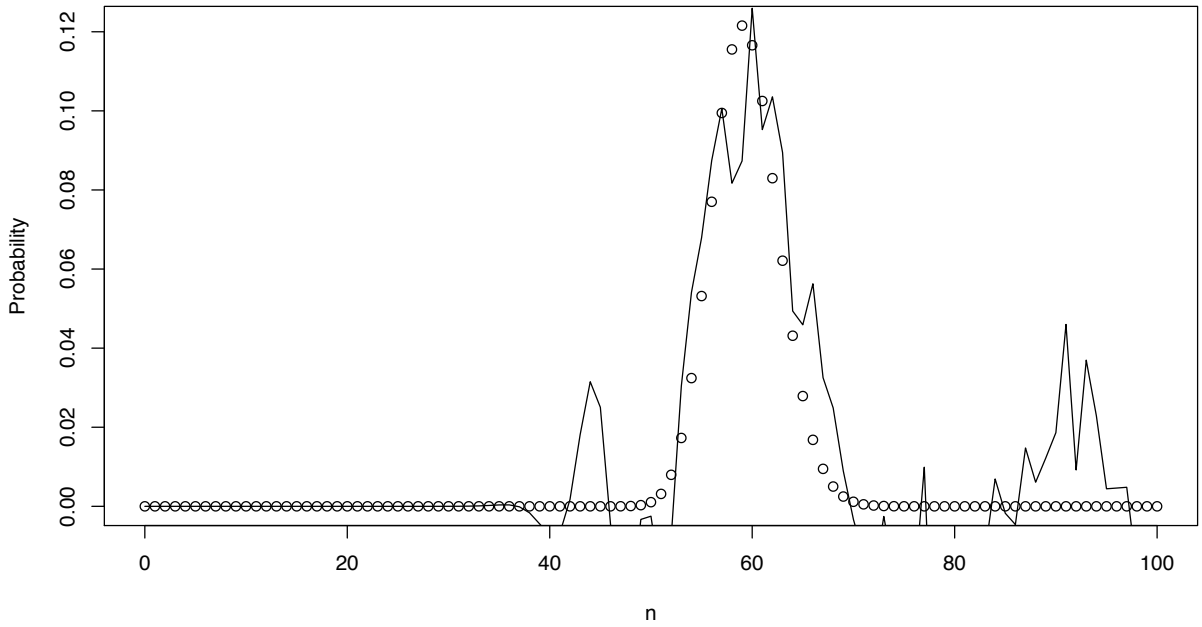


Figure 2.6: Comparison of Moran model transition probabilities $P_{50,n}(t = 0.2)$ computed by two methods with $N = 100$, $\alpha = 210$, $\beta = 20$, $u = 0.002$, and $v = 0$. The open circles correspond with our error-controlled method, and the solid line corresponds with the matrix exponentiation method. This choice of parameters causes wild fluctuations in probabilities reported by the matrix exponentiation method since the stochastic rate matrix becomes nearly singular.

m nucleotides at time 0 and there are n nucleotides at time t later, the probability of this event is $P_{m,n}(t)$.

However, an important aspect of biological sequence evolution is conservation of the structure and biophysical properties of proteins that result from transcription and translation of DNA sequences. After coding DNA is transcribed into RNA, ribosomes translate 3-nucleotide chunks (codons) of the RNA into a single amino acid residue, that is then joined to the end of a growing protein polymer. Insertions or deletions (indels) in a DNA sequence that result in a shift in this triplet code are called "frame-shift" mutations. It is likely that a frame-shift indel occurring in a protein-coding DNA sequence results in a protein that is prematurely terminated or possesses structural and chemical characteristics unlike the ancestral protein. Insertions or deletions whose length is a multiple of three should be more common. We seek to model this behavior in a novel way: suppose the indel process is a BDP similar in spirit to the one presented by Thorne et al (1991), and the rate of insertion and deletion of nucleotides depends on the number of nucleotides already inserted, modulo (mod) 3:

$$\lambda_n = \begin{cases} n\beta_0 & \text{if } n-1 = 0 \mod 3 \\ n\beta_1 & \text{if } n-1 = 1 \mod 3 \\ n\beta_2 & \text{if } n-1 = 2 \mod 3 \end{cases} \quad \text{and} \quad \mu_n = \begin{cases} n\gamma_0 & \text{if } n-1 = 0 \mod 3 \\ n\gamma_1 & \text{if } n-1 = 1 \mod 3 \\ n\gamma_2 & \text{if } n-1 = 2 \mod 3 \end{cases} . \tag{2.59}$$

Here we assume that $\beta_2 > \beta_0, \beta_1$, and $\gamma_1 > \gamma_0, \gamma_2$ so that transitions to state n such that $n - 1 = 0 \mod 3$ occur at a faster rate per nucleotide. The linear-periodic nature of these birth and death rates make solution of the orthogonal polynomials and measure corresponding with this BDP difficult. The approximant method of Murphy and O'Donohoe also fails here for large n. However, using our error-controlled method, numerical results are readily available. Figure 2.7 shows $P_{1,n}(t)$ for $n = 0, \ldots, 50$ at various times t. Note that the distribution of the number of inserted bases has peaks at the integers mod three. Finally, it is worth noting that the dearth of tractable BDPs for indel events has been a major deterrent in statistical sequence alignment and we are actively exploring solutions to this problem using our error-controlled method.

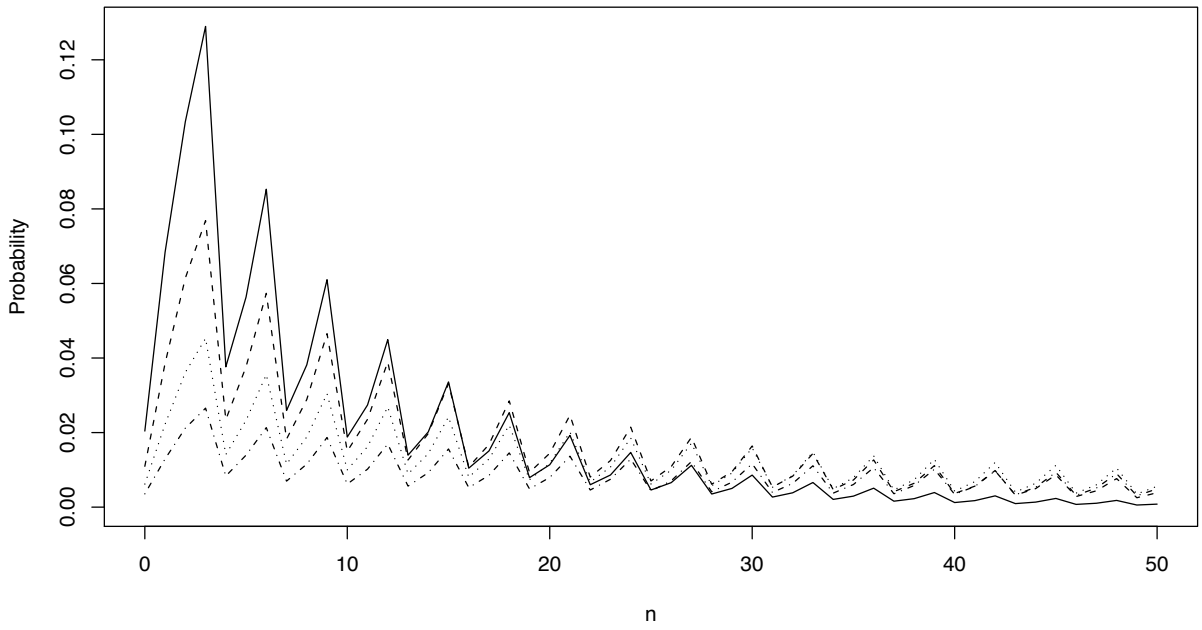


Figure 2.7: Frameshift-aware indel model probability of observing n inserted DNA bases, given starting at $m = 1$. The transition probability $P_{1,n}(t)$ is shown for $t = 5$ (solid line), $t = 7$ (dashed line), $t = 9$ (dotted line), and $t = 11$ (dash-dotted line), with parameters $\beta_0 = 0.3$, $\beta_1 = 1$, $\beta_2 = 4$, $\gamma_0 = 2$, $\gamma_1 = 0.2$, and $\gamma_2 = 0.2$.

2.4 Conclusion

Traditionally the simple BDP with linear rates has dominated modeling applications, since its transition probabilities and other quantities of interest find analytic expressions. However, increasingly sophisticated models in ecology, genetics, and evolution, among other fields, may necessitate more advanced computational methods to handle processes whose birth and death rates do not easily yield analytic solutions. We have demonstrated a flexible method for finding transition probabilities of general BDPs that works for arbitrary sets of birth and death rates $\{\lambda_n\}$ and $\{\mu_n\}$, and does not require additional analytic information. This should prove useful for rapid development and testing of new models in applications. For simple models whose solution is available, we find that our method agrees with known solutions and remains robust for large starting and ending states and long times t. It is our hope that the method presented here will assist researchers in understanding the properties of increasingly rich and realistic models.

2.5 Appendix

2.5.1 Approximant method

Murphy and O'Donohoe (1975) approximate the inverse Laplace transform of (2.15) by first truncating the continued fraction as a rational approximant through a partial fractions sum. To illustrate the pitfalls of this approach, we derive the inversion expressions presented by Murphy and O'Donohoe and analyze their properties. We provide an example to show that this technique can become numerically unstable. We first seek to uncover the truncation error in the time domain of the transition probabilities. If we truncate the continued fractions (2.15) at depth k, we have

$$\begin{aligned} f_{m,n}^{(k)}(s) &= \left(\prod_{j=n+1}^{m} \mu_j\right) \frac{B_n}{B_{m+1}+} \frac{B_m a_{m+2}}{b_{m+2}+} \frac{a_{m+3}}{b_{m+3}+} \cdots \frac{a_{m+k}}{b_{m+k}} \quad \text{for } n \le m, \text{ and} \\ f_{m,n}^{(k)}(s) &= \left(\prod_{j=m}^{n-1} \lambda_j\right) \frac{B_m}{B_{n+1}+} \frac{B_n a_{n+2}}{b_{n+2}+} \frac{a_{n+3}}{b_{n+3}+} \cdots \frac{a_{n+k}}{b_{n+k}} \quad \text{for } n \ge m. \end{aligned} \tag{2.60}$$

For concreteness, suppose in what follows that $n \geq m$. Note that the denominator of the second equation is simply B_{n+k}. Let $A_k^{(n)}$ be the numerator of the continued fraction in the second equation in (2.60), so

$$f_{m,n}^{(k)} = \left(\prod_{j=m}^{n-1} \lambda_j \right) \frac{A_k^{(n)}}{B_{n+k}}, \tag{2.61}$$

where $A_k^{(n)}$ satisfies $A_k^{(0)} = A_k$, $A_1^{(n)} = \prod_{j=1}^{n+1} a_j$, and

$$A_k^{(n)} = a_{n+k} A_{k-2}^{(n)} + b_{n+k} A_{k-1}^{(n)}. \tag{2.62}$$

Note also that the difference between truncated estimates in the Laplace domain (s) is

$$\begin{aligned} \frac{A_{n+k}}{B_{n+k}} - \frac{A_n}{B_n} &= \frac{A_{n+k}B_n - A_n B_{n+k}}{B_{n+k}B_n} \\ &= \frac{(-1)^n A_k^{(n)}}{B_{n+k}B_n}. \end{aligned} \tag{2.63}$$

This yields the generalized determinant formula

$$A_{n+k}B_n - A_n B_{n+k} = (-1)^n A_k^{(n)}, \tag{2.64}$$

and at a root s_i of $B_{n+k}(s)$, we have

$$A_k^{(n)}(s_i) = (-1)^n A_{n+k}(s_i) B_n(s_i). \tag{2.65}$$

Now if $s_1, s_2, \ldots, s_n$ are the roots of $B_n(s)$, we have, using the previous line and a partial fractions decomposition of (2.60), the formula for the Laplace transform of the transition probability $P_{m,n}(t)$, truncated at k,

$$\begin{aligned} f_{m,n}^{(k)}(s) &= \left(\prod_{j=m}^{n-1} \lambda_j \right) \frac{B_m(s) A_k^{(n)}(s)}{B_{n+k}(s)} \\ &= \left(\prod_{j=m}^{n-1} \lambda_j \right) \frac{B_m(s) A_k^{(n)}(s)}{\prod_{i=1}^{n+k}(s - s_i)} \\ &= \left(\prod_{j=m}^{n-1} \lambda_j \right) \sum_{i=1}^{n+k} \frac{B_m(s) B_n(s_i) A_{n+k}(s_i)}{\prod_{j \neq i}(s_j - s_i)} \left(\frac{1}{s - s_i} \right), \end{aligned} \tag{2.66}$$

since we only require the values of $A_{n+k}(s)$ and $B_n(s)$ at the zeros of $B_{n+k}(s)$. Then inverse transforming, an approximate formula for the transition probability $P_{m,n}(t)$ is

$$P_{m,n}^{(k)}(t) \approx \left(\prod_{j=m}^{n-1} \lambda_j \right) \sum_{i=1}^{n+k} \frac{B_m(s_i) B_n(s_i) A_{n+k}(s_i)}{\prod_{j \neq i}(s_j - s_i)} e^{-s_i t}. \tag{2.67}$$

The roots of $B_n(s)$, used in (2.66) and (2.67), are often found numerically as follows. Consider the characteristic polynomial $\det\left(\tilde{B}_n + sI\right)$ of the matrix

$$\tilde{B}_n = \begin{pmatrix} \lambda_0 & 1 & & & & \\ \lambda_0\mu_1 & \lambda_1+\mu_1 & 1 & & & \\ & \lambda_1\mu_2 & \lambda_2+\mu_2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda_{n-3}\mu_{n-2} & \lambda_{n-2}+\mu_{n-2} & 1 \\ & & & & \lambda_{n-2}\mu_{n-1} & \lambda_{n-1}+\mu_{n-1} \end{pmatrix}. \tag{2.68}$$

It is clear that the nth partial denominator $B_n(s) = \det(\tilde{B}_n + sI)$, and this quantity is zero when $-s$ is an eigenvalue of the matrix $\tilde{B}_n$. Therefore, the negatives of the eigenvalues of $\tilde{B}_n$ are the roots of $B_n(s)$. Furthermore, $\tilde{B}_n$ can be transformed into a real symmetric matrix via a similarity transform and hence $B_n(s)$ has precisely n roots, all of which are simple, real, and negative. One usually finds these eigenvalues via the QR algorithm or similar numerical techniques (Press, 2007). However, the iterative eigendecomposition of (2.68) generates small errors in the eigenvalues for large $n + k$. These errors are amplified in the product in the denominator of each summand in (2.67), resulting in a sum with both positive and negative terms that may be very large. Klar et al (2010) encounter similar instability in this algorithm. Their solution is to find the roots of the terms in the numerator and compute each summand as a product of individual numerators and denominators in an attempt to keep roundoff error in the product from accumulating. So, if $z_1, \ldots, z_{n+k}$ are the roots of A_{n+k} then (2.67) becomes

$$P_{m,n}^{(k)}(t) \approx \left(\prod_{j=m}^{n-1} \lambda_j\right) \sum_{i=1}^{n+k} B_m(s_i) B_n(s_i)(z_i - s_i) \prod_{j \neq i} \left(\frac{z_j - s_i}{s_j - s_i}\right) e^{-s_i t}. \tag{2.69}$$

This procedure does improve the numerical stability of the computation, but requires two eigendecompositions of possibly large matrices for every evaluation of $P_{m,n}(t)$, increasing the computational cost and, for large m and n, the roundoff error. In our opinion, it is more advantageous to avoid truncation of the continued fraction (5.5) at a pre-specified index, and instead evaluate the continued fraction until convergence during numerical inversion. Figure 2.1 shows how approximant methods fail for large n.

2.5.2 A power series method

Parthasarathy and Sudhesh (2006a) present exact solutions by transforming continued fractions such as (5.5) into an equivalent power series. Wall (1948) shows that Jacobi fractions of this type can always be represented by an equivalent power series. However, the small radius of convergence of power series expressions for transition probabilities can limit their usefulness for long times or large birth or death rates. Parthasarathy and Sudhesh show that $P_{0,n}(t)$ has a power series representation given by

$$P_{m,n}(t) = \left(\prod_{k=0}^{n-1} a_{2k}\right) \sum_{m=0}^{\infty} (-1)^m A(m, 2n) \frac{t^{m+n}}{(m+n)!}, \tag{2.70}$$

where

$$A(m,n) = \sum_{i_1=0}^{n} a_{i_1} \sum_{i_2=0}^{i_1+1} a_{i_2} \sum_{i_3=0}^{i_2+1} a_{i_3} \cdots \sum_{i_m=0}^{i_{m-1}+1} a_{i_m}, \tag{2.71}$$

with $A(0,n) = 1$ (Parthasarathy and Sudhesh, 2006a,b). Here, $a_{2n} = \lambda_n$ and $a_{2n+1} = \mu_n$ in the notation used in their papers. This approach is unique because it yields an exact analytic expression for the transition probabilities of a general BDP. However, the radius of convergence of the power series depends on the specified rates, and this radius may be quite small. To illustrate the pitfalls of this approach, consider $a_n = (n+1)\lambda$, corresponding to the BDP with $\lambda_n = (2n+1)\lambda$ and $\mu_n = 2n\lambda$ (Parthasarathy and Sudhesh, 2006a, Example 4.6). The power series for the transition probability in this process becomes

$$P_{0,n}(t) = \sum_{m=0}^{\infty} (-1)^m \frac{(2n+2m)!}{m!n!} \frac{(\lambda t/2)^{n+m}}{(n+m)!}. \tag{2.72}$$

Then the radius of convergence R of the power series is given by

$$\begin{aligned} 1/R &= \lim_{m\to\infty} \left| \frac{(2n+2m+2)! \left(\frac{\lambda}{2}\right)^{n+m+1}}{(m+1)!n!(n+m+1)!} \times \frac{m!n!(n+m)!}{(2n+2m)! \left(\frac{\lambda}{2}\right)^{n+m}} \right| \\ &= \lim_{m\to\infty} \frac{(2m+2n+1)(2n+2m+2)}{(m+1)(n+m+1)} \left(\frac{\lambda}{2}\right) \\ &= \lim_{m\to\infty} \frac{2m+2n+1}{m+1} \lambda \\ &= 2\lambda. \end{aligned} \tag{2.73}$$

And so the series diverges when $2\lambda t > 1$. To illustrate the limitations of the power series approach, note that in this process, the transition intensity from 0 to 1 is λ, so the expected

first-passage time from 0 to 1 is $\mathbb{E}(T_{0,1}) = 1/\lambda$. Therefore, we cannot evaluate (2.72) when t is greater than $\mathbb{E}(T_{0,1})/2$. If n is much greater than 1, we may be unable to reliably evaluate $P_{0,n}(t)$ for times near $\mathbb{E}(T_{0,n})$.

CHAPTER 3

Estimation for general birth-death processes

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. The E-step in the EM algorithm is available in closed-form for some linear BDPs, but otherwise previous work has resorted to approximation or simulation. Remarkably, the E-step conditional expectations can also 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 novel and efficient computation of the conditional expectations for all BDPs, eliminating the need for approximation 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. We show that our Laplace convolution technique outperforms competing methods when available and demonstrate a technique to accelerate EM algorithm convergence. Finally, we validate our approach using synthetic data and then apply our methods to estimation of mutation parameters in microsatellite evolution.

3.1 Introduction

A birth-death process (BDP) is a continuous-time Markov chain that models a non-negative 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 et al, 1991; Krone and Neuhauser, 1997; Novozhilov et al, 2006). For example, BDPs can characterize epidemic dynamics, (Bailey, 1964; Andersson and Britton, 2000), speciation and extinction (Nee et al, 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 et al, 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. 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, 1982b). 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 et al, 2007).

Progress in estimating birth and death rates in 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 et al, 1991; Holmes and Bruno, 2001; Rosenberg et al, 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 et al, 1977). In finite state-space Markov chains, the relevant conditional expectations (the E-step of the EM algorithm) can often be computed efficiently, and several researchers have derived EM algorithms for estimating transition rates in this context (Lange, 1995a; Holmes and Rubin, 2002; Hobolth and Jensen, 2005; Bladt and Sorensen, 2005; Metzner et al, 2007). Unfortunately, finding these conditional expectations for general BDPs poses challenges since the joint distribution of the states and waiting times (or its generating function) is usually not available in closed-form. Notably, Holmes and Bruno (2001); Holmes and Rubin (2002) and Doss et al (2010) are 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 BDPs under a variety of sampling sce-

narios. Indeed, more complex and realistic models like those reviewed by Novozhilov et al (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 a framework for deriving EM algorithms for estimating rate parameters of a 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 or repeated simulation that has 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 is faster than competing techniques and validate it using simulated data. Finally, we conclude with an application that analyzes microsatellite evolution and answers an open question in evolutionary genomics.

3.2 General BDPs and their EM algorithms

3.2.1 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. 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 $P_{i,j}(\tau) = \Pr(X(\tau) = j \mid X(0) = i)$ obey the system of ordinary differential equations

$$\begin{aligned} \frac{\mathrm{d}P_{i,0}(\tau)}{\mathrm{d}\tau} &= \mu_1 P_{i,1}(\tau) - \lambda_0 P_{i,0}(\tau), \text{ and} \\ \frac{\mathrm{d}P_{i,j}(\tau)}{\mathrm{d}\tau} &= \lambda_{j-1} P_{i,j-1}(\tau) + \mu_{j+1} P_{i,j+1}(\tau) - (\lambda_j + \mu_j) P_{i,j}(\tau), \end{aligned} \tag{3.1}$$

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

For some simple parameterizations of λ_k and μ_k, closed-form solutions exist for the transition probabilities $P_{i,j}(\tau)$, but this is not possible for most models. Karlin and McGregor (1957b) show 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 et al, 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. Laplace transforming equation (5.2) yields

$$\begin{aligned} s f_{i,0}(s) - \delta_{i0} &= \mu_1 f_{i,1}(s) - \lambda_0 f_{i,0}(s), \\ s f_{i,j}(s) - \delta_{ij} &= \lambda_{j-1} f_{i,j-1}(s) + \mu_{j+1} f_{i,j+1}(s) - (\lambda_j + \mu_j) f_{i,j}(s), \end{aligned} \tag{3.2}$$

where $f_{i,j}(s)$ is the Laplace transform of $P_{i,j}(\tau)$ and $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Letting $i = 0$ and rearranging (3.2), we obtain the recurrence relations

$$\begin{aligned} f_{0,0}(s) &= \frac{1}{s + \lambda_0 - \mu_1 \left(\frac{f_{0,1}(s)}{f_{0,0}(s)} \right)}, \text{ and} \\ \frac{f_{0,j}(s)}{f_{0,j-1}(s)} &= \frac{\lambda_{j-1}}{s + \mu_j + \lambda_j - \mu_{j+1} \left(\frac{f_{0,j+1}(s)}{f_{0,j}(s)} \right)}. \end{aligned} \tag{3.3}$$

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

generalized continued fraction

$$f_{0,0}(s) = \cfrac{1}{s+\lambda_0 - \cfrac{\lambda_0\mu_1}{s+\lambda_1+\mu_1 - \cfrac{\lambda_1\mu_2}{s+\lambda_2+\mu_2-\cdots}}}. \tag{3.4}$$

This is an exact expression for the Laplace transform of the transition probability $P_{0,0}(\tau)$. In (5.5), 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 (5.5) becomes

$$f_{0,0}(s) = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3+\cdots}}}. \tag{3.5}$$

We can write this more compactly as

$$f_{0,0}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\frac{a_3}{b_3+}\cdots. \tag{3.6}$$

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

$$f_{0,0}^{(k)}(s) = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\cdots\frac{a_k}{b_k} = \frac{A_k(s)}{B_k(s)}, \tag{3.7}$$

where $A_k(s)$ and $B_k(s)$ are the numerator and denominator of the rational function $f_{0,0}^{(k)}$. The transition probabilities $P_{i,j}(\tau)$ for $i, j > 0$ can be derived in continued fraction form by combining (3.2) and (5.5) to obtain

$$f_{i,j}(s) = \begin{cases} \left(\displaystyle\prod_{k=j+1}^{i}\mu_k\right)\frac{B_j(s)}{B_{i+1}(s)+}\frac{B_i(s)a_{i+2}}{b_{i+2}+}\frac{a_{i+3}}{b_{i+3}+}\cdots & \text{for } j \leq i, \\ \left(\displaystyle\prod_{k=i}^{j-1}\lambda_k\right)\frac{B_i(s)}{B_{j+1}(s)+}\frac{B_j(s)a_{j+2}}{b_{j+2}+}\frac{a_{j+3}}{b_{j+3}+}\cdots & \text{for } i \leq j, \end{cases} \tag{3.8}$$

(Murphy and O'Donohoe, 1975; Crawford and Suchard, 2011).

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 them 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 et al, 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, 2006a; Crawford and Suchard, 2011).

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 $P_{a,b}(t)$ 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 a and b is unobserved. Second, transition probabilities play an important role in computing conditional expectations of sufficient statistics, as we shall see below.

3.2.2 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 $\mathbf{Y} = (X(0) = a, X(t) = b)$ from a general BDP, we wish to find maximum likelihood estimates 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

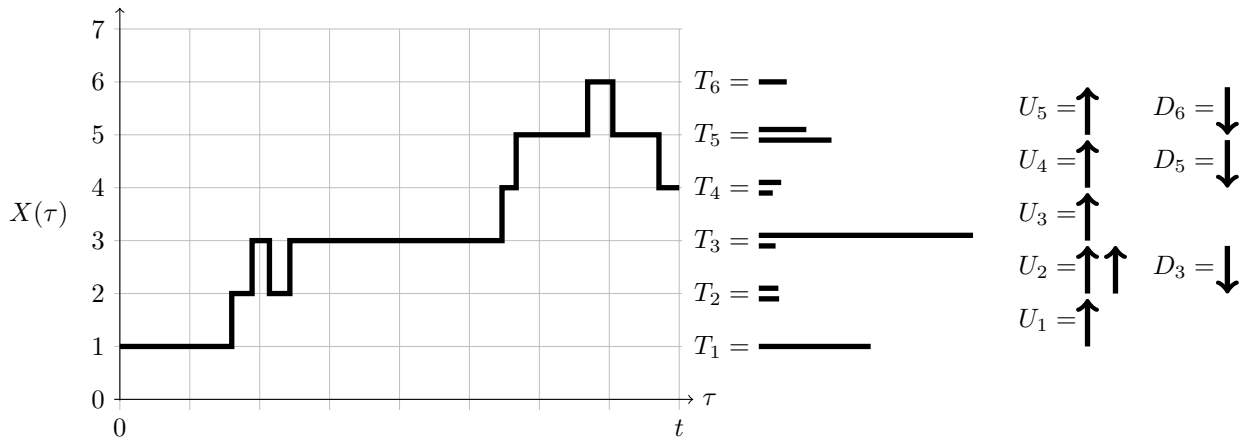


Figure 3.1: A sample path from a birth-death process (BDP) $X(\tau)$. The process starts at state $X(0) = 1$ and is at state $X(t) = 4$ at time t. At right are schematic representations of the time spent in each state T_k, the number of up steps U_k, and the number of down steps D_k. These quantities are the sufficient statistics for estimators of rate parameters in general birth-death processes.

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,

$$T_{\text{particle}} = \int_0^t X(\tau)\, \mathrm{d}\tau = \sum_{k=0}^{\infty} k T_k, \tag{3.9}$$

that counts the amount of time lived by each particle since time $\tau = 0$. Of course, the total elapsed time is $t = \sum_{k=0}^{\infty} T_k$. We demonstrate these concepts schematically in Figure 3.1.

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

$$\ell(\boldsymbol{\theta}) = \sum_{k=0}^{\infty} U_k \log\big[\lambda_k(\boldsymbol{\theta})\big] + D_k \log\big[\mu_k(\boldsymbol{\theta})\big] - \big[\lambda_k(\boldsymbol{\theta}) + \mu_k(\boldsymbol{\theta})\big] T_k. \tag{3.10}$$

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 (3.10) without them.

We therefore appeal to the EM algorithm for iterative maximum likelihood estimation with missing data (Dempster et al, 1977). In the EM algorithm, we define a surrogate objective function Q by taking the expectation of the complete data log-likelihood (3.10), 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 maximum likelihood estimate of $\boldsymbol{\theta}$. Taking the expectation of (3.10) conditional on $\mathbf{Y}$ and $\boldsymbol{\theta}^{(m)}$, we form the surrogate function Q:

$$\begin{aligned} Q\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}\big) &= \mathbb{E}\big[\ell(\boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}^{(m)}\big] \\ &= \sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y}) \log\big[\lambda_k(\boldsymbol{\theta})\big] + \mathbb{E}(D_k|\mathbf{Y}) \log\big[\mu_k(\boldsymbol{\theta})\big] - \mathbb{E}(T_k|\mathbf{Y})\big[\lambda_k(\boldsymbol{\theta}) + \mu_k(\boldsymbol{\theta})\big], \end{aligned} \tag{3.11}$$

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.

3.2.3 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 $\mathbf{Y}$. Doss et al (2010) adopt an approach for linear BDPs that combines analytic results with simulations. For some models, these authors are 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 manipulate this generating function to complete the E-step. For a more complicated linear model, Doss et al resort to approximating the relevant conditional expectations by simulating sample paths, conditional on $\mathbf{Y}$ (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:

$$\mathbb{E}(U_k|\mathbf{Y}) = \frac{\int_0^t P_{a,k}(\tau)\lambda_k P_{k+1,b}(t-\tau)\ \mathrm{d}\tau}{P_{a,b}(t)}, \tag{3.12a}$$

$$\mathbb{E}(D_k|\mathbf{Y}) = \frac{\int_0^t P_{a,k}(\tau)\mu_k P_{k-1,b}(t-\tau)\ \mathrm{d}\tau}{P_{a,b}(t)}, \quad \text{and} \tag{3.12b}$$

$$\mathbb{E}(T_k|\mathbf{Y}) = \frac{\int_0^t P_{a,k}(\tau)P_{k,b}(t-\tau)\ \mathrm{d}\tau}{P_{a,b}(t)}. \tag{3.12c}$$

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

$$\mathbb{E}(U_k|\mathbf{Y}) = \lambda_k \frac{\mathcal{L}^{-1}\Big[f_{a,k}(s)\ f_{k+1,b}(s)\Big](t)}{P_{a,b}(t)}, \tag{3.13a}$$

$$\mathbb{E}(D_k|\mathbf{Y}) = \mu_k \frac{\mathcal{L}^{-1}\Big[f_{a,k}(s)\ f_{k-1,b}(s)\Big](t)}{P_{a,b}(t)}, \quad \text{and} \tag{3.13b}$$

$$\mathbb{E}(T_k|\mathbf{Y}) = \frac{\mathcal{L}^{-1}\Big[f_{a,k}(s)\ f_{k,b}(s)\Big](t)}{P_{a,b}(t)}. \tag{3.13c}$$

where $\mathcal{L}^{-1}$ denotes inverse Laplace transformation. Although these formulas are equivalent to (3.12), 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.

To calculate the numerators of (3.13), we use the Laplace inversion method popularized by Abate and Whitt (1992b, 1995). This involves a Riemann sum approximation of the inverse transform that stabilizes the discretization error and is amenable to series acceleration

methods (Abate and Whitt, 1999; Press, 2007). To evaluate the continued fraction Laplace transforms $f_{a,b}(s)$, we use the modified Lentz method (Lentz, 1976; Thompson and Barnett, 1986; Press, 2007).

3.2.4 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.

3.2.4.1 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

$$Q(\boldsymbol{\theta}) = \sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y}) \log[k\lambda] + \mathbb{E}(D_k|\mathbf{Y}) \log[k\mu] - \mathbb{E}(T_k|\mathbf{Y})k(\lambda + \mu). \tag{3.14}$$

Taking the derivative of (3.14) with respect to the unknown parameters, setting the result to zero, and solving for λ and μ gives the M-step updates

$$\lambda^{(m+1)} = \frac{\mathbb{E}(U|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}, \text{ and} \tag{3.15a}$$

$$\mu^{(m+1)} = \frac{\mathbb{E}(D|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}. \tag{3.15b}$$

These updates correspond to the usual maximum likelihood estimators in the continuously observed process (Reynolds, 1973). Note that the transition probabilities $P_{a,b}(t)$ in the denominators of the expectations in (3.12) cancel out in (3.15a) and (3.15b). When this is the case, transition probabilities are not necessary to derive an EM algorithm.

3.2.4.2 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

$$\ell(\boldsymbol{\theta}) = \sum_{k=0}^{\infty} U_k \log(k\lambda + \nu) + D_k \log(k\mu) - T_k[k(\lambda + \mu) + \nu]. \tag{3.16}$$

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 H such that for all $\boldsymbol{\theta}$, $H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) \leq \ell(\boldsymbol{\theta})$ and $H\big(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}\big) = \ell\big(\boldsymbol{\theta}^{(m)}\big)$ as follows:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &\geq H\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) \\ &= \sum_{k=0}^{\infty} U_k\big[p_k \log\big(p_k\lambda\big) + (1-p_k)\log\big((1-p_k)\nu\big)\big] + D_k \log(\mu) - \big[k(\lambda+\mu)+\nu\big]T_k,\end{aligned} \tag{3.17}$$

where

$$p_k = \frac{k\lambda^{(m)}}{k\lambda^{(m)} + \nu^{(m)}}. \tag{3.18}$$

Then letting $Q\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}\big) = \mathbb{E}\left(H(\boldsymbol{\theta}) \mid \mathbf{Y}, \boldsymbol{\theta}^{(m)}\right)$ 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, 2010b). Maximizing Q with respect to λ and ν yields the updates

$$\lambda^{(m+1)} = \frac{\sum_{k=0}^{\infty} p_k \mathbb{E}(U_k|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}, \text{ and} \tag{3.19a}$$

$$\nu^{(m+1)} = \frac{\sum_{k=0}^{\infty} (1-p_k)\mathbb{E}(U_k|\mathbf{Y})}{t}. \tag{3.19b}$$

Expression (3.19a) is similar to (3.15a), 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 (3.15b).

3.2.4.3 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

$$Q\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}\big) = \sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y})[\log(\lambda) - \beta k] + \mathbb{E}(D_k|\mathbf{Y}) \log(\mu) - \mathbb{E}(T_k|\mathbf{Y})[\lambda k^2 e^{-\beta k} + k\mu]. \tag{3.20}$$

Since λ and β appear together, we opt for a numerical Newton step. The gradient of Q with respect to these parameters is

$$F = \begin{pmatrix} \dfrac{\mathbb{E}(U|\mathbf{Y})}{\lambda} - \displaystyle\sum_{k=0}^{\infty} k^2 e^{-\beta k} \mathbb{E}(T_k|\mathbf{Y}) \\ -\displaystyle\sum_{k=0}^{\infty} \left[k\mathbb{E}(U_k|\mathbf{Y}) + \lambda k^3 e^{-\beta k} \mathbb{E}(T_k|\mathbf{Y}) \right] \end{pmatrix}, \tag{3.21}$$

and the Hessian is

$$H = \begin{pmatrix} -\dfrac{\mathbb{E}(U|\mathbf{Y})}{\lambda^2} & -\displaystyle\sum_{k=0}^{\infty} k^3 e^{-\beta k} \mathbb{E}(T_k|\mathbf{Y}) \\ -\displaystyle\sum_{k=0}^{\infty} k^3 e^{-\beta k} \mathbb{E}(T_k|\mathbf{Y}) & \lambda \displaystyle\sum_{k=0}^{\infty} k^4 e^{-\beta k} \mathbb{E}(T_k|\mathbf{Y}). \end{pmatrix}. \tag{3.22}$$

Then we update these parameters by

$$\begin{pmatrix} \lambda^{(m+1)} \\ \beta^{(m+1)} \end{pmatrix} = \begin{pmatrix} \lambda^{(m)} \\ \beta^{(m)} \end{pmatrix} - H^{-1} F. \tag{3.23}$$

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 (3.15b).

3.2.4.4 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

$$Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) = \sum_{k=0}^{N} \mathbb{E}(U_k|\mathbf{Y}) \log(\beta) + \mathbb{E}(D_k|\mathbf{Y}) \log(\gamma) - \mathbb{E}(T_k|\mathbf{Y})(k(N-k)\beta + k\gamma)/N, \tag{3.24}$$

and the update for β is

$$\beta^{(m+1)} = \frac{N\mathbb{E}(U|\mathbf{Y})}{\sum_{k=0}^{N}(N-k)k\mathbb{E}(T_k|\mathbf{Y})}. \tag{3.25}$$

The update for γ is

$$\gamma^{(m+1)} = \frac{N\mathbb{E}(D|\mathbf{Y})}{\mathbb{E}(T_{\text{particle}}|\mathbf{Y})}. \tag{3.26}$$

3.2.4.5 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 $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})^t$. 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 generalized linear model (GLM) framework. We link

$$\log(\lambda_{ik}) = g(k, \mathbf{z}_i^t\boldsymbol{\theta}_\lambda) \quad \text{and} \quad \log(\mu_{ik}) = h(k, \mathbf{z}_i^t\boldsymbol{\theta}_\mu), \tag{3.27}$$

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:

$$\begin{aligned} Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) = \sum_{i=1}^{N}\sum_{k=0}^{\infty}\Big[&\mathbb{E}(U_k|\mathbf{Y}_i)g(k, \mathbf{z}_i^t\boldsymbol{\theta}_\lambda) + \mathbb{E}(D_k|\mathbf{Y}_i)h(k, \mathbf{z}_i^t\boldsymbol{\theta}_\mu) \\ &- \mathbb{E}(T_k|\mathbf{Y}_i)\left(e^{g(k,\mathbf{z}_i^t\boldsymbol{\theta}_\lambda)} + e^{h(k,\mathbf{z}_i^t\boldsymbol{\theta}_\mu)}\right)\Big]. \end{aligned} \tag{3.28}$$

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 generalized linear model extension of the simple linear BDP in which

$$\log(\lambda_{ik}) = \log(k) + \mathbf{z}_i^t\boldsymbol{\theta}_\lambda, \quad \text{and} \quad \log(\mu_{ik}) = \log(k) + \mathbf{z}_i^t\boldsymbol{\theta}_\mu. \tag{3.29}$$

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

$$\nabla_{\boldsymbol{\theta}_\lambda} Q = \sum_{i=1}^{N} \mathbb{E}(U|\mathbf{Y}_i)\mathbf{z}_i - e^{\mathbf{z}_i^t \boldsymbol{\theta}_\lambda} \mathbb{E}(T_{\text{particle}}|\mathbf{Y}_i)\mathbf{z}_i \tag{3.30}$$

and the second differential (Hessian) of Q is

$$\mathbf{d}^2_{\boldsymbol{\theta}_\lambda} Q = -\sum_{i=1}^{N} e^{\mathbf{z}_i^t \boldsymbol{\theta}_\lambda} \mathbb{E}(T_{\text{particle}}|\mathbf{Y}_i)\mathbf{z}_i \mathbf{z}_i^t. \tag{3.31}$$

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

$$\boldsymbol{\theta}_\lambda^{(m+1)} = \boldsymbol{\theta}_\lambda^{(m)} - \left(\mathbf{d}^2_{\boldsymbol{\theta}_\lambda} Q\right)^{-1} \nabla_{\boldsymbol{\theta}_\lambda} Q. \tag{3.32}$$

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 (3.32) for some problems. Unfortunately there is no quasi-Newton option since in general $\mathbb{E}(T_{\text{particle}}|\mathbf{Y})e^{\mathbf{z}_i^t \boldsymbol{\theta}_\lambda}$ 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.

3.2.5 Implementation

Before presenting simulation results and our application to microsatellite evolution, we briefly outline some implementation details that ease our subsequent analyses.

3.2.5.1 E-step 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

$$\mathbb{E}(U|\mathbf{Y}) = \sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y}) = \frac{\sum_{k=0}^{\infty} \lambda_k \mathcal{L}^{-1}\Big[f_{a,k}(s)\ f_{k+1,b}(s)\Big](t)}{P_{a,b}(t)}. \tag{3.33}$$

Fortunately, the conditional expectations of U_k, D_k, and T_k are usually small for $k \ll \min(a,b)$ and $k \gg \max(a,b)$, so it is possible to replace the infinite sum in (3.33) by a finite one. We find an additional increase in computational efficiency by exchanging the order of Laplace inversion and summation. Then (3.33) becomes

$$\mathbb{E}(U|\mathbf{Y}) \approx \frac{\mathcal{L}^{-1}\left[\sum_{k=k_{\min}}^{k_{\max}} \lambda_k f_{a,k}(s) f_{k+1,b}(s)\right](t)}{P_{a,b}(t)}, \tag{3.34}$$

where we choose $k_{\min}$ to be the largest $k < \min(a,b)$ such that $\lambda_k |f_{a,k}(s) - f_{k+1,a}| < 10^{-8}$ and $k_{\max}$ to be the first $k > \max(a,b)$ such that $\lambda_k |f_{a,k}(s) f_{k+1,b}(s)| < 10^{-8}$. In practice, we rarely need to compute expectations for k less than $\min(a,b) - 10$ or greater than $\max(a,b) + 10$.

3.2.5.2 Quasi-Newton acceleration of EM iterates

EM algorithms are notorious for slow convergence, especially near optima. When appropriate, 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 (Lange, 1995a; Louis, 1982; Meilijson, 1989; Jamshidian and Jennrich, 1993). Figure 3.2 shows 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 (transition probability) for any general BDP (Crawford and Suchard, 2011).

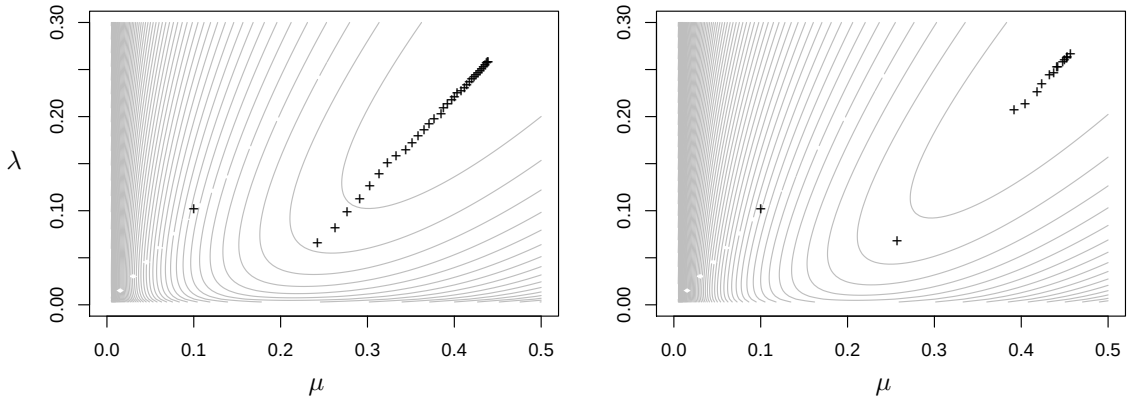


Figure 3.2: Effect of quasi-Newton acceleration on iterates of the expectation-maximization (EM) algorithm for a simple linear BDP with birth rate λ and death rate μ. Contour lines sketch the log-likelihood from $N = 50$ discrete samples. Iterates are shown with the "+" symbol. On the left, ordinary 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.

3.2.5.3 Asymptotic variance of EM estimates

Finding the observed information matrix for an EM estimate can be challenging. Louis (1982) gives formulae for the observed information, which Doss et al (2010) use to derive analytic expressions for the observed information for very simple BDPs. However, analytic expressions for the asymptotic variance are generally hard to find for more complicated models. 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. 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

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P_{a_i,b_i}(t_i), \tag{3.35}$$

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 $\mathbf{H}(\hat{\boldsymbol{\theta}}) = d^2\ell(\hat{\boldsymbol{\theta}})$ is the numerical Hessian evaluated at the estimated value $\hat{\boldsymbol{\theta}}$, then $\hat{\mathbf{I}} \approx -\mathbf{H}(\hat{\boldsymbol{\theta}})$.

3.3 Results

3.3.1 Laplace convolution E-step comparison

To illustrate the computational speedup that the Laplace convolution formulae (3.13) and their acceleration in section 3.2.5.1 achieve over existing methods, we calculate conditional expectations for various BDP models for performing the E-step and report computing times in Table 3.1. The first method in the table employs rejection sampling of trajectories where we condition on the starting state, and reject based on the ending state (Bladt and Sorensen, 2005). The second method adapts an endpoint-conditioned simulation algorithm (Hobolth, 2008; Hobolth and Stone, 2009). The third considers naïve time-domain convolution (Equa-

tion (3.12)) using the `integrate` function in `R`. Finally, we compute the same quantities via the Laplace-domain convolution method outlined in section 3.2.3. In our implementations, we have made every effort to reuse as much shared `R` code as possible, with the aim of making the routines comparable. We consider four different BDPs. For a simple linear BDP and a linear BDP with immigration, we use the data $\mathbf{Y} = (X(0) = 19, X(2) = 27)$. Under a logistic model, the data are $\mathbf{Y} = (X(0) = 10, X(2) = 16)$, and for the SIS model the data are $\mathbf{Y} = (X(0) = 10, X(2) = 31)$. We list all model parameter values in Table 3.1.

As seen in Table 3.1, the Laplace convolution method is often more than 10 times faster than the other methods. In terms of time-performance, the endpoint-conditioned simulation stands as second best, achieving almost comparable speed in the logistic BDP. To interpret this finding, we recall that Hobolth (2008) constructs an endpoint-conditioned simulation for performing the E-step in finite state-space Markov chains. Therefore, to adapt this method we approximate each BDP by a Markov chain with a finite transition rate matrix. To choose the arbitrary dimension of this matrix we truncate the process at the first state $k > \max(a, b)$ such that $P_{a,k}(t) < 10^{-5}$, and the resulting estimates agree substantially with the other methods. We are aware that the size of the rate matrix affects the speed of the simulation routine, so 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. For the logistic model, such a stringent upper bound lies just above the relatively small carrying capacity. However, endpoint-conditioned simulation completely fails for the SIS model, an issue we discuss later. Finally, and quite naturally, the two convolution methods arrived at nearly the same answer for each model; the difference is largely due to very different sources of numerical error, but at disparate computational costs.

3.3.2 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

Model	Quantity	Rejection sampling	ECS	Time-conv	Laplace-conv
Simple linear (3.2.4.1)	$\mathbb{E}(U\|\mathbf{Y})$	1.449	0.741	19.606	0.084
$\lambda = 0.5$, $\mu = 0.3$	$\mathbb{E}(D\|\mathbf{Y})$	1.375	0.743	21.224	0.086
	$\mathbb{E}(T_{\text{particle}}\|\mathbf{Y})$	1.432	0.636	16.488	0.087
Immigration (3.2.4.2)	$\sum_k p_k \mathbb{E}(U\|\mathbf{Y})$	1.192	0.697	15.669	0.085
$\lambda = 0.5$, $\nu = 0.2$	$\mathbb{E}(D\|\mathbf{Y})$	1.324	0.689	21.058	0.086
$\mu = 0.3$	$\mathbb{E}(T_{\text{particle}}\|\mathbf{Y})$	1.319	0.703	14.961	0.089
Logistic (3.2.4.3)	$\mathbb{E}(U\|\mathbf{Y})$	50.810	0.162	21.907	0.102
$\lambda = 0.5$, $\alpha = 0.2$	$\mathbb{E}(D\|\mathbf{Y})$	56.957	0.180	20.851	0.102
$\mu = 0.3$	$\sum_k k^2 e^{-k\alpha} \mathbb{E}(T_k\|\mathbf{Y})$	50.764	0.168	21.623	0.107
SIS (3.2.4.4)	$\mathbb{E}(U\|\mathbf{Y})$	7.880	*	5.295	0.059
$\beta = 0.5$, $\gamma = 0.3$	$\mathbb{E}(D\|\mathbf{Y})$	8.886	*	2.749	0.048
	$\sum_k (N-k) k \mathbb{E}(T_k\|\mathbf{Y})$	8.456	*	4.269	0.053

Table 3.1: Compute times (seconds) to perform various E-steps for four different BDP models. We report text section numbers in which the models are described in parentheses. For each E-step, we consider several methods. In all cases, the Laplace method takes substantially less time. The endpoint-conditioned simulation (ECS) method fails for the susceptible-infectious-susceptible (SIS) infectious disease model.

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 generalized linear model (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 3.2 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.

3.3.3 Application to microsatellite evolution

Microsatellites are short tandem repeats of characters in a DNA sequence (Schlötterer, 2000; Ellegren, 2004; Richard et al, 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 (Whittaker et al, 2003; Calabrese and Durrett, 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; Eckert and Hile, 2009; Kelkar et al, 2008; 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 et al

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

Table 3.2: Point-estimates and their standard errors (SE) for simulated observations under various BDPs. 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.

(2002) study 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 et al (2002) record 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 present an extended application of our BDP inference technique to chimpanzee-human microsatellite evolution, drawing on the data in Table 6 of the supplementary information in Webster et al (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 and yields insight into the complicated process of microsatellite evolution.

3.3.3.1 Evolutionary model

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

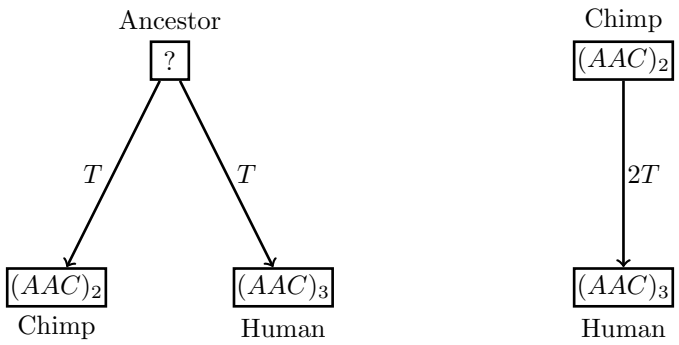


Figure 3.3: Reversibility of the BDP implies that the evolutionary relationship between contemporary chimpanzees and the most recent common ancestor can be inverted. On the left, the most recent common ancestor of chimpanzees and humans lived at time T in the past. At a certain locus, chimpanzees have a microsatellite consisting of 2 repeats of the motif AAC, and at an orthologous locus, humans have 3 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.

the observation $\mathbf{Y} = (C, H)$ is

$$\begin{aligned} \Pr(\mathbf{Y}) &= \sum_{k=0}^{\infty} \pi_k P_{k,C}(t)\ P_{k,H}(t) \\ &= \pi_C \sum_{k=0}^{\infty} P_{C,k}(t)\ P_{k,H}(t) \\ &= \pi_C P_{C,H}(2t), \end{aligned} \tag{3.36}$$

where π_k is the equilibrium probability of the microsatellite having k repeats. The second line follows by reversibility and the third by the Chapman-Kolmogorov equality. Therefore, the log-likelihood of the observation $\mathbf{Y}$ is now $\log \pi_C + \ell(\boldsymbol{\theta}; \mathbf{Y})$. Figure 3.3 shows a schematic representation of this reversibility argument.

3.3.3.2 BDP rates and equilibrium distribution

The observed data for microsatellite i are $\mathbf{Y}_i = (X_i(0), X_i(1))$, 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 et al (2002) dataset, it is evident that microsatellites with small numbers of repeats are not detected. Rose and Falush (1998) argue that there is a minimum number of repeats necessary for microsatellite mutation via polymerase slippage. Sainudiin et al (2004) interpret 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 et al 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

$$\lambda_k = \begin{cases} k\lambda + \lambda & k \geq x_{\min} \\ 0 & k < x_{\min} \end{cases} \qquad \text{and} \qquad \mu_k = \begin{cases} k\mu & k > x_{\min} \\ 0 & k \leq x_{\min}. \end{cases} \tag{3.37}$$

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

$$\pi_k = \begin{cases} \left(1 - \dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^{k - x_{\min} - 1} & k \geq x_{\min} \\ 0 & k < x_{\min}, \end{cases} \tag{3.38}$$

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 (3.38) assigns positive probability to all repeat numbers greater than or equal to $x_{\min}$.

3.3.3.3 Likelihood and surrogate function

Now we augment the log-likelihood with the log-equilibrium probability of observing $X_i(0)$ chimpanzee repeats

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \pi_{X_i(0)} + \ell(\boldsymbol{\theta}; \mathbf{Y}_i), \tag{3.39}$$

where $\ell(\boldsymbol{\theta}; \mathbf{Y}_i)$ is equivalent to (3.10). Including the influence of the equilibrium distribution is similar to imposing a prior distribution on λ and μ. To ensure the existence of the equilibrium distribution (3.38), we must also incorporate the constraint $\lambda < \mu$. To achieve maximization of the augmented log-likelihood (3.39) under this constraint, we impose a barrier term of the form $\gamma \log(\mu - \lambda)$. By iteratively maximizing and sending the barrier penalty $\gamma \to 0$, we can achieve maximization under the inequality constraint. More formally, if we let

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\log \pi_{X_i(0)} + \ell(\boldsymbol{\theta}; \mathbf{Y}_i)\right] + \gamma \log(\mu - \lambda), \tag{3.40}$$

then

$$\operatorname*{argmax}_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) \to \operatorname*{argmax}_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) \tag{3.41}$$

under the constraint $\lambda < \mu$ as $\gamma \to 0$.

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:

$$\mathbf{z}_i = \begin{cases} (1, 0, 0, p_a, p_c, p_t)^t & r_i = 1 \\ (1, 1, 0, p_a, p_c, p_t)^t & r_i = 2 \\ (1, 0, 1, p_a, p_c, p_t)^t & r_i \geq 3 \end{cases} \tag{3.42}$$

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 becomes

$$\log(\lambda_{k,i}) = \log(k+1) + \alpha + \mathbf{z}_i^t\boldsymbol{\theta} \quad \text{and} \quad \log(\mu_{k,i}) = \log(k) + \mathbf{z}_i^t\boldsymbol{\theta}. \tag{3.43}$$

Therefore $(\alpha, \boldsymbol{\theta})^t$ is the 7×1 vector of unknown parameters. Putting all this together, the surrogate function becomes

$$\begin{aligned} Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\big) \propto \Bigg(\sum_{i=1}^{N} X_i(0)\alpha + \log\left(1 - e^{\alpha}\right) + \Bigg[\sum_{k=0}^{\infty} \mathbb{E}(U_k|\mathbf{Y}_i)(\alpha + \mathbf{z}_i^t\boldsymbol{\theta}) + \mathbb{E}(D_k|\mathbf{Y}_i)\mathbf{z}_i^t\boldsymbol{\theta} \\ - \mathbb{E}(T_k|\mathbf{Y}_i)\left((k+1)e^{\alpha+\mathbf{z}_i^t\boldsymbol{\theta}} + ke^{\mathbf{z}_i^t\boldsymbol{\theta}}\right) \Bigg] \Bigg) + \gamma\log(-\alpha), \end{aligned} \tag{3.44}$$

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

Table 3.3 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. 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 and T nucleotides also give rise to higher λ and μ. These conclusions are largely consistent with the descriptive results obtained by Webster et al (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.

3.4 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

Parameter	Covariate	Estimate	SE
$\boldsymbol{\theta}_1$	Intercept	-1.3105	0.1236
$\boldsymbol{\theta}_2$	$r_i = 2$	0.2854	0.0983
$\boldsymbol{\theta}_3$	$r_i \geq 3$	-1.5405	0.1079
$\boldsymbol{\theta}_4$	p_a	0.2207	0.1725
$\boldsymbol{\theta}_5$	p_c	-0.3822	0.0577
$\boldsymbol{\theta}_6$	p_t	0.0477	0.0002
α	birth	-0.0889	0.0039

Table 3.3: Maximum likelihood estimates of parameters in the microsatellite model and their asymptotic standard errors. The first three 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 $\lambda = \exp(\alpha + \mathbf{z}_i^t\boldsymbol{\theta})$ and the death rate is $\mu = \exp(z_i^t\boldsymbol{\theta})$. Estimated birth and death rates are higher for dinucleotide repeats than for mononucleotide repeats or microsatellites whose motifs have 3, 4, or 5 nucleotides. Mircrosatellites whose motif consists, for example, of A nucleotides have higher birth and death rates compared to G nucleotides.

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 (3.35). 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 time-domain numerical integration or simulation of BDP trajectories. As we show in Table 3.1, 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 matrix eigen-decomposition to calculate transition probabilities. As we show for the SIS model, this matrix becomes nearly singular, causing the simulation algorithm to fail, even when we choose parameter values that are not biologically unreasonable. The Laplace convolution in the E-step of our algorithm is more generic with equivalent or better performance. 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, 2010). 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 paper.

If one cannot find analytic parameter updates in the M-step, several options remain available. With a minorizing function as in section 3.2.4.2, an EM-MM algorithm is viable. Further, one or more numerical Newton steps offers an alternative, as in sections 3.2.4.3 and 3.2.4.5. One may employ other gradient-based methods as well. Although the MM update derived for the BDP with immigration (section 3.2.4.2) 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 in order 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 high-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 3.2.4.5), or our microsatellite example. 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 et al, 2010).

With regard to our example, we present a novel way of studying the evolution of microsatellite repeats using a generalized linear model. Previous efforts often ignore the evolutionary relationship between organisms, use incomplete or equilibrium models of repeat numbers, or fit separate models to motifs of different sizes. We treat motif size as a categorical variable and incorporate motif nucleotide composition, allowing us to fit a single model to all the microsatellite observations simultaneously. Though our rate specification (3.37) and resulting equilibrium distribution (3.38) are intended to be somewhat simplistic, more

sophisticated models that are informed by biological considerations may be fruitful. The only requirement in our setup is that the gradient and Hessian of λ_k, μ_k, and π_k be available for any repeat number k. 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 aglorithms 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.

3.5 Conclusion

Previous work on parameter estimation in BDPs almost exclusively confines itself to inference of birth and death rates under the simple linear model. To rectify this situation, we present a flexible and robust framework for deriving EM algorithms to estimate parameters in any general BDP, using discrete observations. We hope that this contribution encourages development of more sophisticated and realistic birth-death models in applied work, since researchers can now estimate parameters using more complicated rate structures, even when the data are observed at discrete times.

CHAPTER 4

Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees

The branching structure of biological evolution confers statistical dependencies on phenotypic trait values in related organisms. For this reason, comparative macroevolutionary studies usually begin with an inferred phylogeny that describes the evolutionary relationships of the organisms of interest. The probability of the observed trait data can be computed by assuming a model for trait evolution, such as Brownian motion, over the branches of this fixed tree. However, the phylogenetic tree itself contributes statistical uncertainty to estimates of other evolutionary quantities, and many comparative evolutionary biologists regard the tree as a nuisance parameter. In this paper, we present a framework for analytically integrating over unknown phylogenetic trees in comparative evolutionary studies by assuming that the tree arises from a continuous-time Markov branching model called the Yule process. To do this, we derive a closed-form expression for the distribution of phylogenetic diversity, which is the sum of branch lengths connecting a set of taxa. We then present a generalization of phylogenetic diversity which is equivalent to the expected trait disparity in a set of taxa whose evolutionary relationships are generated by a Yule process and whose traits evolve by Brownian motion. We derive expressions for the distribution of expected trait disparity under a Yule tree. Given one or more observations of trait disparity in a clade, we perform fast likelihood-based estimation of the Brownian variance for unresolved clades. Our method does not require simulation or a fixed phylogenetic tree. We conclude with a brief example illustrating Brownian rate estimation for thirteen families in the Mammalian order Carnivora, in which the phylogenetic tree for each family is unresolved.

4.1 Introduction

Evolutionary relationships between organisms induce statistical dependencies in their phenotypic traits (Felsenstein, 1985). Closely related species that have been evolving separately for only a short time will generally have similar trait values, and species whose most recent common ancestor is more distant will often have dissimilar trait values (Harvey and Pagel, 1991). However, the origins of phenotypic diversity are still poorly understood (Eldredge and Gould, 1972; Gould and Eldredge, 1977; Ricklefs, 2006; Bokma, 2010). Even simple idealized models of evolutionary change can give rise to highly varying phenotype values (Foote, 1993; Sidlauskas, 2007), and researchers disagree about the relative importance of time, the rate of speciation, and the rate of phenotypic evolution in generating phenotypic diversity (Ricklefs, 2004; Purvis, 2004; Ricklefs, 2006).

Comparative phylogenetic studies seek to explain phenotypic differences between groups of taxa, and stochastic models of evolutionary change have assisted in this task. Researchers often treat phenotypic evolution as a Brownian motion process occurring independently along the branches of a *fixed* macroevolutionary tree (Felsenstein, 1985). In comparative studies, the Brownian motion model of trait evolution has a convenient consequence: given an evolutionary tree topology and branching times, the trait values at the concurrently observed tips of the tree are distributed according to a multivariate normal random variable. Brownian motion on a fixed phylogenetic tree is the basis for the most popular regression-based methods for comparative inference and hypothesis testing (Grafen, 1989; Garland et al, 1992; Martins and Hansen, 1997; Blomberg et al, 2003; O'Meara et al, 2006; Revell, 2010). In the regression approach, inference of evolutionary parameters of interest becomes a two-step process: first, one must infer a phylogenetic tree; then, *conditional on that tree*, one estimates relevant evolutionary parameters, usually by maximizing likelihood of the observed trait data under the model for trait evolution. Unfortunately, the uncertainty involved in estimating the tree propagates into the comparative analysis in a way that is difficult to account for (but see Stone (2011)), and comparative researchers often lack a precise phylogenetic tree on which to base a regression analysis of trait data. Modern techniques for dealing with this issue

generally resort to simulation. Some researchers simulate a large number of possible trees and estimate parameters conditional on a single representative tree, such as the maximum clade credibility tree (see, for example, Alfaro et al (2009)). Alternative approaches that utilize simultaneous simulation of trees and parameters via Bayesian methods are gaining in popularity (Sidlauskas, 2007; Slater et al, 2012; Drummond et al, 2012).

However, simulation methods can be extremely slow and may require assumptions about prior distributions of unknown parameters that are difficult to justify. Indeed, in macroevolutionary studies, the phylogenetic tree is often not of interest *per se*, but must be taken into account in order to accurately model the dependency of the traits under consideration. Many comparative phylogeneticists regard the evolutionary tree as a nuisance parameter in the larger evolutionary statistical model. For this reason, there is increased interest in tree-free methods of comparative analysis that preserve information about the variance of phenotypic values within unresolved clades (Bokma, 2010).

To develop a method for comparative inference in evolutionary studies that does not rely on a particular tree, it is convenient to specify a *generative* model for phylogenetic trees. In the Yule (pure-birth) process, every existing species independently gives birth with instantaneous rate λ; when there are n species, the total rate of speciation is $n\lambda$ (Yule, 1925). The Yule process is widely used as a null model in evolutionary hypothesis testing and can provide a plausible prior distribution on the space of evolutionary trees in Bayesian phylogenetic inference (Nee et al, 1994; Rannala and Yang, 1996; Nee, 2006). One can easily derive finite-time transition probabilities (Bailey, 1964), and efficient methods exist to simulate samples from the distribution of Yule trees, conditional on tree age, number of species, or both (Stadler, 2011). Interestingly, some researchers have pointed out that even the simple Yule process can have unexpected properties that may be relevant in evolutionary theory and reconstruction (Gernhard et al, 2008; Steel and Mooers, 2010).

Due to Yule trees' simple Markov branching structure and analytically tractable transition probabilities, many researchers have made progress in characterizing summary properties of the Yule process – that is, integrating over all Yule tree realizations. For example, Steel and McKenzie (2002) study aspects of the shape of phylogenies under the Yule model,

such as the distribution of the number of edges separating a subset of the extant taxa from the MRCA; Gernhard et al (2008) find distributions of branch lengths; Steel and Mooers (2010) study the expected length of pendant and interior edges of Yule trees; and Steel and McKenzie (2001) and Mulder (2011) study the distribution of the number of internal nodes separating taxa.

One important summary statistic for trees in biodiversity applications is *phylogenetic diversity* (PD), defined as the sum of all branch lengths in the minimum spanning tree connecting a set of taxa (Faith, 1992). Applied researchers in evolutionary biology have found PD to be useful in conservation and biodiversity applications; see, e.g., Webb et al (2002), Moritz (2002), and Turnbaugh et al (2008). PD also has the virtue of being a mathematically tractable statistic for phylogenetic trees, and has attracted interest from researchers interested in its properties. For example, Faller et al (2008) show that the asymptotic distribution (as the number of taxa $n \to \infty$) of PD is normal and give a recursion for computing the distribution of PD where edge lengths are integral. Mooers et al (2011) discuss branch lengths on Yule trees and expected loss of PD in conservation applications. Most importantly for our study, Stadler and Steel (2012) find the moment-generating function for PD conditional on n extant taxa and tree age t under the Yule model. Following on these inspiring results, we seek now to study analytic properties of Yule trees that are useful for comparative evolutionary studies.

In this paper, we present a framework for computing probability distributions related to diversity and quantitative trait evolution over unresolved Yule trees and describe methods for estimating related parameters. We first give a mathematical description of the Yule model of speciation and briefly discuss its properties. Next, we introduce the Markov reward process, a probabilistic method for deriving probability distributions related to the accumulation of diversity under a Yule model. In Theorem 1, we give an expression for the probability distribution of PD under a Yule model, conditional on the number of species n, time to the most recent common ancestor (TMRCA) or tree age t, and speciation rate λ. We then demonstrate an important and previously unappreciated relationship between *trait disparity*, the sample variance for a group of taxa (O'Meara et al, 2006), and PD for traits

evolving on a Yule tree via Brownian motion. Theorem 2 gives an expression for the distribution of expected trait disparity when integrating over the branch lengths of a Yule tree. Next, we describe a statistical method for performing fast maximum likelihood estimation of Brownian variance, given an unresolved clade and observed trait disparity. Our approach does not require fixing a phylogenetic tree or specification of prior probabilities for unknown parameters. The method is simulation-free and does not seek to infer branch lengths or ancestral states. We show empirically that our estimators are asymptotically consistent. We conclude with an application of our method to body size evolution in the Mammalian order Carnivora.

4.2 Mathematical background

To aid in exposition, we briefly establish some notation. Denote the *topology* of a phylogenetic tree by τ. A topology is the shape of a tree, disregarding branch lengths or age. We always condition our calculations on the phylogenetic tree having age t with n extant taxa. Let $t_1, t_2, \ldots$ denote the branching points of a tree, where t_k is the time of branching from k to $k+1$ lineages. We measure time in the forward direction, so at the TMRCA, $t = 0$. This is in keeping with our mechanistic orientation: the Yule process, to be developed below, runs forward in time from 0 to t.

4.2.1 Yule processes

Let $Y(t) \in \{1, 2, \ldots\}$ be a Yule process with birth rate λ that keeps track of the number of species at time t. The transition probabilities $P_{mn}(t) = \Pr(Y(t) = n \mid Y(0) = m)$ satisfy the Kolmogorov forward equations

$$\begin{aligned} \frac{\mathrm{d}P_{m1}(t)}{\mathrm{d}t} &= -\lambda P_{m1}(t), \quad \text{and} \\ \frac{\mathrm{d}P_{mn}(t)}{\mathrm{d}t} &= -\lambda n P_{mn}(t) + (n-1)\lambda P_{m,n-1}(t) \end{aligned} \tag{4.1}$$

for $n \geq 1$. This infinite system of ordinary differential equations can be solved to yield closed forms for the finite-time transition probabilities,

$$P_{mn}(t) = \binom{n-1}{m-1} e^{-m\lambda t}(1 - e^{-\lambda t})^{n-m}, \tag{4.2}$$

which have a negative binomial form (Bailey, 1964). In the Yule process, we are only concerned with the number of species that exist at any moment in time, not their genealogy. That is, we assume that the lineage that branches is chosen uniformly from all extant lineages. The transition probability (4.2) is useful for performing statistical inference: suppose we know the branching rate λ and the age t of a tree, and we observe $Y(t) = n$. Then (4.2) gives the *likelihood* of our observation. Figure 4.1 shows an example realization of a Yule tree, with the corresponding counting process diagram below. In this example, $\lambda = 2$, $Y(0) = 2$, and $Y(t = 1) = 12$.

4.2.2 Markov reward processes

In a Markov reward process, a non-negative reward a_k accrues for each unit of time a Markov process spends in state k (Neuts, 1995). Consider a Yule process $Y(s)$ beginning at $Y(0) = 1$ and ending at $Y(t) = n$. The accumulated reward up to time t is

$$R_t = \int_0^t a_{Y(s)} \, \mathrm{d}s. \tag{4.3}$$

When $Y(s)$ is observed continuously from time 0 to t, the process $a_{Y(s)}$ is a fully-observed step function, and R_t can be easily computed as the area under that function. To illustrate, suppose that the process makes jumps at times $t_1, \ldots, t_{n-1}$, and we define $t_0 = 0$ and $t_n = t$. We assume $Y(s)$ is right-continuous, so $Y(t_i) = i + 1$. Then at time t, the accumulated reward is

$$R_t = \sum_{i=1}^{n} a_{Y(t_{i-1})}(t_i - t_{i-1}) = \sum_{i=1}^{n} a_i(t_i - t_{i-1}). \tag{4.4}$$

When only $Y(0)$ and $Y(t)$ are observed, it can be challenging to compute the distribution of R_t. In our proofs of Theorems 1 and 2, we appeal to the method developed by Neuts (1995) and Minin and Suchard (2008) to find reward probabilities conditional on $Y(0)$ and $Y(t)$.

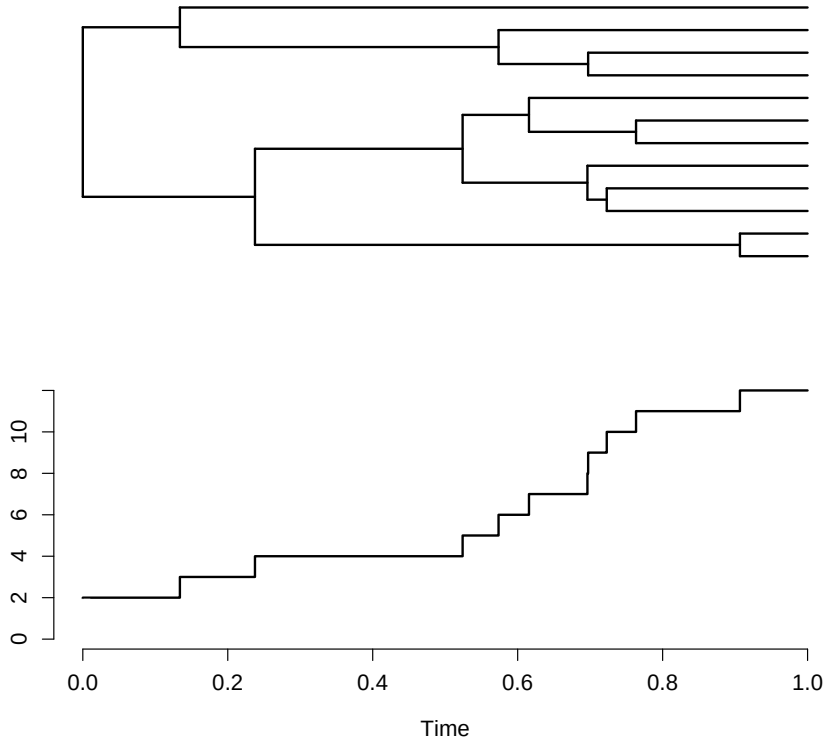


Figure 4.1: Example of a Yule (pure-birth) tree with the corresponding counting process $Y(t)$ below. The birth rate in this example is $\lambda = 2$. In the counting process representation of this realization, we only keep track of the total number of species in existence at each time.

Let

$$v_{mn}(x,t) = \Pr(R_t = x, Y(t) = n \mid Y(0) = m) \tag{4.5}$$

be the joint probability that the reward at time t is x and the process is in state n, given that the process began in state m at time 0. This joint probability formulation is more mathematically convenient than the more natural conditional probability, as we demonstrate in the proofs of the Theorems. However, it is easy to transform $v_{mn}(x,t)$ into the conditional probability via Bayes' rule, as we show below. Appendix 4.7 gives a preliminary lemma deriving a representation for Yule reward processes that will be useful in proving the Theorems that follow.

4.3 The distribution of phylogenetic diversity in a Yule process

The Yule process is a simple and analytically tractable mechanistic model for producing the birth times of a clade. If we assume that the species that undergoes speciation is chosen randomly from the extant species at that time, then the Yule process is also a distribution over bifurcating trees of age t. The Markov rewards framework provides a technique to understand integrals over Yule processes in precisely this context. In this section, we study the distribution of PD in trees generated by a Yule process. PD depends only on the branching times, and not the underlying topology, of the phylogenetic tree, making it a suitable first step in our goal of integrating over trees in comparative studies.

To proceed, let $Y(s)$ be a Yule process with branching rate λ that keeps track of the number of lineages at time s. We seek an expression for PD, the total branch length of the tree, which is equivalent to the area under the trajectory of the counting process $Y(s)$. Define a Markov reward process with $Y(s)$ and $a_k = k$ for $k = 1, 2, \ldots$. Then

$$R_t = \int_0^t a_{Y(s)}\, \mathrm{d}s = \int_0^t Y(s)\, \mathrm{d}s. \tag{4.6}$$

We now state our first Theorem giving an expression for the distribution of R_t in a Yule process.

Theorem 1. *For a Yule process with birth rate λ, starting at $Y(0) = m$ and ending at $Y(t) = n$,*

$$v_{mn}(x,t) = \begin{cases} \delta(x - mt)e^{-m\lambda t} & m = n \\ \dfrac{\lambda^{n-m}e^{-\lambda x}}{(n-m-1)!} \displaystyle\sum_{j=m}^{n} \binom{n-1}{j-1}\binom{j-1}{m-1}(-1)^{j-m}(x-jt)^{n-m-1}H(x-jt) & n > m \end{cases} \tag{4.7}$$

where $\delta(x)$ is the Dirac delta function and $H(x)$ is the Heaviside step function.

The proof of this Theorem is given in Appendix 4.8. There has been disagreement about whether the definition of PD in different contexts includes the root lineage (Faith, 1992; Faith and Baker, 2006; Crozier et al, 2006; Faith, 2006). We do not take a stance on this issue but note that if t is the stem age of an unresolved clade, then taking $a_1 = 1$ in (4.3) includes the root in the distribution of accumulated PD, and $a_1 = 0$ does not. The form of (4.7) will change slightly if $m = 1$ and $a_1 = 0$.

The probability distribution of PD, conditional on $Y(0) = m$ and $Y(t) = n$, is

$$f_Y(x \mid m, n, t, \lambda) = \frac{v_{mn}(x,t)}{P_{mn}(t)}, \tag{4.8}$$

where $P_{mn}(t)$ is the Yule transition probability (4.2). This family of probability distributions has some interesting properties. Figure 4.2 shows $f_Y(x|m, n, t, \lambda)$ for $m = 1$ (with $a_1 = 1$), $n = 1, \ldots, 8$, $\lambda = 1.2$, and $t = 1$. The unusual shape of the distribution for smaller n demonstrates the piecewise nature of the density, apparent in the functional form (4.7). Interestingly, Faller et al (2008) show that the distribution PD tends toward a normal distribution as $n \to \infty$, a fact suggested by the shape of the distributions in Figure 4.2.

These distributions have some practical uses. First, one can predict the PD that will arise under the Yule model from a collection of extant species up to time t in the future. Second, we can calculate the probability that future PD at time t in one group is greater than in the other, conditional on the number of species and diversification rates in both groups; this probability may have uses in conservation applications. Third, conditional on an inferred phylogenetic tree for a set of n taxa, one could compute the resulting PD x and

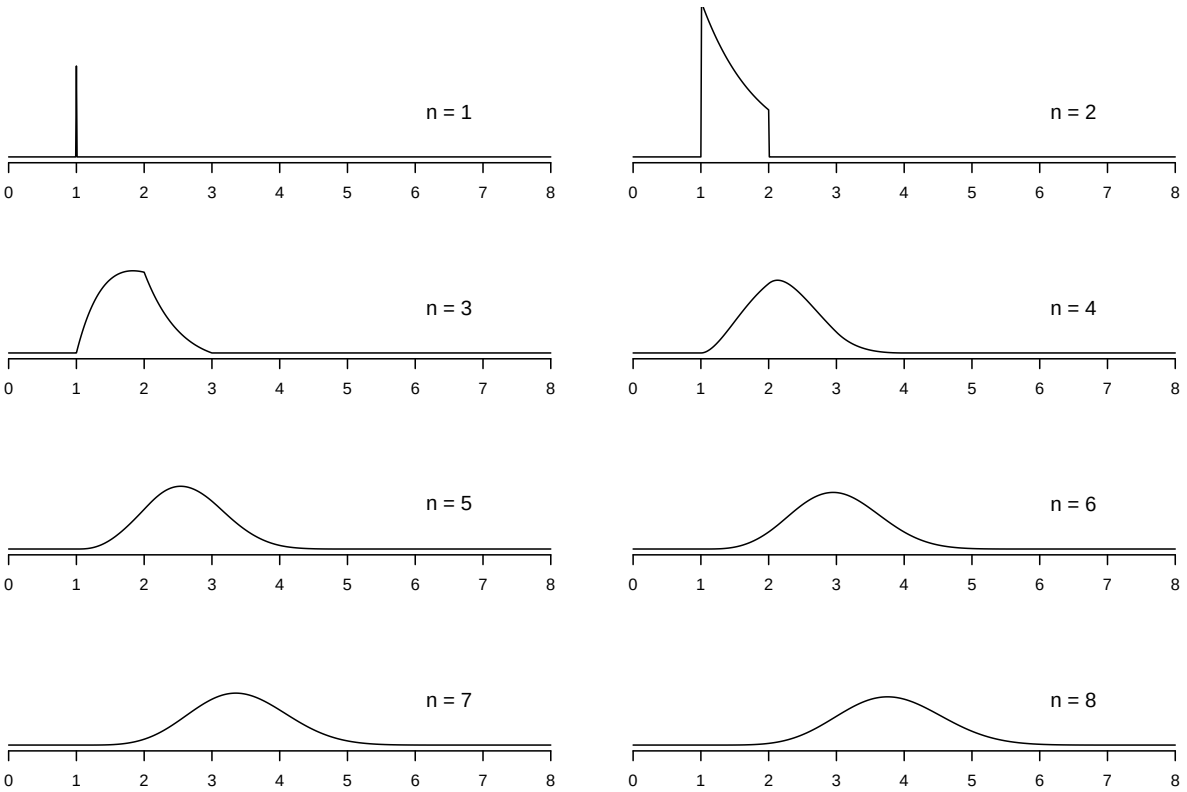


Figure 4.2: Probability densities of phylogenetic diversity (PD), or total branch length, under the Yule process starting at $Y(0) = 1$, ending at $Y(t) = n$ for $n = 1, \ldots, 8$, with $t = 1$ and $\lambda = 1.2$. When $n = 1$, no births have occurred, so the accrued PD must be exactly $x = t = 1$, which we represent here as a point mass at 1. For $n = 2$, the minimum accumulated PD is one, since the process spent at most one unit of time with one species; likewise the maximum accumulated PD is two, since the process spent at most one unit of time with two species. The functional form of (4.7) reveals the piecewise nature of the density, which gradually becomes smoother as n becomes large. The vertical probability axis is the same for all plots.

perform a hypothesis test to evaluate the Yule-PD model using the quantity

$$\Pr(\text{PD} > x) = \int_x^{nt} f_Y(x)\ \mathrm{d}x \tag{4.9}$$

where $f_Y(x)$ is given by (4.8) and nt is the maximum PD that can accumulate in time t, conditional on $Y(t) = n$.

4.4 The distribution of expected phenotypic variance

Since researchers generally do not know the phylogenetic tree for a set of species with certainty, PD is not observable until after a tree has been estimated. Unfortunately, the uncertainty involved in estimating a tree propagates into subsequent estimates of PD based on that tree, and our distributional results may no longer apply. We therefore seek a distribution for an analogous quantity that is observable directly from knowledge of the number of species n and their trait values, bypassing the need to infer a detailed phylogenetic tree. For this, we will need a model for phenotypic trait evolution on the branches of an unknown phylogenetic tree generated by a Yule process.

The simplest and most popular model for evolution of continuous phenotypic traits on phylogenetic trees is Brownian motion (Felsenstein, 1985). Under this model, trait increments over a branch of length t are normally distributed with mean 0 and variance $\sigma^2 t$. The trait values for extant species at the present time are observed as the vector $\mathbf{X} = (X_1, \ldots, X_n)$. For a given topology τ with n taxa and branching times $\mathbf{t} = (t_2, \ldots, t_{n-1})$, with tip data $\mathbf{X}$ generated on the branches of this tree by zero-mean Brownian motion with variance σ^2, the tip data are distributed according to a multivariate normal random variable. More formally,

$$\mathbf{X} \sim N\big(\mathbf{0}, \sigma^2 \mathbf{C}(\tau, \mathbf{t})\big), \tag{4.10}$$

where the entries of the variance-covariance matrix $\mathbf{C}(\tau, \mathbf{t}) = \{c_{ij}\}$ are defined as follows: $c_{ii} = t$, and c_{ij} is the time of shared ancestry for taxa i and j, where $i \neq j$. O'Meara et al (2006) introduces *disparity*, the sample variance of the tip data $\mathbf{X}$,

$$\text{disparity}(\mathbf{X}) = \frac{1}{n}(\mathbf{X} - \bar{X})'(\mathbf{X} - \bar{X}), \tag{4.11}$$

where $\bar{X}$ is the mean of the elements of $\mathbf{X}$. The expectation of the disparity, conditional on the tree topology τ, branching times $\mathbf{t}$, and the Brownian variance σ^2, is

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}(\text{disparity} \mid \tau, \mathbf{t}, \sigma^2) &= \frac{1}{n}\mathbb{E}_{\mathbf{X}}\big((\mathbf{X}-\bar{X})'(\mathbf{X}-\bar{X}) \mid \tau, \mathbf{t}, \sigma^2\big) \\
&= \sigma^2\left[\frac{\operatorname{tr}\big(\mathbf{C}(\tau,\mathbf{t})\big)}{n} - \frac{1}{n^2}\mathbf{1}'\mathbf{C}(\tau,\mathbf{t})\mathbf{1}\right] \\
&= \sigma^2\left[t - \frac{1}{n^2}\mathbf{1}'\mathbf{C}(\tau,\mathbf{t})\mathbf{1}\right] \\
&= \sigma^2\left[t - \frac{1}{n^2}\left(nt + 2\sum_{i=1}^{n}\sum_{j<i} c_{ij}\right)\right] \\
&= \sigma^2\left[\left(1-\frac{1}{n}\right)t - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j<i} c_{ij}\right],
\end{aligned} \tag{4.12}$$

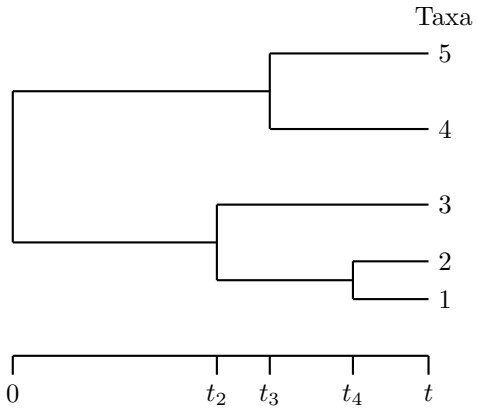
where we use the notation $\mathbb{E}_{\mathbf{X}}$ to indicate that the expectation is taken over realizations of the Brownian process that generates $\mathbf{X}$. The fourth line above arises since the matrix $\mathbf{C}(\tau,\mathbf{t})$ is symmetric and every element on the diagonal is t. However, every entry c_{ij} is either zero or a branching time from the vector $\mathbf{t}$, so the nonzero terms in the sum consist of branching times t_k. Let z_k be the coefficient multiplying the branching time t_k in the last line of (4.12). Then we can express expected disparity as a weighted sum of the branching times,

$$\mathbb{E}_{\mathbf{X}}(\text{disparity} \mid \tau, \mathbf{t}, \sigma^2) = \sigma^2 \sum_{k=2}^{n} z_k t_k. \tag{4.13}$$

Figure 4.3 illustrates how tree topology determines the matrix $\mathbf{C}(\tau,\mathbf{t})$ and expected disparity.

4.4.1 Expected disparity as an accumulated reward

The expected disparity (4.13) has features in common with PD, since it is a scalar quantity that accumulates over the branches of the tree from time 0 to t. The difference is that disparity implicitly incorporates tree-topological factors, which enter (4.13) as weights in the sum of the branch lengths. In addition, a Yule tree accumulates PD even when there is a single lineage, but the same is not true for disparity. We develop these ideas in greater detail in this section.



$$\mathbf{C}(\tau, \mathbf{t}) = \begin{pmatrix} t & t_4 & t_2 & 0 & 0 \\ t_4 & t & t_2 & 0 & 0 \\ t_2 & t_2 & t & 0 & 0 \\ 0 & 0 & 0 & t & t_3 \\ 0 & 0 & 0 & t_3 & t \end{pmatrix}$$

$$\mathbb{E}(\text{disparity}|\tau, \mathbf{t}, \sigma^2) = \sigma^2 \left(\frac{4}{5}t - \frac{4}{25}t_2 - \frac{2}{25}t_3 - \frac{2}{25}t_4 \right)$$

Figure 4.3: How tree topology determines the matrix of Brownian covariances and expected disparity. At left, a tree topology τ of crown age t has 5 taxa and branch times $\mathbf{t} = (t_2, t_3, t_4)$, where t_k is the time of the branch from k to $k+1$ lineages. At right, the corresponding matrix $\mathbf{C}(\tau, \mathbf{t})$ of Brownian covariances. The diagonal entries of $\mathbf{C}(\tau, \mathbf{t})$ are all t. The (i, j)th entry of $\mathbf{C}(\tau, \mathbf{t})$ is the time of shared ancestry between taxa i and j, for $i \neq j$. For example, taxa 1 and 2 share ancestry for time t_4. Expected disparity (using Brownian variance σ^2) is calculated using (4.12). Trait disparity cannot accumulate when there is only one species, so we draw the tree τ beginning with two lineages at time 0.

Our goal is to express (4.13) as a Markov reward process in a form equivalent to (4.4),

$$R_t = \sigma^2 \sum_{k=1}^{n} a_k (t_k - t_{k-1}). \tag{4.14}$$

From (4.12) we see that $a_n = z_n = \left(1 - \frac{1}{n}\right)$, and a_{n-1} can be found using a_n and z_{n-1}, and so on. We can formalize this recursive solution for the rewards by equating (4.13) and (4.14) as follows:

$$\sum_{k=2}^{n} z_k t_k = \sum_{k=1}^{n} a_k (t_k - t_{k-1}) = a_n t_n + \sum_{k=1}^{n-1} (a_k - a_{k+1}) t_k. \tag{4.15}$$

Then recursively solving for the a_k's gives $a_1 = 0$ and

$$a_k = \sum_{j=k}^{n} z_j. \tag{4.16}$$

for $k = 2, \ldots, n$. Now defining $R_t(\mathbf{a})$ to be the Yule reward process with rewards $\mathbf{a} = (a_1, \ldots, a_n)$ under the topology τ, the expected disparity is distributed as

$$\mathbb{E}_{\mathbf{X}}(\text{disparity} \mid \tau, \lambda, \sigma^2, n) \sim \sigma^2 R_t(\mathbf{a}). \tag{4.17}$$

Note that we no longer need to condition on the branch lengths $\mathbf{t} = (t_1, \ldots, t_n)$ in the expected disparity – they have been "integrated out". Therefore, to find the distribution of expected trait variance under a Brownian motion process on a Yule tree with topology τ, we need only find the relevant rewards $\mathbf{a}$ and compute the corresponding distribution of $R_t(\mathbf{a})$.

As a concrete example, consider the five-taxon tree in Figure 4.3. The expected disparity, given this topology τ and arbitrary branch lengths $\mathbf{t} = (t_2, t_3, t_4)$, is

$$\mathbb{E}(\text{disparity}|\tau, \mathbf{t}, \sigma^2) = \sigma^2 \left[\frac{4}{5} t - \frac{4}{25} t_2 - \frac{2}{25} t_3 - \frac{2}{25} t_4 \right]. \tag{4.18}$$

The coefficients are given by

$$\mathbf{z} = \left(-\frac{4}{25},\ -\frac{2}{25},\ -\frac{2}{25},\ \frac{4}{5} \right). \tag{4.19}$$

Solving for the rewards $\mathbf{a}$, we obtain

$$\mathbf{a} = \left(0, \frac{12}{25},\ \frac{16}{25},\ \frac{18}{25},\ \frac{4}{5} \right) \tag{4.20}$$

which is easily verified by hand. This leads us to our second Theorem, which gives an expression for the distribution of $R_t(\mathbf{a})$.

Theorem 2. *In a Yule process with rate λ and arbitrary rewards $\mathbf{a} = (a_1, \ldots, a_n)$, the Laplace transform of $v_{mn}(x, t|\mathbf{a})$ is given by*

$$f_{mn}(r,t) = \begin{cases} e^{-(m\lambda + a_m r)t} & m = n, \text{ and} \\ \lambda^{n-m} \dfrac{(n-1)!}{(m-1)!} \displaystyle\sum_{j=m}^{n} \dfrac{e^{-(j\lambda + a_j r)t}}{\prod_{k \neq j} \left(\lambda(k-j) + r(a_k - a_j)\right)} & n > m. \end{cases} \tag{4.21}$$

The proof of this Theorem is given in Appendix 4.9. To obtain the probability distribution of the accumulated reward, we must invert (4.21),

$$v_{mn}(x,t) = \mathscr{L}^{-1}\big[f_{mn}(r,t)\big](x). \tag{4.22}$$

For $m = n$ and $n = m + 1$, there are simple expressions for the inverse Laplace transform. Under certain conditions on the rewards $\mathbf{a}$, there is a straightforward analytic inversion of (4.21) for general $n > m$, but the rewards computed using the times of shared ancestry in a phylogenetic tree do not always satisfy these conditions. Therefore, it is often easier to numerically invert (4.21); we discuss this issue in much greater detail in Appendix 4.10, and provide a straightforward method for numerical inversion of the Laplace transform (4.21) based on the method popularized by Abate and Whitt (1995).

4.4.2 Approximate likelihood and inference for σ^2

We now describe a statistical procedure for using Theorem 2 to perform statistical for the unknown Brownian variance σ^2. Suppose that in a clade of n species we have a crude tree topology τ (without branch lengths). This topology could be derived from parsimony, distance-based tree reconstruction methods, or one could simply use family/genus/species information to assign a hierarchy of relationships and resolve polytomies randomly. Given the tree topology τ, one can compute the rewards $\mathbf{a}$. Suppose also that we have calculated trait disparity for each of J independent continuous quantitative traits that arise from Brownian motion on the branches of the unknown phylogenetic tree, starting at the root. Let

$$D_n^{(j)} = \frac{1}{n}\left(\mathbf{X}^{(j)} - \bar{X}^{(j)}\right)'\left(\mathbf{X}^{(j)} - \bar{X}^{(j)}\right), \tag{4.23}$$

be the observed disparity for the jth phenotypic trait, where $\mathbf{X}^{(j)}$ is the vector of n trait values for the jth phenotypic trait and $\bar{X}^{(j)}$ is the mean of the elements of $\mathbf{X}^{(j)}$. Then we calculate the mean disparity $\bar{D}_n$ across these J traits:

$$\bar{D}_n = \frac{1}{J}\sum_{j=1}^{J} D_n^{(j)}. \tag{4.24}$$

Note that in order to find $\bar{D}_n$, we do not need the individual trait measurements themselves – only the disparities. Then by the law of large numbers, $\bar{D}_n \to \mathbb{E}(D_n)$ as $J \to \infty$, where D_n is the asymptotic mean disparity across all possible traits. Therefore, we approximate the distribution of $\bar{D}_n$ as follows:

$$\begin{aligned} \bar{D}_n &\approx \mathbb{E}(D_n) \\ &\sim \mathbb{E}_{\mathbf{X}}(\text{disparity} \mid \tau, \sigma^2, \lambda, n, t) \\ &= \sigma^2 R_t(\mathbf{a}). \end{aligned} \tag{4.25}$$

where $\mathbf{a}$ is the vector of rewards obtained from the topology τ. This approximate relation provides the connection between observable mean trait disparity and the probability distribution in Theorem 2 that we need in order to compute the probability of the observed disparities. Suppose the stem age of an unresolved tree is t, and let

$$f_Y(x) = \frac{v_{mn}(x|t, \lambda, \mathbf{a})}{P_{mn}(t)} \tag{4.26}$$

be the distribution of expected disparity in a Yule process with general rewards $\mathbf{a}$, conditional on $Y(0) = m$ and $Y(t) = n$. Here, x is the expected trait disparity, which we approximate by our observed (and therefore fixed) $\bar{D}_n$. From (4.25), we write

$$\frac{\bar{D}_n}{\sigma^2} \sim R_t(\mathbf{a}) \tag{4.27}$$

so the likelihood is approximately

$$f_Y(\bar{D}_n/\sigma^2). \tag{4.28}$$

Finally, we propose the approximate maximum likelihood estimator

$$\hat{\sigma}^2 = \underset{\sigma^2}{\text{argmax}}\ f_Y(\bar{D}_n/\sigma^2). \tag{4.29}$$

To find $\hat{\sigma}^2$, note that in a Yule reward process in which $a_k < a_j$ for $k < j$, the value of the reward is constrained to lie in the interval

$$a_m t \leq R_t(\mathbf{a}) \leq a_n t \tag{4.30}$$

where we have assumed $Y(0) = m$ and $Y(t) = n$. Additionally, when none of the rewards a_k are zero we are justified in dividing $\bar{D}_n$ by the terms in (4.30) to obtain bounds on the possible value of $\hat{\sigma}^2$ that maximizes (4.28),

$$\frac{\bar{D}_n}{a_n t} \leq \hat{\sigma}^2 \leq \frac{\bar{D}_n}{a_m t}. \tag{4.31}$$

When $m = 1$ and $a_1 = 0$, the upper bound is infinity. However, in practice when $m = 1$ and $n \geq 4$, it is safe to assume that

$$\frac{\bar{D}_n}{a_n t} \leq \hat{\sigma}^2 \leq \frac{\bar{D}_n}{a_2 t}. \tag{4.32}$$

We solve (4.29) using the numerical Newton-Raphson method provided by the `R` function `nlm`.

We emphasize that (4.29) is not a traditional maximum likelihood estimator. We have approximated the distribution of $\bar{D}_n$ by the distribution of expected disparity, giving an approximate likelihood (4.28) that may not attain its maximum at the same value of σ^2 as the true likelihood. In addition, the density (4.28) is non-differentiable at several points, and for $n = m + 1$ attains its maximum at the lower boundary $R_t(\mathbf{a}) = a_m t$ (see Figure 4.2 with $m = 1$, $n = 2$). These issues complicate application of traditional asymptotic theory for maximum likelihood estimates, and the classical large sample theory may not apply because the likelihood is only an approximation. One consequence of the violation of these traditional assumptions is that we are unable to provide meaningful standard errors for $\hat{\sigma}^2$ using only the approximate likelihood (4.28).

4.4.3 Simulations

To empirically evaluate the correctness of the analytic distributions we derived in Theorem 2, we simulated trait data via Brownian motion on trees generated by a Yule process with age $t = 1$ and branching rate $\lambda = 1$. For $n = 3, \ldots, 8$, we chose one tree topology and

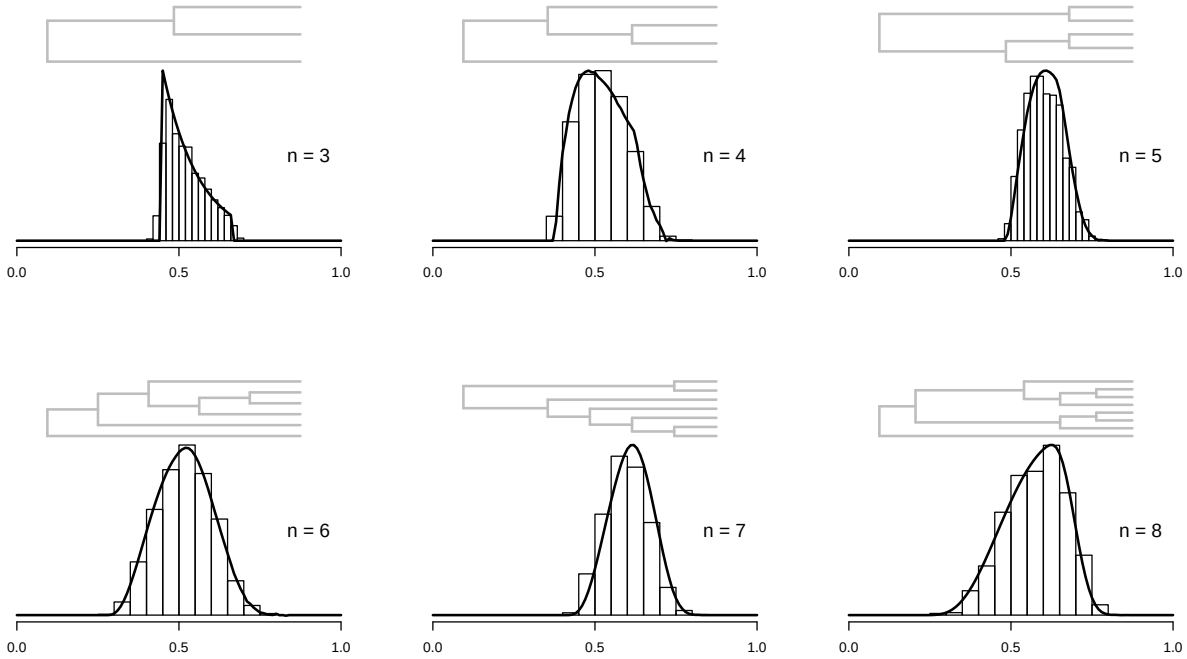


Figure 4.4: Empirical correspondence between the derived expressions for the distribution of expected disparity and simulated mean disparity histograms for trees with different numbers of taxa. For each $n = 3, \ldots, 8$, we simulated a single tree topology (shown above each histogram in gray). We then simulated 2000 sets of branch lengths for this topology under the Yule process. For each set of branch lengths, we calculated the mean disparity from 2000 simulations of zero-mean Brownian motion with variance $\sigma^2 = 1$ on this tree.

simulated $N_{\text{btimes}} = 2000$ sets of branch lengths from a Yule process for n species (Stadler, 2011). For each set of branching times, we simulated $N_{\text{BM}} = 2000$ realizations of Brownian motion with $\sigma^2 = 1$ to generate trait values at the tips of the tree (Paradis et al, 2004). For each of the 2000 sets of tip values, we calculated the mean disparity using (4.12). Figure 4.4 shows histograms of the mean disparities with the analytic distribution $f_Y(x)$ overlaid, with good correspondence. The tree topology (with arbitrary branch lengths for display) is shown in gray above each histogram.

To evaluate our estimation methodology, we take a similar approach, but for each simulated set of mean disparities, we infer $\hat{\sigma}^2$, an approximate maximum likelihood estimate of

σ^2. Figure 4.5 shows estimates of σ^2 for different species richness n under different simulation conditions. For $n = 3, \ldots, 12$, we generated 100 trees, each with $N_{\text{btimes}} = 1, 5, 10, 100$ sets of branching times. For each set of branching times, we evolved $N_{\text{BM}} = 1, 5, 10, 100$ traits by Brownian motion along the branches with rate $\sigma^2 = 1$ and computed the mean disparity. Then, given the N_{btimes} mean disparities, we maximized the approximate likelihood to find $\hat{\sigma}^2$. Each dot in Figure 4.5 represents one estimate, and the dots are jittered slightly to show their density. The variance in the estimator is large when the number of simulated branching time sets and Brownian realizations is small since (4.25) and hence (4.27) become poor approximations to the mean disparity. However, the approximate maximum likelihood estimator for σ^2 appears to have the desirable property of statistical consistency: the deviation of the estimates from the true value $\sigma^2 = 1$ goes to zero as the number of mean disparity observations becomes large.

4.5 Application to evolution of body size in the order Carnivora

To illustrate the usefulness of our method in practical comparative inference, we estimate thirteen family-wise Brownian variance rates for body size evolution in the mammalian order Carnivora using observed log-body size disparities and species richness information (Gittleman, 1986; Gittleman and Purvis, 1998; Slater et al, 2012). Carnivora includes members with very large and small body masses, including wide diversity within individual families (Nowak and Paradiso, 1999). We included only families with 2 or more species, since families with only one species do not reveal useful information about intra-family Brownian variance. The dataset, comprising 284 species, included the families Canidae, Eupleridae, Felidae, Herpestidae, Hyaenidae, Mephitidae, Mustelidae, Otariidae, Phocidae, Prionodontidae, Procyonidae, Ursidae, and Viverridae. Figure 4.6 shows the backbone phylogeny (from Eizirik et al (2010)) and the unresolved clades. Our analysis takes advantage of utilities for manipulating trees and quantitative trait data in the `ape` package (described in Paradis et al (2004)) and simulating trees with the `TreeSim` package (described in Stadler (2011)), using the statistical programming language `R` (CRAN, 2012). We intentionally limit our

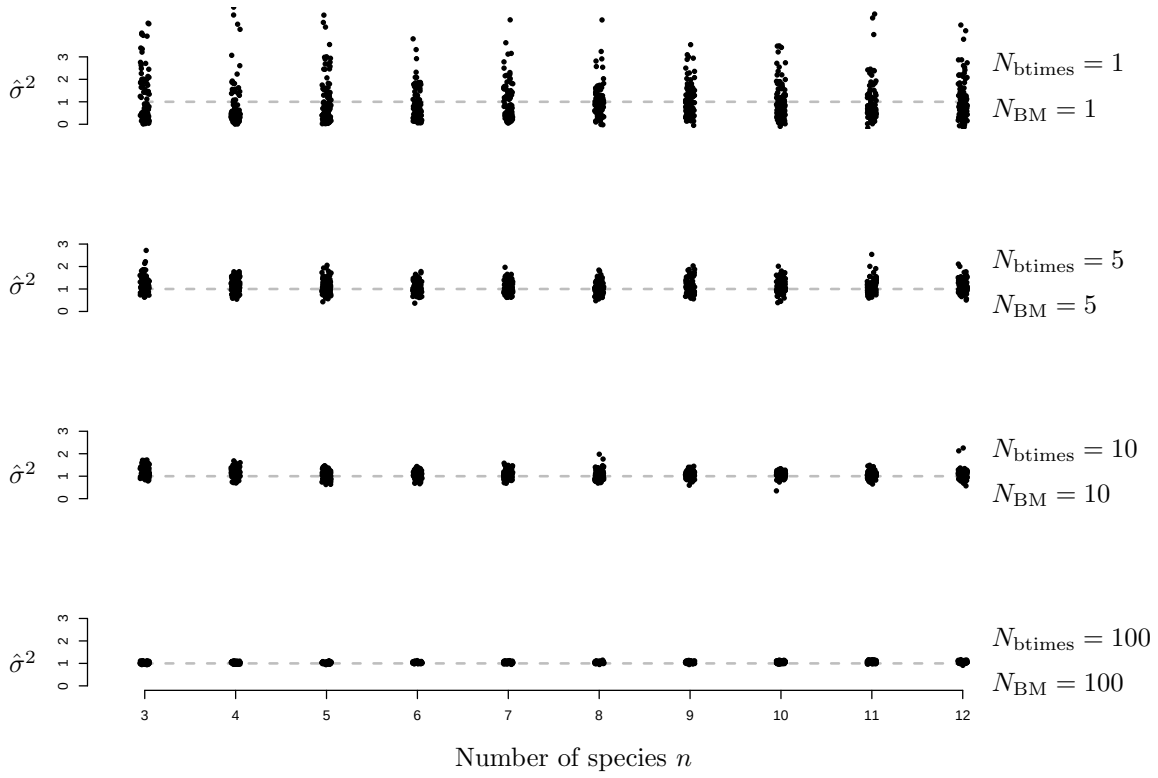


Figure 4.5: Empirical consistency of approximate maximum likelihood estimates of σ^2. The number of species n in the unobserved phylogenetic tree is shown on the horizontal axis. Each set of plots shows 100 estimates of σ^2 from N_{btimes} simulations of different branching times under a Yule process with n species and from N_{BM} independent realizations of Brownian motion used to compute the mean disparity for each set of branching times. The estimates are jittered to show the sampling distribution. The gray dotted line shows the true value $\sigma^2 = 1$.

analysis to the Brownian variance in each family and use only body size disparity in order to demonstrate the simplicity of our approximate method under conditions of very little data.

The first step is to estimate the speciation rate λ from the backbone tree and the unresolved clades. Even though the tree is unobserved within each family, we can still find the exact maximum likelihood estimate of λ for the tree as a whole. On a fully resolved branch of length t, the log-likelihood of λ is $\log(\lambda) - \lambda t$. For an unresolved clade of age t with one species at the crown which grows to include n species, the log-likelihood from (4.2) is proportional to $-\lambda t + (n-1)\log(1-e^{-\lambda t})$. Summing these partial log-likelihoods over the whole tree gives the log-likelihood for λ; maximizing this function, we find that $\hat{\lambda} = 0.069$ per million years. In what follows, we assume that $\lambda = \hat{\lambda}$. To each unresolved family clade we associate the clade disparity for body size. Our analysis will consist of estimating the family-wise Brownian variance σ_j^2 for the jth family, assuming one species at the stem age for each clade shown in Figure 4.6. In this way, we integrate over crown ages for each family under the Yule model. To apply the methodology for modeling expected trait disparity developed in section 4.4, we regard the unresolved clades as "soft polytomies" in which branch lengths are unknown (Purvis and Garland, 1993). We therefore resolve these polytomies randomly without assigning branch lengths.

Our analysis of the Carnivora family-wise Brownian variances takes approximately 30 seconds to run on a laptop computer. However, to evaluate the variability in our estimates, we ran the analysis 100 times with randomly resolved polytomies for each family. Table 4.1 shows the family name, species richness (n), stem age (t), observed body size disparity ($\bar{D}_n$), the mean estimate of σ_j^2, and the approximate standard error of the estimate for each family. We calculated standard errors as the empirical standard deviation of the 100 Brownian variance estimates. Our estimates of σ^2 reveal readily interpretable information about the evolution of body size in each family that is not available directly from observed disparities alone. For example, the families Herpestidae and Viverridae have almost identical species richness, but quite different disparity measurements. Perhaps surprisingly, we have estimated nearly equal Brownian variances for the two families. Why does our method produce such similar estimates of σ^2? The answer lies in the ages of the clades – Viverridae

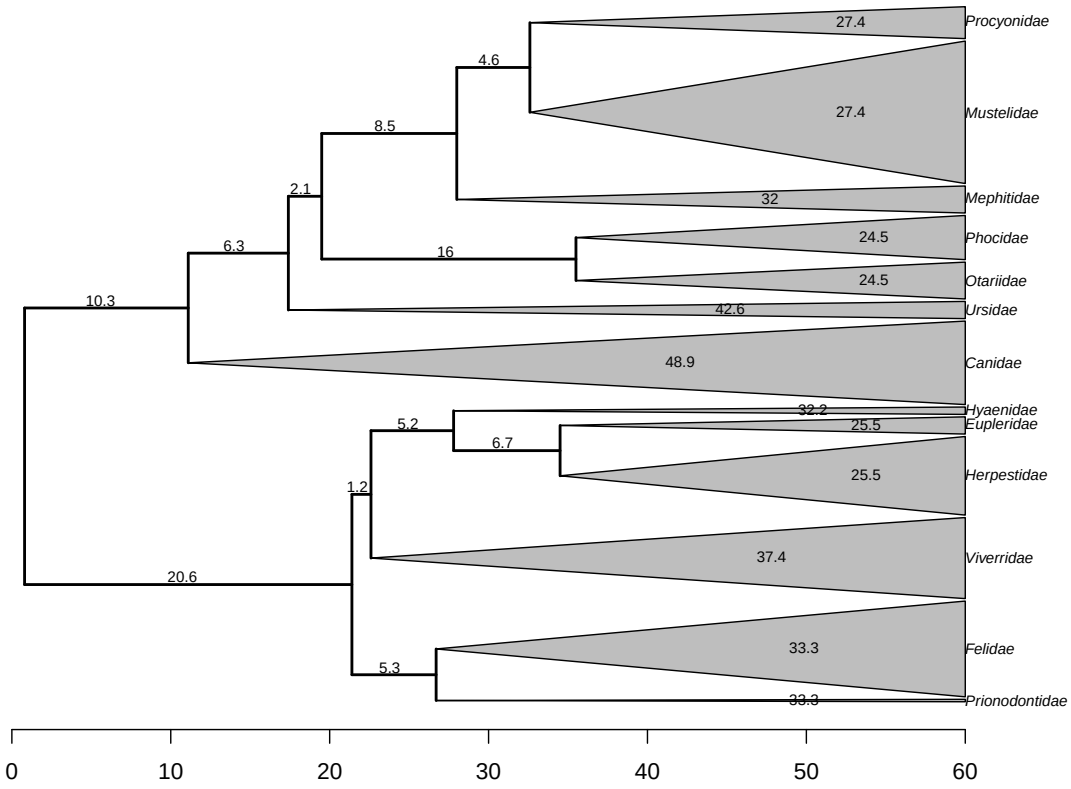


Figure 4.6: A family-level phylogenetic tree for order Carnivora. The phylogeny within each family, shown here as a gray triangle, is not known with certainty. The length of the base of the gray triangles represents the number of species in the family. The "backbone" tree connecting the unresolved clades is assumed fixed. Branch lengths are shown along each branch.

Family	Richness	TMRCA	Disparity	$\hat{\sigma}^2$	SE
Prionodontidae	2	33.30	0.131	0.051	0
Felidae	40	33.30	1.588	0.118	0.094
Viverridae	34	37.40	0.606	0.034	0.018
Herpestidae	33	25.50	0.482	0.035	0.017
Eupleridae	8	25.50	0.916	0.081	0.010
Hyaenidae	4	32.20	0.805	0.084	0.001
Canidae	35	48.90	0.678	0.031	0.012
Ursidae	8	42.60	0.303	0.025	0.002
Otariidae	16	24.50	0.386	0.029	0.005
Phocidae	19	24.50	0.751	0.065	0.030
Mephitidae	12	32.00	0.570	0.038	0.004
Mustelidae	59	27.40	2.263	0.220	0.239
Procyonidae	14	27.40	0.531	0.038	0.006

Table 4.1: Species richness, TMRCA, body size disparity, and estimated Brownian variance $\hat{\sigma}^2$ for each family in the Carnivora dataset. Note that Canidae and Herpestidae have very different disparity measurements, but nearly identical estimates of σ^2. This discrepancy is due to the difference in their ages; we explain the interaction between time, species number and disparity in greater detail in the text.

is almost 50% older than Herpestidae. Two clades with the same richness whose traits evolve by Brownian motion at the same rate can exhibit very different disparity measurements, depending on their ages. This example illustrates how our method provides an approximate way to untangle the complex interaction of time, species, and observed trait variance (in the words of Ricklefs (2006)) for unresolved clades.

4.6 Discussion: Comparative phylogenetics without trees?

In this paper we have outlined a method for integrating over Yule trees. We presented an expression for the distribution of PD in an unresolved tree, conditional on the number of species n and age t. We showed that the expected disparity can be represented in a similar way as PD, since it accumulates along the branches of a phylogenetic tree. We also derived a statistical framework that uses a very small amount of information (n, t, and $\bar{D}_n$) for an unresolved clade to derive a meaningful estimator for the Brownian rate σ^2. It may seem counterintuitive that one can estimate the Brownian rate for an unresolved tree with n taxa, given a single disparity measurement. However, the structure provided by the Yule process allows this inference by providing just enough information about the distribution of branching times that generate the tree to model the average phenotypic disparity under Brownian motion. This permits analytic integration over two random objects: the collection of branching times of the tree and realizations of Brownian motion. Three assumptions make this possible: first, we fix a topology τ without branch lengths; second, we assume that branch lengths come from a Yule process; third, we compute the distribution of expected disparity, which is a scalar quantity that encapsulates the most important information in the covariance matrix $\mathbf{C}(\tau)$. In exchange for these assumptions, we gain what frustrated reviewers of Felsenstein's paper apparently wished for: an estimator that "obviate[s] the need to have an accurate knowledge of phylogeny" (Felsenstein, 1985). Whether these assumptions are warranted depends fundamentally on the scientific questions at hand, and the available data.

Perhaps the most satisfying use of our method is in providing an approximate and model-based answer to the questions posed by Ricklefs (2006) and Bokma (2010), in their similarly-named papers. Our answer is approximate because it substitutes an observation for an expectation in (4.25); it is model-based because we assume that trees arise from a Yule process and traits evolve Brownian motion. Equation (4.25) expresses a heuristic relationship explaining the origins of phenotypic disparity, which we reproduce here for emphasis:

$$\bar{D}_n \sim \sigma^2 R_t(\mathbf{a}). \tag{4.33}$$

On the left-hand side is the observed disparity. On the right-hand side, $R_t(\mathbf{a})$ serves as a scalar summary of tree shape – it depends only on n, clade age t, and tree topology τ. This reveals that even when we restrict our attention to expected disparity under the simplest evolutionary models, the interaction between n, t and the branching structure of the tree in $R_t(\mathbf{a})$ is complex, but the Brownian variance simply scales the tree-topological term. We see that $\bar{D}_n$ scales linearly with σ^2 when n, t, and the topology τ are fixed. However, changing one of n, t, or τ while holding σ^2 constant will induce a nonlinear change in $\bar{D}_n$. We conclude that it is not possible to partition the time-dependent and speciation-dependent influences on the accumulation of trait variance in a simple way as suggested by Ricklefs (2006) under the stochastic models we study in this paper.

As an inducement to spur research on analytic integration over trees, Bokma (2010) offers a monetary reward for an expression for the distribution of sample variance from a birth-death tree with Brownian trait evolution on its branches. We have solved a simpler version of Bokma's challenge by providing the distribution of expected trait variance for a specific topology under a pure-birth process. The expression Bokma (2010) seeks is difficult to find for two reasons: first, it would require analytic integration over discrete tree topologies; second, and more intuitively, integrating over both topologies and Brownian realizations would subsume the Brownian variance σ^2 on the right-hand side of (4.25) into a nonlinear term that depended on n, t, and σ^2 in a very complicated way. Alternatively, simulation-based approaches provide an appealing alternative method to integrate over trees and Brownian motions without requiring approximation of the disparity by its expectation. Indeed, Bayesian methods exist to sample from the distribution of Yule trees, conditional on observed trait values at the tips, thereby providing both estimates of Brownian rates and phylogenies simultaneously, while using all the available trait data (Drummond et al, 2012).

Tree-free comparative evolutionary biology comes at a price – there are several important drawbacks to our approach. First, even under the Yule model for speciation (with the correctly specified branching rate λ) and zero-mean Brownian motion for traits, integrating over all possible Yule trees introduces great uncertainty in estimates of σ^2. Figure 4.5 illustrates this issue: while the estimates of σ^2 eventually converge to the true value as the

number of branch length and trait realizations becomes large, the variance in these estimates can be substantial for smaller datasets. Furthermore, the assumption that $\bar{D}_n \sim \sigma^2 R_t(\mathbf{a})$ may be suspect if the number of traits analyzed is small enough that the mean trait disparity is a poor substitute for the expected disparity.

We conclude with a mixed message about analytic integration over trees. First, it is possible to derive meaningful estimators for parameters of interest under simple evolutionary models, if one is willing to make assumptions about the mean behavior of the models. The estimates are usually reasonable, and may provide valuable insight into the basic properties of evolutionary change under these models – even our simplistic analysis of Carnivora body size evolution reveals the complex interaction of clade age, species number, and evolutionary rate. These estimates may be useful as starting points for more time-consuming simulation analyses. Second, and more pessimistically, sophisticated analytic methods for integrating over trees cannot conjure evolutionary information from the data that is not there already. As evolutionary biologists further refine our knowledge of the tree of life, the number of clades whose phylogeny is truly unknown may diminish, along with interest in tree-free estimation methods.

4.7 Appendix: Markov rewards for Yule processes

In this Appendix, we prove one lemma and the two Theorems presented in the text. In the first proof, we derive a representation of the forward equation for a Yule reward process. Our development follows that given by Neuts (1995).

Lemma 1. *In a Yule reward process $R_t = \int_0^t a_{Y(s)}\, ds$ with arbitrary positive rewards $a_1, a_2, \ldots$, the Laplace-transformed reward probabilities satisfy the ordinary differential equations*

$$\frac{df_{mn}(r,t)}{dt} = -(n\lambda + a_n r) f_{mn}(r,t) + (n-1)\lambda f_{m,n-1}(r,t). \tag{4.34}$$

Proof. Let $V_{mn}(x,t) = \Pr(R_t \leq x, Y(t) = n \mid Y(0) = m)$. We can re-write this quantity in a more useful form by conditioning on the time of departure u from state m, noting that the

accumulated reward is $a_m u$, and then integrating over u. If $m = n$ and no departure occurs, the accumulated reward is $a_m t$. Putting these ideas together, we obtain

$$\begin{aligned} V_{mn}(x,t) &= \Pr(R_t \le x, Y(s) = m \text{ for } 0 \le s \le t) \\ &\quad + \int_0^t \Pr(m \to m+1 \text{ at time } u) V_{m+1,n}(x - a_m u, t-u)\ \mathrm{d}u \\ &= \delta_{mn} e^{-m\lambda t} H(x - a_m t) + \int_0^t m\lambda e^{-m\lambda u} V_{m+1,n}(x - a_m u, t-u)\ \mathrm{d}u. \end{aligned} \tag{4.35}$$

Now consider the Laplace transform $F_{mn}(r,t)$ of $V_{mn}(x,t)$, with respect to the reward variable x,

$$\begin{aligned} F_{mn}(r,t) &= \mathscr{L}\big[V_{mn}(x,t)\big](r) \\ &= \int_0^\infty e^{-rx} V_{mn}(x,t)\ \mathrm{d}x \\ &= \delta_{mn} \frac{e^{-(m\lambda + a_m r)t}}{r} + \int_0^t m\lambda e^{-m\lambda u} \int_{a_m u}^\infty e^{-rx} V_{m+1,n}(x - a_m u, t-u)\ \mathrm{d}x\ \mathrm{d}u. \end{aligned} \tag{4.36}$$

Making the substitution $y = x - a_m u$ in the Laplace integral, we have

$$\begin{aligned} F_{mn}(s,t) &= \delta_{mn} \frac{e^{-(m\lambda + a_m r)t}}{r} + \int_0^t m\lambda e^{-m\lambda u} \int_0^\infty e^{-r(y + a_m u)} V_{m+1,n}(y, t-u)\ \mathrm{d}y\ \mathrm{d}u \\ &= \delta_{mn} \frac{e^{-(m\lambda + a_m r)t}}{r} + \int_0^t m\lambda e^{-(m\lambda + a_m r)u} F_{m+1,n}(r, t-u)\ \mathrm{d}u. \end{aligned} \tag{4.37}$$

Now multiplying both sides by $e^{(m\lambda + a_m r)t}$ and differentiating with respect to t, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}\left[e^{(m\lambda + a_m r)t} F_{mn}(r,t)\right] &= \frac{\partial}{\partial t} \int_0^t m\lambda e^{(m\lambda + a_m r)(t-u)} F_{m+1,n}(r, t-u)\ \mathrm{d}u \\ &= \frac{\partial}{\partial t} \int_0^t m\lambda e^{(m\lambda + a_m r)u} F_{m+1,n}(r,u)\ \mathrm{d}u. \end{aligned} \tag{4.38}$$

Expanding the left-hand side by the product rule and using the fundamental theorem of calculus on the right, we find that

$$e^{(m\lambda + a_m r)t}\left((m\lambda + a_m r) F_{mn}(r,t) + \frac{\partial}{\partial t} F_{mn}(r,t)\right) = e^{(m\lambda + a_m r)t} m\lambda F_{m+1,n}(r,t). \tag{4.39}$$

Cancelling common factors and rearranging, we obtain the Kolmogorov backward equation,

$$\frac{\partial}{\partial t} F_{mn}(r,t) = -(m\lambda + a_m r) F_{mn}(r,t) + m\lambda F_{m+1,n}(r,t). \tag{4.40}$$

However,

$$rF_{mn}(r,t) = \mathscr{L}\left[\frac{\partial}{\partial x}V_{mn}(x,t)\right](r) = \mathscr{L}\big[v_{mn}(x,t)\big](r) = f_{mn}(r,t). \tag{4.41}$$

Plugging $rF_{mn}(r,t) = f_{mn}(r,t)$ into (4.40), we find that the $f_{mn}(r,t)$ satisfy the same system of ordinary differential equations,

$$\frac{\partial}{\partial t}f_{mn}(r,t) = -(m\lambda + a_m r)f_{mn}(r,t) + m\lambda f_{m+1,n}(r,t). \tag{4.42}$$

These are the *backward equations* for the Laplace transformed reward process. To solve (4.42), we note that any solution to the forward equations is a solution to the backward equations in a birth process (Grimmett and Stirzaker, 2001). Therefore, (4.42) is equivalent to the forward system

$$\frac{\partial f_{mn}(r,t)}{\partial t} = -(n\lambda + a_n r)f_{mn}(r,t) + (n-1)\lambda f_{m,n-1}(r,t) \tag{4.43}$$

for $n = m, m+1, m+2, \ldots$ This completes the proof. □

4.8 Appendix: Proof of Theorem 1

Proof. Lemma 1 with $a_k = k$ for $k = 0, 1, \ldots$ gives

$$\frac{\partial f_{mn}(r,t)}{\partial t} = -n(\lambda + r)f_{mn}(r,t) + (n-1)\lambda f_{m,n-1}(r,t). \tag{4.44}$$

Define $g_{mn}(r,s)$ to be the Laplace transform of $f_{mn}(r,t)$ with respect to the time variable t. Transforming (4.44) gives

$$sg_{mn}(r,s) - \delta_{mn} = -n(\lambda + r)g_{mn}(r,s) + (n-1)\lambda g_{m,n-1}(r,s). \tag{4.45}$$

Letting $m = n$, we find that

$$g_{mm}(r,s) = \frac{1}{s + m(\lambda + r)}. \tag{4.46}$$

Next, we form a recurrence and solve for $g_{mn}(r,s)$ to obtain

$$\begin{aligned} g_{mn}(r,s) &= \frac{(n-1)\lambda}{s + n(\lambda + r)}g_{m,n-1}(r,s) \\ &= \frac{(n-1)\cdots m\lambda^{n-m}}{\prod_{j=m+1}^{n}\big[s + j(\lambda + r)\big]}g_{m,m+1}(r,s) \\ &= \frac{(n-1)!}{(m-1)!}\frac{\lambda^{n-m}}{\prod_{j=m}^{n}\big[s + j(\lambda + r)\big]}. \end{aligned} \tag{4.47}$$

We proceed via a partial fractions decomposition of the product in the denominator above,

$$\begin{aligned} g_{mn}(r,s) &= \lambda^{n-m}\frac{(n-1)!}{(m-1)!}\sum_{j=m}^{n}\left(\prod_{k\neq j}(\lambda+r)(k-j)\right)^{-1}\frac{1}{s+j(\lambda+r)} \\ &= \frac{(n-1)!}{(m-1)!}\lambda^{n-m}\sum_{j=m}^{n}\frac{\left[\left(\prod_{k=m}^{j-1}(k-j)\right)\left(\prod_{k=j+1}^{n}(k-j)\right)\right]^{-1}}{(\lambda+r)^{n-m}}\frac{1}{s+j(\lambda+r)} \\ &= \frac{(n-1)!}{(m-1)!}\lambda^{n-m}\sum_{j=m}^{n}\frac{\left[(-1)^{j-m}(j-m)!(n-j)!\right]^{-1}}{(\lambda+r)^{n-m}}\frac{1}{s+j(\lambda+r)} \\ &= \lambda^{n-m}\sum_{j=m}^{n}\binom{n-1}{j-1}\binom{j-1}{m-1}\frac{(-1)^{j-m}}{(\lambda+r)^{n-m}}\frac{1}{s+j(\lambda+r)}. \end{aligned} \tag{4.48}$$

when $n > m$. Inverse transforming with respect to s, we obtain

$$f_{mm}(r,t) = e^{-m(\lambda+r)t} \tag{4.49}$$

and

$$f_{mn}(r,t) = \lambda^{n-m}\sum_{j=m}^{n}\binom{n-1}{j-1}\binom{j-1}{m-1}\frac{(-1)^{j-m}}{(\lambda+r)^{n-m}}e^{-j(\lambda+r)t} \tag{4.50}$$

when $n > m$. Again inverse transforming (4.50), this time with respect to the Laplace reward variable r, we find that for $m = n$,

$$v_{mm}(x,t) = \delta(x-mt)e^{-m\lambda t} \tag{4.51}$$

which is a point mass at $x = mt$. For $n > m$,

$$v_{mn}(x,t) = \frac{\lambda^{n-m}e^{-\lambda x}}{(n-m-1)!}\sum_{j=m}^{n}\binom{n-1}{j-1}\binom{j-1}{m-1}(-1)^{j-m}(x-jt)^{n-m-1}H(x-jt). \tag{4.52}$$

This completes the proof. □

4.9 Appendix: Proof of Theorem 2

Proof. Lemma 1 with arbitrary rewards a_k, $k = 1, 2, \ldots$, gives

$$\frac{\mathrm{d}f_{mn}(r,t)}{\mathrm{d}t} = -(n\lambda + a_n r)f_{mn}(r,t) + (n-1)\lambda f_{m,n-1}(r,t). \tag{4.53}$$

To solve the system, apply the Laplace transform with respect to time t. First note that the transform of $f_{mm}(r,t)$ is

$$g_{mm}(r,s) = \frac{1}{s + m\lambda + a_m r}. \tag{4.54}$$

Transforming the nth equation, and recalling that $f_{mn}(r,0) = 0$ for $n > m$,

$$\begin{aligned}
sg_{mn}(r,s) - f_{mn}(r,0) &= -(n\lambda + a_n r)g_{mn}(r,s) + (n-1)\lambda g_{m,n-1}(r,s) \\
g_{mn}(r,s)(s + n\lambda + a_n r) &= (n-1)\lambda g_{m,n-1}(r,s) \\
g_{mn}(r,s) &= \frac{(n-1)\lambda}{s + n\lambda + a_n r} g_{m,n-1}(r,s) \\
&= \frac{(n-1)!}{(m-1)!} \frac{\lambda^{n-m}}{\prod_{j=m+1}^{n}(s + j\lambda + a_j r)} g_{mm}(r,s) \\
&= \frac{(n-1)!}{(m-1)!} \frac{\lambda^{n-m}}{\prod_{j=m}^{n}(s + j\lambda + a_j r)}.
\end{aligned} \tag{4.55}$$

We expand the denominator by partial fractions to find

$$g_{mn}(r,s) = \lambda^{n-m} \frac{(n-1)!}{(m-1)!} \sum_{j=m}^{n} \frac{\prod_{k \neq j} \left[\lambda(k-j) + r(a_k - a_j)\right]^{-1}}{s + j\lambda + a_j r}. \tag{4.56}$$

Transforming back to the time domain, we have, for $m = n$,

$$f_{mm}(r,t) = e^{-(m\lambda + a_m r)t}. \tag{4.57}$$

When $n > m$,

$$f_{mn}(r,t) = \lambda^{n-m} \frac{(n-1)!}{(m-1)!} \sum_{j=m}^{n} \frac{e^{-(j\lambda + a_j r)t}}{\prod_{k \neq j} \big(\lambda(k-j) + r(a_k - a_j)\big)}. \tag{4.58}$$

This completes the proof. □

4.10 Appendix: Analytic and numerical inversion

Analytic inversion of (4.21) in Theorem 2 is possible, but unfortunately depends on the structure of the tree topology in unexpected ways. One convenient property of the rewards $\mathbf{a} = (a_1, \ldots, a_n)$ is that $a_i < a_{i+1}$ for all $1 \leq i \leq n-1$, a fact apparent from (4.12). Therefore $a_i \neq a_j$ for distinct i and j. When $m = n$, no speciation events have taken place, and we have

$$v_{mm}(x,t) = \delta(x - a_m t)e^{-m\lambda t}. \tag{4.59}$$

For $n = m + 1$, there is only one distinct topology, so

$$v_{m,m+1}(x,t) = \frac{m\lambda e^{-m\lambda t}}{a_{m+1} - a_m}\left[\exp\left(-\frac{\lambda(x - a_m t)}{a_{m+1} - a_m}\right) H(x - a_m t) \right. \\ \left. - e^{\lambda t}\exp\left(-\frac{\lambda(x - a_{m+1} t)}{a_{m+1} - a_m}\right) H(x - a_{m+1} t)\right]. \tag{4.60}$$

In general, when

$$\frac{\ell - j}{a_\ell - a_j} - \frac{k - j}{a_k - a_j} \neq 0 \tag{4.61}$$

for any l, k, or j in $1, \ldots, n$, then (4.58) becomes

$$\begin{aligned} f_{mn}(r,t) &= \lambda^{n-m}\frac{(n-1)!}{(m-1)!}\sum_{j=m}^{n}\frac{e^{-(j\lambda + a_j r)t}}{\prod_{k\neq j}(a_k - a_j)\left(\frac{\lambda(k-j)}{a_k - a_j} + r\right)} \\ &= \lambda^{n-m}\frac{(n-1)!}{(m-1)!}\sum_{j=m}^{n}\frac{e^{-(j\lambda + a_j r)t}}{\prod_{k\neq j}(a_k - a_j)}\sum_{k\neq j}\frac{\prod_{\substack{\ell\neq k \\ \ell\neq j}}\left(\frac{\lambda(\ell-j)}{a_\ell - a_j} - \frac{\lambda(k-j)}{a_k - a_j}\right)^{-1}}{\frac{\lambda(k-j)}{a_k - a_j} + r}, \end{aligned} \tag{4.62}$$

and so the full probability density for $n > m$ is

$$v_{mn}(x,t) = \lambda^2\frac{(n-1)!}{(m-1)!}\sum_{j=m}^{n}\frac{e^{-j\lambda t}H(x - a_j t)}{\prod_{k\neq j}(a_k - a_j)}\sum_{k\neq j}\frac{\exp\left[-\frac{\lambda(k-j)}{a_k - a_j}(x - a_j t)\right]}{\prod_{\substack{\ell\neq k \\ \ell\neq j}}\left(\frac{\ell - j}{a_\ell - a_j} - \frac{k-j}{a_k - a_j}\right)} \tag{4.63}$$

However, the rewards **a** for many topologies do not satisfy (4.61). This can be seen in Figure 4.7, where $m = 2$ and $n = 4$. Then we see that when $j = 2$, $k = 4$, and $\ell = 3$ in (4.63),

$$\frac{4 - 2}{a_4 - a_2} = \frac{4 - 2}{0.75 - 0.5} = 8 \tag{4.64}$$

and

$$\frac{3 - 2}{a_3 - a_2} = \frac{3 - 2}{0.75 - 0.625} = 8, \tag{4.65}$$

so the denominator in the second sum in (4.63) is zero. Unfortunately this happens whenever there is symmetry in the tree so that more than one pair of taxa have the same time of shared ancestry. Note also that (4.63) does not reduce to (4.21) in Theorem 2 when $a_k = k$ since the denominator in the summand of (4.63) is zero.

Despite the difficulty in writing a general inversion to obtain $f_{mn}(x,t)$ for any topology, numerical inversion to arbitrary precision remains straightforward. Abate and Whitt (1995)

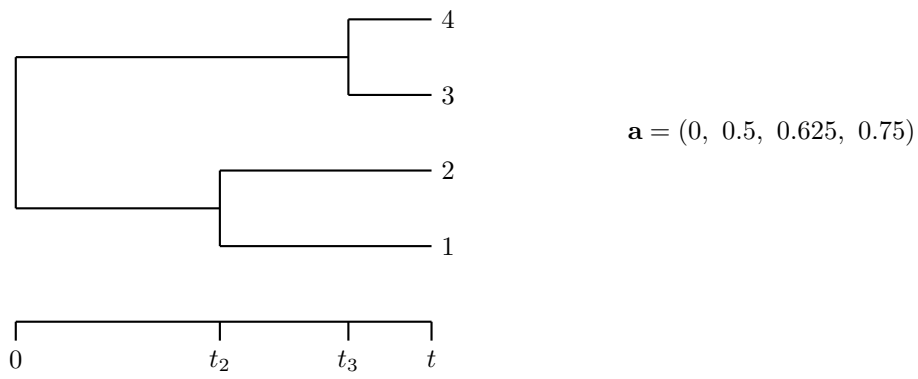


Figure 4.7: Demonstration of a problematic reward vector $\mathbf{a}$ computed for the symmetric four-taxon tree. In this case, the analytic inversion formula (4.63) cannot be applied, since the denominator in the sum becomes zero.

describe a numerical method for inverting the Laplace transform of probability densities by a discrete Riemann sum using the trapezoidal rule with step size h:

$$v_{mn}(x,t) \approx \frac{e^{A/2}}{2x}\mathrm{Re}\left[f_{m,n}\left(\frac{A}{2x},t\right)\right] + \frac{e^{A/2}}{x}\sum_{k=1}^{\infty}(-1)^k\mathrm{Re}\left[f_{m,n}\left(\frac{A+2k\pi i}{2x},t\right)\right], \qquad (4.66)$$

where we choose $A = 20$.

CHAPTER 5

Sex, lies, and self-reported counts: Bayesian analysis of longitudinal heaped integer data

Respondents to surveys are often asked to report numeric quantities, and sometimes they do not report those numbers acurrately. They may round up or down to the nearest integer, decimal place, or multiple of 5 or 10. This phenomenon is called grouping, heaping, coarsening, or digit preference, and the error inherent in these self-reported numbers can seriously bias estimation. Heaping is a well-known problem in many survey settings, and inference for heaped data is a major unsolved problem in statistical inference for both continuous quantities and counts. Heaping models often must make use of covariates or longitudinal sampling, which complicates the inference task substantially. We present a mixture model characterization of heaping: each subject i draws his or her true count X_i from a common distribution F, and then reports the possibly different value $Y_i|X_i$ according to a reporting distribution $G(X_i)$. We describe a novel parameterization of the reporting distribution whose parameters are readily interpretable as rates of over- and under-reporting and rounding to multiples of 5 or 10 by characterizing G using a general birth-death process. We present a Bayesian hierarchical model for longitudinal samples with covariates to infer both the unobserved true distribution of counts and the parameters that underly the heaping process. Finally, we apply our methods to longitudinal self-reported counts of sex acts in a study of high-risk behavior in HIV-positive youth.

5.1 Introduction

When survey respondents report numeric quantities, they often recall those numbers with error. In addition, respondents sometimes round up or down, for example to the nearest integer, decimal place, or multiple of 5 or 10. Several competing terms describe these errors. *Grouped* observations can only be resolved up to an interval that contains the true value; rounding to the nearest integer is an example of grouping. In a generalization of this concept, *coarsened* data can be resolved only up to a subset of the sample space (Heitjan and Rubin, 1991). *Heaped* data arise when there are various levels of coarsening. Heaping is a well-known problem in many survey settings, and robust inference for heaped data is a major unsolved problem in statistical inference for both continuous quantities and counts (Heitjan, 1989; Wang and Heitjan, 2008; Wright and Bray, 2003; Crockett and Crockett, 2006; Schneeweiss et al, 2010).

Reporting errors are well-studied for a variety of quantities, including self-reported age (Myers, 1954; Stockwell and Wicks, 1974; Myers, 1976); height and weight (Rowland, 1990; Schneeweiss and Komlos, 2009); elapsed time (Huttenlocher et al, 1990); and household purchases (Browning et al, 2003). Respondents may be even more inclined to misreport when the survey addresses topics that seem private, embarrassing, or culturally taboo. For example, there may be significant misreporting in studies of drug use (Klovdahl et al, 1994; Roberts and Brewer, 2001); cigarette use (Brown et al, 1998; Wang and Heitjan, 2008); or number of sex acts or sexual partners (Westoff, 1974; Golubjatnikov et al, 1983; Wiederman, 1997; Weinhardt et al, 1998; Fenton et al, 2001; Ghosh and Tu, 2009). Roberts and Brewer (2001) outline possible explanations and cognitive procedures that might give rise to heaping in self-reported data, including simple recall error, digit preference (e.g. counts ending in 0 or 5), or computing an approximate rate of an event and multiplying by time to obtain a count. Sometimes reporting errors arise though other mechanisms. For example, surveys that seek the number of lifetime sexual partners may induce respondents to misreport if they regard their true count as being embarrassingly high or low.

In practice, when errors due to inaccurate recall or reporting bias are a problem, the

researchers conducting a survey can use a variety of techniques to increase the accuracy of reports, including combinations of written questionnaires, in-person and telephone interviews, and continuous self-monitoring (Weinhardt et al, 1998). However, since researchers are limited by time and funds, no survey methodology can guarantee error-free results. This means that substantial inaccuracies may remain in self-reported data, and there is a clear need for statistical methodology that can assist researchers in learning from data that contain reporting errors.

Several authors have proposed statistical techniques to correct variance estimates under heaping of continuous or count data (Sheppard, 1897; Schneeweiss and Komlos, 2009; Schneeweiss et al, 2010; Schneeweiss and Augustin, 2006), and approximations and corrections to the maximum likelihood equations for estimation using grouped data (Tallis, 1967; Lindley, 1950). Others have explored smoothing techniques for heaped data on the grounds that smoothing may have the effect of "spreading out" grouped responses (Hobson, 1976; Singh et al, 1994). Roberts and Brewer (2001) discuss tests for detecting heaping in discrete distributions.

Notably, Heitjan (1989) provides a comprehensive review of previous inference strategies for grouped continuous data. Heitjan and Rubin (1991) introduce the concept of coarsening, in which we observe only a subset of the complete data sample space. They outline an integral representation of the likelihood for coarsened data that is useful in a variety of applied scenarios. Heitjan and Rubin (1990) apply these ideas to age heaping using a data imputation technique, and Wang and Heitjan (2008) study the impact of a drug treatment on smoking, where counts of cigarettes are heaped. Jacobsen and Keiding (1995) discuss extensions of the concept of coarse data to more general sample spaces than those considered by Heitjan and Rubin (1991). In another useful development, Wright and Bray (2003) interpret a dataset of reported (and evidently heaped) nuchal translucency measurements explicitly as samples from a mixture model. They propose a Gibbs sampling scheme to draw from the joint distribution of the true counts and unknown rounding parameters.

Many of these approaches make use of the mixture model paradigm, which provides a convenient framework for attacking heaping problems. To make this clear in the context of

heaping in count data, suppose respondents to a survey are asked to report a non-negative integer quantity. We imagine the reporting process as follows: respondent i draws his or her true count X_i from a distribution $F(\boldsymbol{\phi})$ taking values on the non-negative integers. The respondent then reports the possibly different value $Y_i|X_i$ from a reporting distribution $G(X_i, \boldsymbol{\theta})$, which depends on the true count X_i. Here, $\boldsymbol{\theta}$ is a vector of parameters underlying the heaping distribution G. In this way, we can interpret the reported count Y_i as a mixture of possibly different counts, where F determines the mixing proportions. The likelihood contribution of an observed count y becomes

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\phi}; y) &= \int_{\mathbb{N}} g(y|x, \boldsymbol{\theta}) \ \ \mathrm{d}F(x; \boldsymbol{\phi}) \\ &= \sum_{x=0}^{\infty} g(y|x, \boldsymbol{\theta}) \ f(x|\boldsymbol{\phi}) \end{aligned} \tag{5.1}$$

where $g(y|x, \boldsymbol{\theta})$ is the probability mass function (PMF) of y given x and $\boldsymbol{\theta}$, and $f(x|\boldsymbol{\phi})$ is the PMF of F. The quantities of greatest interest are often the true counts X_i or the parameters $\boldsymbol{\phi}$ underlying the true count distribution $F(\boldsymbol{\phi})$. The mixture model paradigm provides the tools we need to attack this problem. The distribution of Y_i given $F(\boldsymbol{\phi})$ can be regarded as a mixture of distributions $G(X_i, \boldsymbol{\theta})$, where F determines the mixing proportions. Figure 5.1 shows two graphical representations of this mixture model for heaped counts.

Applied researchers often have significant insight into the general properties or shape of $G(X_i, \boldsymbol{\theta})$ through methodological research into reducing error in survey sampling; see, for example, Weinhardt et al (1998) and Fenton et al (2001). These heuristic descriptions of reporting error for specific survey questions can guide statisticians in constructing rigorous probability models that capture important subjective qualities of heaping behavior. Perhaps surprisingly, a useful way to parameterize the reporting distribution $G(X_i, \boldsymbol{\theta})$ is to imagine that the true count X_i is the starting point of a continuous-time Markov chain on the non-negative integers known as a general birth-death process (BDP). Jumps from state x to $x + 1$ or $x - 1$ occur with instantaneous rates λ_x and μ_x, respectively. We specify $\mu_0 = 0$ to keep the process on the non-negative integers. Grunwald et al (2011) and Lee and Weiss (2011) model under- and over-dispersion in count data using a limited linear BDP with $\lambda_x = \lambda x$ and $\mu_x = \mu x$, but do not explicitly address heaping. In addition to modeling

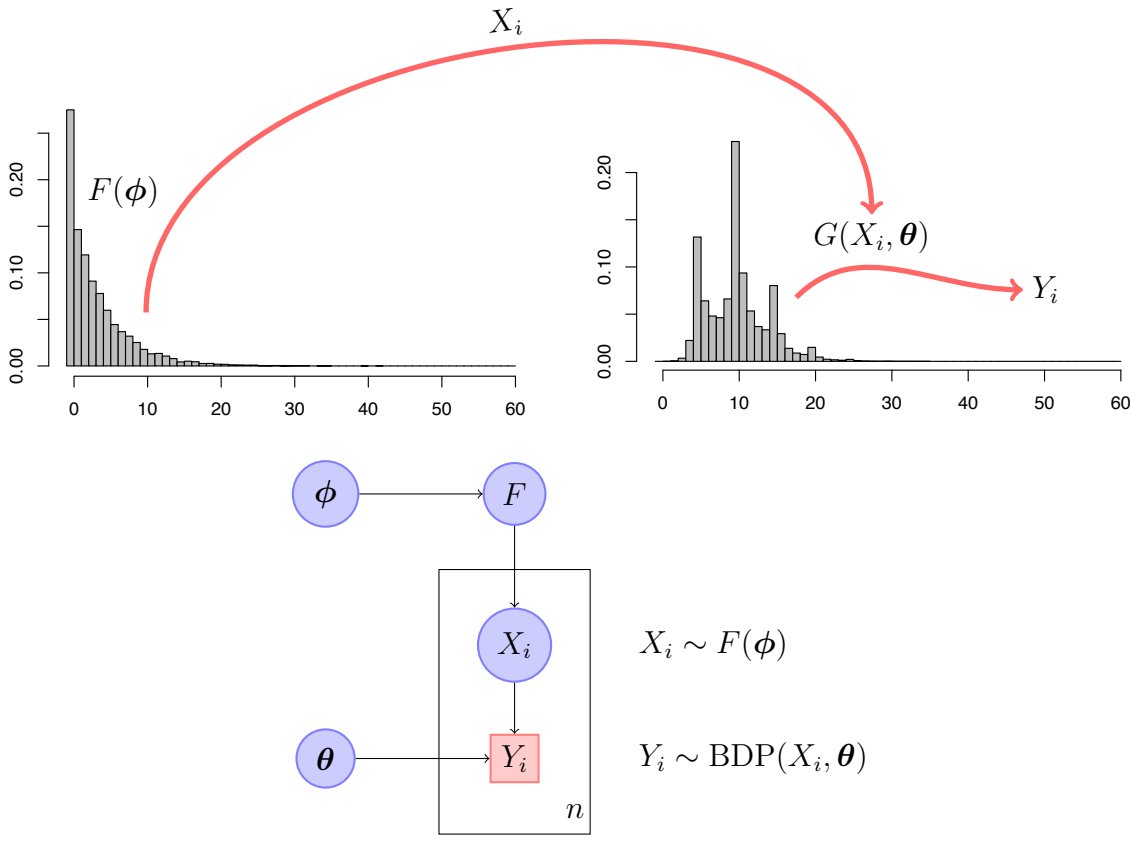


Figure 5.1: Two equivalent representations of the mixture model for heaped count data. The top panel shows a schematic diagram of the idealized reporting process: subject i chooses his or her true count X_i from the distribution F, then reports the possibly different count Y_i, drawn from the distribution $G(X_i, \boldsymbol{\theta})$. The bottom panel shows a graphical representation of the same process.

over-dispersion, BDPs can be very useful in parameterizing families of probability measures on the non-negative integers (Klar et al, 2010).

In this paper, we present an innovative framework for analyzing heaped longitudinal count data with covariates using a Bayesian hierarchical model and a novel characterization of the reporting distribution. In section 5.2 we describe a flexible class of reporting distributions that arise naturally from a carefully specified BDP: We model $Y_i|X_i$ as the state of a BDP after a short deterministic time t. By giving the jumping rates a certain parametric form, we develop a family of distributions $G(X_i, \boldsymbol{\theta})$ which are equivalent to the transition probability distributions of the BDP. Next, in section 5.3, we outline a Bayesian hierarchical model for longitudinal counts and a Gibbs-Metropolis scheme for sampling from the joint posterior distribution of the unknown parameters. We are interested in learning about the parameters $\boldsymbol{\phi}$ underlying the true counts, the inferred true counts X_i themselves, and the parameters $\boldsymbol{\theta}$ that govern the reporting/heaping process. Finally, in section 5.4, we demonstrate our method on longitudinal self-reported counts of sexual acts obtained during a study on the behaviors of HIV-positive youth.

5.2 Parameterizing the reporting distributions

In our formulation, G represents a family of reporting distributions indexed by the true count X. To parameterize $G(X, \boldsymbol{\theta})$ so that heaping can occur, we let $Y|X$ represent the outcome of an unbounded continuous-time Markov random walk, taking values on the non-negative integers, starting at the true count X, and evolving for a finite arbitrary time. The general BDP is a well-studied stochastic process that we can adapt to produce just such a family of reporting distributions by computing its finite-time transition probabilities $\Pr(Y = y|X, \boldsymbol{\theta})$ of this process, which determine the reporting distribution G. We specify the jumping rates of the BDP in a novel way so that the process is attracted to states on which heaping is expected to occur. One of the benefits of this approach is that we need only three parameters to define an infinite family of reporting distributions for heaped count data. We first present some background on the general BDP and show how to obtain

transition probabilities through a continued fraction expansion of the Laplace transform of the transition probabilities of the process.

5.2.1 General birth-death processes

A general BDP is a continuous-time Markov random walk on the non-negative integers (Feller, 1971). Let $X(t)$ be the location of the walk at time t. Define the transition probability $P_{ab}(t) = \Pr(X(t) = b \mid X(0) = a)$ to be the probability that the process is in state b at time t, given that it started at state a at time 0. A general BDP obeys the Kolmogorov forward equations

$$\frac{\mathrm{d}P_{ab}(t)}{\mathrm{d}t} = \lambda_{b-1}P_{a,b-1}(t) + \mu_{b+1}P_{a,b+1}(t) - (\lambda_b + \mu_b)P_{ab}(t) \tag{5.2}$$

where $P_{ab}(0) = 1$ if $a = b$ and zero otherwise, and we specify $\mu_0 = \lambda_{-1} = 0$. The forward equations are an infinite sequence of ordinary differential equations describing the probability flow into and out of state b at time t. Karlin and McGregor (1957b) provide a detailed derivation of properties of general BDPs. Unfortunately, it remains notoriously difficult to find analytic expressions for the transition probabilities in almost all general BDPs, and often one must resort to numerical techniques (Novozhilov et al, 2006; Renshaw, 2011).

A useful method for finding the transition probabilities $P_{ab}(t)$ is to apply the Laplace transform to both sides of the forward equations (5.2). This has the effect of turning the infinite system of differential equations into a recurrence relation, which can be solved to give a single expression for the Laplace transform of the transition probability. To illustrate, denote the Laplace transform of the transition probability $P_{ab}(t)$ as

$$h_{ab}(s) = \int_0^\infty e^{-st}P_{ab}(t)\ \mathrm{d}t. \tag{5.3}$$

Then (5.2) becomes

$$\begin{aligned} sh_{00}(s) - P_{00}(0) &= \mu_1 h_{01}(s) - \lambda_0 h_{00}(s), \text{ and} \\ sh_{0b}(s) - P_{0,b}(0) &= \lambda_{b-1}h_{0,b-1}(s) + \mu_{b+1}h_{0,b+1}(s) - (\lambda_b + \mu_b)h_{0b}(s) \end{aligned} \tag{5.4}$$

for $b \geq 1$. Rearranging (5.4) and forming a recurrence, we find a continued fraction repre-

sentation for $h_{00}(s)$,

$$h_{00}(s) = \cfrac{1}{s + \lambda_0 - \cfrac{\lambda_0\mu_1}{s + \lambda_1 + \mu_1 - \cfrac{\lambda_1\mu_2}{s + \lambda_2 + \mu_2 - \cdots}}}. \tag{5.5}$$

This is the Laplace transform of the transition probability $P_{00}(t)$. From this equation, it is possible to derive similar continued fraction representations for $h_{ab}(s)$, for any $X(0) = a$ and $X(t) = b$ (Murphy and O'Donohoe, 1975). Only recently, Crawford and Suchard (2011) give a robust numerical method for inverting the Laplace transforms $h_{ab}(s)$ to compute the transition probabilities in any general BDP with arbitrary jumping rates $\{\lambda_k\}_{k=0}^{\infty}$ and $\{\mu_k\}_{k=0}^{\infty}$. In our heaping parameterization, in which we fix $t = 1$ and regard $P_{ab}(1)$ as a function of an unknown parameter vector $\boldsymbol{\theta}$ controlling the jumping reates, we imagine that a birth-death process takes the true count $X_i = x$ to the reported count $Y_i = y$. Therefore, our family of reporting distributions is given by

$$g(y|x, \boldsymbol{\theta}) = P_{xy}(1; \boldsymbol{\theta}), \tag{5.6}$$

where the transition probability on the right side is computed using the parameters $\boldsymbol{\theta}$.

5.2.2 Specifying the jumping rates for heaping

Now that we can compute transition probabilities for general BDPs, we must specify the jumping rates $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\mu_k\}_{k=1}^{\infty}$ in the process so that the desired heaping behavior occurs. In our application, we assume that heaping occurs at multiples of 5. We therefore design a BDP in which the random walk is attracted to nearby multiples of 5. We further assume that misremembering increases with increased counts. This seems intuitively resonable: subjects whose true number of sex acts is greater than 100 may be less able to accurately recall this number than subjects whose true count is less than 10. For this reason, and to keep the variance in the reporting distribution from growing unreasonably large, we specify that the propensity for over- or under-reporting scales roughly with the logarithm of the true count.

Note that if the true count is k, the quantity k mod 5 is the (positive) deviation from the greatest multiple of 5 that is less than k. Likewise, $5-(k \bmod 5)$ is the (negative) deviation from the next multiple of 5 that is greater than k. Let $\boldsymbol{\theta} = (\theta_{\text{up}},\ \theta_{\text{down}},\ \theta_{\text{round}})$ be the three-element vector of heaping parameters. Since the rates of a BDP must be non-negative, we restrict the elements of $\boldsymbol{\theta}$ to be non-negative as well. Now consider a BDP with jumping rates

$$\begin{aligned} \lambda_k &= \theta_{\text{up}} \log(k+1) + \theta_{\text{round}}(k \bmod 5) \\ \mu_k &= \theta_{\text{down}} \log(k+1) + \theta_{\text{round}}\big(5 - (k-1 \bmod 5)\big) \end{aligned} \tag{5.7}$$

Although the parametric form of (5.7) is somewhat complicated, the three-parameter vector $\boldsymbol{\theta}$ has a straightforward interpretation: θ_{up} and θ_{down} represent the propensity to over- and under-report; θ_{round} is the propensity of rounding up or down to multiples of 5.

Figure 5.2 shows reporting distributions for some specific values of the true count x. For low true counts x, the distribution of the reported count y is centered very close to x. As the true count x becomes larger, the distribution of $y|x$ becomes more dispersed, and peaks appear at nearby multiples of 5. We emphasize that there is no closed-form solution for the transition probability $P_{xy}(1|\boldsymbol{\theta}) = g(y|x, \boldsymbol{\theta})$ for the BDP defined by (5.7). Fortunately, numerical evaluation of these probabilities is fast and robust, even for very large x and y (Crawford and Suchard, 2011).

It is also important to note that we do not suggest that each respondent mentally executes a random walk, starting at his or her true count, to arrive at the reported count. Likewise, we do not mean that our inference scheme involves simulation of a birth-death Markov chain. The BDP is simply a convenient way to derive an infinite family of reporting distributions that is 1) indexed by the true count X, 2) controlled by a small number of parameters that are readily interpretable, and 3) can be computed quickly to provide the likelihood of the reported count Y given the true count X. We use a BDP to parameterize the reporting process because we know of no other probabilistic mechanism that can give rise to such a rich family of probability distributions with the desired properties that requires so little analytical and computational effort.

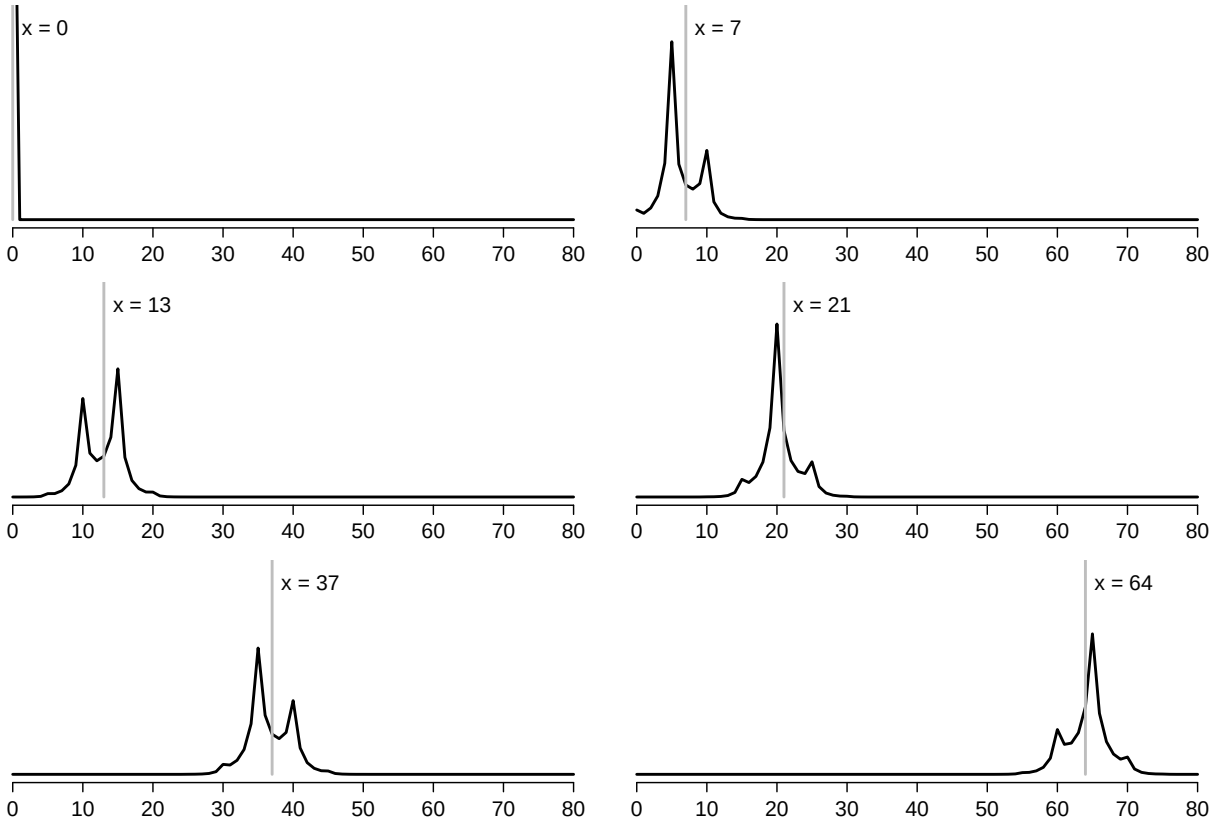


Figure 5.2: Reporting probabilities $g(y|x, \boldsymbol{\theta})$ for $x = 0, \ldots, 80$ for the heaping model. The vertical probability axis is the same for each plot. Note that for small true counts x, the reporting distribution $g(y|x)$ is centered near x, but there is a nonzero probability that the reported count is zero. For larger values of x, the reporting distribution becomes more dispersed around x and has peaks at multiples of 5. In this figure, we use $\boldsymbol{\theta} = (0.8, 0.6, 0.9)$ to demonstrate the distinctness of the peaks that emerge when rounging is significant.

5.3 A hierarchical model for longitudinal counts

In this section, we combine a generalized linear mixed model (GLMM) for longitudinal data with the BDP heaping model for self-reported counts. Label subjects $i = 1, \ldots, N$, with each subject's self-reported count Y_{it} at each of n_i time-points $t_{i1}, \ldots, t_{in_i}$. In addition, we record covariates W_{it} and Z_{it} for each subject at each time point. Consider the following hierarchical model for the reported count Y_{it}:

$$\begin{aligned} Y_{it} &\sim \text{BDP}(X_{it}, \boldsymbol{\theta}), \\ X_{it} &\text{ has distribution } F_{it} = \text{GLMM}(\nu_{it}, \omega), \\ \log \nu_{it} &= \mathbf{W}_{it}\boldsymbol{\alpha} + \mathbf{Z}_{it}\boldsymbol{\beta}_i, \text{ and} \\ \boldsymbol{\beta}_i &\sim \text{Normal}(\mathbf{0}, \mathbf{D}_\beta) \end{aligned} \tag{5.8}$$

where ν_{ti} and ω are the subject-timepoint-specific mean and common variance of the GLMM. Reasonable choices include the geometric, Poisson, negative binomial, or power-law distributions. The mean ν_{it} is determined by a fixed effect α multiplying a vector of covariates and a subject-specific random effect $\boldsymbol{\beta}_i$ with variance $\mathbf{D}_\beta$. The parameters $\boldsymbol{\theta}$ underlying the reporting distribution via the BDP are assumed constant across all subjects and timepoints. Figure 5.3 shows a graphical representation of this hierarchical model for longitudinal counts.

To complete our Bayesian hierarchical model for longitudinal studies, we specify prior distributions as follows:

$$\begin{aligned} \boldsymbol{\alpha} &\sim N(\boldsymbol{\alpha}_0, \mathbf{D}_\alpha), \\ \boldsymbol{\theta}_j &\sim \text{Gamma}(\gamma_j, k_j) \quad \text{for } j = 1, 2, 3 \\ \mathbf{D}_\beta &\sim \text{Inverse-Wishart}(\boldsymbol{\Omega}_\beta, \mathbf{m}_\beta), \text{ and} \\ \omega &\sim \text{Inverse-Gamma}(a_\omega, b_\omega). \end{aligned} \tag{5.9}$$

We restrict attention in this paper to the challenge of modeling self-reported counts. Our objective is not to explore prior elicitation or model selection; we only wish to illustrate the usefulness of our novel reporting distribution and hierarchical model. A principled exploration of prior specification, as performed by Lee and Weiss (2011), may be necessary in general.

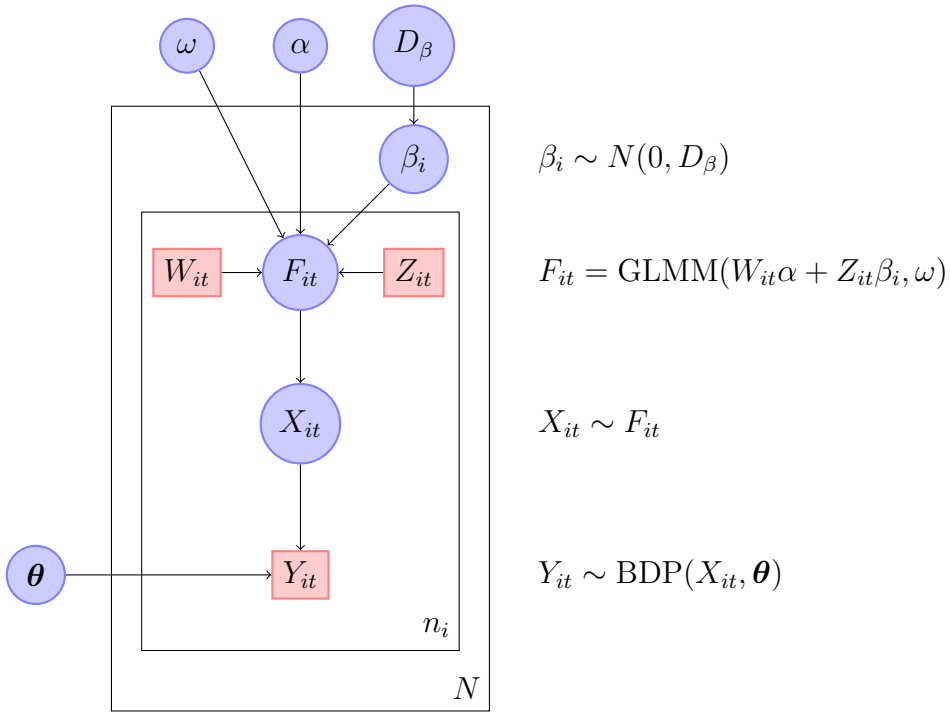


Figure 5.3: Graphical representation of the longitudinal model for the count reported by subject i at time t. Constant quantities (observations Y_{it} and covariates W_{it} and Z_{it}) appear in red boxes; random quantities (unknown parameters and imputed true counts X_{it}) appear in blue circles. The diagram does not show dependencies on hyperparameters via prior distributions.

5.3.1 Sampling from the posterior

To learn about the true count distribution and the parameters underlying the heaping process, we must be able to estimate the joint posterior distribution of the unobserved true counts and the unknown parameters. To accomplish this, we resort to posterior simulation via Markov chain Monte Carlo. We describe standard Gibbs and Metropolis-Hastings samplers for the full conditional distributions of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, ω, and $\mathbf{D}_\beta$ in the Appendix. Sampling from the posterior distribution of true counts, conditional on these parameters and the observed reported counts is more challenging, and we briefly describe our approach in this section.

Let $\mathbf{X}$ and $\mathbf{Y}$ be the collection of all true and reported counts for each subject and timepoint. Likewise, let $\mathbf{Z}$ and $\mathbf{W}$ be the collection of all subject- and timepoint-specific covariate vectors. The posterior density is given by

$$\begin{aligned}\Pr(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \omega, \mathbf{D}_{\boldsymbol{\beta}}, \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}) &\propto \Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \Pr(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mathbf{W}, \mathbf{Z}, \mathbf{D}_{\boldsymbol{\beta}}) \\ &\quad \times \Pr(\boldsymbol{\theta}) \Pr(\boldsymbol{\alpha}) \Pr(\omega) \Pr(\boldsymbol{\beta} \mid \mathbf{D}_{\boldsymbol{\beta}}) \Pr(\mathbf{D}_{\boldsymbol{\beta}}) \\ &= \left[\prod_{i=1}^{N} \prod_{j=1}^{n_i} \Pr(Y_{ij} \mid X_{ij}, \boldsymbol{\theta}) \Pr(X_{ij} \mid \mathbf{Z}_{ij}, \mathbf{W}_{ij}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i, \omega) \right] \\ &\quad \times \Pr(\boldsymbol{\theta}) \Pr(\boldsymbol{\alpha}) \Pr(\omega) \Pr(\boldsymbol{\beta} \mid \mathbf{D}_{\boldsymbol{\beta}}) \Pr(\mathbf{D}_{\boldsymbol{\beta}}).\end{aligned} \tag{5.10}$$

We calculate the reporting probability $\Pr(Y_{ij} \mid X_{ij}, \boldsymbol{\theta})$ using the method outlined in section 5.2.1; in general no closed analytic form exists. The lack of conjugacy between $F = \Pr(X_{it} | \mathbf{Z}_{it}, \mathbf{W}_{it}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i, \omega)$ and $G = \Pr(Y_{it} | X_{it}, \boldsymbol{\theta})$, and between F and the prior distributions complicates matters significantly. Fortunately, the discrete nature of the count data makes possible some simplifications. The conditional distribution of the unobserved true counts $\mathbf{X}$ is given by

$$\begin{aligned}\Pr(X_{it} = x \mid Y_{it} = y, \mathbf{Z}_{it}, \mathbf{W}_{it}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i, \omega) &\propto \Pr(Y_{it} = y \mid X_{it} = x, \boldsymbol{\theta}) \Pr(X_{it} \mid \mathbf{Z}_{it}, \mathbf{W}_{it}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i, \omega) \\ &= g(y|x, \boldsymbol{\theta}) f(x|\boldsymbol{\phi})\end{aligned} \tag{5.11}$$

for each possible x. Although in principle the true count x could range from zero to infinity, this is clearly unrealistic for real-life counts. In practice, the distribution of $X_{it}|Y_{it}$ is unimodal, decays quickly for $x \ll y$ and $x \gg y$, and is centered near Y_{it}. Therefore, if $g(y|x)f(x) \approx 0$ for $x < x_{\min}$ and $x > x_{\max}$, then we sample X_{it} from

$$\{x_{\min},\ x_{\min}+1,\ \ldots,\ x_{\max}\} \tag{5.12}$$

with probability proportional to

$$\big\{g(y|x_{\min})f(x_{\min}),\quad g(y|x_{\min}+1)f(x_{\min}+1),\quad \ldots\quad ,\quad g(y|x_{\max})f(x_{\max})\big\} \tag{5.13}$$

where the dependence of f and g on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ respectively has been omitted for clarity. Sampling the true count X_{it} for each observed count Y_{it} is computationally feasible because we only need to compute the set of reporting probabilities $g(y|x, \boldsymbol{\theta})$ in (5.13) once for each *unique* reported count, since the parameters $\boldsymbol{\theta}$ that underly the reporting distributions are the same for all subjects and timepoints. Sampling a true count for each observation is then a matter of computing the subject- and timepoint-specific prior probabilities corresponding to the observed data, and sampling a single count from (5.13). This scheme is a more formal Bayesian version of the multiple imputation approach introduced in the context of heaped or coarsened data by Heitjan and Rubin (1991).

5.4 Application to self-reported counts of sex acts

To illustrate the effectiveness of our mixture model and BDP parameterization of G, we analyze a dataset derived from the CLEAR study of sexual behavior of HIV-positive youth (Rotheram-Borus et al, 2001). Respondents (175, interviewed up to 5 times over several years) reported the number of sex acts they engaged in during the previous three months. Figure 5.4 summarizes the data. The top left panel shows longitudinal reported counts of sex acts for each subject. The top right panel shows the distribution of interview times, sorted by maximum follow-up time after baseline. The bottom left panel shows a histogram of aggregate reported counts of sex acts across all subjects and timepoints, and the bottom right panel shows the same histogram in greater detail. There are several striking features

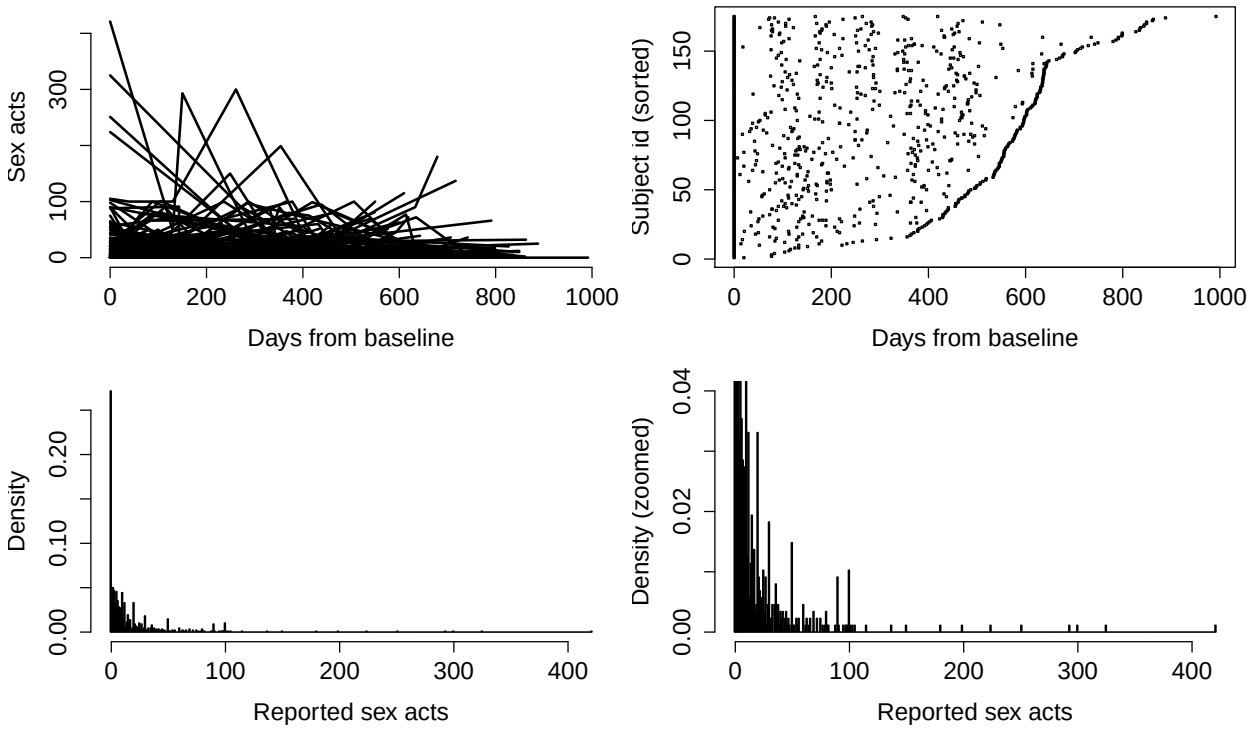


Figure 5.4: Summary of longitudinal self-reported counts of sex acts. Top left: reported sex acts over time, from baseline for all subjects. Top right: illustration of longitudinal samples (black squares) for each subject, at each time point. The subjects are sorted by maximum follow-up time. Bottom left: histogram of aggregate reported sex acts in the previous three months, for all subjects, at all time points. The bin size is one. Bottom right: the same histogram in greater detail. Note the peaks at multiples of 5 and 10.

of the reported counts: 1) A large proportion of the counts are zero; 2) the histogram clearly shows peaks at integer multiples of 5 and 10; and 3) a few counts are very large: 14 are over 100. Researchers are typically interested in the true distribution of counts of sexual acts and insight into reporting errors (by way of $\boldsymbol{\theta}$). In addition, insight into the heaping/rounding process may provide useful prior information that can be used to mitigate the effects of heaping in future studies.

The unique features of the dataset make traditional analysis (for example, classical Poisson random effects regression) unappealing, since the histogram of reported counts in Figure 5.4 appears both multimodal and overdispersed. Instead, we use the longitudinal hierarchical framework outlined in the previous section to learn about the underlying true counts and the heaping process. To do this, we need to specify and justify the reporting distribution $G(X, \boldsymbol{\theta})$. We must make several assumptions, informed by the results of relevant public health studies (Westoff, 1974; Golubjatnikov et al, 1983; Wiederman, 1997; Weinhardt et al, 1998; Fenton et al, 2001; Ghosh and Tu, 2009). Based on the histogram of aggregate counts in Figure 5.4, we assume that heaping occurs at multiples of 5, and we use the BDP rate model in (5.7). We also find it intuitively reasonable that the variance of $Y_{it}|X_{it}$ increases as X_{it} gets larger. We therefore assume heuristically that the propensity to misremember increases roughly with the logarithm of the true count, as shown in (5.7).

We let W_{it} in (5.8) be a 8×1 vector of covariates for subject i at time t as follows: age, gender, an indicator for men who have sex with men (MSM), an indicator for injection drug use, time since baseline interview, indicators for intervention via telephone and in-person, and an indicator for use of methamphetamine or other stimulant drugs. Most of these covariates remain unchanged over the time points. Some, such as time since baseline interview, use of drugs, and intervention types, do change with t. We let $Z_{it} = 1$, making β_i a scalar; this provides a subject-specific random intercept. We specify the underlying GLMM to be a negative binomial distribution with mean ν_{it} and common variance ω to allow for over-dispersion.

We specify prior distributions as given in (5.9), with the exception of one: since the subject-specific intercept β_i is now a scalar, its prior distribution is now Inverse-Gamma,

and we have

$$\beta_i \sim \text{Inverse-Gamma}(\Omega_\beta, m_\beta). \tag{5.14}$$

We specify hyperparameters as follows: for the fixed effects $\boldsymbol{\alpha}$, $\alpha_0 = \mathbf{0}$ and $\mathbf{D}_\alpha = 10\mathbf{I}$ where $\mathbf{I}$ is the identity matrix; for the heaping parameters $\boldsymbol{\theta}$, $\gamma_j = 2$ $k_j = 4$ for $j = 1, 2, 3$; for the subject-specific random effects β_i, $\Omega_\beta = 1$ and $m_\beta = 200$; and for the overdispersion ω, $a_\omega = 5$, and $b_\omega = 20$.

5.4.1 Results

To evaluate the usefulness of our heaping method, we fit a Bayesian hierarchical model with and without heaping. In the model without heaping, we did not infer the true counts X_{it} or rounding parameters $\boldsymbol{\theta}$. To fit each model, we sampled from the posterior distributions of $\mathbf{X}$ and the unknown parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ using the scheme outlined in section 5.3.1. Estimates of the fixed effects $\boldsymbol{\alpha}$ for both models are summarized in Figure 5.5, with results for the model without heaping shown on the left and results for the model with heaping shown on the right. The whiskers denote the 2.5th and 75th percentiles, and the black dot shows the median. Subject age and study timepoint were normalized. Subjects who received either telephone or in person intervention showed higher true counts in relation to controls, and stimulant use was associated with higher true counts. To a lesser extent, trading sex was associated with higher true counts. Interestingly, male subjects were associated with higher true counts under the model with heaping, but not in the model without heaping.

Figure 5.6 shows marginal posterior boxplots of the reporting parameters $\boldsymbol{\theta}$. These estimates offer a straightforward interpretation. The first element of $\boldsymbol{\theta}$, marked "Up" indicates substantial over-reporting with wide variability. The second element of $\boldsymbol{\theta}$, marked "Down" indicates low levels of under-reporting. In every sample from the posterior distribution of $\boldsymbol{\theta}$, we found that $\theta_{\text{up}} > \theta_{\text{down}}$, so we conclude that the posterior probability $\Pr(\theta_{\text{up}} > \theta_{\text{down}}) \approx 1$. The third element of $\boldsymbol{\theta}$ is the heaping/rounding parameter, which indicates that subjects engage in rounding to nearby multiples of five, a conclusion that may be self-evident from the histogram of reported counts in Figure 5.4. It is important to note that these estimates

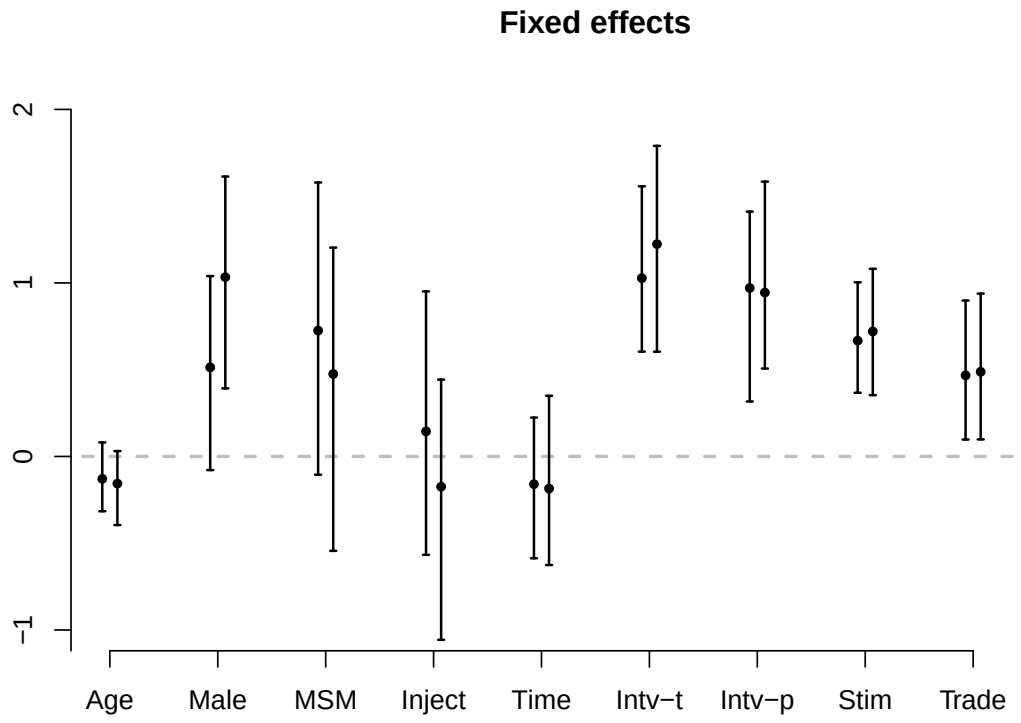


Figure 5.5: Marginal posterior estimates of the fixed effects parameters $\boldsymbol{\alpha}$. The results for the model without heaping are shown on the left, and those for the model with heaping are shown on the right. The black dot denotes the median and the whiskers mark the 2.5 and 97.5 percentiles. Perhaps surprisingly, both telephone (Intv-t) and in-person interivewing (Intv-p) is associated with higher true counts of sex acts. Methamphetamine/stimulant use (Stim), and to a lesser extent trading sex (Trade) are also associated with higher counts. Age, sexual preference (MSM), time on the study (Time). In general, estimates for the fixed effects are very similar for both models under comparison, but under the heaping model, male subjects have increased true counts.

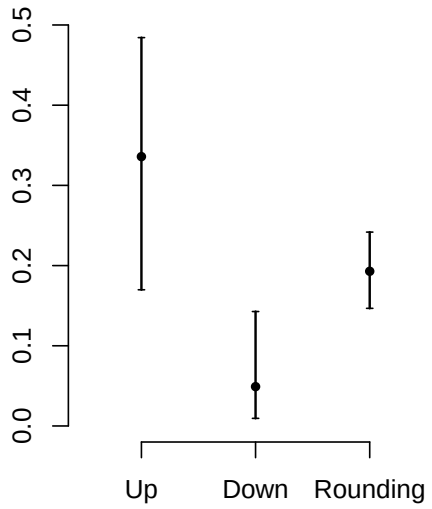


Figure 5.6: Marginal posterior estimates of heaping parameters $\boldsymbol{\theta}$. There is substantial and varying over-reporting (Up), little under-reporting (Down), and moderate rounding for the CLEAR sex act dataset, under our model for the underlying true counts. These parameters represent the reporting behavior that best explains the difference between the estimated true counts X_{it} and the observed reported counts Y_{it}.

do not necessarily reflect the true misreporting behavior of the subjects in the study. Rather, the estimates indicate the misreporting behavior that best explains the difference between the inferred true counts X_{it} and the observed reported counts Y_{it}.

The inferred distribution of true counts is shifted toward lower values than the reported counts. In addition, the variance of imputed true counts increases as the reported count becomes larger. To make this clearer, Figure 5.7 shows examples of the conditional distribution of $X_{it}|Y_{it}$ for several subjects, whose reported counts are shown by a gray line. Note that the covariates W_{it} can substantially impact the estimation for X_{it} in different subjects.

Figure 5.8 shows two-dimensional histograms of posterior predictive residuals as a function of reported count and normal quantile-quantile plots of residuals for both models. On

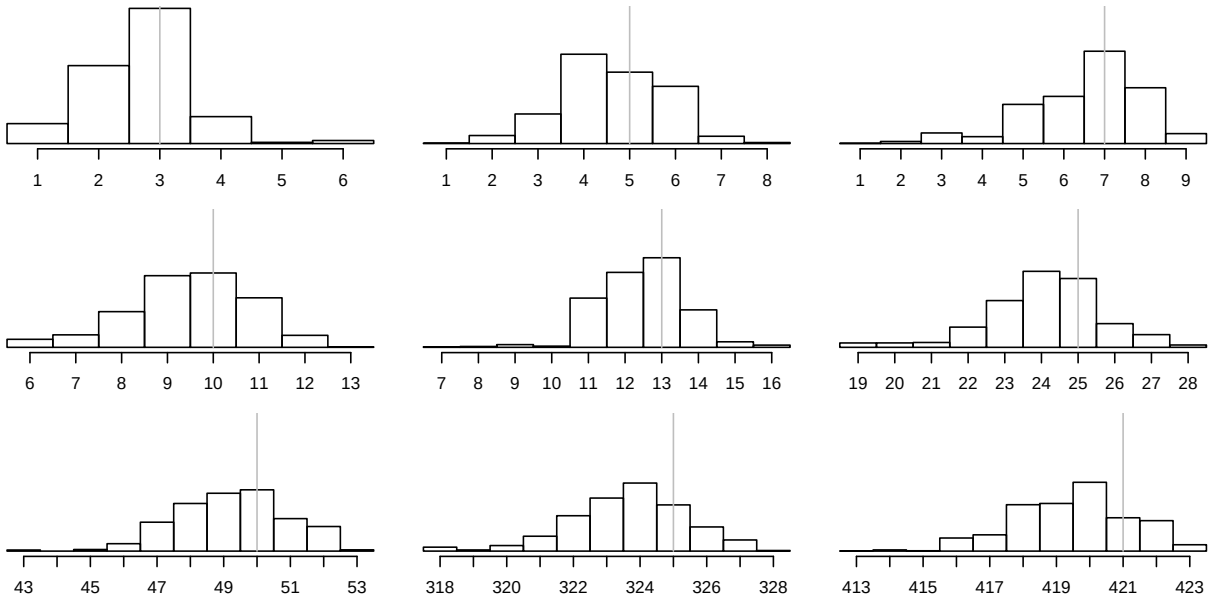


Figure 5.7: Examples of the inferred marginal posterior distributions of true counts X_{it} for individual subjects under the heaping model. A gray vertical line denotes the reported count.

the left, the model without heaping fits well for very small counts, but residuals quickly become very large for the sparser high reported counts. On the right, the model with heaping has small residuals clustered near zero for most reported counts. The 5%, 50% and 95% quantiles are shown overlaid on the density plots. The normal Q-Q plots reveal the substantial deviation from normality in the residuals for the model without heaping. Note the different spans of the sample quantiles in the vertical axes. The heaping model has fairly normal residuals, which appear binned in the Q-Q plot because they are integer counts.

5.5 Discussion

In this paper, we have illustrated how researchers can infer true integer counts from inaccurate reported counts by formulating an explicit model for heaping. Furthermore, we have demonstrated how a mixture model approach can disentangle true counts from the rounding/heaping/misremembering process that generates the reported counts. To accomplish these goals, we have employed an established Bayesian hierarchical model for longitudinal

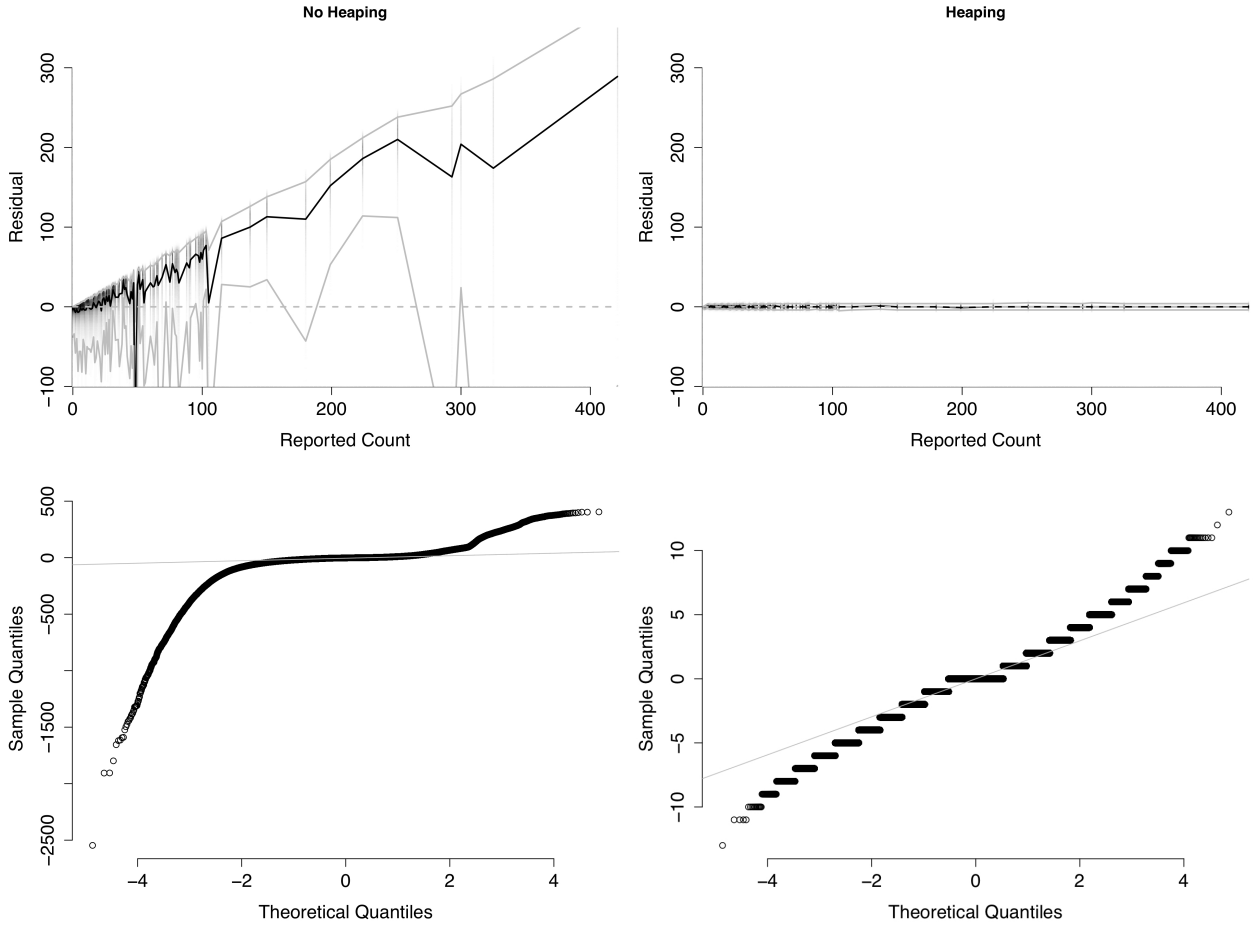


Figure 5.8: Posterior predictive residual densities for the model without heaping (left) and with heaping (right). The top panels show two-dimensional histograms of posterior predictive residual density by corresponding reported count. In these two-dimensional histograms, darker color indicates greater density. The 5%, 50%, and 95% quantile lines are overlaid. At bottom, normal quantile-quantile plots of posterior predictive residuals. The residuals for the model without heaping are highly non-normal. The residuals for the model with heaping are so small that the integral nature of the counts can be seen in the quantiles.

observations. Our most substantial innovation is the novel reporting distribution $G(X, \boldsymbol{\theta})$ based on a birth-death Markov chain with special jumping rates. Use of a birth-death process to model overdispersion or reporting error has been proposed before (Grunwald et al, 2011; Lee and Weiss, 2011). However, we have substantially expanded the possibilities for general birth-death models of reporting error to explicitly incorporate both overdispersion and heaping, while providing a computational method to evaluate likelihoods and sample from the posterior distribution of the true counts. This approach has the benefit of providing a sophisticated and highly configurable family of reporting distributions indexed by the true count X. In addition, we accomplish this using only three parameters that have intuitive meanings that should appeal to applied researchers conducting future surveys. In our view, the most compelling reason for incorporating heaping into the hierarchical model is illustrated in Figure 5.8. In exchange for more sophisticated modeling and increased computational time, the posterior predictive residuals are far better behaved and more normal. The model with heaping clearly fits the observed data better than the unheaped version.

In our application, the results for the fixed effects $\boldsymbol{\alpha}$ and random effects $\boldsymbol{\beta}$ are similar for the models with and without heaping, but under the heaping model, male subjects were associated with higher true counts. We consider the mixture model approach to estimating $\mathbf{X}$ and $\boldsymbol{\theta}$ to be a useful way of understanding residuals. That is, our inferences of $\boldsymbol{\theta}$ represent the heaping behavior required to explain the differences between the observed counts $\mathbf{Y}$ and the inferred true counts $\mathbf{X}$, under F_{it}. In this way, we argue that incoroprating heaping into a hierarchical model for counts does not substantially complicate Bayesian inference, and yeilds a wealth of information on the rounding behavior that subjects must exhibit to reconcile the observed counts with the inferred model for the true counts.

In order to apply our method, modelers must use *a priori* knowledge about the topics addressed by the survey, and the characteristics of the survey population, to devise a meaningful model for reporting errors. It remains unclear exactly how researchers can incorporate meaningful prior information about reporting error into new studies. Applied and methodological research in public health offers some clues. Researchers in this field often address the problem of reporting error in surveys related to sexuality and other taboo topics. Wang

and Heitjan (2008) discuss validation of reported counts of cigarettes smoked by measuring tobacco products in the blood. Other survey methods are possible, including using diary-like surveys or repeated questionairres to assess reporting error. We propose that studies like these can provide useful prior information about rounding parameters $\boldsymbol{\theta}$ in our model. Armed with prior information about rounding propensities, perhaps stratified by personal attributes such as gender, age, or sexual orientation, public health researchers could proceed with a Bayesian analysis similar to the one outlined in this paper to investigate the distribution of true counts $\mathbf{X}$, in addition to the more traditional quantities of interest ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in this paper).

One benefit of abandoning conjugacy assumptions about $f(x|\boldsymbol{\phi})$ and $g(y|x, \boldsymbol{\theta})$ is that we are free to choose a realistic parametric form for the distribution of true counts. For example, many studies have observed that count data often follow power-law type distributions with heavy tails (see, for example, Clauset et al (2007)), which could be easily incorporated into our model for counts. Substantial increases in computational time usually accompany choice of nonconjugate hierarchical models. In our case, we have tried to mitigate the increased computational cost by incorporating efficient sampling routines. However, the heaping model clearly increases computational time, which may not be appropriate for very large datasets or time-sensitive analyses.

Finally, we caution that care must be taken to avoid applying a model for heaping when it is inappropriate – apparent digit preference in a survey does not necessarily indicate heaping or rounding. For example, a smoker may limit his or her daily consumption to a single pack (usually 20 cigarettes). If many respondents to a questionnaire about smoking habits report multiples of 20 cigarettes, modelers may wrongly assume that the subjects have rounded their true counts, when in reality cigarette pack size imposes a natural unit of consumption for smokers. Wang and Heitjan (2008) call this phenomenon "self-rationing". Designing a model that can accommodate various assumptions about both the mechanism generating the true counts, and the cognitive process that gives rise to the reported counts, can be challenging. Therefore, nonparametric techniques that can infer the locations of heaping points in a parsimonious way might prove valuable. We are actively exploring this topic.

5.6 Appendix

5.6.1 Sampling $\boldsymbol{\alpha}$

The full conditional distribution of $\boldsymbol{\alpha}$ is

$$\begin{aligned}\Pr(\boldsymbol{\alpha} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mathbf{D}_{\boldsymbol{\beta}}) &= \Pr(\boldsymbol{\alpha} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\beta}, \omega) \\ &\propto \Pr(\mathbf{X} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega) \Pr(\boldsymbol{\alpha}) \\ &= \left[\prod_{i=1}^{N} \prod_{j=1}^{n_i} \Pr(X_{ij} \mid \mathbf{Z}_{ij}, \mathbf{W}_{ij}, \boldsymbol{\alpha}, \beta_i, \omega)\right] \Pr(\boldsymbol{\alpha}).\end{aligned} \tag{5.15}$$

We sample $\boldsymbol{\alpha}$ using a Metropolis-Hastings step with a multivariate normal proposal density centered at the current state with an adaptive variance term. Letting $\boldsymbol{\alpha}^{(j)}$ be the jth sample of $\boldsymbol{\alpha}$, the $(j+1)$th proposal is

$$\boldsymbol{\alpha}^{(j+1)} \sim \text{Normal}(\boldsymbol{\alpha}^{(j)}, \epsilon_j \mathbf{I}) \tag{5.16}$$

where $\epsilon_1 = 0$ and

$$\epsilon_{j+1} = \epsilon_j + \frac{\frac{1}{j}\#(\text{successes}) - 0.6}{1 + \log(j)}. \tag{5.17}$$

Here, #(successes) is the number of accepted proposals, of the previous j proposals. This adaptive variance term converges to a value that achieves the target acceptance probability of 0.6 on average.

5.6.2 Sampling β_i

Similarly, the full conditional distribution of β_i is

$$\begin{aligned}\Pr(\beta_i \mid \mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\alpha}, \omega, \mathbf{D}_{\boldsymbol{\beta}}) &= \Pr(\beta_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\alpha}, \mathbf{D}_{\boldsymbol{\beta}}) \\ &\propto \Pr(\mathbf{X}_i \mid \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\alpha}, \omega, \beta_i) \Pr(\beta_i \mid \mathbf{D}_{\boldsymbol{\beta}}) \\ &= \left[\prod_{j=1}^{n_i} \Pr(X_{ij} \mid \mathbf{Z}_{ij}, \mathbf{W}_{ij}, \boldsymbol{\alpha}, \beta_i, \omega)\right] \Pr(\beta_i \mid \mathbf{D}_{\boldsymbol{\beta}}).\end{aligned} \tag{5.18}$$

We also sample each β_i using a Metropolis-Hastings step with a normal proposal density centered at the current estimate using the adaptive variance procedure described above for $\boldsymbol{\alpha}$.

5.6.3 Sampling $\boldsymbol{\theta}$

The conditional posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}\Pr(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mathbf{D}_{\boldsymbol{\beta}}) &\propto \Pr(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y}) \\ &\propto \Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) \\ &= \left[\prod_{i=1}^{N} \prod_{j=1}^{n_i} \Pr(Y_{ij} \mid X_{ij}, \boldsymbol{\theta})\right] \Pr(\boldsymbol{\theta}).\end{aligned} \tag{5.19}$$

Since the elements of $\boldsymbol{\theta}$ are non-negative, we use a proportional scaling proposal. At step j, the proposal for the ℓth element of $\boldsymbol{\theta}$ is

$$\theta_{\ell}^{(j+1)} = \theta_{\ell}^{(j)} \exp\left[\epsilon_j \left(U - \frac{1}{2}\right)\right] \tag{5.20}$$

where U is a random variable uniformly distributed between 0 and 1, $\epsilon_1 = 1$, and the adaptive scaling ϵ_j is defined as above. Then using the change of variables formula, the proposal ratio is

$$\frac{\Pr\left(\theta_{\ell}^{(j)} \mid \theta_{\ell}^{(j+1)}\right)}{\Pr\left(\theta_{\ell}^{(j)} \mid \theta_{\ell}^{(j)}\right)} = \exp\left[\epsilon_j \left(U - \frac{1}{2}\right)\right]. \tag{5.21}$$

5.6.4 Sampling ω

The full conditional density of ω is

$$\begin{aligned}\Pr(\omega \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{D}_{\boldsymbol{\beta}}) &= \Pr(\omega \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\propto \Pr(\mathbf{X} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega) \Pr(\omega) \\ &= \left[\prod_{i=1}^{N} \prod_{j=1}^{n_i} \Pr(X_{ij} \mid \mathbf{Z}_{ij}, \mathbf{W}_{ij}, \boldsymbol{\alpha}, \beta_i, \omega)\right] \Pr(\omega).\end{aligned} \tag{5.22}$$

We sample ω using a Metropolis-Hastings step with a proportional scaling proposal analagous to the one illustrated above for the elements of $\boldsymbol{\theta}$.

5.6.5 Sampling $\mathbf{D}_{\beta}$

The full conditional density of $\mathbf{D}_{\beta}$ is

$$\Pr(D_{\beta} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \omega, \mathbf{D}_{\boldsymbol{\beta}}) = \Pr(D_{\beta} \mid \boldsymbol{\beta}). \tag{5.23}$$

Recall that $D_\beta \sim$ Inverse-Wishart(Ω_β, m_β). Conditional on $\boldsymbol{\beta}$, we have

$$D_\beta \sim \text{Inverse-Wishart}\left(\boldsymbol{\beta}\boldsymbol{\beta}^t + \Omega_\beta, N + m_\beta\right). \tag{5.24}$$

where N is the number of subjects. Therefore Gibbs sampling of $\mathbf{D}_\beta$ is accomplished directly.

Bibliography

Abate J, Whitt W (1992a) The Fourier-series method for inverting transforms of probability distributions. Queueing Syst 10:5–87

Abate J, Whitt W (1992b) Numerical inversion of probability generating functions. Oper Res Lett 12:245–251

Abate J, Whitt W (1995) Numerical inversion of Laplace transforms of probability distributions. ORS J Comput 7(1):36–43

Abate J, Whitt W (1999) Computing Laplace transforms for numerical inversion via continued fractions. INFORMS J Comput 11(4):394–405

Alfaro M, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky D, Carnevale G, Harmon L (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. P Natl A Sci USA 106(32):13,410–13,414

Allee WC, Emerson AE, Park O (1949) Principles of Animal Ecology. Saunders Philadelphia

Amos W (2010) Mutation biases and mutation rate variation around very short human microsatellites revealed by human-chimpanzee-orangutan genomic sequence alignments. J Mol Evol 71:192–201

Andersson H, Britton T (2000) Stochastic Epidemic Models and their Statistical Analysis. Lecture notes in statistics, Springer New York

Anscombe FJ (1953) Sequential estimation. J Roy Stat Soc B 15(1):1–29

Bailey NTJ (1964) The Elements of Stochastic Processes with Applications to the Natural Sciences. Wiley New York

Bankier JD, Leighton W (1942) Numerical continued fractions. Am J Math 64(1):653–668

Bhargava A, Fuentes F (2010) Mutational dynamics of microsatellites. Mol Biotechnol 44:250–266

Bladt M, Sorensen M (2005) Statistical inference for discretely observed Markov jump processes. J Roy Stat Soc B Met 67(3):395–410

Blanch G (1964) Numerical evaluation of continued fractions. SIAM Rev 6(4):383–421

Blomberg S, Garland T, Ives A (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717–745

Bokma F (2010) Time, species, and separating their effects on trait variance in clades. Syst Biol 59(5):602–607

Bordes G, Roehner B (1983) Application of Stieltjes theory for S-fractions to birth and death processes. Adv Appl Probab 15(3):507–530

Brown RA, Burgess ES, Sales SD, Whiteley JA, Evans DM, Miller IW (1998) Reliability and validity of a smoking timeline follow-back interview. Psychol Addict Behav 12:101–112

Browning M, Crossley TF, Weber G (2003) Asking consumption questions in general purpose surveys. Econ J 113(491):F540–F567

Calabrese P, Durrett R (2003) Dinucleotide repeats in the drosophila and human genomes have complex, length-dependent mutation processes. Mol Biol Evol 20(5):715–725

Chakraborty R, Kimmel M, Stivers D, Davison L, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. P Natl Acad Sci USA 94(3):1041–1046

Clauset A, Shalizi C, Newman M (2007) Power-law distributions in empirical data. Arxiv preprint arxiv:07061062

Cotton JA, Page RDM (2005) Rates and patterns of gene duplication and loss in the human genome. Proc R Soc B 272:277–283

CRAN (2012) The comprehensive R archive network. URL http://cran.r-project.org

Craviotto C, Jones WB, Thron WJ (1993) A survey of truncation error analysis for Padé and continued fraction approximants. Acta Appl Math 33:211–272

Crawford FW, Suchard MA (2011) Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. J Math Biol – In press

Crawford FW, Suchard MA (2012) Estimation for Markov counting processes in evolution

Crawford FW, Minin VN, Suchard MA (2012a) Estimation in the general birth-death process. Under review

Crawford FW, Weiss RE, Suchard MA (2012b) Sex, lies, and self-reported counts: Bayesian longitudinal analysis of heaped integer data. In progress

Crockett A, Crockett R (2006) Consequences of data heaping in the British religious census of 1851. Hist Method 39(1):24–46

Crozier R, Agapow P, Dunnett L (2006) Conceptual issues in phylogeny and conservation: a reply to faith and baker. Evol Bioinform Online 2:197–199

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

Darwin JH (1956) The behaviour of an estimator for a simple birth and death process. Biometrika 43(1):23–31

Dauxois J (2004) Bayesian inference for linear growth birth and death processes. J Stat Plan Infer 121(1):1–19

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39(1):1–38

Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PLoS ONE 1(1):e85

Dennis B (2002) Allee effects in stochastic populations. Oikos 96(3):389–401

Donnelly P (1984) The transient behaviour of the Moran model in population genetics. Math Proc Cambridge 95(02):349–358

Doss CR, Suchard MA, Holmes I, Kato-Maeda M, Minin VN (2010) Great expectations: EM algorithms for discretely observed linear birth-death-immigration processes. ArXiv e-prints 1009.0893

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution

Eckert KA, Hile SE (2009) Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol Carcinogen 48(4):379–388

Eizirik E, Murphy W, Koepfli K, Johnson W, Dragoo J, Wayne R, O'Brien S (2010) Pattern and timing of diversification of the mammalian order carnivora inferred from multiple nuclear gene sequences. Mol Phylogenet Evol 56(1):49–63

Eldredge N, Gould S (1972) Punctuated equilibria: an alternative to phyletic gradualism, Freeman, Cooper and Company, pp 82–115

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5(6):435–445

Faith D (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1):1–10

Faith D (2006) The role of the phylogenetic diversity measure, pd, in bio-informatics: getting the definition right. Evolutionary bioinformatics online 2:277

Faith D, Baker A (2006) Phylogenetic diversity (pd) and biodiversity conservation: some bioinformatics challenges. Evol Bioinform Online 2:121

Faller B, Pardi F, Steel M (2008) Distribution of phylogenetic diversity under random extinction. J Theor Biol 251(2):286–296

Feller W (1971) An Introduction to Probability Theory and its Applications. Wiley series in probability and mathematical statistics, Wiley New York

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125(1):1–15

Fenton KA, Johnson AM, McManus S, Erens B (2001) Measuring sexual behaviour: methodological challenges in survey research. Sex Transm Infect 77(2):84–92

Flajolet P, Guillemin F (2000) The formal theory of birth-and-death processes, lattice path combinatorics and continued fractions. Adv Appl Probab 32(3):750–778

Foote M (1993) Contributions of individual taxa to overall morphological disparity. Paleobiology 19(4):403–419

Garland T, Harvey P, Ives A (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41(1):18–32

Gernhard T, Hartmann K, Steel M (2008) Stochastic properties of generalised Yule models, with biodiversity applications. J Math Biol 57(5):713–735

Ghosh P, Tu W (2009) Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. J Am Stat Assoc 104(486):474–485

Gittleman J (1986) Carnivore life history patterns: allometric, phylogenetic, and ecological associations. Am Nat 127(6):744–771

Gittleman J, Purvis A (1998) Body size and species–richness in carnivores and primates. P Roy Soc Lond B Bio 265(1391):113–119

Golubjatnikov R, Pfister J, Tillotson T (1983) Homosexual promiscuity and the fear of AIDS. Lancet 322:681

Gould S, Eldredge N (1977) Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3(2):115–151

Grafen A (1989) The phylogenetic regression. Phil T Roy Soc B 326(1233):119–157

Grassmann W (1977a) Transient solutions in Markovian queues : An algorithm for finding them and determining their waiting-time distributions. Eur J Oper Res 1(6):396–402

Grassmann WK (1977b) Transient solutions in Markovian queueing systems. Comput Oper Res 4(1):47–53

Grimmett G, Stirzaker D (2001) Probability and random processes. Oxford University Press

Grunwald GK, Bruce SL, Jiang L, Strand M, Rabinovitch N (2011) A statistical model for under- or overdispersed clustered and longitudinal count data. Biometrical J 53(4):578–594

Guillemin F, Pinchon D (1999) Excursions of birth and death processes, orthogonal polynomials, and continued fractions. J Appl Probab 36(3):752–770

Harvey P, Pagel M (1991) The comparative method in evolutionary biology. Oxford university press

Heitjan DF (1989) Inference from grouped continuous data: A review. Stat Sci 4(2):164–179

Heitjan DF, Rubin DB (1990) Inference from coarse data via multiple imputation with application to age heaping. J Am Stat Assoc 85(410):304–314

Heitjan DF, Rubin DB (1991) Ignorability and coarse data. Ann Stat 19(4):2244–2253

Hobolth A (2008) A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-depending substitution rates. J Comput Graph Stat 17(1):1–25

Hobolth A, Jensen JL (2005) Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat Appl Genet Mol 4(1):1–19

Hobolth A, Stone EA (2009) Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. Ann Appl Stat 3(3):1024–1231

Hobson R (1976) Properties preserved by some smoothing functions. J Am Stat Assoc 71(355):763–766

Holmes I, Bruno WJ (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics 17(9):803–820

Holmes I, Rubin G (2002) An expectation maximization algorithm for training hidden substitution models. J Mol Biol 317(5):753–764

Huttenlocher J, Hedges LV, Bradburn NM (1990) Reports of elapsed time: Bounding and rounding processes in estimation. J Exp Psychol Learn 16(2):196–213

Ismail MEH, Letessier J, Valent G (1988) Linear birth and death models and associated Laguerre and Meixner polynomials. J Approx Theory 55(3):337–348

Jacobsen M, Keiding N (1995) Coarsening at random in general sample spaces and random censoring in continuous time. Ann Stat 23(3):774–786

Jamshidian M, Jennrich RI (1993) Conjugate gradient acceleration of the EM algorithm. J Am Stat Assoc 88(421):221–228

Kalbfleisch JD, Lawless JF (1985) The analysis of panel data under a Markov assumption. J Am Stat Assoc 80(392):863–871

Karlin S, McGregor J (1957a) The classification of birth and death processes. Trans Am Math Soc 86(2):366–400

Karlin S, McGregor J (1957b) The differential equations of birth-and-death processes, and the Stieltjes moment problem. Trans Am Math Soc 85(2):589–646

Karlin S, McGregor J (1958a) Linear growth, birth and death processes. J Math Mech 7(4):643–662

Karlin S, McGregor J (1958b) Many server queueing processes with Poisson input and exponential service times. Pacific J Math 8(1):87–118

Karlin S, McGregor J (1962) On a genetics model of Moran. Math Proc Cambridge 58(02):299–311

Karlin S, Taylor HM (1975) A First Course in Stochastic Processes. Academic Press

Keiding N (1974) Estimation in the birth process. Biometrika 61(1)

Keiding N (1975) Maximum likelihood estimation in the birth-and-death process. Ann Stat 3(2):363–372

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18(1):30–38

Kendall DG (1948) On the generalized "birth-and-death" process. Ann Math Stat 19(1):1–15

Kendall DG (1949) Stochastic processes and population growth. J Roy Stat Soc B Met 11(2):230–282

Kimmel M, Axelrod D (2002) Branching Processes in Biology. Interdisciplinary applied mathematics: Mathematical biology, Springer

Kingman JFC (1982a) The coalescent. Stat Proc Appl 13(3):235–248

Kingman JFC (1982b) On the genealogy of large populations. J Appl Probab 19:27–43

Klar B, Parthasarathy PR, Henze N (2010) Zipf and Lerch limit of birth and death processes. Probab Eng Inform Sc 24(01):129–144

Klovdahl A, Potterat J, Woodhouse D, Muth J, Muth S, Darrow W (1994) Social networks and infectious disease: The Colorado Springs study. Soc Sci Med 38(1):79 – 88

Krone SM, Neuhauser C (1997) Ancestral processes with selection. Theor Popul Biol 51:210–237

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. P Natl Acad Sci USA 95(18):10,774–10,778

Lange K (1995a) A gradient algorithm locally equivalent to the EM algorithm. J Roy Stat Soc B Met 57(2):425–437

Lange K (1995b) A quasi-Newton acceleration of the EM algorithm. Stat Sinica 5:1–18

Lange K (2010a) Applied Probability, 2nd edn. Springer texts in statistics, Springer New York

Lange K (2010b) Numerical Analysis for Statisticians (Statistics and Computing), 2nd edn. Springer New York

Lee J, Weiss R (2011) Using a birth-death process to account for reporting errors in longitudinal self-reported counts of behavior. Submitted

Lenin RB, Parthasarathy PR (2000) A birth-death process suggested by a chain sequence. Comput Math Appl 40(2-3):239–247

Lentz WJ (1976) Generating Bessel functions in Mie scattering calculations using continued fractions. Appl Opt 15(3):668–671

Levin D (1973) Development of non-linear transformations for improving convergence of sequences. Int J Comput Math 3(B):371–388

Lindley DV (1950) Grouping corrections and maximum likelihood equations. Math Proc Cambridge 46(01):106–110

Liu H, Beckett LA, DeNardo GL (2007) On the analysis of count data of birth-and-death process type: with application to molecularly targeted cancer therapy. Statist Med 26:11141135

Lorentzen L, Waadeland H (1992) Continued Fractions with Applications. North-Holland, Amsterdam

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J Roy Stat Soc B Met 44(2):226–233

Martins E, Hansen T (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Mayrose I, Barker MS, Otto SP (2010) Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst Biol 59(2):132–144

Mederer M (2003) Transient solutions of Markov processes and generalized continued fractions. IMA J Appl Math 68(1):99–118

Meilijson I (1989) A fast improvement to the EM algorithm on its own terms. J Roy Stat Soc B Met 51(1):127–138

Meng XL, Rubin DB (1991) Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. J Am Stat Assoc 86(416):899–909

Metzner P, Dittmer E, Jahnke T, Schütte C (2007) Generator estimation of Markov jump processes. J Comput Phys 227:353–375

Minin V, Suchard M (2008) Counting labeled transitions in continuous-time markov models of evolution. J Math Biol 56(3):391–412

Mohanty S, Montazer-Haghighi A, Trueblood R (1993) On the transient behavior of a finite birth-death process with an application. Comput Oper Res 20(3):239–248

Mooers A, Gascuel O, Stadler T, Li H, Steel M (2011) Branch lengths on birth-death trees and the expected loss of phylogenetic diversity. Syst Biol 61(2):195–203

Moran PAP (1951) Estimation methods for evolutive processes. J Roy Stat Soc B Met 13(1):141–146

Moran PAP (1953) The estimation of the parameters of a birth and death process. J Roy Stat Soc B Met 15(2):241–245

Moran PAP (1958) Random processes in genetics. Math Proc Cambridge 54(01):60–71

Moritz C (2002) Strategies to protect biological diversity and the evolutionary processes that sustain it. Syst Biol 51(2):238–254

Mulder W (2011) Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. J Theor Biol 280(1):139–145

Murphy JA, O'Donohoe MR (1975) Some properties of continued fractions with applications in Markov processes. IMA J Appl Math 16(1):57–71

Murray J (2002) Mathematical Biology: An Introduction, Interdisciplinary applied mathematics, vol 1. Springer, New York

Myers RJ (1954) Accuracy of age reporting in the 1950 united states census. J Am Stat Assoc 49(268):826–831

Myers RJ (1976) An instance of reverse heaping of ages. Demography 13(4):577–580

Nee S (2006) Birth-death models in macroevolution. Annu Rev Ecol Evol S 37:1–17

Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos T Roy Soc B 344(1309):305–311

Neuts MF (1995) Algorithmic Probability: A Collection of Problems (Stochastic Modeling Series). Chapman and Hall/CRC

Novozhilov AS, Karev GP, Koonin EV (2006) Biological applications of the theory of birth-and-death processes. Brief Bioinform 7(1):70–85

Nowak R, Paradiso J (1999) Walker's mammals of the world. Cambridge Univ Press

Numerical Recipes Software (2007) Derivation of the Levin transformation. Numerical recipes webnote No 6 URL http://www.nr.com/webnotes?6

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60(5):922–933

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290

Parthasarathy PR, Sudhesh R (2006a) Exact transient solution of a state-dependent birth-death process. J Appl Math Stoch Anal 82(6):1–16

Parthasarathy PR, Sudhesh R (2006b) A formula for the coefficients of orthogonal polynomials from the three-term recurrence relations. Appl Math Lett 19(10):1083–1089

Parthasarathy PR, Lenin RB, Schoutens W, Assche WV (1998) A birth and death process related to the Rogers-Ramanujan continued fraction. J Math Anal Appl 224(2):297–315

Press WH (2007) Numerical Recipes: the Art of Scientific Computing. Cambridge University Press New York

Purvis A (2004) Evolution: How do characters evolve? Nature 432(7014):Published online

Purvis A, Garland T (1993) Polytomies in comparative analyses of continuous characters. Syst Biol 42(4):569–575

Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43(3):304–311

Renshaw E (1993) Modelling Biological Populations in Space and Time. Cambridge Studies in Mathematical Biology, Cambridge University Press

Renshaw E (2011) Stochastic Population Processes: Analysis, Approximations, Simulations. Oxord University Press

Revell L (2010) Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1(4):319–329

Reynolds JF (1973) On estimating the parameters of a birth-death process. Aust J Stat 15(1):35–43

Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72(4):686–727

Ricklefs R (2004) Cladogenesis and morphological diversification in passerine birds. Nature 430(6997):338–341

Ricklefs R (2006) Time, species, and the generation of trait variance in clades. Syst Biol 55(1):151–159

Roberts JM, Brewer DD (2001) Measures and tests of heaping in discrete quantitative distributions. J Appl Stat 28(7):887–896

Rose O, Falush D (1998) A threshold size for microsatellite expansion. Mol Biol Evol 15(5):613–615

Rosenberg NA, Tsolaki AG, Tanaka MM (2003) Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in Mycobacterium tuberculosis. Theor Popul Biol 63(4):347–363

Rosenlund SI (1978) Transition probabilities for a truncated birth-death process. Scand J Stat 5(2):119–122

Rotheram-Borus M, Lee M, Murphy D, Futterman D, Duan N, Birnbaum J, Lightfoot M (2001) Efficacy of a preventive intervention for youths living with HIV. Am J Public Health 91(3):400–405

Rowland M (1990) Self-reported weight and height. Am J Clin Nutr 52(6):1125–1133

Sainudiin R, Durrett RT, Aquadro CF, Nielsen R (2004) Microsatellite mutation models. Genetics 168(1):383–395

Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365–371

Schneeweiss H, Augustin T (2006) Some recent advances in measurement error models and methods. Allgemeines Statistisches Archiv 90(1):183–197

Schneeweiss H, Komlos J (2009) Probabilistic rounding and sheppard's correction. Stat Methodol 6(6):577 – 593

Schneeweiss H, Komlos J, Ahmad AS (2010) Symmetric and asymmetric rounding: a review and some new results. AStA Adv Stat Anal 94(3):247–271

Sharma OP, Dass J (1988) Multi-server Markovian queue with finite waiting space. Sankhya Ser B 50(3):428–431

Sheppard WF (1897) On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. P Lond Math Soc s1-29(1):353–380

Sidlauskas B (2007) Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. Evolution 61(2):299–316

Singh K, Suchindran C, Singh R (1994) Smoothed breastfeeding durations and waiting time to conception. Soc Biol 41(3-4):229–39

Slater G, Harmon L, Wegmann D, Joyce P, Revell L, Alfaro M (2012) Fitting models of continuous trait evolution to incompletely sampled comparative data using Approximate Bayesian Computation. Evolution 66:752–762

Stadler T (2011) Simulating trees with a fixed number of extant species. Syst Biol 60(5):676–684

Stadler T, Steel M (2012) Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J Theor Biol 297:33–40

Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by Yule-type speciation models. Math Biosci 170(1):91–112

Steel M, McKenzie A (2002) The 'shape' of phylogenies under simple random speciation models. Biological Evolution and Statistical Physics 585:162–180

Steel M, Mooers A (2010) The expected length of pendant and interior edges of a Yule tree. Appl Math Lett 23(11):1315–1319

Stockwell EG, Wicks JW (1974) Age heaping in recent national censuses. Soc Biol 21(2):163–167

Stone E (2011) Why the phylogenetic regression appears robust to tree misspecification. Syst Biol 60(3):245–260

Tallis GM (1967) Approximate maximum likelihood estimates from grouped data. Technometrics 9(4):599–606

Tan WY, Piantadosi S (1991) On stochastic growth processes with application to stochastic logistic growth. Stat Sinica 1:527–540

Taylor H, Karlin S (1998) An Introduction to Stochastic Modeling. Academic Press San Diego

Thompson IJ, Barnett AR (1986) Coulomb and Bessel functions of complex arguments and order. J Comput Phys 64:490–509

Thorne J, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol 33(2):114–124

Turnbaugh P, Hamady M, Yatsunenko T, Cantarel B, Duncan A, Ley R, Sogin M, Jones W, Roe B, Affourtit J, et al (2008) A core gut microbiome in obese and lean twins. Nature 457(7228):480–484

Vowles EJ, Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Mol Biol Evol 23(3):598–607

Wall HS (1948) Analytic Theory of Continued Fractions. University Series in Higher Mathematics, D. Van Nostrand Company, Inc. New York

Wallis J (1695) Opera Mathematica Volume 1. Oxoniae e Theatro Shedoniano, reprinted by Georg Olms Verlag, Hildeshein, New York, 1972

Wanek LA, Goradia TM, Elashoff RM, Morton DL (1993) Multi-stage Markov analysis of progressive disease applied to melanoma. Biom J 35(8):967–983

Wang H, Heitjan DF (2008) Modeling heaping in self-reported cigarette counts. Statistics in Medicine 27(19):3789–3804

Webb C, Ackerly D, McPeek M, Donoghue M (2002) Phylogenies and community ecology. Annu Rev Ecol Syst 33:475–505

Webster MT, Smith NGC, Ellegren H (2002) Microsatellite evolution inferred from human and chimpanzee genomic sequence alignments. P Natl Acad Sci USA 99(13):8748–8753

Weinhardt LS, Forsyth AD, Carey MP, Jaworski BC, Durant LE (1998) Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. Arch Sex Behav 27:155–180

Westoff CF (1974) Coital frequency and contraception. Fam Plann Perspect 6(3):136–141

Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM (2003) Likelihood-based estimation of microsatellite mutation rates. Genetics 164(2):781–787

Wiederman MW (1997) The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. J Sex Res 34(4):375–386

Wolff RW (1965) Problems of statistical inference for birth and death queuing models. Oper Res 13(3):343–357

Wright DE, Bray I (2003) A mixture model for rounded data. J Roy Stat Soc D 52(1):3–13

Yule (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos T R Soc Lon B 213:21–87

Zhou H, Lange K, Suchard M (2010) Graphics processing units and high-dimensional optimization. Stat Sci 25(3):311–324