Likelihood-Based Inference for Partially Observed Multi-Type Markov Branching Processes

Jason Xu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Vladimir N. Minin, Chair

Peter Guttorp

Jon Wakefield

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Likelihood-Based Inference for Partially Observed Multi-Type Markov Branching Processes

Jason Xu

Chair of the Supervisory Committee:
Vladimir N. Minin
Department of Statistics and Department of Biology

Markov branching processes are a class of continuous-time Markov chains (CTMCs) frequently used in stochastic modeling with ubiquitous applications. Bivariate or multi-type processes are necessary to model phenomena such as competition, predation, or infection, but often feature large or uncountable state spaces, rendering many general CTMC techniques impractical. We present spectral techniques to compute the transition probabilities and related quantities discretely and unevenly observed multi-type branching processes, enabling likelihood-based inference. Our technique reduces these calculations to low dimensional integration, and analogously enables calculation of related terms such as expected sufficient statistics within an expectation maximization (EM) algorithm. We rigorously assess our EM algorithm in several simulation studies applied to a birth-death-shift (BDS) model, and apply it to estimate intrapatient time evolution of IS*6110* transposable element, a genetic marker frequently used during epidemiological studies of *Mycobacterium tuberculosis*. Further, we incorporate our methods for computing transition probabilities within a compressed sensing framework, demonstrating scalability in the presence of sparsity. Finally, we extend these ideas to loss function estimation of *in vivo* hematopoietic rates from single-cell lineage tracking data, and develop efficient Bayesian methods for fitting general stochastic epidemic models to discretely observed time series data and partially observed incidence data.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Above all, I would like to thank my committee members, who have been a source of guidance and inspiration throughout my graduate education. Peter's influence and oftentimes daunting breadth of knowledge have been formative in shaping my research path since the beginning of my time here—I can only hope to one day become such an inventive thinker. Jon taught me most of what I know during my first year of coursework with good humor, and I've continued to benefit from his expertise and to endure his wit over the years through research and the occasional heavy metal concert. Vladimir has been the best advisor possible within the limits of my imagination, and it is indeed difficult to imagine that I would have come remotely close to producing this thesis without his lead. He has been impossibly patient and wise, maintaining a delicate balance between allowing me a generous amount of creative freedom, and carefully guiding me to ensure that I charted a fine course. Vlad is a truly kind and dedicated mentor with a beautiful mind, and I will always admire and do my best to emulate his intrepid eagerness to seek out new problems in new fields, and to fashion wildly imaginative approaches toward their solution.

Each of you has been a role model of mine; you have brought a deep humanity to my understanding of the discipline that I hope to carry with me throughout the profession. I could not have picked a better crew to help me navigate this PhD.

There are a number of others at the University of Washington I must thank for making it such a wonderful and stimulating place to work. My life would have quickly fallen apart without the tremendous help of Ellen Reynolds and the seemingly magical departmental staff. The faculty members have been an invaluable resource, and it has been a privilege to have grown under their direction. I am indebted to June Morita for mentoring me through my first teaching experience, and making it a deeply enjoyable and rewarding one. Jan Abkowitz taught me how to communicate across disciplines, and showed me the joy of collaboration through our work on hematopoietic modeling. I am grateful to Emily Fox, who has greatly influenced the way I think about data and pushed me as a researcher through our work in scalable machine learning, and Hari Narayanan, who introduced me to new and elegant areas of statistics through our work on sample complexity and function interpolation. Cindy Dunbar, Marc Suchard, Forrest Crawford, Lam Ho, Dillon Laird, and Sam Koelle have been an absolute pleasure to work with, and I look forward to continuing our conversations in the future.

I wish to thank my past mentors and teachers for their encouragement even when I truly knew nothing, and for instilling in me their own passion and curiosity. In particular, Mark Huber and Kevin Lin are responsible for introducing me to the beauty of stochastic processes and probabilistic reasoning. I owe an extra serving of gratitude to Janet Munro and Susan Williams, still bigger influences in my education than they know. You have inspired me.

My colleagues and friends in the department have provided an overwhelmingly warm and welcoming community. I would not have made it through these years without Alex Tank and Kitty Mohammed, and am grateful to have shared this journey with such kindred companions. I must also single out Rebecca Ferrell, one of the most brilliant people I know, with whom I was lucky to share a rare friendship as well as an office. There are far too many others I must thank, and to each of you, I owe more than I can say. I'm proud to have spent some of our best years together—it's been golden.

Outside of the department, my less statistical friends have been no less a constant source of support, even from afar. I owe thanks to Brennan Vincent for showing me how to properly use a computer, and to Chris Fowler for being a fine friend over many climbs and ensuing ales. I want to thank Lutèce "Lettuce" Ragueneau, my partner in crime throughout the dissertation writing process, for both her support and much-needed distractions. I'm very fond of the times and conversations I've shared with Wes Jackson, Nick Driscoll, Abe Engle, Matts Junge and Robinson, Yunqi Bu, Kunal Mangal, Sam Nerenberg, Jay Garlapati, Cameron Solem, and Mark Pearcy, who have sustained me through the lows and dead ends. Special thanks to Howard Cheng, Alan Mackey, Gleb Zhelezov, and Juan Carlos Ramirez for enduring me and many technical and mathematical discussions, even at odd hours and over the most inane of messaging platforms.

Last and largest, this is dedicated to my family, whose inexhaustible love and faith in me are the foundation for everything I've done.

DEDICATION

To Mom, Dad, Chris

Chapter 1

OVERVIEW OF CONTRIBUTIONS

Branching processes have a rich history in modeling the reproductive behavior of individuals or particles in a population over time. Originally motivated as a model to study the extinction probability of aristocratic family lines across generations, branching processes have since been refined and rediscovered in many different settings. Even the original application of simple Galton-Watson branching processes to model the survival of family surnames was considered in at least two famous cases separately—they were named after work by Francis Galton together with Henry William Watson [Harris, 1963], who independently rediscovered results known three decades earlier in 1845 by the French statistician Iréné-Jules Bienaymé; see [Heyde and Seneta, 1972, Guttorp et al., 1995] for more details on the history. Branching processes are simple to specify—each particle acts independently, and reproduces or dies according to a fixed probability distribution—yet they have proven to be a powerful and flexible modeling tool despite this simplicity. As the range of applications grew to include scientific fields ranging from cellular and nuclear kinetics to genetics and disease dynamics, many elaborations on the simple Galton-Watson-Bienaymé process, in which unit intervals of time are measured discretely as generations and all particles are of a single type, have been considered. A more rigorous characterization of the mathematical and probabilistic foundations of these systems has developed along the way, and has naturally lead to quantitative questions about the behavior of the process over time. For instance, denoting the process $\mathbf{X}(t)$, one may ask: what is the probability that the population goes extinct by a given time, $\Pr(\mathbf{X}(t) = 0)$, or the probability of ultimate extinction, $\lim_{t\to\infty} \Pr(\mathbf{X}(t) = 0)$? What are the expected size and variance of the population $\mathbf{X}(t)$? What is the equilibrium distribution of $\mathbf{X}(t)$? Such questions have been well-studied with an aim of characterizing

the dependence of a model's behavior given a set of parameters $\boldsymbol{\theta}$ that specify the branching process. Mathematical expressions for these quantities are available in many cases; most of these results concern mean behavior, asymptotic properties, critical thresholds, and the like.

However, we are often faced with the problem of estimating the most probable model parameters $\hat{\boldsymbol{\theta}}$, or a range of parameters that can plausibly explain how a branching population evolves, given an observation of the process. In many cases, the observations provide an incomplete picture of the underlying process, which evolves continuously through time: data may be available only at unevenly spaced times corresponding to clinical sampling times in an epidemiological study, or at a discrete set of measurement times corresponding to a limitation in monitoring frequency given an experimental design or technology. This inverse problem can be significantly more challenging. The known quantities mentioned above have limited use toward inference. Many statistically relevant quantities are difficult to derive and have no known solutions, or are infeasible to compute from a numerical or algorithmic standpoint. In particular, closed forms for finite-time *transition probabilities*, the conditional probability that a trajectory ends at a given state, given a starting state and time interval, are unavailable. These transition probabilities are central to many inferential approaches, comprising the observed likelihood function when data from the process are available at a set of discrete times. Related quantities such as endpoint-conditioned moments and steady-state distributions pose similar challenges.

Recent work has resolved some of these questions in one-dimensional cases such as general birth-death processes and linear birth-death processes with immigration, but it is often necessary to model systems with more than one species — bivariate or other multi-type processes are commonly used to model phenomena such as competition, predation, or infection. In this thesis, we focus on developing tools for fitting data to multi-type branching processes that evolve in continuous time, taking values in a discrete state space. In practice, the data provide only snapshots of the process in time, and significant obstacles arise in statistical inference of the process rates $\boldsymbol{\theta}$ in the partially observed setting. While these branching processes are a class of continuous-time Markov chains (CTMCs) and inherit their nice math-

ematical properties, the models often feature large or uncountable state spaces, rendering general CTMC techniques such as matrix exponentiation and simulation-based approaches impractical. Fortunately, the additional branching structure can be further exploited to derive new methods and algorithms when these general-purpose tools fall short. We present novel techniques to evaluate the aforementioned quantities in this setting, enabling mainstays of statistical inference for missing data to be applied in both frequentist and Bayesian frameworks.

To this end, we introduce spectral techniques to compute transition probabilities comprising the observed data likelihood for discretely and unevenly observed multi-type branching processes, enabling likelihood-based inference. Our technique reduces these calculations to low dimensional integration, and analogously enables calculation of related terms such as expected sufficient statistics within an expectation maximization (EM) algorithm. We assess robustness, accuracy and efficiency of our EM algorithm in several simulation studies applied to a birth-death-shift (BDS) model, and apply it to estimate intrapatient time evolution of transposable elements, genetic markers with important applications to molecular epidemiology. We then demonstrate that orders of magnitude increases in computational efficiency can be achieved in the presence of sparsity, and scale our methods using a compressed sensing framework. We demonstrate these gains in the context of the BDS model, as well as a two-type stochastic compartmental model. Finally, we consider applications of branching process techniques to single-cell lineage barcoding data and partially observed data from disease epidemics. In these contexts, we develop a method of moments estimator when likelihood-based inference is still intractable, and pursue new approximations that faithfully mimic nonlinear dynamics in stochastic epidemic models, yielding closed form finite-time transition densities and efficient proposal densities for *exact* Bayesian inference.

In the following chapter, we briefly introduce notation and background on continuous-time Markov chains (CTMCs), and provide an overview of likelihood-based inference in the fully observed and partially observed settings. Working in this framework, we define Markov branching processes and introduce relevant classic and modern techniques in Chapter 3,

providing a sense of the body of work that the following chapters will build upon. Having established technical background and context within the larger literature, Chapter 4 develops new methodology to compute the discretely observed likelihood of independent realizations from a multi-type branching process. Our approach applies spectral techniques to the generating function of the process, enabling maximum likelihood (ML) inference as well as an expectation-maximization algorithm that outperforms direct maximization in the *panel data* setting, where data are observed at discrete and possibly irregularly spaced time points, and rates of each independent realization can be parametrized by multiple process-specific covariates. We rigorously assess performance in a series of simulation studies, demonstrating that our method is significantly more robust than prior work relying on rigid model simplifications. This method is general, and is fully implemented in an `R` package `bdsem`; we focus exposition on the BDS model of transposon evolution, and apply the algorithm to estimate intrapatient time evolution of IS*6110* transposable element, a genetic marker used in epidemiological studies of *Mycobacterium tuberculosis*.

While these techniques enable likelihood based inference when classical matrix-based CTMC methods or simulation approaches are impractical, they also suffer limitations when population sizes are large. This bottleneck motivates the material in Chapter 5, where we investigate accelerating transition probability computations in the presence of sparsity. We demonstrate that enormous gains in efficiency can be achieved by incorporating our spectral technique within a compressed sensing paradigm, enabling an order of magnitude increase in the population sizes we can consider by imposing only a generic sparsity assumption. These results are assessed on the BDS model as well as a two-type stochastic compartmental model of hematopoiesis, the complex mechanism of blood cell production, that has been analyzed in a series of statistical studies.

Despite the improvements over existing methods presented in these chapters, computational limitations still exclude their application to very large, high-throughput datasets from recently emerged experimental technologies. Such data have already produced significant scientific insights, but lack any quantitative framework for rigorous statistical tasks such as

parameter inference or model selection. Chapter 6 is devoted to developing such a framework for multi-compartment branching process models of hematopoiesis. We present an efficient loss function estimator based on the method of moments for single-cell lineage tracking data from genetic barcoding experiments, enabling cell fate decision rate inference in a much wider class of branching models than the previous two-compartment model. This method not only deals with the discretely observed aspect of the data, but overcomes the challenge of *partial observations* based on noisy sampling from a latent branching process. We close by discussing possible future work in adapting this estimation procedure within model selection frameworks.

Chapter 7 introduces a new two-type branching process approximation to the general stochastic epidemic model, also called the Susceptible-Infected-Removed (SIR) model. The SIR model mechanistically describes infectious disease dynamics over time as a nonlinear stochastic system, and past work has considered branching birth-death approximations for early stages of an epidemic, as well as coupling strategies relying on branching process techniques to derive asymptotic results. Our multi-type branching representation provides an improved approximation that closely mimics the nonlinear interactions between infected and susceptible individuals, despite only allowing linear, independent interaction and thereby retaining mathematical simplicity. We show that remarkably, this formulation admits closed forms for transition probabilities, allowing for very fast likelihood computations for discretely observed epidemic data that do not require more advanced spectral techniques. Furthermore, we apply the method to Bayesian analysis of data from the Plague of Eyam, and demonstrate that estimated parameters under this approximation are very similar to those recovered using computationally intensive exact methods, and preferable to cruder approximations in previous studies. Next, we consider the two-type branching approximation for *partially* observed incidence data, the case where only new infections are observed. The branching process approximation leads to an efficient way to propose complete data trajectories conditional on such partial incidence data, and we develop a data augmentation Markov chain Monte Carlo algorithm that scales to very large epidemics and targets the exact joint posterior of

the parameters governing disease dynamics. We close this chapter by exploring extensions to include covariate-dependent rates and stratified populations. Finally, Chapter 8 summarizes the work presented in this thesis with a discussion and future directions.

Chapter 2

PRELIMINARIES OF CTMC INFERENCE

This chapter establishes the notation, terminology, and basic modeling framework we will use throughout this thesis. We begin by providing an overview of CTMCs, important quantities that characterize them, and standard methods for computing these quantities. Next, we review existing methods for statistical inference in both the complete data and partially observed data settings and note the challenges that arise.

2.1 Continuous-time Markov chains

Continuous-time Markov chains are popular models used to describe the time-evolution of processes arising in many scientific disciplines, in part due to their nice mathematical properties. As branching processes are a special type of this class, we begin with a brief review.

2.1.1 Mathematical formulation

A CTMC, also called a continuous-time Markov process, is a random vector $\mathbf{X}(t)$ taking values in a discrete state space Ω that obeys the *Markov property*: the future behavior of the chain depends only on its current state, and not on any past history of the process. Put simply, this property asserts that the chain is memoryless. That is, for any set of times $0 \leq s_1 < s_2 < \ldots < s_n < s$ and any possible set of states $j, k, i_1, \ldots, i_n \in \Omega$, at time $t \geq 0$ we have

$$\Pr(\mathbf{X}(t+s) = k | \mathbf{X}(s) = j, \mathbf{X}(s_n) = i_n, \ldots, \mathbf{X}(s_1) = i_1) = \Pr(\mathbf{X}(t+s) = k | \mathbf{X}(s) = j).$$

We are interested in *time-homogeneous* CTMCs, where transition probabilities do not change over time so that

$$\Pr(\mathbf{X}(t+s) = k | \mathbf{X}(s) = j) = \Pr(\mathbf{X}(t) = k | \mathbf{X}(0) = j).$$

In this case, we abbreviate notation for these finite-time transition probabilities as

$$\Pr(\mathbf{X}(t) = k | \mathbf{X}(0) = j) = p_{jk}(t).$$

The collection of these probabilities $\mathbf{P}(t) = \{p_{jk}(t)\}_{j,k\in\Omega}$ is called the *transition probability matrix*, a central quantity in describing the CTMC. Note this matrix is a function of t, and that $\mathbf{P}(0) = \mathbf{I}$ is the identity matrix. We will posit that the matrix is well-behaved in the sense that $\lim_{t\to 0^+} \mathbf{P}(t) = \mathbf{I}$: under this assumption, there always exists an *intensity matrix* $\mathbf{Q} = \{q_{jk}\}_{j,k\in\Omega}$, also called the *infinitesimal generator*, parametrizing the CTMC which does *not* depend on t. The entries q_{jk} now correspond to the instantaneous rates of transitions in the chain:

$$q_{jk} := \lim_{t\to 0^+} \frac{p_{jk}(t)}{t}.$$

We restrict our attention to processes $\mathbf{X}(t)$ that are *stable* and *conservative*. Stability is a regularity assumption avoiding explosive behavior; formally, $0 \leq q_{jk} \leq \infty$; conservative chains satisfy $q_j := -q_{jj} = \sum_{k\in\Omega} q_{jk}$, so that q_j represents the total instantaneous rate of leaving state j, and is balanced as the sum of instantaneous jump rates from j to all other states $k \in \Omega$.

Under this construction, the dwell times or waiting times of the CTMC are exponentially distributed: if $\mathbf{X}(t) = j$, then the time until $\mathbf{X}(t)$ moves to a new state is distributed $\text{Exp}(q_j)$, and at this instant the process jumps to state $k \neq j$ with probability q_{jk}/q_j. Note that this property gives us a straightforward way to simulate from the model: draw exponential random variables sequentially to determine the time of next events, and then sample the jump probability to determine the type of event that occurs. We also remark that this exponential

waiting time property highlights the way that CTMCs are a natural generalization of the simpler discrete-time Markov chain: instead of unit times between transitions, the times between events in a CTMC are independent exponential random variables.

To relate the transition probability matrix and infinitesimal generator, we will use the Chapman-Kolmogorov equation, an important property that follows straightforwardly from the Markov and homogeneity assumptions that states $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$ for all $t, s > 0$. Entrywise, this means for any $j, k \in \Omega$,

$$p_{jk}(t+s) = \sum_{i \in \Omega} p_{ji}(t) p_{ik}(s),$$

which expresses the probability as a sum over possible intermediate states i we may pass through starting from state j after any length of time t, and then reaching the final state l after the remaining time s. Applying this relation to the derivative of the transition probability matrix,

$$\mathbf{P}'(t) = \lim_{h \to 0^+} \frac{\mathbf{P}(t+h) - \mathbf{P}(t)}{h} = \lim_{h \to 0^+} \frac{\mathbf{P}(t)\mathbf{P}(h) - \mathbf{P}(t)}{h} = \lim_{h \to 0^+} \frac{\mathbf{P}(t)[\mathbf{P}(h) - \mathbf{I}]}{h} = \mathbf{P}(t)\mathbf{Q},$$

where the last equality follows if we recall $\mathbf{P}(0) = \mathbf{I}$. This result is known as the Kolmogorov forward equation, also known as the master equation. The relation $\mathbf{P}'(t) = \mathbf{Q}\mathbf{P}(t)$ also holds by exchanging the order of multiplication above and is called the Kolmogorov backward equation. Either of these matrix differential equations, together with initial condition $\mathbf{P}(0) = \mathbf{I}$, can be solved to yield the unique solution $\mathbf{P}(t) = \exp(\mathbf{Q}t)$. Thus, we see that the finite-time transition probabilities of a CTMC are related to the infinitesimal generator via matrix exponentiation.

Together with the distribution over the initial state of the chain if $\mathbf{X}(0)$ is uncertain, $\mathbf{Q}$ completely characterizes a CTMC, and as such is the usual target of inference from data. Because $\mathbf{Q}$ is often a large (or infinite) matrix, it is often parametrized by a lower-dimensional vector $\boldsymbol{\theta}$, which instead becomes the objective for inference. In the following subsections, we

will review statistical inference using the likelihood in both the fully observed and partially observed cases.

2.1.2 Completely observed processes

In the complete data case, suppose we fully observe the trajectory of a homogeneous CTMC over a time interval $[0, T]$. Since the trajectory of $\mathbf{X}(t)$ is simply a sequence of jumps between states at random event times, equivalently this means we have recorded all N_T transitions or jumps that $\mathbf{X}(t)$ makes for $t \in [0, T]$, and have recorded the times of each transition $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{N_T})$ as well as their corresponding states, along with the initial state, $\mathbf{z} = (z_0, z_1, \ldots, z_{N_T})$. The Markov property yields a likelihood that factors nicely: recalling the exponential waiting times between events from our construction sheds intuition on the likelihood for the fully observed data:

$$L_c(\boldsymbol{\tau}, \mathbf{z}; \mathbf{Q}) = q_{z_0} e^{-q_{z_0}\tau_1} q_{z_1} e^{-q_{z_1}(\tau_2 - \tau_1)} \cdots q_{z_{N_T-1}} e^{-q_{z_{N_T-1}}(\tau_{N_T} - \tau_{N_T-1})} e^{-q_{z_{N_T}}(T - \tau_{N_T})} \prod_{i=1}^{N_T} \frac{q_{z_{i-1}z_i}}{q_{z_{i-1}}}.$$

This further simplifies in terms of the sufficient statistics of the trajectory, given by the total number of transitions between states $j, k \in \Omega$ $N_T(j, k)$ and the total time spent in each state j $d_T(j)$ in the interval $[0, T]$. Regrouping terms based on these quantities, the likelihood and log-likelihood $\ell_c(\boldsymbol{\tau}, \mathbf{z}; \mathbf{Q})$ are given by

$$\begin{aligned} L_c(\boldsymbol{\tau}, \mathbf{z}; \mathbf{Q}) &= \prod_{j \neq k} \left(q_{jk}^{N_T(j,k)} \right) e^{\sum_j d_T(j) q_j}, \\ \ell_c(\boldsymbol{\tau}, \mathbf{z}; \mathbf{Q}) &= \sum_j \sum_{k \neq j} N_T(j, k) \ln q_{ij} - \sum_k \sum_{k \neq j} d_T(j) q_{jk}. \end{aligned}$$

We see that the complete data log-likelihood of a CTMC has a convenient exponential family form, and for instance yields analytic expressions for derivatives in terms of its sufficient statistics. From here, it is straightforward to differentiate and arrive at the maximum likeli-

hood estimates for $\mathbf{Q}$:

$$\hat{q}_{jk} = \frac{N_T(j,k)}{d_T(j)} \qquad \text{and} \qquad \hat{q}_j = \frac{\sum_{j \neq k} N_T(j,k)}{d_T(j)}.$$

Furthermore, the complete-data likelihood is amenable to straightforward Bayesian inference. If we assume independent gamma priors on transition rates $q_{jk} \sim \text{Gamma}(\alpha, \beta)$, it is straightforward to show that the likelihood is conjugate to these priors, and the rates are also independently gamma distributed *a posteriori*:

$$q_{jk} | \mathbf{N}_T, \mathbf{d}_T \sim \text{Gamma}(N_T(j,k) + \alpha, d_T(j) + \beta).$$

Unfortunately, it is virtually never the case that we completely observe $\mathbf{X}(t)$ in practice. Nonetheless, these complete-data likelihoods are important quantities that show up within data augmentation and imputation algorithms even when the fully observed process is unavailable, as we will see later. The following section addresses the predominant setting for statistical analysis of CTMC data, where $\mathbf{X}(t)$ is only observed at discrete snapshots in time.

2.2 Discretely observed CTMCs

Suppose now that the process $\mathbf{X}(t)$ is observed only at a set of non-informative random times $t_1, \ldots, t_n$, and the corresponding observed states at these times are denoted $\mathbf{x} = [x_1, \ldots, x_n]$. The discretely observed likelihood for these data is then a product of the finite-time transition probabilities between observations, and its (log) form is given by

$$\ell_o(x_1, \ldots, x_n; \boldsymbol{\theta}) = \sum_{i=2}^{n} \log p_{x_{i-1}, x_i}(t_i - t_{i-1}; \boldsymbol{\theta}). \tag{2.1}$$

When multiple independent realizations of the process are observed, the joint log likelihood is simply the product of these quantities corresponding to each trajectory. Thus, likelihood computations require calculating these finite-time transition probabilities $p_{jk}(h)$ for any $j, k \in$

Ω and $h > 0$; we suppress the notational dependence on $\boldsymbol{\theta}$ for simplicity. While the formula $\mathbf{P}(h) = \exp(\mathbf{Q}h)$ always holds, its computation can be unwieldy or impossible in many cases. Formally, this matrix exponential is an infinite series

$$e^{\mathbf{Q}h} = \sum_{n=0}^{\infty} \frac{(\mathbf{Q}h)^n}{n!}.$$

When $\mathbf{Q}$ itself is an infinite matrix, this definition is not much help; otherwise, truncating the infinite sum can lead to wild inaccuracies, and even a more stable truncation via uniformization may be numerically unsatisfying. Typically, this matrix exponential is computed via numerical eigenvalue decomposition, and thus its computational complexity $\mathcal{O}(|\Omega|^3)$ is cubic in the size of the state space; see [Moler and Loan, 2003] for a review of such numerical methods. Directly maximizing this discrete data likelihood can therefore be challenging in all but relatively simple scenarios where $|\Omega|$ is modestly sized and $\boldsymbol{\theta}$ is low-dimensional. Similar challenges arise analogously in Bayesian approaches—the likelihood must be evaluated to determine acceptance ratios within a Metropolis-Hastings Markov Chain Monte Carlo algorithm, for instance. We will see in later sections how the additional structure of branching processes leads us to pursue alternate techniques for computing these finite-time transition probabilities $p_{jk}(h)$ appearing in the discrete data likelihood. In lieu of a computable discrete data likelihood function, researchers must resort to less statistically efficient methods that do not fully utilize the data, or simulation-based approaches bypassing likelihood computations that can be difficult to implement or computationally infeasible in complex models with large state spaces.

Below, we discuss two generic frameworks for imputing missing data that are in many cases preferable to directly maximizing the partially observed likelihood. Both take us back to the realm of complete data at the expense of considering some auxiliary information, and will be relevant to later chapters of the thesis.

2.2.1 EM algorithm

The celebrated expectation-maximization (EM) algorithm [Dempster et al., 1977] is a method that seeks to maximize the observed log-likelihood (2.1) indirectly. At each iteration p, the algorithm relies on the recursion

$$\boldsymbol{\theta}_{p+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}_p}[\ell_c(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\tau})|\mathbf{x}],$$

where we see the expectation of the complete data log-likelihood appears, and the new parameters are updated by maximizing this expectation. Instead of directly maximizing ℓ_o, the algorithm makes use of this expectation which acts as a more tractable surrogate function. The EM algorithm alternates between computing this expectation under current parameter settings (E-step), and then setting the new parameters at the next iteration by maximizing this expectation (M-step). When an appropriate EM algorithm is nontrivial, the difficulty usually lies in computing the E-step; the M-step is then typically accomplished using standard optimization routines such as Newton-Raphson updates. One can show that this procedure maintains an ascent property that implies convergence to a *local* maximum of ℓ_o [Lange, 1995].

Recall that the complete-data counterpart to the discretely-observed likelihood in Equation (2.1) is given by

$$\ell_c(\boldsymbol{\tau}, \mathbf{z}; \mathbf{Q}) = \sum_{i=2}^{n} \left[\sum_j \sum_{k \neq j} N_i(j,k) \ln q_{jk} \right] - \sum_{i=2}^{n} \left[\sum_j \sum_{k \neq j} d_i(j) q_{jk} \right],$$

where $N_i(j,k)$ is the number of transitions from state j to k in the ith time interval, and similarly $d_i(j)$ is the total time the process spends in state j during the ith time interval. The Markov property yields the nice form above that breaks down computations by interval, and we see that computing the expectation $\mathrm{E}[\ell_c]$ reduces to computing expected sufficient statistics. Further, by homogeneity, we can shift any interval to begin at 0, and only need to compute the endpoint-conditioned expectations for any states $j, k, l, m \in \Omega$ and any time

$t \geq 0$:

$$\mathrm{E}[N(j,k)|\mathbf{X}(0) = l, \mathbf{X}(t) = m], \qquad \mathrm{E}[d(j)|\mathbf{X}(0) = l, \mathbf{X}(t) = m].$$

Using Chapman-Kolmogorov arguments, one can derive integral forms for these expressions:

$$\mathrm{E}[N(j,k)|\mathbf{X}(0) = l, \mathbf{X}(t) = m] = \frac{1}{p_{lm}(t)} \int_0^t p_{lj}(w) q_{jk} p_{km}(w) \, \mathrm{d}w,$$

and similarly for the expected dwell time

$$\mathrm{E}[d(j)|\mathbf{X}(0) = l, \mathbf{X}(t) = m] = \frac{1}{p_{lm}(t)} \int_0^t p_{lj}(w) p_{jm}(t-w) \, \mathrm{d}w.$$

Again we see that while formulae are available, it is not always clear how to compute these endpoint-conditioned moments—in fact, we see that transition probabilities appear in the expressions, so that these too rely on a practical method to compute transition probabilities.

2.2.2 Markov chain Monte Carlo and data augmentation

Both the complete-data likelihood and discrete-data likelihood can be useful toward developing Markov chain Monte Carlo (MCMC) routines in a Bayesian framework, which we briefly describe here. Recall that the Bayesian approach toward inference of $\boldsymbol{\theta}$ targets the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$, where as before $\mathbf{x} = (x_1, \ldots, x_n)$ denotes a discrete set of observations of a CTMC $\mathbf{X}(t)$. Bayes rule yields that

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta}, \mathbf{x})}{\pi(\mathbf{x})} = \frac{\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x})},$$

where $\pi(\boldsymbol{\theta})$ denotes the prior distribution for $\boldsymbol{\theta}$, and $\pi(\mathbf{x})$ is the marginal distribution of the data $\mathbf{x}$. In particular, note that the the discrete-data likelihood $\pi(\mathbf{x}|\boldsymbol{\theta})$ appears here, whose logarithm we have previously denoted as $\ell_o(\mathbf{x}; \boldsymbol{\theta})$.

In some cases, the posterior distribution has convenient closed forms—we have seen that the complete data log-likelihood of a CTMC forms a conjugate family with gamma priors.

More often than not, especially in complex models, $\pi(\boldsymbol{\theta}|\mathbf{x})$ does not admit a closed form, often due to intractability of its normalizing constant, so that generating samples from $\pi(\boldsymbol{\theta}|\mathbf{x})$ or integrating functions over $\pi(\boldsymbol{\theta}|\mathbf{x})$ to compute moments of interest is nontrivial. Various Monte Carlo approaches such as importance sampling and rejection sampling have been proposed as a general-purpose way to generate independent samples from such posteriors [Rubinstein and Kroese, 2011]. These samples form an empirical distribution $\hat{\pi}_n(\boldsymbol{\theta}|\mathbf{x})$ approximating the posterior, which can be used to compute desired quantities—any definite integral can be numerically approximated by its corresponding empirical Monte Carlo sum, and the law of large numbers ensures that they agree in the limit as the number of samples n tends to infinity. In high-dimensional settings, however, these procedures for obtaining independent samples are often too difficult to be of any practical use.

Markov chain Monte Carlo (MCMC) is a method that instead samples *correlated* elements from the posterior by constructing an ergodic Markov chain whose stationary distribution is equal to $\pi(\boldsymbol{\theta}|\mathbf{x})$. As this Markov chain performs a random walk over its stationary distribution, its values $\{\boldsymbol{\theta}_i\}$ at each iteration i comprise a correlated sample from the target posterior. Despite correlation, the ergodic theorem tells us that their empirical mean converges to the theoretical mean of any integrable function f:

$$\lim_{M\to\infty} \sum_{i=1}^{M} f(\boldsymbol{\theta}_i) \to_p \mathrm{E}_{\pi(\boldsymbol{\theta}|\mathbf{x})}[f(\boldsymbol{\theta})|\mathbf{x}].$$

The Metropolis-Hastings algorithm, arguably the most important MCMC algorithm, gives us a simple but powerful recipe to construct such a Markov chain [Hastings, 1970]. Beginning from an initial parameter value $\boldsymbol{\theta}_0$ (which for instance can be proposed from the prior), the algorithm iteratively proposes new values according to a distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)$. Whether the Markov chain will move to the proposed value $\boldsymbol{\theta}^*$ depends on an acceptance ratio

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{x})q(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_i|\mathbf{x})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}\right\} = \min\left\{1, \frac{\pi(\mathbf{x}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*)}{\pi(\mathbf{x}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}\right\}.$$

We accept the proposal and set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ with probability α; otherwise, the chain does not move and we set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$. The form of the acceptance step guarantees that the stationary distribution of this chain is exactly equal to $\pi(\boldsymbol{\theta}|\mathbf{x})$. The proposal distribution q is essentially arbitrary from this point of view, although it has important practical implications—whether the chain explores the parameter space (*mixes*) efficiently or gets stuck in one region by either rejecting too many unlikely proposals or accepting too many similar proposals can crucially depend on choice of q. Note that this scheme cleverly bypasses the need to compute the normalizing constant $\pi(\mathbf{x})$, which cancels itself out in the ratio. Finally, we remark that this technique is likelihood-based as the terms $\pi(\mathbf{x}|\boldsymbol{\theta})$ or equivalently $\exp(\ell_o(\mathbf{x};\boldsymbol{\theta}))$ appear in computing the acceptance ratio. Therefore, when the data $\mathbf{x}$ are observations of a CTMC, the same computational challenges in computing this likelihood must again be considered here. There is an entire literature devoted to developing MCMC routines bypassing likelihood evaluations via various simulation-based approaches—particle filtering or sequential Monte Carlo algorithms [Doucet et al., 2000], their related pseudo-marginal approaches [Andrieu et al., 2010], and approximate Bayesian computation (ABC) [Marjoram et al., 2003, Toni et al., 2009] have been successfully applied to fitting CTMCs to data. We will not discuss these methods in detail here, but remark that they all have their own computational limitations in systems with large populations or numbers of events per interval.

2.2.3 Auxiliary variable augmentation

We've seen that the complete-data log likelihood ℓ_c is a much more mathematically convenient expression than its discrete-data counterpart ℓ_o. This advantage motivates a data-augmented Markov chain Monte Carlo (DA-MCMC) approach [Tanner and Wong, 1987], which bases inference and computations around ℓ_c instead of ℓ_o by introducing auxiliary or latent variables $\mathbf{Z}$. These variables are unobserved, but together with the data $\mathbf{x}$ contain all the information appearing in ℓ_c, and their inclusion thereby completes or *augments* the observed data. Recall, for instance, that $\mathbf{Z} = \{\mathbf{z}, \boldsymbol{\tau}\}$ the complete set of jumps and their event times, respectively, for a CTMC. The target distribution in this case is the joint posterior

distribution over parameters and auxiliary variables

$$\pi(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{Z})}{\pi(\mathbf{x})} = \frac{\pi(\mathbf{x}|\mathbf{Z}, \boldsymbol{\theta})\pi(\mathbf{Z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x})}.$$

In this setting, the complete data likelihood $\pi(\mathbf{Z}|\boldsymbol{\theta})$ appears in place of the observed data likelihood, along with the term $\pi(\mathbf{x}|\mathbf{Z}, \boldsymbol{\theta})$. This term is an indicator function equal to 1 if the augmented data is consistent with observations: that is,

$$\pi(\mathbf{x}|\mathbf{Z}, \boldsymbol{\theta}) = \mathbf{1}_{\{\mathbf{x} \text{ agrees with } \mathbf{Z}\}} = \prod_{i=1}^{n} \mathbf{1}_{\{x_i = \mathbf{Z}(t_i)\}}.$$

This augmentation technique is not only relevant for imputing hidden events between discrete observations, but can also be used in the *partially* observed case, where the data consist of noisy observations of the process $\mathbf{X}(t)$ rather than its exact states at a set of discrete times. For instance, when $\mathbf{X}(t)$ is the latent Markov process of a hidden Markov model (HMM) [Rabiner, 1989], these probabilities $\pi(\mathbf{x}|\mathbf{Z}, \boldsymbol{\theta})$ are the *emission probabilities* of the sampling distribution instead of indicator functions. For brevity, we will not discuss in detail the case of partially observed processes here.

In the data-augmented setting, it is necessary to choose proposal distributions for parameters $q_{\boldsymbol{\theta}}$ as well as for the augmented data $q_{\mathbf{Z}}$. Although auxiliary variables $\mathbf{Z}$ enable simpler likelihood computations, the difficulty now translates to designing a good way to propose them efficiently and consistently with the observed data—an effective choice of $q_{\mathbf{Z}}$ is the cornerstone of a good DA-MCMC algorithm.

Proposing consistent trajectories amounts to the ability to simulate a Markov process over an interval $[0, s]$ conditional on endpoints $\mathbf{X}(0), \mathbf{X}(s)$. Crudely forward simulating from $\mathbf{X}(0)$ for time t and rejecting all those that do not agree with the right-endpoint works only for simple, low-dimensional settings and small t. In some special cases, one can derive direct sampling methods conditional on endpoints or uniformed sampling based on a discrete-time Markov chain skeleton of the CTMC [Hobolth and Stone, 2009]. After obtaining a set of

samples from the joint posterior $\{\mathbf{Z}_i, \boldsymbol{\theta}_i\}$ via the Metropolis-Hastings algorithm or any other MCMC method, obtaining samples from the marginal posterior distribution of interest $\pi(\boldsymbol{\theta}|\mathbf{x})$ is trivial: marginalizing out the distribution of auxiliary variables from the joint posterior is accomplished by simply ignoring or discarding the values $\{\mathbf{Z}_i\}$.

Chapter 3

BRANCHING PROCESSES

Having established the basics of inference for continuous-time Markov chain models, we are ready to present branching processes as a special class within this framework. We will see many analogs between branching process techniques and the tools used toward studying CTMCs, and will discuss how their additional properties allow us to revisit the computational tools discussed in the context of CTMCs from new perspectives. This enables us to develop computational and mathematical approaches that provide tractable alternatives when general-purpose CTMC techniques are impractical.

3.1 Definition and notation

In a branching process, a collection of *independently acting* individuals, or particles, can reproduce and die according to a probability distribution. We focus on the continuous-time, multi-type case. In this setting, each particle type can have a distinct mean lifespan and reproductive probabilities, and lives for an exponentially distributed length of time. At time of death, a particle can give rise to particles of its own type as well as other types according to the aforementioned distribution.

Consistent with our CTMC notation, denote a linear, multi-type branching process by a random vector $\mathbf{X}(t)$ taking values in a discrete state space Ω, with entries $X_i(t)$ denoting the number of type i particles present at time $t \geq 0$. For exposition and notational simplicity, we will focus on the case where there are two types of particles. Each of the two type i particles produces k type 1 particles and l type 2 particles with *instantaneous progeny rates*

$a_i(k,l)$ upon completion of its lifespan, and the rates of no event occurring are defined as

$$\alpha_1 := a_1(1,0) = -\sum_{(k,l)\neq(1,0)} a_1(k,l), \qquad \alpha_2 := a_2(0,1) = -\sum_{(k,l)\neq(0,1)} a_2(k,l),$$

so that $\sum_{k,l} a_i(k,l) = 0$ for $i = 1, 2$ (recall the conservative assumption for CTMCs). These reproduction rates characterize the process, analogously to how the infinitesimal rates in $\mathbf{Q}$ characterize a CTMC— indeed, these rates $a_i(k,l)$ parametrize the generator $\mathbf{Q}$ of a branching process.

The assumption of particle independence implies *linearity*, the property that overall instantaneous rates are multiplicative in the total number of particles present at that instant. For example, the infinitesimal probability of jumping to k type 1 and l type 2 particles beginning with j type 1 particles over a short interval of time h is

$$\Pr(\mathbf{X}(h) = (k,l) | \mathbf{X}(0) = (j,0)) = j \cdot a_1(k,l) \cdot h + o(h).$$

Subsequently, offspring of each particle evolve according to the same set of instantaneous rates, and we assume that these rates $a_j(k,l)$ do not change over time; i.e. $a_j(k,l;t) = a_j(k,l)$ at all times $t \geq 0$—this assumption means the branching process is *time-homogeneous*. Notice that while per-particle event rates are fixed through time, the cumulative rates of any given event are inhomogeneous, as they are linear in the population sizes $\mathbf{X}(t)$ which evolves stochastically. Together, these assumptions imply that each type i particle has an exponentially distributed lifespan with rate $-\alpha_i$. Due to memorylessness of the exponential distribution in this construction, the process is Markovian, and $\mathbf{X}(t)$ evolves over time as a CTMC [Guttorp, 1995].

To concretely tie this back to CTMCs and highlight the contrasts, we will proceed with an illustrative example using the birth-death process, arguably the most prevalent of single-type branching processes. In a homogeneous, linear birth-death process, only two kinds of "branching" events are permitted: not surprisingly, births and deaths. We denote the per-

particle birth rate λ and the death rate μ; using notation analogous to the two-type branching processes $a_j(k,l)$, the instantaneous rates corresponding to possible events are given by $a(2) = \lambda, a(0) = \mu$, since a birth event corresponds to an instant at which one particle is replaced by two, and in the event of a death, the particle is replaced by zero particles (we don't use the subscripts in the above notation since this population only considers one population type). The only remaining nonzero rate is $a(1) = \sum_k a(k) = -\lambda - \mu$. This process takes values in the state space of natural numbers $\Omega = \mathbb{N}$; note 0 is an absorbing state corresponding to extinction.

Recall that every CTMC can be characterized by its infinitesimal generator $\mathbf{Q}$, which contains its instantaneous rates in the usual notation for an arbitrary CTMC. From the above specification together with the linearity assumption, we see that the process has instantaneous rates

$$q_{ij} = \begin{cases} i\lambda & j = i+1, \\ i\mu & j = i-1, \\ -i(\mu+\lambda) & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

As Ω is unbounded for the linear birth-death process, $\mathbf{Q}$ is an infinite matrix:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 2\mu & -2(\lambda+\mu) & 2\lambda & 0 & 0 & \cdots \\ 0 & 0 & 3\mu & -3(\lambda+\mu) & 3\lambda & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \end{pmatrix}.$$

The prescient reader may already feel wary of specifying a relatively simple process with such an unwieldy mathematical object—results relying on matrix exponentiation cannot be applied as there is effectively an infinity of equalities at play due to the Kolmogorov forward or backward system. This will generally be the case when the components $X_i(t)$ track the

size of a population of type i species that is allowed to grow without a specified limit; the multi-type case is only more complicated from this perspective as $\mathbf{Q}$ is less likely to have such a transparent structure. In contrast, viewing the birth-death process as a branching process only requires specifying two parameters $a(0)$ and $a(2)$, and we will see that this infinite set of equations is "wrapped up" and translated into an infinite sum via generating functions, which can be shown to satisfy a now *finite* system of Kolmogorov equations.

3.2 Generating functions of branching processes

While the theory of branching processes draws from many disparate branches of mathematics, the use of generating functions is perhaps most central. The probability generating function of a branching process is intricately related to its fundamental probabilistic characteristics, and in many ways serves a similar role in characterizing the process as the infinitesimal generator $\mathbf{Q}$ characterizes an arbitrary CTMC. Much like the the connection between $\mathbf{Q}$ and the probabilistic evolution of a CTMC via matrix exponentiation, the relationships between the properties of a branching process and its various generating functions also arise from the Kolmogorov equations.

Denoting the components of a two-type process $\mathbf{X}(t) = \{X_1(t), X_2(t)\}$, its *probability generating function* (PGF) is defined as

$$\phi_{jk}(t, s_1, s_2) := \mathrm{E}\left(s_1^{X_1(t)} s_2^{X_2(t)} \mid X_1(0) = j, X_2(0) = k\right) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{(j,k),(l,m)}(t) s_1^l s_2^m.$$

Immediately, we notice that the finite-time transition probabilities appear in this expression. Indeed, they are related to the PGF via partial differentiation:

$$p_{(jk),(lm)}(t) = \frac{1}{l!\,m!}\frac{\partial^l}{\partial s_1}\frac{\partial^m}{\partial s_2}\phi_{jk}(t)\bigg|_{s_1=s_2=0}.$$

Furthermore, the extinction probability of the process at time t, beginning with j type 1 and k type 2 particles, is given by $\phi_{jk}(t, \mathbf{0})$. Moments can also be obtained in a similar fashion

from the PGF: for example, the mean behavior satisfies

$$\mathrm{E}[X_i(t)|\mathbf{X}(0) = (j,k)] = \frac{\partial}{\partial s_i}\phi_{jk}(t,\theta,s_1,s_2)|_{s_1=s_2=1},$$

although it may be easier to derive such quantities by making use of alternative generating functions. For instance, substituting $s_1 = e^x, s_2 = e^y$ yields the corresponding moment generating function

$$M_{jk}(t,x,y) = \phi_{jk}(t,e^x,e^y) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{(j,k),(l,m)}(t)e^{lx+my}$$

and the cumulant generating function is given by

$$K_{jk}(t,x,y) = \log M_{jk}(t,x,y).$$

Expanding these generating functions in powers of x, y provides their more familiar definitions in terms of joint moments or cumulants of the distribution:

$$M(t,x,y) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} \mu_{j,k}^{(l,m)}(t)\frac{x^l y^m}{l!\,m!},$$
$$K(t,x,y) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} \kappa_{j,k}^{(l,m)}(t)\frac{x^l y^m}{l!\,m!},$$

where the coefficients now correspond to the (l,m)th cross-moments or cross-cumulants of a process beginning at $\mathbf{X}(0) = (j,k)$. Depending on the desired quantities and specific process at hand, some generating functions may be more useful or readily solvable than others; see [Bailey, 1964] for details. One can often additionally construct and manipulate similar generating functions, for instance governing the total number of particles across different types, enabling the consideration of a number of system effects, such as immigration. Similar relationships to those mentioned above therefore extend to many settings, enabling us to consider their use toward arbitrary marginal distributions, their products, and their

coefficients—the excellent review by Dorman et al. [2004] provides further details.

These powerful properties hold generally, and the Kolmogorov equations governing such generating functions are usually straightforward to write down, typically appearing as partial differential equations (PDEs) or systems of ordinary differential equations (ODEs). The choice whether to use the forward or backward version of the Kolmogorov equations largely depends on the problem, and may make all the difference between welcoming familiar forms and encountering wildly intractable expressions. As alluded to previously, there is also an art to choosing or devising the proper generating function for the task. In almost all cases, the difficulty arises in obtaining closed form solutions, as the resulting systems are typically nonlinear and impossible to solve analytically. The challenges we face are therefore largely computational and numerical in nature.

3.2.1 Deriving their relation to Kolmogorov equations

That the PGF of a branching process satisfies the Kolmogorov equations is a classical result; we derive the backward equations for the two-type case here. Begin by introducing the *pseudo-generating functions*, defined

$$u_i(s_1, s_2) := \sum_k \sum_l a_i(k,l) s_1^k s_2^l$$

for $i = 1, 2$, and notice their relationship to the probability generating function:

$$\begin{aligned}
\phi_1(t, s_1, s_2) &= E\left[s_1^{X_1(t)} s_2^{X_2(t)} \mid X_1(0) = 1, X_2(0) = 0 \right] \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} p_{(1,0),(k,l)}(t) s_1^k s_2^l \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \left[\mathbf{1}_{k=1,l=0} + a_1(k,l)t + o(t) \right] s_1^k s_2^l \\
&= s_1 + u_1(s_1, s_2)t + o(t). \qquad (3.1)
\end{aligned}$$

An analogous expression for $\phi_2(t, s_1, s_2)$ is obtained similarly. We see from (3.1) that ϕ_i and u_i satisfy

$$\frac{d\phi_1(t, s_1, s_2)}{dt}\Big|_{t=0} = u_1(s_1, s_2), \qquad \frac{d\phi_2(t, s_1, s_2)}{dt}\Big|_{t=0} = u_2(s_1, s_2).$$

It can be straightforwardly shown that the particle independence assumption yields a convenient factorization of the PGF $\phi_{i,j} = \phi_1^i \phi_2^j$. Therefore, it suffices to work with only ϕ_1, ϕ_2 throughout.

Now, we show that the generating functions ϕ_1, ϕ_2 themselves satisfy a Chapman-Kolmogorov relation:

$$\begin{aligned}
\phi_1(t+h, s_1, s_2) &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} p_{(1,0),(k,l)}(t+h) s_1^k s_2^l \\
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_{(1,0),(i,j)}(t) p_{(i,j),(k,l)}(h)\right] s_1^k s_2^l \quad \text{(by Chapman-Kolmogorov eqn.)} \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_{(1,0),(i,j)}(t)\left[\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} p_{(i,j),(k,l)}(h) s_1^k s_2^l\right] \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_{(1,0),(i,j)}(t)\phi_{ij}(h, s_1, s_2) \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_{(1,0),(i,j)}(t)\phi_1(h, s_1, s_2)^i \phi_2(h, s_1, s_2)^j \quad \text{(by particle independence)} \\
&= \phi_1(t, \phi_1(h, s_1, s_2), \phi_2(h, s_1, s_2)).
\end{aligned}$$

This form is useful for deriving the forward equation for ϕ_1; because the argument was symmetric in t and h, exchanging their roles allows the last equality to equivalently be written instead as

$$\begin{aligned}
\phi_1(t+h, s_1, s_2) &= \phi_1(t, \phi_1(h, s_1, s_2), \phi_2(h, s_1, s_2)) \\
&= \phi_1(h, \phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2)).
\end{aligned}$$

Now, to derive the backward equations, we begin by expanding $\phi_1(t+h, s_1, s_2)$ around t and

apply the above relation:

$$\begin{aligned}\phi_1(t+h,s_1,s_2) &= \phi_1(t,s_1,s_2) + \frac{d\phi_1(t+h,s_1,s_2)}{dh}\Big|_{h=0} h + o(h) \\ &= \phi_1(t,s_1,s_2) + \frac{d\phi_1(h,\phi_1(t,s_1,s_2),\phi_2(t,s_1,s_2))}{dh}\Big|_{h=0} h + o(h) \\ &= \phi_1(t,s_1,s_2) + u_1(\phi_1(t,s_1,s_2),\phi_2(t,s_1,s_2))h + o(h).\end{aligned}$$

Since an exactly analogous argument applies for ϕ_2, we arrive at the system

$$\begin{cases} \frac{d}{dt}\phi_1(t,s_1,s_2) = u_1(\phi_1(t,s_1,s_2),\phi_2(t,s_1,s_2)), \\ \frac{d}{dt}\phi_2(t,s_1,s_2) = u_2(\phi_1(t,s_1,s_2),\phi_2(t,s_1,s_2)), \end{cases}$$

subject to initial conditions

$$\phi_1(0,s_1,s_2) = s_1, \quad \phi_2(0,s_1,s_2) = s_2.$$

A similar argument leads to the forward equations, which we do not discuss in detail here as the following chapters will make use of backward equation arguments.

The forms of the pseudo-generating functions u_i can immediately be written down given the set of non-zero instantaneous progeny rates $a_i(k,l)$ that specify the branching process. For a concrete example, we revisit the linear birth-death process and consider the analogous single-type notation: its pseudo-generating function is

$$u(s) = \sum_k a(k)s^k = \mu + (-\lambda-\mu)s + \lambda s^2 = (\lambda s - \mu)(s-1).$$

The PGF of the birth-death process thus satisfies the Kolmogorov backward equation

$$\frac{d}{dt}\phi(t,s) = [\lambda\phi(t,s) - \mu][\phi(t,s) - 1].$$

In the multi-type case, we would instead arrive at a system of m ODEs equations if m is the

number of particle types. The PGF also satisfies the forward equation: in this example,

$$\frac{\partial}{\partial t}\phi(t,s) = (\lambda s - \mu)(s-1)\frac{\partial}{\partial s}\phi(t,s).$$

Notice the forward equation yields a PDE: in the multi-type case, we typically obtain a single but more complicated PDE.

3.3 Computing coefficients of a generating function

We have seen that the PGF of a branching process satisfies the Kolmogorov equations, and thus transition probabilities can be obtained by computing its coefficients as an alternative to matrix exponentiation. We've also seen that the coefficients of the PGF and of any arbitrary generating function can be expressed via partial differentiation of the generating function. However, it is rare that the PGF admits an analytic solution, and taking its higher order partials is not necessarily straightforward even in the few cases that it does. Instead, the PDE or ODE systems are commonly evaluated numerically via standard stepwise methods, e.g. Runge-Kutta methods [Butcher, 1987]. All such methods are prone to a degree of discretization error, but this error is typically minor enough that numerical solutions are considered reliable in all but very long time intervals t or in pathological, "stiff" systems.

While this gives us a way to numerically evaluate the PGF $\phi_{jk}(t, \mathbf{s})$ at any arguments $t, \mathbf{s}$, we still need to compute its partial derivatives. In contrast with the discretization error arising from solving an ODE once, the accumulation of such errors combined with those resulting from numerical partial differentiation *do* cause concern. In many cases, we need to take hundreds or thousands of numerical derivatives in order to recover a single coefficient, often at each step within an iterative algorithm. In addition to runtime considerations, repeatedly computing numerical derivatives quickly compounds error, leading to imprecision and instability.

Instead, Lange [1982] provides a clever alternative that pulls these coefficients out with the finite Fourier transform (FFT), making use of integration rather than differentiation.

The technique is generic to univariate generating functions $\phi(s) = \sum_{k=0}^{\infty} c_k s^k$ corresponding to a discrete random variable X, where each coefficient $c_k = \Pr(X = k)$, and was motivated by targeting coefficients of a generating function that form the equilibrium distribution of a deleterious gene.

Begin by considering a change of variables $s = e^{2\pi i w}$, so that the domain becomes an equally spaced set of points along the unit circle in the complex plane. The generating function in turn becomes a periodic function

$$\phi(e^{2\pi i w}) = \sum_{k=0}^{\infty} c_k e^{2\pi i w \cdot k}.$$

From this point of view, the coefficients of interest c_k are now simply the kth Fourier coefficients of a Fourier series, and hence we may recover them by the Fourier inversion formula:

$$c_k = \int_0^1 \phi(e^{2\pi i w}) e^{-2\pi i w k} \, \mathrm{d}t.$$

Practically, this integral can now be evaluated using any standard discrete sum approximation to integration: for instance, as a Riemann sum

$$c_k \approx \frac{1}{N-1} \sum_{j=0}^{N-1} \phi(e^{2\pi i j/N}) e^{-2\pi i j k/N} \tag{3.2}$$

with larger choice of N yielding a finer grid and thus more precise approximation. Recall that numerical solutions to the generating function ϕ can reliably be computed using standard numerical differential equation solvers, and under this transformation, the problematic higher-order partial derivatives no longer appear in expressing the coefficients of ϕ. In particular, the FFT provides a very fast way to compute *all* such coefficients c_k for $0 \leq k \leq N-1$ simultaneously, given solutions to ϕ at the complex arguments appearing in equation (3.2).

3.3.1 Connection to transition probabilities of a branching process

Recent work by Doss et al. [2013] adapts this method to analyze linear birth-death processes with immigration. The authors seek to develop an Expectation Maximization (EM) algorithm to fit such models to *panel data*, settings where the data are assumed to be generated from multiple independent, continuous-time processes observed at discrete and possibly irregularly spaced time points, whose rates can be a function of many process-specific covariates. The infinite state space of the process prohibits previously considered EM algorithms and methods for fitting CTMCs. Instead, Doss et al. [2013] construct a joint generating function $H(t, \mathbf{s})$ whose partial derivatives $\frac{\partial}{\partial s_i} H(t, \mathbf{s})$ are themselves power series with meaningful coefficients. These coefficients correspond to transition probabilities and other necessary quantities appearing in E-step calculations of an EM algorithm, and can be extracted using the aforementioned trick of Lange [1982]. Without going into detail, the authors tie this idea with the original solution to the forward PDE governing the PGF of a linear birth-death process [Kendall, 1948], generalizing Kendall's argument to show that the joint generating function H_i too satisfies a similar quasi-linear forward PDE. The resulting equation takes the form of a Cauchy problem, and the authors derive an analytic solution using the method of characteristics.

Extending this solution technique to the multi-type setting is nontrivial, and it is not clear whether a solution to the multivariate version of Kendall's PDE exists. However, these ideas and their use toward computing transition probabilities lay a foundation for the material in the following chapter, dedicated to generalizing the methods to the case of discretely observed, multi-type branching processes. We show that this framework does nicely apply to multivariate generating functions. In addition to easily computing transition probabilities, we present a set of new generating functions that analogously admit related quantities such as restricted moments and particle dwell times, sufficient statistics of the complete likelihood. We derive a computationally tractable solution to their system of backward Kolmogorov equations, leading to the development of a novel EM algorithm applicable to panel data. This

allows for robust and efficient maximum likelihood inference in high dimensional parameter settings when process rates can be parametrized as functions of multiple covariates, and is empirically demonstrated to be preferable over direct maximization of the observed-data likelihood.

Chapter 4

SPECTRAL TECHNIQUES AND EM ALGORITHM FOR LIKELIHOOD-BASED INFERENCE IN PARTIALLY OBSERVED BIRTH-DEATH-SHIFT MODELS

Originally introduced as a mathematical model for the survival of family surnames, the tools from branching process theory have since found a breadth of applications including biology, genetics, epidemiology, quantum optics, and nuclear fission [Jagers, 1975, Guttorp, 1991, Kimmel and Axelrod, 2002, Renshaw, 2011]. One of the most widely used classes of branching processes are birth-death (BD) processes, a simple yet flexible model for single-species population dynamics. The popularity of BD processes is in part attributable to their well-understood mathematical properties. To accurately model behavior in many applications, however, it is often necessary to consider systems with more than one species — bivariate or other multi-type processes are commonly used to model phenomena such as competition, predation, or infection [Neyman et al., 1956, Alsmeyer, 1993, Renshaw, 2011]. Multi-type branching processes form one class of models that can accommodate populations with multiple types, but pose considerable computational challenges for statistical inference. We introduce new methods to overcome these challenges, enabling likelihood-based inference in partially observed, multi-type branching processes.

Many statistically relevant quantities are available in closed form for the linear, homogeneous BD process and several of its variants, including transition probabilities, stationary distributions, and moments [Bailey, 1964, Keiding, 1975]. Further, analytical expressions of transition probabilities as series of orthogonal polynomials are known for general, nonlinear BD processes [Karlin and McGregor, 1958] and can be conveniently computed numerically [Murphy and O'Donohoe, 1975, Crawford and Suchard, 2012]. The ability to compute finite-

time transition probabilities enables likelihood-based inference for partially observed BD processes, since the observed likelihood is a function of these transition probabilities. Recent work by Doss et al. [2013] and Crawford et al. [2014] introduces techniques to additionally compute conditional moments of BD sufficient statistics for linear and general birth-death-immigration processes, enabling calculation of the expected complete-data likelihood necessary in an expectation-maximization (EM) algorithm [Dempster et al., 1977].

Unfortunately, methods to evaluate finite-time transition probabilities and conditional moments are not known in the multi-type setting, and generalizing the techniques available in the single-species case is nontrivial. Solutions to the Kolmogorov equations in multi-type settings are available only for several linear, closed systems such as the immigration-death-shift process, but simple modifications such as the presence of birth events significantly complicate analysis [Puri, 1968, Renshaw, 2011]. Without these quantities, likelihood-based estimation is limited to simulation-based inference via Monte Carlo EM or MCMC [Golinelli, 2000, Golinelli et al., 2006] and asymptotic approximations, such as moment-based estimating equations [Catlin et al., 2001]. However, these approaches have shortcomings. MCMC approaches require augmenting the state space by high-dimensional latent variables and become computationally prohibitive when the state space is large. Moment-based methods are statistically less efficient than likelihood-based approaches and thus often inappropriate for smaller datasets, requiring a large number of observations to produce meaningful standard errors and confidence intervals.

Here we extend the analysis of Doss et al. [2013], deriving previously unavailable numerical solutions to transition probabilities and conditional moments for discretely observed, multi-type branching processes. Modifying ideas introduced by Kendall [1948], we simplify the systems of backward equations for several relevant generating functions, and then apply the spectral approach of Lange [1982] to extract expected sufficient statistics and transition probabilities. This enables us to evaluate the observed likelihood, as well as to reduce the challenging computation of expected complete-data log-likelihood necessary in an EM algorithm to efficient evaluation of expected sufficient statistics by low-dimensional integration.

Our EM algorithm can be applied in settings where the data are assumed to be generated from independent, continuous-time multi-type branching processes, observed at discrete and possibly irregularly spaced time points, whose rates can be a function of many process-specific covariates. Medical applications, for instance, commonly feature such panel data, where rates corresponding to the transmission of a disease or growth of a cell may depend on patient-specific characteristics. While similar methods have been explored for fitting continuous-time finite state-space Markov chains to panel data [Jackson, 2011, Kalbfleisch and Lawless, 1985, Lange, 1995, Lange and Minin, 2013], our method allows for multivariate and potentially infinite state-space processes.

Though our methodology applies broadly to any linear multi-type branching process, we focus attention to estimating the rates of a birth-death-shift (BDS) process. The BDS process adds the possibility of shift events to the standard BD framework, and is useful for modeling systems that allow for elements to switch locations or types — a shift is essentially a simultaneous birth and death. For example, in epidemiological applications, interaction between infected and susceptible populations can be captured as a shift event, involving a simultaneous increase and decrease in the respective populations. Spatial BDS processes have also been studied to improve Metropolis-Hastings algorithms for perfect sampling [Huber, 2012] relevant to a range of spatial statistical applications; see Illian et al. [2008] for an overview. Our motivation stems from the BDS process proposed by Rosenberg et al. [2003] to model evolution of transposons — mobile genetic elements that can replicate, die, or shift locations along the genome. Specifically, Rosenberg et al. [2003] study the within-host evolution of the IS*6110* transposon in the *Mycobacterium tuberculosis* genome via a BDS model. Accurately estimating the rates of these events is important in molecular epidemiology [Tanaka and Rosenberg, 2001].

Rosenberg et al. [2003] infer the birth, death, and shift rates of the BDS model of IS*6110* evolution from an ongoing database of *M. tuberculosis* patients from San Francisco, but their rate estimates rely on approximate maximum likelihood estimators (MLEs) under a rigid assumption that at most one event occurs per observation interval. This study was

revisited in [Doss et al., 2013], in which the authors derive an EM algorithm for inference in discretely observed birth-death-immigration (BDI) processes amenable to high-dimensional optimization. Although this approach more realistically allows multiple events to occur per interval, the methodology in [Doss et al., 2013] is limited to the single-species setting, effectively replacing the BDS model with a simpler model that ignores particle locations and shift events. Our methodology extends the analysis of Doss et al. [2013], allowing for both the possibility of multiple events per observation interval as well as the consideration of shift events. We show that the dynamics of the BDS model can be captured in a two-type branching process framework in that transition probabilities of both processes are nearly identical. We then derive an EM algorithm for discretely observed multi-type branching processes, and rigorously assess its performance in several simulation studies. Finally, we revisit the San Francisco tuberculosis dataset, applying our algorithm to estimate rates of the IS*6110* transposon as a function of relevant covariates.

4.1 Birth-death-shift model for transposable elements

The birth-death-shift process proposed by Rosenberg et al. [2003] has been used to model evolutionary dynamics of transposable elements or *transposons*, genomic mobile sequence elements. Each transposon can (1) duplicate, with the new copy moving to a new genomic location; (2) shift to a different genomic position; or (3) be removed and lost from the genome, independently of all other transposons. These events occur at instantaneous rates proportional to the total transposon copy number at that time. Thus, transposons evolve according to a linear birth-death-shift (BDS) process in continuous time.

The process of transposon evolution within a host is observable by serially genotyping the organism of interest, e.g., *Mycobacterium tuberculosis* as in Rosenberg et al. [2003]. *M. tuberculosis* genome typically has between 0 and 25 copies of the IS6110 element. The number and chromosomal position of the IS*6110* element can be visualized using restriction fragment length polymorphism (RFLP). This technique entails restriction endonuclease digestion of the *M. tuberculosis* DNA which is run in an agarose gel, southern blotting and probing with

a peroxidase labeled IS*6110* probe. Birth, death, and shift events are thus detectable via changes in the number and size of the bands where the IS6110 elements are located.

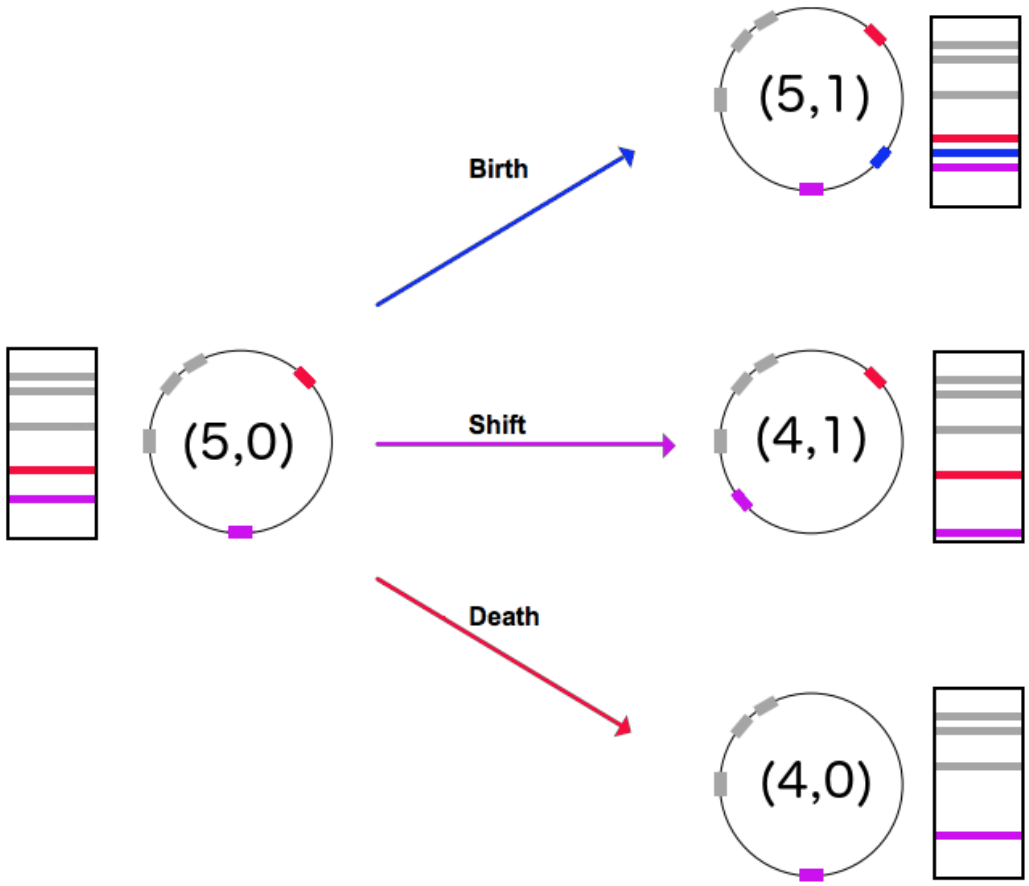


Figure 4.1: Illustration of the three types of transposition—birth, death, shift—along a genome, represented by circles. Transposons, depicted by filled rectangles along the circles/genomes, correspond to observable gel bands, denoted by horizontal lines in the rectangles next to each circle diagram. Numbers within each circle represent each configuration $\mathbf{X}(t)$ in the notation introduced in section 4.1.2. More specifically, we call the gel band on the left our initial configuration and set the number of particles of type 1 to the number of bands, 5, and the number of particles of type 2 to 0. On the right set of diagrams, a birth event keeps the number of type 1 particles intact and increments the number of type 2 particles by one, a death event changes the number of type 1 particles from five to four and keeps the number of type 2 particles at zero, and finally a shift event decreases the number of type 1 particles by one and increases the number of type 2 particles by one.

Estimating the rates based on observed changes at genotyping times in this experimental setup corresponds to inference in a discretely observed linear BDS process. That is, we assume each element behaves independently, and that overall rates of each event are proportional to total copy number k. Together with the time-homogeneity assumption, waiting times until occurrence of an event are distributed exponentially with rate $k\theta$, where $\theta = \lambda + \mu + \nu$. When an event occurs, the probability that it is birth, death, or shift is given by $\lambda/\theta, \nu/\theta$, and μ/θ respectively. The BDS process is therefore a continuous-time Markov chain (CTMC).

The states in our process $\widetilde{\mathbf{x}} \in \{0,1\}^S := \widetilde{\Omega}$ can be represented as binary vectors, where S is the number of possible locations transposons may occupy along the genome, 0's denote unoccupied sites, and 1's correspond to sites occupied by a transposon. Now, denote the $2^S \times 2^S$ rate matrix or infinitesimal generator corresponding to this CTMC as $\mathbf{Q} = \{q_{\widetilde{\mathbf{x}}_1,\widetilde{\mathbf{x}}_2}\}$, where $q_{\widetilde{\mathbf{x}}_1,\widetilde{\mathbf{x}}_2}$ denotes the instantaneous rate of jumping to $\widetilde{\mathbf{x}}_2$ beginning from $\widetilde{\mathbf{x}}_1$ with $\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2 \in \widetilde{\Omega}$. To write down the entries of $\mathbf{Q}$, first define $C^+(\widetilde{\mathbf{x}})$ as the set of all configurations with one additional site occupied relative to $\widetilde{\mathbf{x}}$. Thus, $C^+(\widetilde{\mathbf{x}})$ contains states corresponding to one birth event beginning with $\widetilde{\mathbf{x}}$. Similarly $C^{\rightarrow}(\widetilde{\mathbf{x}})$ contains states where one additional site is occupied and one originally occupied site is no longer occupied, and $C^-(\widetilde{\mathbf{x}})$ contains states where one originally occupied site in $\widetilde{\mathbf{x}}$ is no longer occupied. Then $|C^+(\widetilde{\mathbf{x}}_1)| = S - k$, $|C^-(\widetilde{\mathbf{x}}_1)| = k$, and $|C^{\rightarrow}(\widetilde{\mathbf{x}}_1)| = |C^+(\widetilde{\mathbf{x}}_1)| \times |C^-(\widetilde{\mathbf{x}}_1)|$, and finally the entries of the generator $\mathbf{Q}$ are given by

$$q_{\widetilde{\mathbf{x}}_1,\widetilde{\mathbf{x}}_2} = \frac{\lambda}{|C^+(\widetilde{\mathbf{x}}_1)|}\mathbf{1}_{\{\widetilde{\mathbf{x}}_2 \in C^+(\widetilde{\mathbf{x}}_1)\}} + \frac{\nu}{|C^{\rightarrow}(\widetilde{\mathbf{x}}_1)|}\mathbf{1}_{\{\widetilde{\mathbf{x}}_2 \in C^{\rightarrow}(\widetilde{\mathbf{x}}_1)\}} + \frac{\mu}{|C^-(\widetilde{\mathbf{x}}_1)|}\mathbf{1}_{\{\widetilde{\mathbf{x}}_2 \in C^-(\widetilde{\mathbf{x}}_1)\}}. \tag{4.1}$$

4.1.1 BDS process with covariates

We are interested in inference when the data consist of m independent processes $\left\{\widetilde{\mathbf{X}}^p(t)\right\}$, $p = 1, \ldots, m$, each discretely observed at times $0 = t_{p,0} < t_{p,1} < \ldots < t_{p,n(p)}$. We assume each $\{\widetilde{\mathbf{X}}^p(t)\}$ process evolves according to a linear BDS model with per-particle instantaneous birth rate $\lambda_p \geq 0$, shift rate $\nu_p \geq 0$, and death rate $\mu_p \geq 0$. The data, observations from each

process, are points in the previously defined state space, with $\widetilde{\mathbf{X}}^p(t) \in \widetilde{\Omega}$ for any fixed p and t. For example, in transposon evolution, each patient p is genotyped at $n(p)+1$ observation times, and at each given time, the 1's present in the data vector correspond to locations in the gel currently occupied by transposons. The observed data corresponding to a given process $\{\widetilde{\mathbf{X}}^p(t)\}$ can thus be collected in a $S \times [n(p)+1]$ matrix with columns corresponding to observation times, and the full observed dataset can be collected into a $S \times \sum_{p=1}^{m}[n(p)+1]$ matrix: $\mathbf{Y} = (\widetilde{\mathbf{X}}^1(t_{1,0}), \ldots, \widetilde{\mathbf{X}}^1(t_{1,n(1)}), \ldots, \widetilde{\mathbf{X}}^m(t_{m,0}), \ldots, \widetilde{\mathbf{X}}^m(t_{m,n(m)}))$.

The rates of each process are determined by a vector of c covariates $\mathbf{z}_p = (z_{p,1}, z_{p,2}, \ldots, z_{p,c}) \in \mathbb{R}^c$ through a log-linear model:

$$\log(\lambda_p) = \boldsymbol{\beta}^{\lambda} \cdot \mathbf{z_p}, \quad \log(\nu_p) = \boldsymbol{\beta}^{\nu} \cdot \mathbf{z_p}, \quad \log(\mu_p) = \boldsymbol{\beta}^{\mu} \cdot \mathbf{z_p}, \tag{4.2}$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}^{\lambda}, \boldsymbol{\beta}^{\nu}, \boldsymbol{\beta}^{\mu})$ are the regression coefficients and $\cdot$ represents a vector product. For instance, in an epidemiological study, these covariates may contain patient-specific disease process and demographic information.

The observed data log-likelihood is obtained by summing over transition terms of observations for each process, and summing over all processes:

$$\widetilde{\ell}_o(\mathbf{Y}; \boldsymbol{\beta}) = \sum_{p=1}^{m} \sum_{j=0}^{n(p)-1} \log \widetilde{p}_{\widetilde{\mathbf{X}}^p(t_{p,j}), \widetilde{\mathbf{X}}^p(t_{p,j+1})}(t_{p,j+1} - t_{p,j}; \lambda_p, \nu_p, \mu_p), \tag{4.3}$$

where $\widetilde{p}_{\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2}(t; \lambda, \nu, \mu) = \Pr_{\lambda,\nu,\mu}(\widetilde{\mathbf{X}}(t) = \widetilde{\mathbf{x}}_2 \mid \widetilde{\mathbf{X}}(0) = \widetilde{\mathbf{x}}_1)$ denotes a transition probability of the BDS process. We are interested in computing the maximum likelihood estimates (MLEs) of parameters $\boldsymbol{\beta}$ of the BDS process. Notice that if the transition probabilities were available for given λ, ν, μ, and t values, one could maximize the likelihood in (4.3) using standard off-the-shelf optimization procedures. However, due to the large state space of all possible configurations of occupied sites, analysis of these transition probabilities is intractable. To approximate the BDS model likelihood above, we introduce a two-type branching process such that computationally tractable transition probabilities of this process

are numerically close to the transition probabilities of the BDS model over any observation interval. The following sections detail the correspondence between the BDS model and the two-type branching process, and develops methodology for inference in the branching process framework.

4.1.2 Reducing the state space

The size of the original state space $|\widetilde{\Omega}| = 2^S$ quickly becomes unmanageable as S grows so that analysis using the rate matrix defined in (4.1) becomes unwieldy for all but small values of S. Previous work by Doss et al. [2013] addresses this issue by collapsing the state space to one dimension, distilling the data to counts of copy numbers at each observation time. In this simplified setting, they develop tools for inference in a discretely observed birth-death-immigration framework. However, this approximate model ignores particle shifts which do not affect the total copy number, rendering the shift rate unidentifiable. Further, collapsing the state space in this way violates the Markov assumption in the BDS model. In particular, waiting times between birth and death events are exponentially distributed under the model in Doss et al. [2013], but under the BDS model with shift events, the waiting time between a birth and death no longer follows an exponential distribution.

Instead of ignoring shifts, we propose a reduction of the state space into a two-dimensional representation $\Omega \in \mathbb{N} \times \mathbb{N}$. Elements of this reduced space are pairs $\mathbf{X}(t) = (x_{old}, x_{new}) \in \Omega$ tracking the number of originally occupied and newly occupied sites at the end of each observation interval. As an example, assume six particles are present initially at time t_0, and a shift and a birth occur before the first observation t_1, and a death occurs before a second observation at t_2. When considering the first observation interval $[t_0, t_1)$, we have $\{\mathbf{X}(t_0) = (6, 0), \mathbf{X}(t_1) = (5, 2)\}$. When computing the next transition probability over $[t_1, t_2)$, we now have $\{\mathbf{X}(t_1) = (7, 0), \mathbf{X}(t_2) = (6, 0)\}$, since all seven of the particles at t_1, now the left endpoint of the observation interval, now become the initial population. This seemingly inconsistent definition of the state at $\mathbf{X}(t_1)$ is not a problem: we will see that all necessary computations occur separately on disjoint intervals, so that our reduced representation of

the original process needs only to be defined consistently for any given pair of consecutive observations.

Formally, this state space transformation is a mapping $\psi : \widetilde{\Omega} \times \widetilde{\Omega} \to \Omega \times \Omega$ on consecutive pairs of observations in $\widetilde{\Omega}$ to the reduced state space that can be computed $\psi : \left\{\widetilde{\mathbf{X}}(t_1), \widetilde{\mathbf{X}}(t_2)\right\} \mapsto \{(a,0),(b,c)\} = \{\mathbf{X}(t_1), \mathbf{X}(t_2)\}$, where $a = \sum_{j=1}^{S} \widetilde{X}_j(t_1)$ is the total number of initially occupied sites in $\widetilde{\mathbf{X}}(t_1)$, $b = \sum_{j=1}^{S} \mathbf{1}_{\{\widetilde{X}_j(t_2)=\widetilde{X}_j(t_1)\}} \widetilde{X}_j(t_1)$ is the number of initially occupied sites that remain occupied, and $c = \sum_{j=1}^{S} \mathbf{1}_{\{\widetilde{X}_j(t_2)-\widetilde{X}_j(t_1)=1\}}$ is the number of newly occupied sites in $\widetilde{\mathbf{X}}(t_2)$ not present in $\widetilde{\mathbf{X}}(t_1)$.

Note that while ψ significantly reduces the size of the state space, the mapping discards information about specific particle locations, which is uninformative to inferring birth, death, and shift rates due to symmetry induced by particle independence. The number of changes in locations between observations — the data relevant to our estimation task — is preserved in the image of ψ.

4.2 Methodology

4.2.1 Modeling via two-type branching process

Working now in the space Ω, we can treat x_{old} and x_{new} as particle types in a two-type branching process. Let $a_j(k,l)$ be the rate of producing k type 1 particles and l type 2 particles, beginning with one type j particle, $j = 1, 2$. Then the nonzero rates defining the two-type branching process corresponding to the birth-death-shift model are given by

$$\begin{aligned} &a_1(1,1) = \lambda, \quad a_1(0,1) = \nu, \quad a_1(0,0) = \mu, \quad a_1(1,0) = -(\lambda+\nu+\mu), \\ &a_2(0,2) = \lambda, \quad a_2(0,1) = -(\lambda+\mu), \quad a_2(0,0) = \mu. \end{aligned} \tag{4.4}$$

This characterization enables us to apply a generating function approach to calculate transition probabilities of the process. Defining $X_j(t)$, the number of particles of type j at time

t, we consider the generating function

$$\phi_{jk}(t, s_1, s_2) = \mathrm{E}\left(s_1^{X_1(t)} s_2^{X_2(t)} \mid X_1(0) = j, X_2(0) = k\right) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{(j,k),(l,m)}(t) s_1^l s_2^m. \quad (4.5)$$

Using the Kolmogorov backward equations, we derive equations and a closed form solution for ϕ_{jk}, detailed in Appendix A. Although an analytical expression is available, it involves special functions that are in practice are often unstable. Instead, we simplify the backward equations so that evaluating ϕ_{jk} only requires solving a single linear ordinary differential equation, which is easily accomplished using standard Runge-Kutta methods [Butcher, 1987].

With ϕ_{jk} available, transition probabilities are related to the PGF (5.1) via differentiation:

$$p_{(jk),(lm)}(t) = \frac{\partial^l}{\partial s_1}\frac{\partial^m}{\partial s_2}\phi_{jk}(t)\bigg|_{s_1=s_2=0}. \quad (4.6)$$

This repeated differentiation is computationally intensive and numerically unstable for large l, m, but following Lange [1982], we can map the domain $s_1, s_2 \in [0,1]\times[0,1]$ to the boundary of the complex unit circle, instead setting $s_1 = e^{2\pi i w_1}, s_2 = e^{2\pi i w_2}$. The generating function becomes a Fourier series whose coefficients are the desired transition probabilities

$$\phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) = \sum_{l,m=0}^{\infty} p_{(jk),(lm)}(t) e^{2\pi i l w_1} e^{2\pi i m w_2}$$

Applying a Riemann sum approximation to the Fourier inversion formula, we can now compute the transition probabilities via integration instead of differentiation:

$$\begin{aligned} p_{(jk),(lm)}(t) &= \int_0^1\int_0^1 \phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) e^{-2\pi i l w_1} e^{-2\pi i m w_2} dw_1 dw_2 \\ &\approx \frac{1}{N^2}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) e^{-2\pi i l u/N} e^{-2\pi i m v/N}. \end{aligned} \quad (4.7)$$

Choice of a larger N leads to a finer and thus more accurate Riemann sum approximation of the integral, and also allows us to compute transition probabilities to and from a

larger total particle population of either type. The Fast Fourier transform (FFT) enables efficient computation of these coefficients [Henrici, 1979], and in our application and simulation studies, we find that a grid size as small as $N = 16$ yields accurate results. With transition probabilities available, we may closely approximate $\widetilde{\ell}_o(\mathbf{Y}; \boldsymbol{\beta})$ by the branching process likelihood

$$\ell_o(\mathbf{Y}; \boldsymbol{\beta}) = \sum_{p=1}^{m} \sum_{j=0}^{n(p)-1} \log p_{\mathbf{X}^p(t_{p,j}), \mathbf{X}^p(t_{p,j+1})}(t_{p,j+1} - t_{p,j}; \lambda_p, \nu_p, \mu_p), \tag{4.8}$$

so that maximizing the observed likelihood in (4.3) gives approximately the same parameter estimates as maximizing (4.8).

4.2.2 EM algorithm for the BDS process

With transition probabilities of the process available, it is already possible to produce MLEs of the covariate effects associated with birth, death, and shift rates by numerical maximization of the observed likelihood. However, an EM algorithm approach often outperforms off-the-shelf optimization procedures in missing data problems, offering a significantly faster and more robust solution. Let $\ell_c(\mathbf{X}, \boldsymbol{\beta})$ denote the complete data log-likelihood, $\mathbf{X}$ the complete data, and $\mathbf{Y}$ the available observations. The EM algorithm begins with an initial parameter estimate $\boldsymbol{\beta}_0$, and then at each j^{th} iteration, updates the estimate by setting

$$\boldsymbol{\beta}_j = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \, \mathrm{E}_{\boldsymbol{\beta}_{j-1}} \left[\ell_c(\mathbf{X}, \boldsymbol{\beta}) \mid \mathbf{Y} \right]. \tag{4.9}$$

Each iteration involves a computation of the expectation term called the *E-step*, followed by a maximization of the expectation called the *M-step*.

E-step

The fully observed BDS process is a continuous-time Markov chain, so its complete-data log-likelihood can be written as

$$\ell_c(\mathbf{X};\boldsymbol{\beta}) = \sum_{p=1}^{m}\left[b_p \log\lambda_p + f_p\log\nu_p + d_p\log\mu_p - (\lambda_p+\mu_p+\nu_p)\sum_{k=0}^{\infty}k\tau_p(k) + \sum_{k=0}^{\infty}\log\tau_p(k)\right], \tag{4.10}$$

where $\tau_p(k)$ is the total time process $\mathbf{X}^p(t)$ spends with total copy number $X_1^p(t)+X_2^p(t)=k$, b_p is the total number of births, f_p the number of shifts, and d_p the number of deaths for each patient $p=1,\ldots,m$ — these quantities are the complete data sufficient statistics [Guttorp, 1995]. Notice the final term in (4.10) is constant with respect to the parameters. We see that in order to obtain the expected complete-data log-likelihood, we need to calculate only expected births — $\mathrm{E}_{\boldsymbol{\beta}}[b_p \mid \mathbf{Y}]$, shifts — $\mathrm{E}_{\boldsymbol{\beta}}[f_p \mid \mathbf{Y}]$, deaths — $\mathrm{E}_{\boldsymbol{\beta}}[d_p \mid \mathbf{Y}]$, and particle time — $\mathrm{E}_{\boldsymbol{\beta}}[R_p \mid \mathbf{Y}]$, where the last quantity is defined as

$$\mathrm{E}_{\boldsymbol{\beta}}[R_p \mid \mathbf{Y}] := \mathrm{E}_{\boldsymbol{\beta}}\left[\int_{t_{p,0}}^{t_{p,n(p)}} X_1(s)+X_2(s)ds \mid \mathbf{Y}\right] = \mathrm{E}_{\boldsymbol{\beta}}\left[\sum_{k=0}^{\infty}k\tau_p(k) \mid \mathbf{Y}\right].$$

By independence of the p processes and linearity of expectations, each expectation breaks into sums of expectations over the observation intervals. Further, by homogeneity, it suffices to be able to calculate the quantities

$$e^{+}_{jk,lm}(t) = \mathrm{E}\left[b_{p,t} \mid \mathbf{X}^p(0)=(j,k), \mathbf{X}^p(t)=(l,m)\right],$$

$$e^{\rightarrow}_{jk,lm}(t) = \mathrm{E}\left[f_{p,t} \mid \mathbf{X}^p(0)=(j,k), \mathbf{X}^p(t)=(l,m)\right],$$

$$e^{-}_{jk,lm}(t) = \mathrm{E}\left[d_{p,t} \mid \mathbf{X}^p(0)=(j,k), \mathbf{X}^p(t)=(l,m)\right],$$

$$e^{*}_{jk,lm}(t) = \mathrm{E}\left[R_{p,t} \mid \mathbf{X}^p(0)=(j,k), \mathbf{X}^p(t)=(l,m)\right],$$

for all non-negative integers j, k, l, m. Dependence of these quantities on rates λ_p, ν_p, μ_p is suppressed in the notation here for simplicity. As noticed by Minin and Suchard [2008] and Doss et al. [2013], it is easier to work via the restricted moments

$$
\begin{aligned}
m^{+}_{jk,lm}(t) &= \mathrm{E}\left[b_{p,t}1_{\{\mathbf{X}^p(t)=lm\}} \mid \mathbf{X}^p(0) = (j,k)\right] = \sum_{n=0}^{\infty} n q^{+}_{jk,lm}(n,t), \\
m^{\rightarrow}_{jk,lm}(t) &= \mathrm{E}\left[f_{p,t}1_{\{\mathbf{X}^p(t)=lm\}} \mid \mathbf{X}^p(0) = (j,k)\right] = \sum_{n=0}^{\infty} n q^{\rightarrow}_{jk,lm}(n,t), \\
m^{-}_{jk,lm}(t) &= \mathrm{E}\left[d_{p,t}1_{\{\mathbf{X}^p(t)=lm\}} \mid \mathbf{X}^p(0) = (j,k)\right] = \sum_{n=0}^{\infty} n q^{-}_{jk,lm}(n,t), \\
m^{*}_{jk,lm}(t) &= \mathrm{E}\left[R_{p,t}1_{\{\mathbf{X}^p(t)=lm\}} \mid \mathbf{X}^p(0) = (j,k)\right] = \int_{x=0}^{\infty} x \mathrm{d} q^{*}_{jk,lm}(x,t),
\end{aligned}
$$

where

$$
\begin{aligned}
q^{*}_{jk,lm}(x,t) &= \Pr[R_{p,t} \leq x, \mathbf{X}^p(t) = (l,m) \mid \mathbf{X}^p(0) = (j,k)], \\
q^{+}_{jk,lm}(n,t) &= \Pr[b_{p,t} = n, \mathbf{X}^p(t) = (l,m) \mid \mathbf{X}^p(0) = (j,k)],
\end{aligned}
$$

and $q^{\rightarrow}, q^{-}$ are defined analogously. The conditional expectations can then be recovered after dividing by transition probabilities, i.e.

$$
e^{+}_{jk,lm}(t) = m^{+}_{jk,lm}(t)/p_{jk,lm}(t).
$$

These restricted moments can be computed with a similar approach used to obtain transition probabilities. We begin by defining the pseudo-generating functions: for expected births, let

$$
g^{+}_{jk,lm}(r,t) = \sum_{n=0}^{\infty} q^{+}_{jk,lm}(n,t) r^n.
$$

Ignoring notational dependence on individual patients for simplicity, we define the joint generating function

$$H^+_{jk}(r,s_1,s_2,t) = \mathrm{E}\left[r^{b_t}s_1^{X_1(t)}s_2^{X_2(t)} \mid \mathbf{X}(0)=(j,k)\right]$$
$$= \sum_l\sum_m\sum_n \Pr\left[b_t=n, \mathbf{X}(t)=(k,l) \mid \mathbf{X}(0)=(j,k)\right] r^n s_1^l s_2^m = \sum_l\sum_m g^+_{jk,lm}(r,t)s_1^l s_2^m.$$

Pseudo-generating functions for shifts and deaths are defined analogously, and the pseudo-generating function for particle time is defined as

$$H^*_{jk}(r,s_1,s_2,t) = \sum_l\sum_m \int_{x=0}^{\infty} e^{-rx}\mathrm{d}q^*_{jk,lm}(x,t)s_1^l s_2^m := \sum_l\sum_m V_{jk,lm}(r,t)s_1^l s_2^m,$$

where $V_{jk,lm}(r;t) = \int_0^\infty e^{-rx}\mathrm{d}q^*_{jk,lm}(x;t)$ is the Laplace-Stieltjes transform of $q^*_{jk,lm}(x;t)$. In each case we can define series whose coefficients are our quantities of interest by partial differentiation:

$$G^+_{jk}(s_1,s_2,t) = \frac{d}{dr}H^+_{jk}(r,s_1,s_2,t)\Big|_{r=1} = \sum_l\sum_m\left[\sum_n nq^+_{jk,lm}(n,t)\right]s_1^l s_2^m = \sum_l\sum_m m^+_{jk,lm}(t)s_1^l s_2^m. \tag{4.11}$$

$G^{\rightarrow}_{jk}$ and G^-_{jk} are defined analogously, and the expression for particle time is instead differentiated at $r=0$:

$$G^*_{jk}(s_1,s_2,t) = \frac{d}{dr}H^*_{jk}(r,s_1,s_2,t)\Big|_{r=0} = \sum_l\sum_m\left[\int_{x=0}^{\infty} x\mathrm{d}q^*_{jk,lm}(x,t)\right]s_1^l s_2^m = \sum_l\sum_m m^*_{jk,lm}(t)s_1^l s_2^m. \tag{4.12}$$

We see that given expressions for H^+_{jk}, $H^{\rightarrow}_{jk}$, H^-_{jk}, and H^*_{jk}, the coefficients corresponding to moments $m^+_{jk,lm}$, $m^{\rightarrow}_{jk,lm}$, $m^-_{jk,lm}$, $m^*_{jk,lm}$ can then be numerically computed using FFT analogously to (4.7) by replacing ϕ_{jk} with the corresponding G_{jk} functions. For notational simplicity, we use G_{jk} when referring collectively to G^+_{jk}, $G^{\rightarrow}_{jk}$, G^-_{jk}, and G^*_{jk}, and similarly define H_{jk}.

Having reduced our task to computing H_{jk}, we define $H_1 := H_{10}(r,s_1,s_2,t)$ and $H_2 :=$

$H_{01}(r, s_1, s_2, t)$. By independence of particles in the branching process, we have $H_{jk} = H_1^j H_2^k$. In all four cases, H_2 is analytically available, and we derive an ordinary differential equation for H_1, summarized in the theorem below. We present the result for a branching process with rates corresponding to the birth-death-shift model, but such systems of equations are available for an arbitrary time-homogeneous multi-type branching process.

Theorem 4.2.1 *Let $\{X_t\}$ be a two-type branching defined by the rates in* (5.15)*. Denote particle time and the number of births, shifts, and deaths over the interval $[0, t)$ by R_t, b_t, f_p, and d_t respectively. Define the generating functions corresponding to births as*

$$H_1^+(r, s_1, s_2, t) = E\left[r^{b_t} s_1^{X_1(t)} s_2^{X_2(t)} \mid \mathbf{X}(0) = (1, 0)\right] \text{ and}$$
$$H_2^+(r, s_1, s_2, t) = E\left[r^{b_t} s_1^{X_1(t)} s_2^{X_2(t)} \mid \mathbf{X}(0) = (0, 1)\right].$$

Then

$$H_2^+ = y_b + \left[\frac{-\lambda r}{2\lambda r y_b - \lambda - \mu} + \left(\frac{1}{s_2 - y_b} + \frac{\lambda r}{2\lambda r y_b - \lambda - \mu}\right) e^{-(2y_b\lambda r - \lambda - \mu)t}\right]^{-1},$$

where $y_b = (\lambda + \mu + \sqrt{\lambda^2 + 2\lambda\mu + \mu^2 - 4\lambda\mu r})/(2\lambda r)$, and H_1^+ satisfies the following differential equation:

$$\frac{d}{dt} H_1^+(t, s_1, s_2, r) = \lambda r H_1^+ H_2^+ + \nu H_2^+ + \mu - (\lambda + \mu + \nu) H_1^+, \tag{4.13}$$

subject to initial condition $H_1(r, s_1, s_2, 0) = s_1$.

The analogous generating functions for shifts, deaths, and particle time satisfy the fol-

lowing equations:

$$H_2^-(t,s_1,s_2,r) = y_d + \left[\frac{-\lambda}{2\lambda y_d - \lambda - \mu} + \left(\frac{1}{s_2 - y_d} + \frac{\lambda}{2\lambda y_d - \lambda - \mu}\right) e^{-(2y_d\lambda - \lambda - \mu)t}\right]^{-1},$$
$$H_2^{\rightarrow}(t,s_1,s_2,r) = 1 + \left[\frac{\lambda}{\mu - \lambda} + (\frac{1}{s_2 - 1} + \frac{\lambda}{\lambda - \mu})e^{(\mu-\lambda)t}\right]^{-1},$$
$$H_2^*(t,s_1,s_2,r) = y_* + \left[\frac{-\lambda}{2\lambda y_* - \lambda - \mu - r} + \left(\frac{1}{s_2 - y_*} + \frac{\lambda}{2\lambda y_* - \lambda - \mu - r}\right) e^{-(2y_*\lambda - \lambda - \mu - r)t}\right]^{-1},$$

$$\frac{d}{dt}H_1^-(t,s_1,s_2,r) = \lambda H_1^- H_2^- + \nu H_2^- + \mu r - (\lambda + \mu + \nu)H_1^-,$$
$$\frac{d}{dt}H_1^{\rightarrow}(t,s_1,s_2,r) = \lambda H_1^{\rightarrow} H_2^{\rightarrow} + \nu r H_2^{\rightarrow} + \mu - (\lambda + \mu + \nu)H_1^{\rightarrow},$$
$$\frac{d}{dt}H_1^*(t,s_1,s_2,r) = \lambda H_1^* H_2^* + \nu H_2^* + \mu - (\lambda + \mu + \nu + r)H_1^*,$$

where $y_d = (\lambda+\mu+\sqrt{\lambda^2 + 2\lambda\mu + \mu^2 - 4\lambda\mu r})/(2\lambda)$, $y_* = (\lambda+\mu+r+\sqrt{(\lambda+\mu+r)^2 - 4\lambda\mu})/(2\lambda)$, *and* $H_1^-(r,s_1,s_2,0) = H_1^{\rightarrow}(r,s_1,s_2,0) = H_1^*(r,s_1,s_2,0) = s_1$.

Proof The derivations are included in the Appendix B, using the birth equations (B-1) as a detailed example. The other systems follow analogous derivations.

This theorem shows that for each of the necessary sufficient statistics, computations for H_{jk} are essentially reduced to solving a single ordinary differential equation. As discussed for transition probabilities, this is easily accomplished using Runge-Kutta methods, which in practice offer more numerical stability than working with solutions obtained analytically by integrating the ODE. Because we can evaluate H_{jk}, we can also easily differentiate H_{jk} numerically, yielding access to numerical solutions to G_{jk}.

To summarize, with $H_{jk}^+, H_{jk}^-, H_{jk}^{\rightarrow}, H_{jk}^*$ now available, we may obtain the restricted moments by computing the coefficients in the power series $G_{jk}^+, G_{jk}^-, G_{jk}^{\rightarrow}, G_{jk}^*$. These coefficients are recovered using a Riemann approximation to the Fourier inversion formula analogous to

formula (4.7). For instance,

$$m^{+}_{(jk),(lm)}(t) = \int_0^1 \int_0^1 G^{+}_{jk}\left(t, e^{2\pi i t_1}, e^{2\pi i t_2}\right) e^{-2\pi i l t_1} e^{-2\pi i m t_2} dt_1 dt_2$$
$$\approx \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} G^{+}_{jk}\left(t, e^{2\pi i u/N}, e^{2\pi i v/N}\right) e^{-2\pi i l u/N} e^{-2\pi i m v/N}.$$

We are thus able to compute all necessary quantities appearing in the expected complete-data log-likelihood $\mathrm{E}_{\widetilde{\boldsymbol{\beta}}}\left[\ell_c(\mathbf{X}, \boldsymbol{\beta}) \mid \mathbf{Y}\right]$. Recall that sufficient statistics for each patient b_p, f_p, d_p, and R_p break up over intervals: i.e. the total number of births b_p is equal to the sum of the number of births over each disjoint interval $[t_{p,j-1}, t_{p,j})$, with $j = 1, \ldots, n(p)$. Further, by the Markov property, the conditional expectation of the number births over an interval $[t_1, t_2)$ given $\mathbf{Y}$ depends only on the states of the process at the endpoints of the interval:

$$\mathrm{E}\left[b_{p,t_2-t_1} \mid \mathbf{Y}\right] = \mathrm{E}\left[b_{p,t_2-t_1} | \mathbf{X}^p(t_1), \mathbf{X}^p(t_2)\right] = e^{+}_{\mathbf{X}^p(t_1),\mathbf{X}^p(t_2)}(t_2 - t_1) = \frac{m^{+}_{\mathbf{X}^p(t_1),\mathbf{X}^p(t_2)}(t_2 - t_1)}{p_{\mathbf{X}^p(t_1),\mathbf{X}^p(t_2)}(t_2 - t_1)}, \tag{4.14}$$

and the same is true for the other sufficient statistics. Therefore, for each process p,

$$\begin{aligned}
\mathrm{E}_{\widetilde{\boldsymbol{\beta}}}\left[b_p \mid \mathbf{Y}\right] &= \sum_{i=1}^{n(p)} e^{+}_{\mathbf{X}^p(t_{p,i-1}),\mathbf{X}^p(t_{p,i})}(t_{p,i-1} - t_{p,i}; \widetilde{\lambda}_p, \widetilde{\nu}_p, \widetilde{\mu}_p), \\
\mathrm{E}_{\widetilde{\boldsymbol{\beta}}}\left[f_p \mid \mathbf{Y}\right] &= \sum_{i=1}^{n(p)} e^{\rightarrow}_{\mathbf{X}^p(t_{p,i-1}),\mathbf{X}^p(t_{p,i})}(t_{p,i-1} - t_{p,i}; \widetilde{\lambda}_p, \widetilde{\nu}_p, \widetilde{\mu}_p), \text{ and} \\
\mathrm{E}_{\widetilde{\boldsymbol{\beta}}}\left[d_p \mid \mathbf{Y}\right] &= \sum_{i=1}^{n(p)} e^{-}_{\mathbf{X}^p(t_{p,i-1}),\mathbf{X}^p(t_{p,i})}(t_{p,i-1} - t_{p,i}; \widetilde{\lambda}_p, \widetilde{\nu}_p, \widetilde{\mu}_p),
\end{aligned} \tag{4.15}$$

with $\log\left(\widetilde{\lambda}_p\right) = \widetilde{\boldsymbol{\beta}}^{\lambda} \cdot \mathbf{z_p}, \log\left(\widetilde{\nu}_p\right) = \widetilde{\boldsymbol{\beta}}^{\nu} \cdot \mathbf{z_p}, \log\left(\widetilde{\mu}_p\right) = \widetilde{\boldsymbol{\beta}}^{\mu} \cdot \mathbf{z_p}$ similarly to equation (4.2). Finally, combining (4.15), (4.14), and (4.10), the expected complete-data log likelihood up

to a constant is equal to

$$\begin{aligned}\mathrm{E}_{\widetilde{\boldsymbol{\beta}}}\left[\ell_c(\mathbf{X},\boldsymbol{\beta}) \mid \mathbf{Y}\right] \propto \sum_{p=1}^{m}\Bigg\{\sum_{j=1}^{n(p)}\Bigg[&\frac{m^{+}_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}{p_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}\log\lambda_p\\ &+\frac{m^{\rightarrow}_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}{p_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}\log\nu_p\\ &+\frac{m^{-}_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}{p_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}\log\mu_p\\ &-\frac{m^{*}_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}{p_{\mathbf{X}^p(t_{p,j-1}),\mathbf{X}^p(t_{p,j})}(t_{p,j}-t_{p,j-1};\widetilde{\lambda}_p,\widetilde{\nu}_p,\widetilde{\mu}_p)}(\lambda_p+\mu_p+\nu_p)\Bigg]\Bigg\}.\end{aligned} \tag{4.16}$$

M-step

To complete an M-step, we use an efficient Newton-Raphson algorithm to maximize the expectation $g(\boldsymbol{\beta}) = E_{\widetilde{\boldsymbol{\beta}}}\left[\ell_c(\mathbf{X},\boldsymbol{\beta}) \mid \mathbf{Y}\right]$. Each Newton-Raphson step recursively updates parameters using the following equation:

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{cur} - \left[\mathbf{H}g(\boldsymbol{\beta}_{cur})\right]^{-1}\nabla g(\boldsymbol{\beta}_{cur}), \tag{4.17}$$

where ∇g denotes the gradient vector and $\mathbf{H}g$ denotes the Hessian matrix of $g(\boldsymbol{\beta})$. Fortunately, compact analytical forms for these quantities are available. First, we collect complete data sufficient statistics across processes into the following vectors:

$$\mathbf{U}^T = \left(E_{\widetilde{\boldsymbol{\beta}}}\left[b_{1,t_{1,n(1)}}|\mathbf{Y}\right],\ldots,E_{\widetilde{\boldsymbol{\beta}}}\left[b_{m,t_{m,n(m)}}|\mathbf{Y}\right)\right], \quad \mathbf{V}^T = \left(E_{\widetilde{\boldsymbol{\beta}}}\left[f_{1,t_{1,n(1)}}|\mathbf{Y}\right],\ldots,E_{\widetilde{\boldsymbol{\beta}}}\left[f_{m,t_{m,n(m)}}|\mathbf{Y}\right)\right],$$
$$\mathbf{D}^T = \left(E_{\widetilde{\boldsymbol{\beta}}}\left[d_{1,t_{1,n(1)}}|\mathbf{Y}\right],\ldots,E_{\widetilde{\boldsymbol{\beta}}}\left[d_{m,t_{m,n(m)}}|\mathbf{Y}\right)\right], \quad \mathbf{P}^T = \left(E_{\widetilde{\boldsymbol{\beta}}}\left[R_{1,t_{1,n(1)}}|\mathbf{Y}\right],\ldots,E_{\widetilde{\boldsymbol{\beta}}}\left[R_{m,t_{m,n(m)}}|\mathbf{Y}\right)\right].$$

If we aggregate covariate vectors for each process in a $c \times p$ matrix $\mathbf{Z} = (\mathbf{z_1},\ldots,\mathbf{z_m})$ and process-specific rates into vectors $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m), \boldsymbol{\nu} = (\nu_1,\ldots,\nu_m), \boldsymbol{\mu} = (\mu_1,\ldots,\mu_m)$, then

the gradient and Hessian can be expressed as

$$\nabla g(\boldsymbol{\beta}) = \left(-\mathbf{Z}^T\left[\text{diag}(\mathbf{P})\boldsymbol{\lambda} + \mathbf{U}\right], -\mathbf{Z}^T\left[\text{diag}(\mathbf{P})\boldsymbol{\nu} + \mathbf{V}\right], -\mathbf{Z}^T\left[\text{diag}(\mathbf{P})\boldsymbol{\mu} + \mathbf{D}\right]\right), \tag{4.18}$$

$$\mathbf{Hg}(\boldsymbol{\beta}) = \begin{pmatrix} -\mathbf{Z}^T\text{diag}(\mathbf{P})\text{diag}(\boldsymbol{\lambda})\mathbf{Z} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Z}^T\text{diag}(\mathbf{P})\text{diag}(\boldsymbol{\nu})\mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Z}^T\text{diag}(\mathbf{P})\text{diag}(\boldsymbol{\mu})\mathbf{Z} \end{pmatrix}. \tag{4.19}$$

The derivation of these expressions is parallel to those presented in [Doss et al., 2013]. In our experience the M-step generally converges in fewer than ten Newton-Raphson steps. Availability of closed form solutions (4.18) and (4.19) yields very fast execution of each Newton-Raphson step, making the computational cost of the M-step negligible compared to the E-step. Note that computing the M-step using only one Newton-Raphson step rather than iterating until convergence is sufficient to guarantee the ascent property of the EM algorithm [Lange, 1995], but we execute multiple steps because this strategy does not slow down our algorithm.

Accelerating E-step calculations for intervals with no change

In our birth-death-shift application, we may avoid the relatively costly E-step calculations for some intervals by approximating the probability of observing no changes with the probability that no event occurs in the underlying complete process. This approximation is not necessary in our algorithm, but can lead to gains in computational efficiency in settings such as our application where many intervals feature no observed changes.

It is very unlikely that events occur in a time interval $[t_1, t_2)$ yet no change is observed so that $\mathbf{X}(t_1) = \mathbf{X}(t_2)$. For instance, if 12 elements are present initially and a death followed by a birth occur, then we almost always observe $\mathbf{X}(t_1) = (12, 0), \mathbf{X}(t_2) = (11, 1)$ unless the element added by the birth occupies the *exact* location that was previously occupied by the element that dies. This scenario would leave the observed state unchanged, $\mathbf{X}(t_1) = \mathbf{X}(t_2) = (12, 0)$,

but has exceedingly low probability: the already small but non-negligible probability that more than one event occurs is then multiplied by $1/(S-11)$, the probability of the birth occurring in a specific location (recall S is very large). Therefore, it is numerically accurate to treat intervals with no observed changes as if no changes in the latent continuous-time process occur. In this case, the transition probability is easily calculated, given by the tail of an exponential distribution

$$p_{(12,0),(12,0)}(t_2 - t_1) = e^{-12(\lambda+\mu+\nu)(t_2-t_1)}. \tag{4.20}$$

In addition to efficient closed-form transition probability calculation, the expected sufficient statistics necessary for the E-step are known in this setting. If no events occur, we know that $e^{+}_{(k,0),(k,0)}(t) = e^{\rightarrow}_{(k,0),(k,0)}(t) = e^{-}_{(k,0),(k,0)}(t) = 0$, and that the expected particle time is $e^{*}_{(k,0),(k,0)}(t) = kt$. This is not only faster computationally but also more numerically stable, avoiding the division of numerically calculated restricted moments by numerically calculated transition probabilities. We verify this efficient implementation in our simulation studies, as illustrated in Figure 4.4.

Equation (4.20) is the same formula that Rosenberg et al. [2003] used to compute the probability of no event under their frequent monitoring (FM) method, but it is important to note that this approximation is *not* accurate for the other probabilities under FM; see Figure 4.2. Continuing our previous example, consider the probability of transitioning from $(12, 0)$ to $(11, 1)$. Under the FM approximation, such a transition can only happen through a shift event with probability $p_{(12,0),(11,1)}(t_2-t_1) = (\nu)/(\lambda+\nu+\mu)e^{-k(\lambda+\mu+\nu)(t_2-t_1)}$. However, as we have discussed above, a birth followed by death also leads to observing $\mathbf{X}(t_2) = (11, 1)$ in almost all cases. FM assigns zero probability density to such event histories, while our method does not ignore non-negligible contribution of these trajectories to $p_{(12,0),(11,1)}(t_2-t_1)$. Additionally, probabilities $p_{(j,k),(l,m)}(t_2 - t_1)$ are set to zero under FM when for all j, k, l, m where $|j - k| > 1$ or $|l - m| > 1$, preventing the use of all available data during inference. This is later illustrated in Figure 4.3.

4.2.3 Implementation

We implement our algorithm in the form of R package `bdsem`, available at `https://github.com/jasonxu90/bdsem` [Xu and Minin, 2014]. The EM algorithm implementation relies on numerical solutions to differential equations in package `deSolve`, and accommodates panel data settings with unevenly spaced discrete observations. Our package also includes functions for MLE inference using other methods, as well as code for simulating from the BDS process. The software is accompanied by a vignette that steps through simplified versions of all simulation studies.

4.3 Results

4.3.1 Comparison with frequent monitoring

We begin with several simulation experiments assessing the validity of our algorithms. The first simulation study checks whether transition probabilities calculated using our generating function method for the two-type branching process as described in (4.7) coincide with those of the BDS model. We compare these computations to Monte Carlo estimates of these probabilities obtained from simulated trajectories from the birth-death-shift model, and also include a comparison to the FM method presented in [Rosenberg et al., 2003].

The FM model allows at most one event to occur per interval. Thus, over an observation interval $[t_i, t_{i+1})$ beginning with k particles, the probabilities of a birth, death, and shift have closed forms $(\lambda/\theta)e^{-k\theta(t_{i+1}-t_i)}$, $(\mu/\theta)e^{-k\theta(t_{i+1}-t_i)}$, and $(\nu/\theta)e^{-k\theta(t_{i+1}-t_i)}$ respectively, where $\theta = \lambda + \nu + \mu$. The probability of no event occurring is given by $e^{-k\theta(t_{i+1}-t_i)}$, and all other transition probabilities are zero under the FM assumption. Because it becomes more likely that multiple events occur as the length of time between observations, dt, increases, we expect probabilities computed under FM to diverge substantially from the Monte Carlo estimates as we increase dt.

We compute Monte Carlo approximations of transition probabilities from 2000 realizations of a BDS process without covariates, with rates $\lambda = 0.0188, \mu = 0.0147, \nu = 0.00268$.

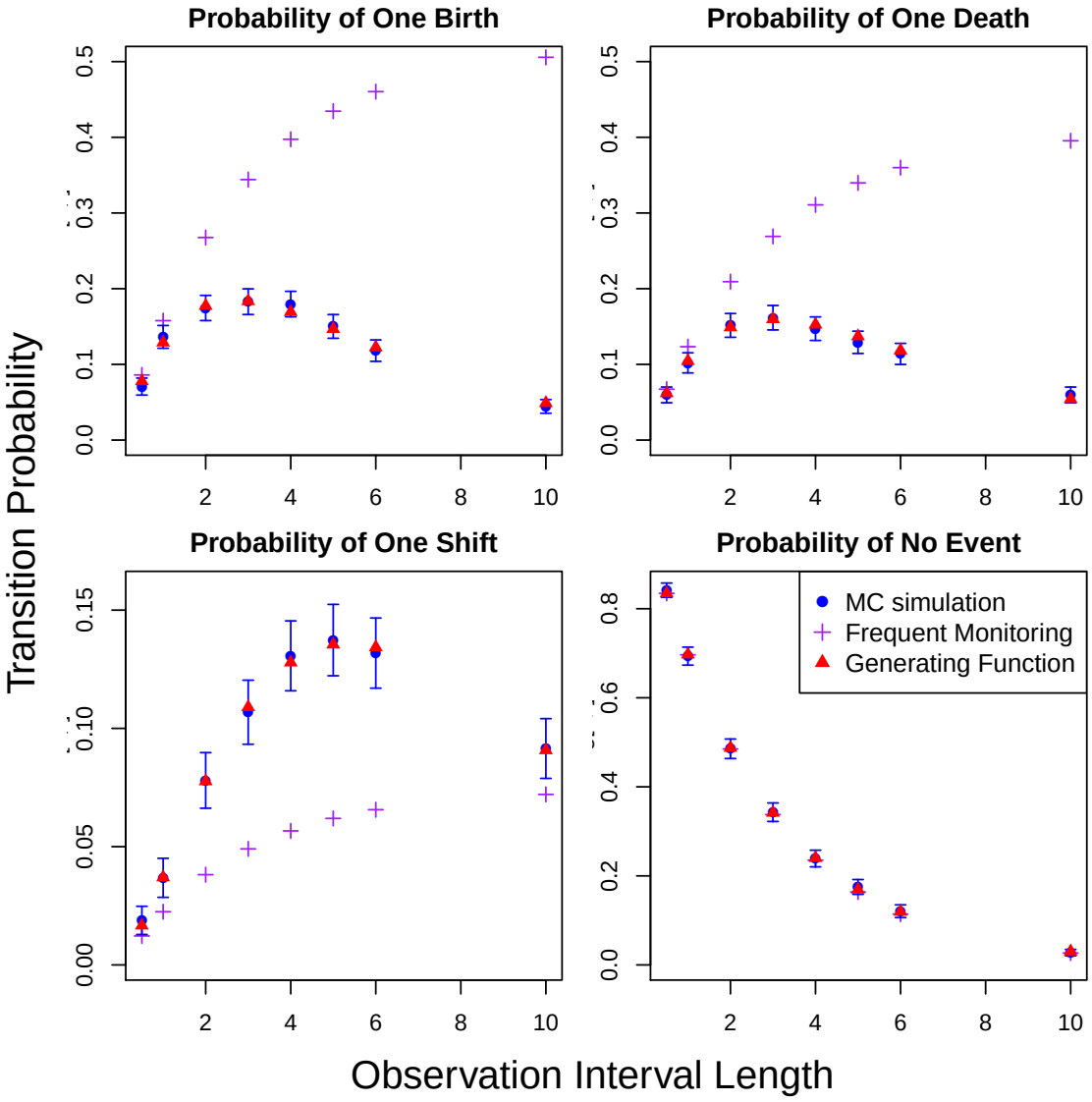


Figure 4.2: Transition probability approximations. BDS transition probabilities are approximated with two methods — the FM method, shown with magenta crosses, and the generating function method, shown with red triangles. We depict Monte Carlo estimates of the BDS transition probabilities with blue circles; vertical blue segments indicate their corresponding Monte Carlo confidence intervals.

These rates are equal to estimates of a transposable element birth, death, and shift rates obtained by Rosenberg et al. [2003] using the FM method. We begin each simulation with an initial population size of 10, and record the state of the process after simulating for dt units of time, varying dt from 0.5 to 10. The approximate transition probability $\hat{p}_{(10,0),(k,l)}(t)$ is then empirically computed by dividing the number of realizations ending in state $\mathbf{X}(t) = (k, l)$ by the total number of simulated processes.

In Figure 4.2, we see that as the length of an observation interval increases, FM approximations become inaccurate, while those obtained using our method remain within the narrow Monte Carlo confidence intervals. However, notice that the probability that no event occurs remains accurate even under the FM approximation, supporting the efficient implementation of our EM algorithm described in Section 4.2.2. Figure C-1 in the Appendix C demonstrates that our method also reliably calculates other transition probabilities that are set to 0 by the FM method, and these computations remain accurate as we vary the rates of the process.

Further, the discrepancies in numerical transition probabilities between methods indeed translate to differences in estimated rates. To see this, we generate a partially observed dataset and infer rates using both methods. We simulate from the BDS process with parameters $\lambda = 0.07, \mu = 0.12, \nu = 0.02$ to resemble the dynamics of the real dataset we will analyze in the next section, and record 200 discretely observed states of the process evenly spaced dt time units apart. Each simulated interval begins with an initial population size drawn uniformly between 1 and 15, and this data generating process is repeated three times, producing three datasets corresponding to inter-observation intervals of lengths $dt = (0.2, 0.4, 0.6)$. We infer the MLE rates for each of the three discretely observed datasets using the generating function method and under the frequent modeling assumption. This entire procedure is then repeated over 200 trials. In the top row of Figure 4.3, we see that our generating function approach successfully recovers the MLE estimates, and coverage of 95% confidence intervals remains close to 0.95 as we increase the length of time intervals between observations. The FM method performs somewhat reasonably for shorter observation intervals, but the bias in these approximate MLEs becomes stark as dt increases, with 95%

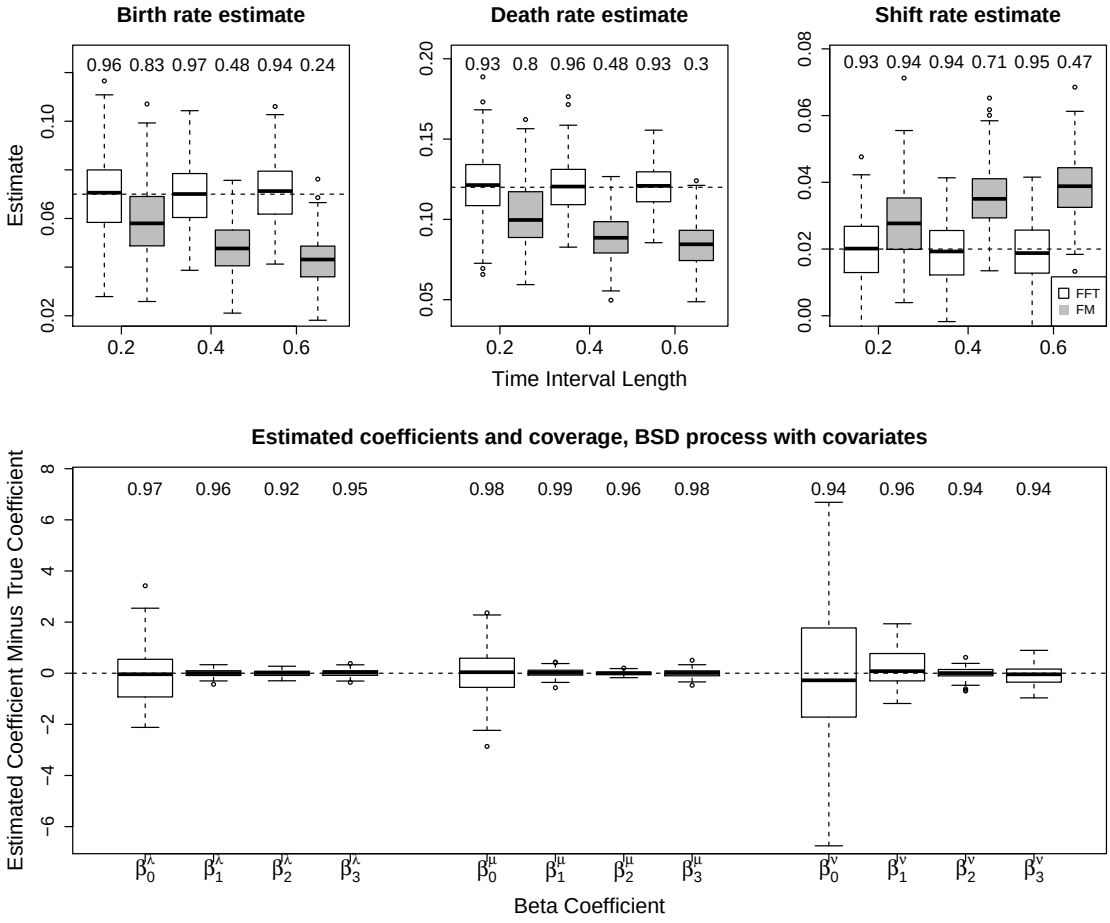


Figure 4.3: MLE parameter estimates on simulated data. The top row displays estimates of global birth, death, and shift rates in the simple BDS for three datasets, each with observation interval lengths $dt = (0.2, 0.4, 0.6)$. True parameter values used to initialize simulations marked by horizontal dashed line, and results using the FM method are included in gray. Monte Carlo coverage probabilities for 95% confidence intervals are displayed above box plots. The bottom row displays estimated coefficients using EM in the BDS process with covariates, shifted by true values.

confidence interval coverage probability dropping as low as 0.24.

Similarly to this transition probability experiment, we check the accuracy of restricted moment computations via simulation by verifying the equality

$$\mathrm{E}(N_t^+ \mid X_0 = i, j) = \sum_{k,l} \mathrm{E}(N_t^+, 1_{x_t=kl} \mid x_0 = i, j)$$

for expected births, and analogous expressions for other expected sufficient statistics. The left hand side is empirically approximated by a Monte Carlo average of the number of births over many realizations of the process, while the restricted moments on the right hand side are the quantities computed via our generating function approach (see Appendix C, Figure C-2).

4.3.2 Estimation of parameters in the BDS model with covariates

With accurate transition probabilities and restricted moments available, we are ready to infer coefficients in the BDS model with covariate-dependent rates using the EM algorithm. We begin by generating a simulated dataset resembling the real data consisting of observations corresponding to 100 "patients," each with three covariates $z_{p,1}, z_{p,2}, z_{p,3} \sim \mathrm{Unif}\{(0,2) \times (6,10) \times (4,6)\}$. An illustration of the format of these data, which reflects the format of the real dataset we later analyze, is provided in Table 4.1. We then simulate patient-specific BDS processes, beginning with rates λ_p, ν_p, μ_p log-linearly related to a true vector of coefficients $\boldsymbol{\beta}$ as defined in (4.2). We collect between 2 and 7 observations per patient, each spaced $dt = 0.4$ apart. Each simulated observation interval begins with an initial number of particles uniformly drawn between 2 and 14. Finally, we set true values of the effect sizes

$$\begin{aligned}
\boldsymbol{\beta}^{\lambda} &= [\log(7.5), \log(0.5), \log(0.3), \log(3)]\,, \\
\boldsymbol{\beta}^{\nu} &= [\log(0.5), \log(8), \log(0.5), \log(0.9)]\,, \\
\boldsymbol{\beta}^{\mu} &= [\log(4), \log(0.3), \log(0.8), \log(0.9)]\,,
\end{aligned}$$

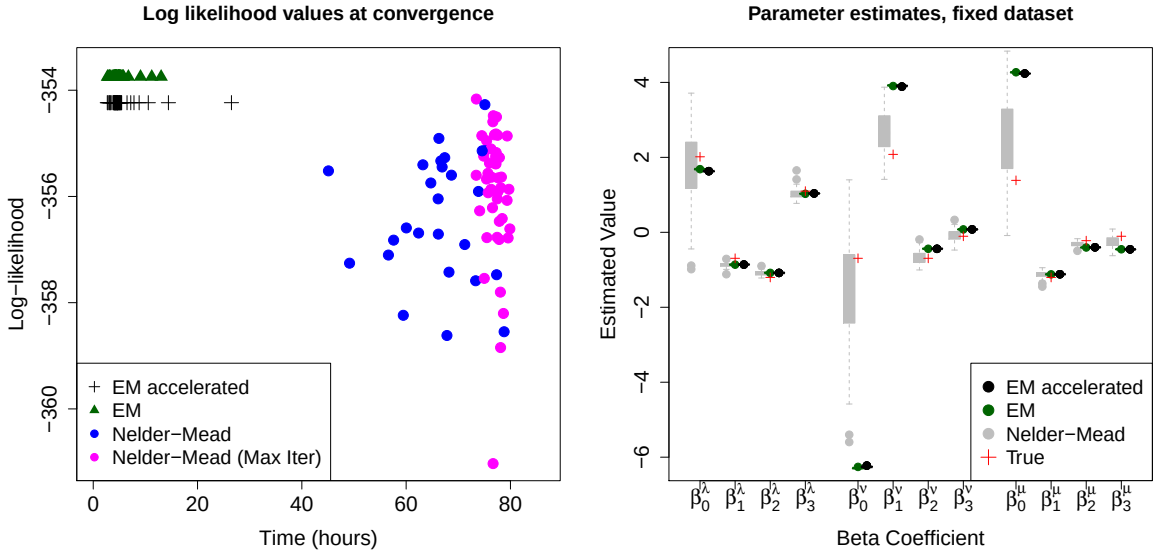
Patient	Time	# Bands	Shift	z_1	z_2	z_3
1	0	9	no	1.3	6.3	4.2
	0.4	9	no	1.3	6.3	4.2
	0.8	10	no	1.3	6.3	4.2
	1.2	10	no	1.3	6.3	4.2
	1.6	10	yes	1.3	6.3	4.2
	2.0	10	no	1.3	6.3	4.2
2	0	14	no	0.7	9.1	5.5
	0.4	14	no	0.7	9.1	5.5
	0.8	13	no	0.7	9.1	5.5
3	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.1: Visualization of data format with covariates z_i, $i = 1, 2, 3$.

chosen so that averaging over patients, the overall birth, shift, and death rates of the process resemble those in previous studies [Rosenberg et al., 2003, Doss et al., 2013].

The EM algorithm is initialized with $\boldsymbol{\beta}_0 \sim N(\boldsymbol{\beta}, \text{diag}(0.5\boldsymbol{\beta}))$, and the entire procedure of generating the dataset and inferring rates via EM is repeated 150 times. In the bottom row of Figure 4.3, we see that the MLEs are again unbiased estimates of the true values, with corresponding confidence interval coverage staying close to 95%.

Having verified that our EM algorithm successfully recovers the true parameters, we turn to a performance comparison with generic optimization via the Nelder-Mead (NM) algorithm implemented in the `optim` package [Nelder and Mead, 1965]. We choose NM as the method for comparison as it proved to be the most robust among the methods available via the `optim` function in R; a similar choice of NM for comparison to EM implementations is motivated in [Lange and Minin, 2013]. In this experiment, we generate one dataset as described in the procedure above from the BDS model with covariates. Fixing these data, we initialize each method with identical initial parameter values and convergence criteria, using a relative tolerance of $\epsilon = 1 \times 10^{-6}$, and repeat this procedure over 100 sets of initial conditions.



(a) (b)

Figure 4.4: The left plot shows converged log-likelihood values using EM, accelerated EM, and Nelder-Mead optimization. The right plot shows parameter estimates produced by the EM, accelerated EM, and Nelder-Mead algorithms, with true parameters values shown as red crosses.

Figure 4.4 displays the log-likelihood values achieved by each algorithm at convergence, as well as values in which Nelder-Mead terminated at an iteration limit set at 2000 steps. We see that in every case, the EM algorithm is significantly faster and finds a better optimum than NM. Further, the wide range of converged log-likelihood values suggests that NM is sensitive to initial conditions — an undesirable feature in this fixed data setting. We also verify that accelerating our EM algorithm according to Section 4.2.2 does not affect the log-likelihood value at convergence up to numerical precision, with a total difference in log-likelihood less than 0.5 accumulated over more than 400 observation intervals. Finally, we note that the comparison between EM and accelerated EM here is included to illustrate that they arrive at the same log-likelihood value and estimates at convergence. The increase in efficiency is not seen here: in these simulated examples, the generating function computations are always performed and cached at each iteration, rather than bypassed for candidate intervals

described in Section 4.2.2. In our application to the real dataset in the next section, we find that accelerating EM runs roughly six times faster than its unaccelerated counterpart.

Our EM approach is not only more stable in terms of the maximized log-likelihood, but also in terms of parameter estimates. The right panel of Figure 4.4 shows that estimates for each coefficient differ by no more than 0.01 across disparate initial conditions under both EM implementations, while a range of estimates are produced by the Nelder-Mead algorithm.

Notice that for some coefficients, estimates produced by NM appear to lie closer to the "true" parameters used to generate the synthetic data. We believe this to be an artifact of centering initial parameter values for both algorithms around the true parameters. Indeed, MLEs corresponding to the likelihood surface of a given *fixed* dataset generally do not exactly coincide with the "true" parameters used to simulate the data. The fact that EM consistently finds a better optimum in terms of log-likelihood demonstrates that this is the case.

4.3.3 Mycobacterium tuberculosis transposable element evolution

We apply our EM algorithm to infer covariate-dependent birth, death, and shift rates of the *M. tuberculosis* transposon IS*6110*, a frequently used marker to track *M. tuberculosis* in the community [McEvoy et al., 2007]. The marker serves as a DNA fingerprint, and in community-based studies patients that share the same or similar *M. tuberculosis* genotypes are considered as part of the same transmission chain [Van Embden et al., 1993, Kato-Maeda et al., 2011]. However, such inference relies on a fairly precise understanding of within-host evolutionary dynamics: for instance, if a DNA marker changes very rapidly, isolates from the same source will be strongly differentiated, and the severity of outbreaks would be underestimated without accounting for the high change rate. Understanding the rates of change of IS*6110*-based genotypes is thus critical toward the interpretation and design of such studies [Tanaka and Rosenberg, 2001], which in turn provide important information toward designing policy decisions such as control and intervention programs.

We analyze data from an ongoing study of the transmission and pathogenesis of *M. tuberculosis* patients in a community study in San Francisco [Cattamanchi et al., 2006,

Suwanpimolkul et al., 2013]. The database includes all culture positive tuberculosis cases reported to the San Francisco Department of Public Health. We included patients with more than one *M. tuberculosis* isolate from specimens sampled more than 10 days apart, genotyped with IS*6110* restriction fragment length polymorphism (RFLP) analysis. We assume that changes in the bands marking RFLP patterns evolve according to a linear birth-death-shift process, and assume that patients are not reinfected with a new strain between observations. Our dataset contains 252 observation intervals corresponding to 196 unique patients observed at 452 time points. Average time between sampling times is 0.35 years, with the longest interval being 2.35 years. Of the 252 intervals, 29 feature end points with distinct genotypes.

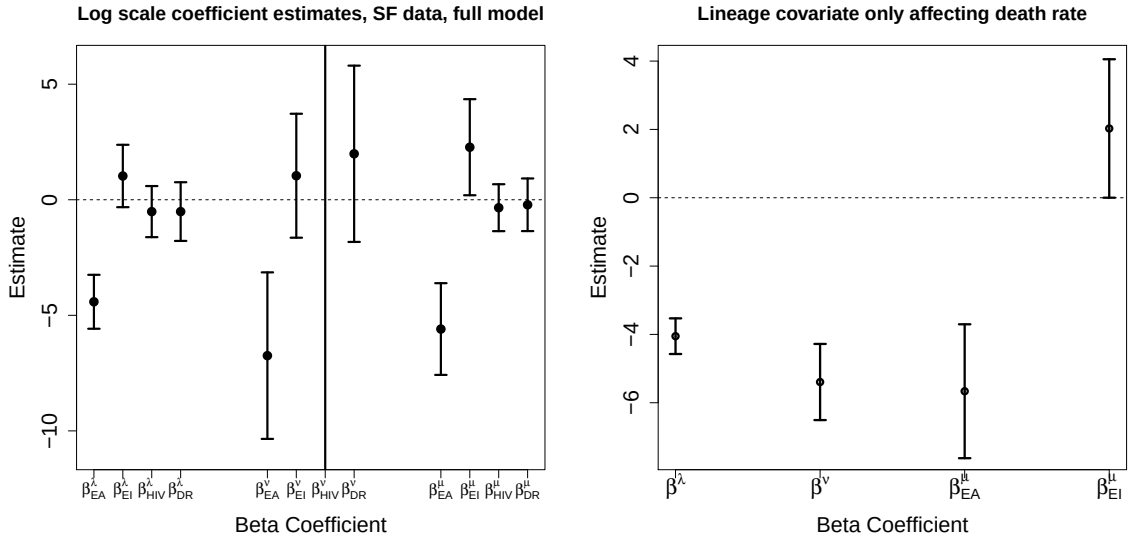
This dataset was analyzed by Rosenberg et al. [2003] under the FM assumption, but these authors necessarily discarded all intervals with more than one change in RFLP bands, as these intervals with "complex changes" are not possible under their restricted model. A later investigation by Doss et al. [2013] relaxes this assumption, allowing for multiple births or deaths to occur, but ignores RFLP band locations entirely, working instead only with total copy numbers evolving under a linear birth-death process. Under this birth-death model, the shift rate becomes unidentifiable, and the study instead infers covariate effects of birth and death rates. Our new method allows for a more principled, complete analysis, utilizing the full dataset without compromising any original modeling assumptions.

We begin by applying our EM algorithm to the simple BDS model with a single birth, death, and shift rate of IS*6110* for all patients. We estimate the MLE rates $\hat{\lambda} = 0.0156$, $\hat{\nu} = 0.00426$, $\hat{\mu} = 0.0187$, with associated 95% confidence intervals $(0.00929, 0.0251)$, $(0.00145, 0.0125)$, and $(0.0177, 0.0301)$ respectively. Starting the algorithm from a range of initial parameter values did not affect these results. These estimates are interpretable as the change rate of IS*6110* per copy, per year, and our results are consistent with previous estimates in the literature: for all rates, confidence intervals overlap those obtained in the frequent monitoring approach in [Rosenberg et al., 2003] as well as those obtained in the BD model [Doss et al., 2013]. Similarly to Doss et al. [2013] which estimates $\mu = 0.0207$, we find that our estimate of death rate μ is higher when allowing for multiple events between observations, compared

to $\mu = 0.0147$ obtained under the FM assumption. This is to be expected, as there are three intervals in which IS*6110* count drops by more than 1 in the dataset. Although confidence intervals overlap, our estimate of the shift rate is noticeably higher than the previous finding $\nu = 0.00268$ under FM, with the upper end of our confidence interval almost twice as large as the upper end of the 95% FM confidence interval $[0, 0.00654)$. Again, our analysis allows inclusion of several intervals that can be explained by at least two genotype changes that were either omitted in earlier studies or interpreted as a single birth event. Our EM algorithm approach is the first method to our knowledge that is able to accurately estimate the shift rate and produce reliable confidence intervals in the BDS model.

In addition to estimating the BDS rates globally, Doss et al. [2013] investigated rates as functions of several covariates in a panel data setting, and their findings in the birth-death framework suggest that *M. tuberculosis* lineage [Gagneux et al., 2006] may have a statistically significant effect on the rates of the process. We reexamine the effect of lineage on the rates in the full BDS model, considering 109 patients infected with Euro-American (EU) lineage strains, 54 patients with East-Asian (EA) strains, and 25 patients with Indo-Oceanic (IO) strains. We combine EU and IO lineages, because Doss et al. [2013] found that the number of IO samples was not sufficient to recover rates for this lineage. Following Doss et al. [2013], we also include HIV infection status of each patient (HIV) and drug resistance status of the *M. Tuberculosis* strain (DR). These attributes are coded as binary covariates: $\mathrm{EI}_p = 1$ if patient p is infected with the EU or IO strain and 0 otherwise, so that intercept terms $\beta_0^\lambda, \beta_0^\mu, \beta_0^\nu$ correspond to the EA strain. The variable $\mathrm{HIV}_p = 1$ if patient p is infected with HIV and 0 otherwise, and $\mathrm{DR}_p = 1$ if patient p is infected with a drug-resistant strain, and 0 otherwise. Covariates are log-linearly related to birth, death, and shift rates: $\log \lambda_p = \beta_0^\lambda + \beta_1^\lambda \mathrm{EI}_p + \beta_2^\lambda \mathrm{HIV}_p + \beta_3^\lambda \mathrm{DR}_p$, $\log \mu_p = \beta_0^\mu + \beta_1^\mu \mathrm{EI}_p + \beta_2^\mu \mathrm{HIV}_p + \beta_3^\mu \mathrm{DR}_p$, $\log \nu_p = \beta_0^\nu + \beta_1^\nu \mathrm{EI}_p + \beta_2^\nu \mathrm{HIV}_p + \beta_3^\nu \mathrm{DR}_p$.

We estimate coefficients in the full log-linear model described above, as well as in several simpler models, using the EM algorithm. The simpler models differ from the full model by either excluding the HIV and DR covariates, or excluding all covariates for specified global



(a) (b)

Figure 4.5: Coefficient estimates and 95% confidence intervals in full model and best model according to BIC. Notice intervals corresponding to β^{μ}_{EI} do not contain 0.

or "simple" rates. For instance, the model labeled "Lineage only, simple ν" in Table 4.2 has five parameters $\boldsymbol{\beta} = (\beta^{\lambda}_0, \beta^{\lambda}_1, \beta^{\mu}_0, \beta^{\mu}_1, \beta^{\nu})$, and rates defined as

$$\log \lambda_p = \beta^{\lambda}_0 + \beta^{\lambda}_1 \mathrm{EI}_p, \quad \log \nu_p = \log \nu = \beta^{\nu}, \quad \log \mu_p = \beta^{\mu}_0 + \beta^{\mu}_1 \mathrm{EI}_p.$$

In all cases, estimates obtained using the accelerated EM algorithm and regular implementation coincide, and neither is sensitive to initial conditions.

A summary and model comparison via the Bayesian Information Criterion (BIC) [Schwarz, 1978] is included in Table 4.2, which selects the model including only the lineage covariate for modeling death rate μ. Coefficient estimates are displayed graphically for the full model as well as the best model selected by BIC in Figure 4.5. While we choose not to report coefficient estimates from each model for brevity, in *all* models, the confidence interval for β^{μ}_{EI} does not contain zero, indicating that strain lineage has a statistically significant effect

Model	# Params	Log-likelihood	BIC
Full, separate EU, IO lineages	15	-119.845	330.01
Full	12	-120.498	313.25
Full, simple ν	9	-122.455	299.10
Lineage covariate only	6	-123.649	293.42
Lineage only, simple ν	5	-123.717	277.54
Lineage only, simple $\boldsymbol{\lambda}, \boldsymbol{\nu}$	**4**	**-124.472**	**273.02**
Simple λ, ν, μ	3	-127.914	273.90

Table 4.2: Model comparison via BIC $\approx -2\hat{L} + k \ln n$. We also fit the log-linear model in Doss et al. [2013], which includes separate indicator variables for Euro-American and Indo-Oceanic lineages. Models described as "lineage only" do not include HIV, DR covariates, and rates described as "simple" are global to all patients, not influenced by covariates in the model.

on the death rate. The estimate $\beta_{EI}^{\hat{\mu}} = 2.028$ under the best model indicates that in Euro-American and Indo-Oceanic lineages loss of IS*6110* element occurs $\exp(2.028) = 7.599$ times faster than in their East-Asian counterpart. Our analysis affirms the result suggested by Doss et al. [2013] in the simpler BD framework: *M. tuberculosis* lineage needs to be taken into consideration when studying disease transmission using IS*6110* genotypes.

4.4 Discussion

We have developed an EM algorithm for inference in a discretely observed, multi-type branching process framework. We focus our attention on fitting BDS processes to panel data, driven by the problem of estimating evolutionary dynamics of IS*6110* — a genetic marker that plays an important role in DNA fingerprinting of *M. tuberculosis*. Our method allows for birth, death, and shift log-rates to be linear combinations of many patient-specific covariates, and is flexible enough to capture the full range of dynamics between observation times by approximating the BDS process with a two-type branching process. To our knowledge, there is no other method of comparable accuracy for fitting BDS processes in this setting.

The generating functions we derive and numerical techniques that use these functions to

calculate previously unavailable transition probabilities and restricted moments are helpful tools toward probabilistic characterization of such processes more generally. We demonstrate how our generating function approach leads to maximum likelihood estimation and evaluation of expected complete-data log-likelihood within an EM algorithm, but note that these calculations also arise in a variety of other statistical techniques for prediction and estimation. For example, availability and tractability of the likelihood via our methods allows for their use in Bayesian inference.

Several problems associated with our numerical methods remain open. First, although we have strong empirical evidence that our branching process approximation to the discretely observed BDS likelihood is very accurate, a rigorous characterization of this approximation is lacking. Filling this theoretical gap is an interesting avenue for future research. Second, our method has potential numerical limitations in settings with high population counts. Computing transition probabilities to population sizes up to N of any particle type typically requires N^p differential equations to be solved, where p is the number of particle types. Although efficient numerical solvers are available and each ODE evaluation can be accomplished in only fractions of a second, requiring millions of evaluations becomes prohibitive, especially within an iterative algorithm. However, because the support of transition probabilities is often concentrated unless observation intervals are very long, we may attempt to harness this sparsity to reduce computational cost. We investigate efficiency gains by leveraging sparsity in the following chapter.

We apply our method to analyze within-host evolution of the transposon IS*6110*, an important marker in genetic fingerprinting of *M. tuberculosis*. We obtain confidence intervals for global birth, death, and shift rates λ, μ, ν that overlap with those obtained by Rosenberg et al. [2003] and with those for λ and μ obtained by Doss et al. [2013]. Thus, our estimates are consistent with previous results. While this suggests that the restrictive model assumptions in earlier approaches are not unreasonable in this application, we draw attention to our significantly higher estimate of ν. Indeed, while the frequent monitoring study by Rosenberg et al. [2003] suggests that the global shift rate ν is an order of magnitude smaller than

the birth and death rates, our method reveals that the shift rate is in fact comparable to the baseline Euro-American death rate after accounting for strain lineage. This is clearly illustrated in Figure 4.5, and suggests that the non-negligible shift event should not be omitted from the model as it was in [Doss et al., 2013]. This novel observation was not possible using existing methodology — our approach is the first to accurately estimate the birth, death, and shift rates as functions of covariates in this discretely monitored setting without compromising model assumptions.

Our covariate-specific rate analysis reaffirms previous indication in the simplified BD framework that strain lineage has a significant effect on the death rate [Doss et al., 2013], although the large confidence intervals suggest that this lineage effect is somewhat marginal. Indeed, more data would be required to be certain in the result, but our principled analysis is assuring in that any spurious result can now be attributed to limited, noisy data rather than to model misspecification. The possibility of differences in rates of genetic marker evolution across lineages is important in epidemiological studies. For example, similar IS*6110* genotypes across multiple individuals infected with EA lineage of *M. tuberculosis* do not provide strong evidence of these individuals belonging to the same transmission chain, because of the slow change rate of IS*6110* in the EA lineage. Failing to account for this may lead to inferring false relationships among genotypically similar clusters of patients.

The BDS model we consider is general enough so that our methods can be applied to studying evolution of any transposable element. Such studies are not limited to infectious disease surveillance—studying evolution of transposable elements in eukaryotes is also of great interest [Biémont, 2010]. Beyond the BDS framework, the tools we develop for fitting branching processes are transferable to many settings. For example, our methodology is applicable to compartmental models, a class of well-known multi-type branching processes that finds applications in modeling cancerous growth, bacterial evolution, and cellular differentiation in systems such as hematopoiesis [Gibson and Renshaw, 1998, Golinelli et al., 2006].

Chapter 5

LEVERAGING SPARSITY TO ACCELERATE GENERATING FUNCTION TECHNIQUES TO COMPUTE TRANSITION PROBABILITIES VIA COMPRESSED SENSING

Computation of transition probabilities is a usual bottleneck in model-based inference using CTMCs [Hajiaghayi et al., 2014], requiring a marginalization over the infinite set of possible end-point conditioned paths. Classically, this marginalization is accomplished by computing the matrix exponential of the infinitesimal generator of the CTMC. However, this procedure has cubic runtime complexity in the size of the state space, becoming prohibitive even for state spaces of moderate sizes. Alternatives also have their shortcomings: *uniformization* methods use a discrete-time "skeleton" chain to approximate the CTMC but rely on a restrictive assumption that there is a uniform bound on all rates [Grassmann, 1977, Rao and Teh, 2011]. Typically, practitioners resort to sampling-based approaches via Markov chain Monte Carlo (MCMC). Specifically, particle-based methods such as sequential Monte Carlo (SMC) and particle MCMC [Doucet et al., 2000, Andrieu et al., 2010] offer a complementary approach whose runtime depends on the number of imputed transitions rather than the size of the state space. However, these SMC methods have several limitations— in many applications, a prohibitively large number of particles is required to impute waiting times and events between transitions, and degeneracy issues are a common occurrence, especially in longer time series. A method by Hajiaghayi et al. [2014] accelerates particle-based methods by marginalizing holding times analytically, but has cubic runtime complexity in the number of imputed jumps between observations and is recommended for applications with fewer than one thousand events occurring between observations.

While the spectral generating function methods we introduced in Chapter 4 provide a

powerful alternative to simulation and avoids costly matrix operations in the context of branching processes, the Riemann approximation to the Fourier inversion formula requires $\mathcal{O}(N^b)$ probability generating function (PGF) evaluations, where b is the number of particle types and N is the largest population size at endpoints of desired transition probabilities. This complexity is no worse than linear in the size of the state space, but can also be restrictive: a two-type process in which each population can take values in the thousands would require millions of PGF evaluations to produce transition probabilities over an observation interval. This can amount to hours of computation in standard computing architectures, because evaluating PGFs for multi-type branching processes involves numerically solving systems of ordinary differential equations (ODEs). Such computations become infeasible within iterative algorithms.

In this chapter, we focus our attention on the efficient computation of transition probabilities in the presence of sparsity, presenting a novel compressed sensing generating function (CSGF) algorithm that dramatically reduces the computational cost of inverting the PGF. We apply our algorithm to a branching process model used to study hematopoiesis as well as a birth-death-shift process with applications to molecular epidemiology, and see that the sparsity assumption is valid for scientifically realistic rates of the processes obtained in previous statistical studies. We compare performance of CSGF to transition probability computations without taking advantage of sparsity, demonstrating a high degree of accuracy while achieving significant improvements in runtime.

5.0.1 Recap: generating function methods

Matrix exponentiation is cubic in $|\Omega|$ and thus prohibitive in many applications, but we may take an alternate approach by exploiting properties of the branching process. Recalling the techniques presented in Chapter 4, the probability generating function (PGF) for a two-type

process is defined

$$\begin{aligned}\phi_{jk}(t, s_1, s_2; \boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{\theta}}(s_1^{X_1(t)} s_2^{X_2(t)} | X_1(0) = j, X_2(0) = k) \\ &= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} p_{(jk),(lm)}(t; \boldsymbol{\theta}) s_1^l s_2^m; \end{aligned} \tag{5.1}$$

this definition extends analogously for any m-type process. We suppress dependence on $\boldsymbol{\theta}$ for notational convenience. As discussed in Chapter 2, Bailey [1964] provides a general technique to write a system of differential equations governing ϕ_{jk} using the Kolmogorov forward or backward equations given the instantaneous rates $a_j(k, l)$. These solutions can be computed numerically using standard algorithms if not analytically, and with ϕ_{jk} available, transition probabilities are related to the PGF (5.1) via differentiation:

$$p_{(jk),(lm)}(t) = \frac{\partial^l}{\partial s_1} \frac{\partial^m}{\partial s_2} \phi_{jk}(t) \bigg|_{s_1 = s_2 = 0}. \tag{5.2}$$

We then compute transition probabilities by applying a Riemann sum approximation to the Fourier inversion of (5.2) evaluated at complex arguments:

$$\begin{aligned} p_{(jk),(lm)}(t) &= \int_0^1 \int_0^1 \phi_{jk}(t, e^{2\pi i w_1}, e^{2\pi i w_2}) e^{-2\pi i l w_1} e^{-2\pi i m w_2} dw_1 dw_2 \\ &\approx \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) e^{-2\pi i l u/N} e^{-2\pi i m v/N}. \end{aligned} \tag{5.3}$$

In practice, this set of transition probabilities $S = \{p_{(jk),(lm)}(t)\}$ for all $l, m = 0, \ldots, N$, given initial values of (j, k), can be obtained via the Fast Fourier Transform (FFT), described in Section 5.2. It is necessary to choose $N > l, m$, since exponentiating the roots of unity can yield at most N distinct values

$$e^{-2\pi i m v/N} = e^{-2\pi i (mv \bmod N)/N};$$

this is related to the Shannon-Nyquist criterion [Shannon, 2001], which dictates that the number of samples required to recover a signal must match its highest frequency. Thus, calculating "high frequency" coefficients— when l, m take large values—requires $\mathcal{O}(N^2)$ numerical ODE solutions, which becomes computationally expensive for large N.

5.0.2 Sparsity:

Given an initial state $\mathbf{X}(0) = (j, k)$, the support of transition probabilities is often concentrated over a small range of (l, m) values. For example, if $\mathbf{X}(t) = (800, 800)$, then the probability that the entire process becomes extinct, $\mathbf{X}(t+s) = (0, 0)$, is effectively zero unless particle death rates are very high or s is a very long time interval. In many realistic applications, $p_{(800,800),(l,m)}(s)$ has non-negligible mass on a small support, for instance only over l, m values between 770 and 820. Visually, a similar example is provided in Figure 5.0.2. While their values can be computed using Equation (5.3) for a choice of $N > 820$, requiring N^2 ODE evaluations toward computing only $(820 - 770)^2$ nonzero probabilities seems wasteful. To exploit the sparsity information in such a setting, we bridge aforementioned branching process techniques to the literature of *compressed sensing*.

5.1 Compressed sensing

Originally developed in an information theoretic setting, the principle of compressed sensing (CS) states that an unknown sparse signal can be recovered accurately and often perfectly from significantly fewer samples than dictated by the Shannon-Nyquist rate at the cost of solving a convex optimization problem [Donoho, 2006, Candès, 2006]. CS is a robust tool to collect high-dimensional sparse data from a low-dimensional set of measurements and has been applied to a plethora of fields, leading to dramatic reductions in the necessary number of measurements, samples, or computations. In our setting, the transition probabilities play the role of a target sparse signal of Fourier coefficients. The data reduction made possible via CS then translates to reducing computations to a random subsample of PGF evaluations, which play the role of measurements used to recover the signal.

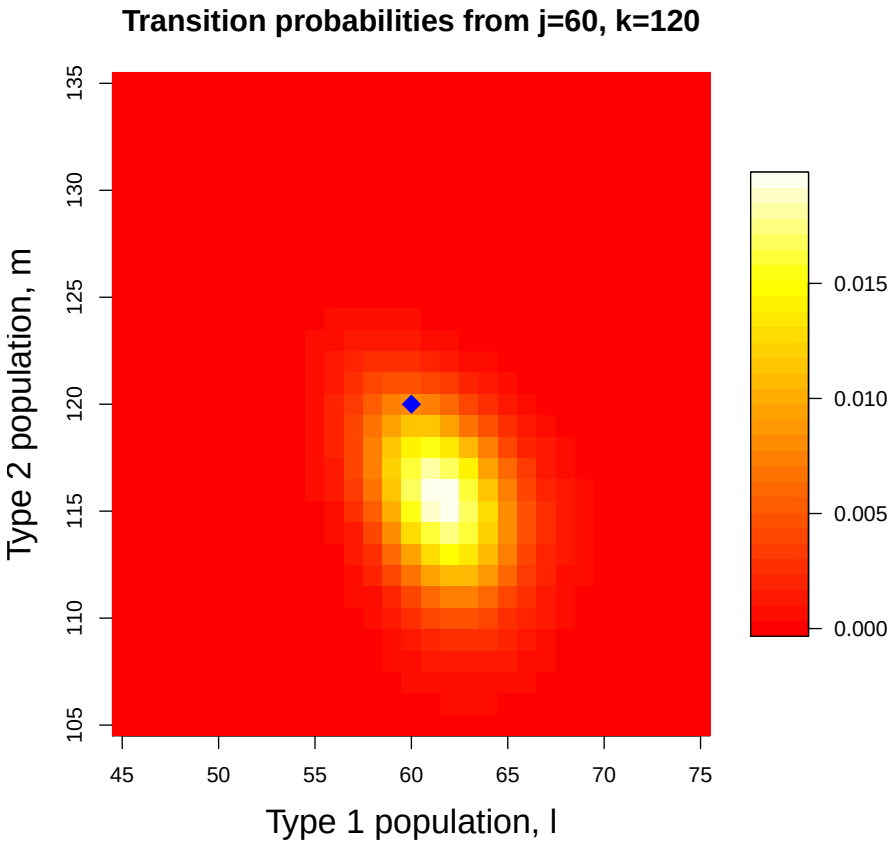


Figure 5.1: In many applications, transition probabilities over a finite time interval are *sparse*. This example illustrates sparse transitions in a two-type hematopoiesis model using scientifically realistic rates. Beginning at $\mathbf{X}(0) = (60, 120)$, denoted by blue diamond, the support of transition probabilities $p_{(60,120),(l,m)}(0.5)$ is only nonzero in a small region

5.1.1 Overview

In the CS framework, the unknown signal is a vector $\mathbf{x} \in \mathbb{C}^N$ observed through a measurement $\mathbf{b} = \mathbf{V}\mathbf{x} \in \mathbb{C}^M$ with $M << N$. Here $\mathbf{V}$ denotes an $M \times N$ *measurement matrix* or sensing matrix. Since $M < N$, the system is underdetermined and inversion is highly ill-posed—the space of solutions is an infinite affine subspace, but CS theory shows that recovery can be accomplished under certain assumptions by seeking the *sparsest* solution. Let $\boldsymbol{\psi}$ be an orthonormal basis of $\mathbb{C}^N$ that allows a K-sparse representation of $\mathbf{x}$: that is, $\mathbf{x} = \boldsymbol{\psi}\mathbf{s}$ where $\mathbf{s}$ is a sparse vector of coefficients such that $||\mathbf{s}||_0 < K$. Candès [2006] proves that recovery can then be accurately accomplished by finding the sparsest solution

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmin}} ||\mathbf{s}||_0 \quad s.t. \quad \mathbf{A}\mathbf{s} = \mathbf{b}, \tag{5.4}$$

where $\mathbf{A} = \mathbf{V}\boldsymbol{\psi}$ is the composition of the measurement and sparsifying matrices. In practice, this non-convex objective is combinatorially intractable to solve exactly, and is instead solved by proxy via ℓ_1-relaxation, resulting in a convex optimization program. In place of Equation (5.4), we optimize the unconstrained penalized objective

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{A}\mathbf{s} - \mathbf{b}||_2^2 + \lambda||\mathbf{s}||_1, \tag{5.5}$$

where λ is a regularization parameter enforcing sparsity of $\mathbf{s}$. The signal $\mathbf{x}$, or equivalently $\mathbf{s}$, can be recovered perfectly using only $M = CK \log N$ measurements for some constant C when $\mathbf{A}$ satisfies the *Restricted Isometry Property* (RIP) [Candès and Tao, 2005, Candès, 2008]—briefly, this requires that $\mathbf{V}$ and $\boldsymbol{\psi}$ to be *incoherent* so that rows of $\mathbf{V}$ cannot sparsely represent the columns of $\boldsymbol{\psi}$ and vice versa. Coherence between $\mathbf{V}, \boldsymbol{\psi}$ is defined as

$$\mu(\mathbf{V}, \boldsymbol{\psi}) = \sqrt{n} \max_{i,j} |\langle \mathbf{V}, \boldsymbol{\psi}_j \rangle|,$$

and low coherence pairs are desirable. It has been shown that choosing random measurements $\mathbf{V}$ satisfies RIP with overwhelming probability [Candès, 2008]. Further, given $\boldsymbol{\psi}$, it is often possible to choose a known ideal distribution from which to sample elements in $\mathbf{V}$ such that $\mathbf{V}$ and $\boldsymbol{\psi}$ are maximally incoherent.

5.1.2 Higher dimensions

CS theory extends naturally to higher-dimensional signals [Candès, 2006]. In the 2D case which will arise in our applications (Section 5.3), the sparse solution $\mathbf{S} \in \mathbb{C}^{N\times N}$ and measurement

$$\mathbf{B} = \mathbf{A}\mathbf{S}\mathbf{A}^T \in \mathbb{C}^{M\times M} \tag{5.6}$$

are matrices rather than vectors, and we solve

$$\hat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{A}\mathbf{S}\mathbf{A}^T - \mathbf{B}||_2^2 + \lambda||\mathbf{S}||_1. \tag{5.7}$$

This can always be equivalently represented in the vector-valued framework: vectorizing

$$\operatorname{vec}(\mathbf{S}) = \widetilde{\mathbf{s}} \in \mathbb{C}^{N^2}, \quad \operatorname{vec}(\mathbf{B}) = \widetilde{\mathbf{b}} \in \mathbb{C}^{M^2},$$

we now seek $\widetilde{\mathbf{b}} = \widetilde{A}\widetilde{\mathbf{s}}$ as in Equations (5.4), (5.5), where $\widetilde{\mathbf{A}} = \mathbf{A} \otimes \mathbf{A}$ is the Kronecker product of $\mathbf{A}$ with itself. In practice, it can be preferable to solve (5.7), since the number of entries in $\widetilde{\mathbf{A}}$ grows rapidly and thus the vectorized problem requires a costly construction of $\widetilde{\mathbf{A}}$ and can be cumbersome in terms of memory.

5.2 CSGF method

We propose an algorithm that allows for efficient PGF inversion within a compressed sensing framework. We focus our exposition on two-type models: linear complexity in $|\Omega|$ is less often a bottleneck in single-type problems, and all generating function methods as well as compressed sensing techniques we describe extend to higher dimensional settings.

We wish to compute the transition probabilities $p_{jk,lm}(t)$ given any $t > 0$ and $\mathbf{X}(0) = (j, k)$. These probabilities can be arranged in a matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ with entries

$$\left\{\mathbf{S}\right\}_{l,m} = p_{jk,lm}(t).$$

Without the CS framework, these probabilities are obtained following Equation (5.3) by first computing an equally sized matrix of PGF solutions

$$\widetilde{\mathbf{B}} = \left\{ \ \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) \right\}_{u,v=0}^{N-1} \in \mathbb{C}^{N \times N}. \tag{5.8}$$

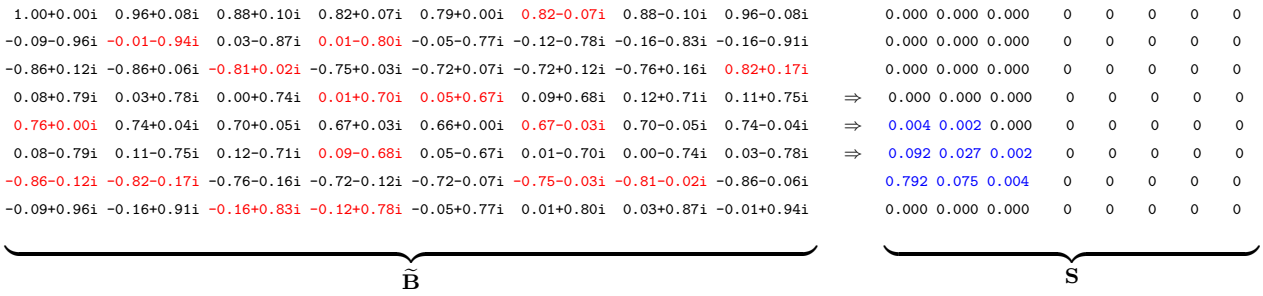
For large N, obtaining $\widetilde{\mathbf{B}}$ is computationally expensive, and our method seeks to bypass this step. When $\widetilde{\mathbf{B}}$ is computed, transition probabilities are then recovered by taking the fast Fourier transform $\mathbf{S} = \text{fft}(\widetilde{\mathbf{B}})$. To better understand how this fits into the CS framework, we can equivalently write the fast Fourier transform in terms of matrix operations $\mathbf{S} = \mathbf{F}\widetilde{\mathbf{B}}\mathbf{F}^T$, where $\mathbf{F} \in \mathbb{C}^{N \times N}$ denotes the discrete Fourier transform matrix (see Appendix). Thus, the sparsifying basis $\boldsymbol{\psi}$ is the Inverse Discrete Fourier Transform (IDFT) matrix $\boldsymbol{\psi} = \mathbf{F}^*$ given by the conjugate transpose of $\mathbf{F}$, and we have $\widetilde{\mathbf{B}} = \boldsymbol{\psi}\mathbf{S}\boldsymbol{\psi}^T$.

When the solution matrix $\mathbf{S}$ is expected to have a sparse representation, our CSGF method seeks to recover $\mathbf{S}$ without computing the full matrix $\widetilde{\mathbf{B}}$, instead beginning with a much smaller set of PGF evaluations $\mathbf{B} \in \mathbb{C}^{M \times M}$ corresponding to random entries of $\widetilde{\mathbf{B}}$ selected uniformly at random. Denoting randomly sampled indices $\mathcal{I}$, this smaller matrix is a projection $\mathbf{B} = \mathbf{A}\mathbf{S}\mathbf{A}^T$ in the form of Equation (5.6) where $\mathbf{A} \in \mathbb{C}^{M \times N}$ is obtained by selecting a subset of rows of $\boldsymbol{\psi}$ corresponding to $\mathcal{I}$. Uniform sampling of rows corresponds to multiplying by a measurement matrix encoding the *spike basis* (or standard basis): formally, this fits into the framework described in Section 5.1.1 as $\mathbf{A} = \mathbf{V}\boldsymbol{\psi}$, with measurement matrix rows $\mathbf{V}_j(l) = \delta(j - l)$. The spike and Fourier bases are known to be *maximally incoherent* in any dimension, so uniformly sampling indices $\mathcal{I}$ is optimal in our setting.

Now in the compressed sensing framework, computing the reduced matrix $\mathbf{B}$ only requires

a logarithmic proportion $|\mathbf{B}| \propto K \log |\widetilde{\mathbf{B}}|$ of PGF evaluations necessary in Equation (5.8). Computing the transition probabilities $\mathbf{S}$ is thus reduced to a signal recovery problem, solved by optimizing the objective in Equation (5.7).

The following example may help to illustrate the idea of the proposed CSGF method:



The sparse solution $\mathbf{S}$ on the right is usually obtained by taking the finite Fourier transform of all entries in $\widetilde{\mathbf{B}}$ on the left. Using compressed sensing, we now bypass computing the full matrix $\widetilde{\mathbf{B}}$, which usually requires solving N^2 ODEs, and instead seek to recover $\mathbf{S}$ using only a sample $\mathbf{B} \subset \widetilde{\mathbf{B}}$, denoted by the red entries of $\widetilde{\mathbf{B}}$. In this illustrative example, $\mathbf{S}$ is only $N = 8$ by 8, but in practice the number of "red entries" we need to compute grows only logarithmically. When N is large, $\mathbf{B}$ is significantly smaller than $\widetilde{\mathbf{B}}$, leading to large gains in computational efficiency.

5.2.1 Solving the ℓ_1 problem

There has been extensive research on algorithms for solving the ℓ_1 regularization objective in Equation (5.5) and related problems [Tibshirani, 1996, Beck and Teboulle, 2009a]. As mentioned previously, vectorizing the problem so that it can be represented in the form (5.5) requires wasteful extra memory; instead we choose to solve the objective in Equation (5.7) using a *proximal gradient descent* (PGD) algorithm.

PGD is useful for solving minimization problems with objective of the form $f(x) = g(x) + h(x)$ with g convex and differentiable, and h convex but not necessarily differentiable. Letting

$$g(\mathbf{S}) = \frac{1}{2}||\mathbf{A}\mathbf{S}\mathbf{A}^T - \mathbf{B}||_2^2, \quad h(\mathbf{S}) = \lambda||\mathbf{S}||_1,$$

we see that Equation (5.7) satisfies these conditions. A form of generalized gradient descent, PGD iterates toward a solution with

$$x_{k+1} = \underset{z}{\operatorname{argmin}}[g(x_k) + \nabla g(x_k)^T(z - x_k) + \frac{1}{2L_k}||z - x_k||_2^2 + h(z)], \tag{5.9}$$

where L_k is a step size that is either fixed or determined via line-search. This minimization has known closed-form solution

$$x_{k+1} = \text{softh}(x_k - L_k \nabla g(x_k), L_k \lambda), \tag{5.10}$$

where softh is the soft-thresholding operator

$$[\text{softh}(x, \alpha)]_i = \text{sgn}(x_i) \max(|x_i| - \alpha, 0). \tag{5.11}$$

Alternating between these steps results in an *iterative soft-thresholding algorithm* that solves the convex problem (5.7) with rate of convergence $\mathcal{O}(1/k)$ when L_k is fixed. The softh() operation is simple and computationally negligible, so that the main computational cost is in evaluating $\nabla g(x_k)$. We derive a closed form expression for the gradient in our setting

$$\nabla g(\mathbf{S}) = -\mathbf{A}^*(\mathbf{B} - \mathbf{A}\mathbf{S}\mathbf{A}^T)\overline{\mathbf{A}}, \tag{5.12}$$

where $\overline{\mathbf{A}}, \mathbf{A}^*$ denote complex conjugate and conjugate transpose of $\mathbf{A}$ respectively. In practice, the inner term $\mathbf{A}\mathbf{S}\mathbf{A}^T$ is obtained as a subset of the inverse fast Fourier transform of $\mathbf{S}$ rather than by explicit matrix multiplication. The computational effort in computing $\nabla g(\mathbf{S})$ therefore involves only the two outer matrix multiplications.

We implement a fast variant of PGD using momentum terms [Beck and Teboulle, 2009b] based on an algorithm introduced by Nesterov, and select step sizes L_k via a simple line-search subroutine [Beck and Teboulle, 2009a]. The accelerated version includes an *extrapolation*

step, where the soft-thresholding operator is applied to a momentum term

$$y_{k+1} = x_k + \omega_k(x_k - x_{k-1})$$

rather than to x_k; here ω_k is an extrapolation parameter for the momentum term. Remarkably, the accelerated method still only requires one gradient evaluation at each step as y_{k+1} is a simple linear combination of previously computed points, and has been proven to achieve the optimal worst-case rate of convergence $\mathcal{O}(1/k^2)$ among first order methods [Nesterov, 1983]. Similarly, the line-search procedure involves evaluating a bound that also only requires one evaluation of ∇g (see Appendix).

Algorithm 1 provides a summary of the CSGF method in pseudocode.

Algorithm 1 `CSGF` algorithm.

Input: initial sizes $X_1 = j, X_2 = k$, time interval t, branching rates $\boldsymbol{\theta}$, signal size $N > j, k$, measurement size M, penalization constant $\lambda > 0$, line-search parameters L, c.
Uniformly sample M indices $\mathcal{I} \subset [0, \dots N-1]/N$
Compute $\mathbf{B} = \left\{ \phi_{jk}(t, e^{2\pi i u/N}, e^{2\pi i v/N}) \right\}_{u,v \in \mathcal{I} \times \mathcal{I}}$
Define $\mathbf{A} = \boldsymbol{\psi}_{\mathcal{I}\cdot}$ the $\mathcal{I}$ rows of IDFT matrix $\boldsymbol{\psi}$
Initialize: $\mathbf{S}_1 = \mathbf{Y}_1 = \mathbf{0}$
for $k = 1, 2, \dots,$ {max iterations} **do**
    Choose $L_k = \texttt{line-search}(L, c, \mathbf{Y}_k)$
    Update extrapolation parameter $\omega_k = \frac{k}{k+3}$
    Update momentum $\mathbf{Y}_{k+1} = \mathbf{S}_k + \omega_k(\mathbf{S}_k - \mathbf{S}_{k-1})$
    Compute $\nabla g(\mathbf{Y}_{k+1})$ according to (5.12)
    Update $\mathbf{S}_{k+1} = \text{soft}(\mathbf{S}_k - L_k \nabla g(\mathbf{Y}_{k+1}), L_k \lambda)$
end for
return $\hat{\mathbf{S}} = \mathbf{S}_{k+1}$

5.3 Examples

We will examine the performance of CSGF in two applications: a stochastic two-compartment model used in statistical studies of *hematopoiesis*, the process of blood cell production, and a birth-death-shift model that has been used to study the evolution of *transposons*, mobile

genetic elements.

5.3.1 Two-compartment hematopoiesis model

Hematopoiesis is the process in which self-sustaining primitive hematopoietic stem cells (HSCs) specialize, or *differentiate*, into progenitor cells, which further specialize to eventually produce mature blood cells. In addition to far-reaching clinical implications — stem cell transplantation is a mainstay of cancer therapy — understanding hematopoietic dynamics is biologically interesting, and provides critical insights of general relevance to other areas of stem cell biology [Orkin and Zon, 2008]. A two-compartment stochastic model depicted in Figure 5.2 has enabled estimation of hematopoietic rates in mammals from data in several studies [Catlin et al., 2001, Golinelli et al., 2006, Fong et al., 2009]. Without the ability to compute transition probabilities, the estimating equation approach by Catlin et al. [2001] is statistically inefficient, resulting in uncertain estimated parameters with very wide confidence intervals. Nonetheless, biologically sensible rates are inferred. Golinelli et al. [2006] observe that transition probabilities are unknown for a linear birth-death process (compartment 1) coupled with an inhomogeneous immigration-death process (compartment 2), motivating their computationally intensive reversible jump MCMC implementation.

However, we can equivalently view the model as a two-type branching process. Under such a representation, it becomes possible to compute transition probabilities via Equation (5.3). The type one particle population X_1 corresponds to hematopoietic stem cells (HSCs), and X_2 represents progenitor cells. With parameters as denoted in Figure 5.2, the nonzero instantaneous rates defining the process are

$$\begin{aligned} a_1(2,0) &= \rho & a_1(0,1) &= \nu & a_1(1,0) &= -(\rho+\nu) \\ a_2(0,0) &= \mu & a_2(0,1) &= -\mu. \end{aligned} \tag{5.13}$$

Having specified the two-type branching process, we derive solutions for its PGF, defined in Equation (5.1), with details in the Appendix:

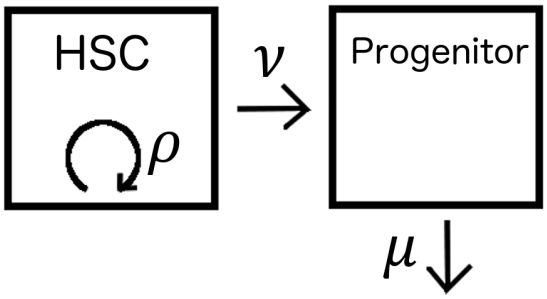


Figure 5.2: HSCs can self-renew, producing new HSCs at rate ρ, or differentiate into progenitor cells at rate ν. Further progenitor differentiation is modeled by rate μ.

Proposition 5.3.1 *The generating function for the two-type model described in* (5.13) *is given by* $\phi_{jk} = \phi_{1,0}^j \phi_{0,1}^k$, *where*

$$\begin{cases} \phi_{0,1}(t, s_1, s_2) = 1 + (s_2 - 1)e^{-\mu t} \\ \frac{d}{dt}\phi_{1,0}(t, s_1, s_2) = \rho\phi_{1,0}^2(t, s_1, s_2) - (\rho + \nu)\phi_{1,0}(t, s_1, s_2) \\ \qquad\qquad\qquad +\nu\phi_{0,1}(t, s_1, s_2). \end{cases} \tag{5.14}$$

We see that $\phi_{0,1}$ has closed form solution so that evaluating ϕ_{jk} only requires solving one ODE numerically, and with the ability to compute ϕ_{jk}, we may obtain transition probabilities using Equation (5.3). In this application, cell populations can easily reach thousands, motivating the CSGF approach to accelerate transition probability computations.

5.3.2 Birth-death-shift model for transposons

Our second application revisits the birth-death-shift (BDS) process described in Chapter 4. Recall the nonzero rates defining the two-type branching process representation of the BDS

model are given by

$$\begin{aligned}
&a_1(1,1) = \beta, && a_1(0,1) = \sigma, && a_1(0,0) = \delta,\\
&a_1(1,0) = -(\beta+\sigma+\delta), && a_2(0,2) = \beta,\\
&a_2(0,1) = -(\beta+\delta), && a_2(0,0) = \delta.
\end{aligned} \tag{5.15}$$

and its PGF is governed by the following system:

$$\begin{cases}
\phi_{0,1}(t,s_1,s_2) = 1 + \left[\frac{\beta}{\delta-\beta} + (\frac{1}{s_2-1} + \frac{\beta}{\beta-\delta})e^{(\delta-\beta)t}\right]^{-1}\\
\frac{d}{dt}\phi_{1,0}(t,s_1,s_2) = \beta\phi_{1,0}\phi_2 + \sigma\phi_{0,1} + \delta - (\beta+\sigma+\delta)s_1.
\end{cases} \tag{5.16}$$

5.4 Results

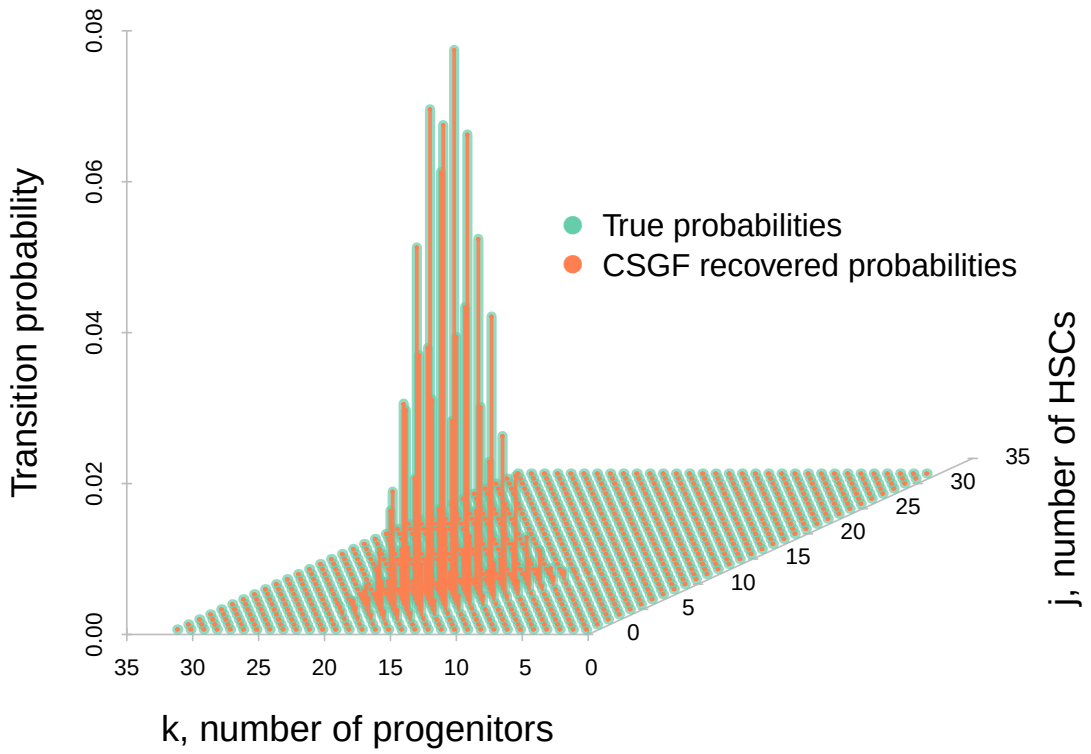


Figure 5.3: Illustrative example of recovered transition probabilities in hematopoiesis model described in Section 5.3. Beginning with 15 HSCs and 5 progenitors over a time period of one week, the CSGF solution $\hat{\mathbf{S}} = \left\{\hat{p}_{(15,5),(j,k)}(1)\right\}$, $j, k = 0, \ldots, 31$, perfectly recovers transition probabilities $\mathbf{S}$, using fewer than half the measurements.

To compare the performance of CSGF to the computation of Equation (5.3) without considering sparsity, we first compute sets of transition probabilities $\mathbf{S}$ of the hematopoiesis model using the full set of PGF solution measurements $\widetilde{\mathbf{B}}$ as described in Equation (5.8). These "true signals" are compared to the signals computed using CSGF $\hat{\mathbf{S}}$, recovered using only a random subset of measurements $\mathbf{B}$ following Algorithm 1. Figure 5.3 provides an illustrative example with small cell populations for visual clarity— we see that the support of transition

probabilities is concentrated (sparse), and the set of recovered probabilities $\hat{\mathbf{S}}$ is visually identical to the true signal.

In each of the aforementioned applications, we calculate transition probabilities $\mathbf{S} \in \mathbb{R}^{N\times N}$ for maximum populations $N = 2^7, 2^8, \ldots 2^{12}$, given rate parameters $\boldsymbol{\theta}$, initial population $\mathbf{X}(0)$, and time intervals t. Each computation of $\mathbf{S}$ requires N^2 numerical evaluations of the ODE systems (5.14), (5.16). For each value of N, we repeat this procedure beginning with ten randomly chosen sets of initial populations $\mathbf{X}(0)$ each with total size less than N. We compare the recovered signals $\hat{\mathbf{S}}$ computed using CSGF to true signals $\mathbf{S}$, and report median runtimes and measures of accuracy over the ten trials, with details in the following sections.

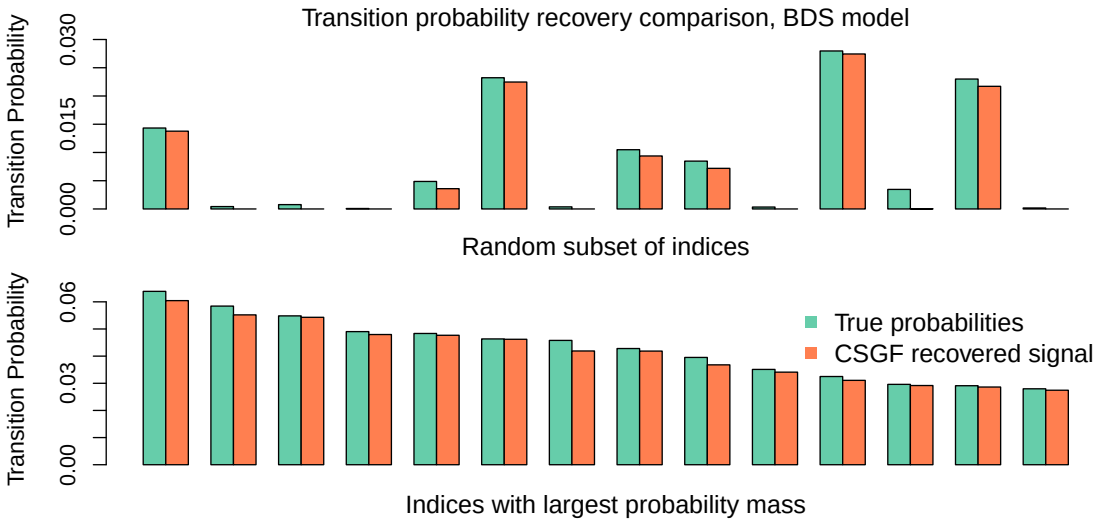


Figure 5.4: Randomly selected probabilities and largest probabilities recovered using CSGF are nearly identical to their true values. Probabilities displayed here correspond to a randomly selected BDS model trial with N=512; transition probabilities $\hat{\mathbf{S}}$ via CSGF are recovered from a sample $\mathbf{B}$ requiring fewer than 2% of ODE computations used to compute $\mathbf{S} = \text{fft}(\widetilde{\mathbf{B}})$.

Parameter settings: In the hematopoiesis example, we set per-week cell fate decision rates $\boldsymbol{\theta}_{\text{hema}} = (0.125, 0.104, 0.147)$ and observation period $t = 1$ week, based on biologically

Table 5.1: Runtimes and error, birth-death-shift model.

N	M	Time (sec), $\widetilde{\mathbf{B}} \in \mathbb{C}^{N\times N}$	Time (sec), $\mathbf{B} \in \mathbb{C}^{M\times M}$	Time (sec), PGD	$\varepsilon_{\max} = \lvert\hat{p}_{ij,kl} - p_{ij,kl}\rvert_{\max}$	$\varepsilon_{\text{rel}} = \varepsilon_{\max}/\lvert p_{ij,kl}\rvert_{\max}$
128	25	39.7	2.3	1.0	5.27×10^{-3}	2.77×10^{-2}
256	33	150.2	3.8	7.8	4.86×10^{-3}	4.71×10^{-2}
512	45	895.8	7.8	25.3	2.71×10^{-3}	4.68×10^{-2}
1024	68	2508.9	18.6	58.2	1.41×10^{-3}	5.12×10^{-2}
2048	101	9788.3	26.1	528.3	8.10×10^{-4}	4.81×10^{-2}
4096	150	40732.7	57.4	2234.7	4.01×10^{-4}	5.32×10^{-2}

Table 5.2: Runtimes and error, hematopoiesis model

N	M	Time (sec), $\widetilde{\mathbf{B}} \in \mathbb{C}^{N\times N}$	Time (sec), $\mathbf{B} \in \mathbb{C}^{M\times M}$	Time (sec), PGD	$\varepsilon_{\max} = \lvert\hat{p}_{ij,kl} - p_{ij,kl}\rvert_{\max}$	$\varepsilon_{\text{rel}} = \varepsilon_{\max}/\lvert p_{ij,kl}\rvert_{\max}$
128	43	108.6	9.3	0.64	9.41×10^{-4}	2.25×10^{-2}
256	65	368.9	22.1	2.1	9.44×10^{-4}	4.73×10^{-2}
512	99	922.1	44.8	8.5	3.23×10^{-4}	3.60×10^{-2}
1024	147	5740.1	118.1	41.9	2.27×10^{-4}	5.01×10^{-2}
2048	217	12754.8	145.0	390.0	1.29×10^{-4}	5.10×10^{-2}
4096	322	58797.3	310.7	2920.3	9.43×10^{-5}	6.13×10^{-2}

sensible rates and observation time scales of data from previous studies of hematopoiesis in mammals [Catlin et al., 2001, Golinelli et al., 2006, Fong et al., 2009]. For the BDS application, we set per-year event rates $\boldsymbol{\theta}_{\text{bds}} = (0.0156, 0.00426, 0.0187)$ estimated in [Xu et al., 2015], and $t = .35$ years, the average length between observations in the San Francisco tuberculosis dataset [Cattamanchi et al., 2006].

In each case, we computed $M^2 = 3K \log N^2$ total random measurements to obtain $\mathbf{B}$ for CSGF, and we set the regularization parameters $\lambda_{\text{hsc}} = \sqrt{\log M}$, $\lambda_{\text{bds}} = \log M$, with more regularization in the BDS application as lower rates and a shorter observation interval leads us to expect more sparsity. While careful case-by-case tuning to choose λ, M would lead to optimal results, we set them in this simple manner across *all* trials to demonstrate a degree of robustness, still yielding promising performance results. In practice one may apply standard cross-validation procedures to select λ, M, and because the target solution is a set

of transition probabilities, checking that entries in the recovered solution $\hat{\mathbf{S}}$ sum close to 1 offers a simpler available heuristic. Finally, though one may expedite convergence of PGD by supplying an informed initial guess with positive values near values $\mathbf{X}(0)$ in practice, we initialize PGD with an uninformative initial value $\mathbf{S}_1 = \mathbf{0}$ in all cases.

5.4.1 Accuracy:

In both models and for all values of N, each signal was reconstructed very accurately. Errors are reported in Tables 5.1 and 5.2 for the BDS and hematopoiesis models respectively. Maximum absolute errors for each CSGF recovery

$$\varepsilon_{\max} = \max_{kl} |\{\hat{\mathbf{S}}\}_{kl} - \{\mathbf{S}\}_{kl}| = \max_{kl} |\hat{p}_{ij,kl}(t) - p_{ij,kl}(t)|$$

are on the order of 10^{-3} at worst. We also report a measure of relative error, and because $\varepsilon_{\max}$ is typically attained at large probabilities, we include the maximum absolute error relative to the largest transition probability

$$\varepsilon_{\text{rel}} = \frac{\varepsilon_{\max}}{\max_{kl} \{S\}_{kl}},$$

providing a more conservative measure of accuracy. We still see that ε_{rel} is on the order of 10^{-2} in all cases. Visually, the accuracy of CSGF is stark: Figure 5.4 provides a side-by-side comparison of randomly selected transition probabilities recovered in the BDS model for $N = 2^9$.

5.4.2 Running Times:

Tables 5.1 and 5.2 show dramatic improvements in runtime using CSGF, reducing the number of ODE computations logarithmically. For instance, with $N = 4096$, we see the time spent on PGF evaluations necessary for CSGF is less than 0.1% of the time required to compute $\mathbf{S}$ in the BDS model, and around 0.5% of computational cost in the less sparse

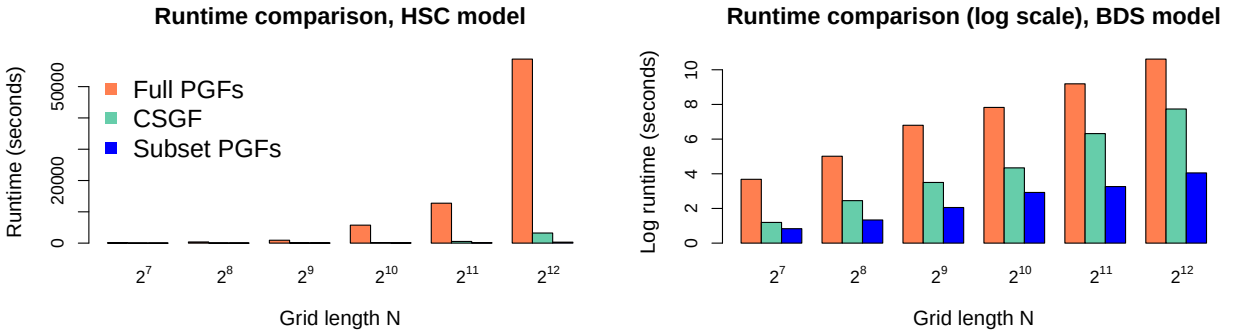


Figure 5.5: Blue bars indicate runtime to compute subset PGFs $\mathbf{B}$; coral bars additionally include PGD runtime. While gains are already significant, an optimized implementation may further reduce total CSGF runtime.

hematopoiesis application. Including the time required for solving Equation (5.7) via PGD, we see that computing $\hat{\mathbf{S}}$ using CSGF reduces runtime by two orders of magnitude, requiring less than 6% of total computational time spent toward computing $\mathbf{S}$ in the worst case. We remark that ODE solutions are computed using a C implementation of efficient solvers via package `deSolve`, while we employ a naive R implementation of PGD. We emphasize the logarithmic reduction in required numerical ODE solutions; an optimized implementation of PGD reducing R overhead will yield further real-time efficiency gains. A visual comparison is provided in Figure 5.4.2.

5.5 Discussion

We have shown that the spectral generating function techniques to compute transition probabilities developed in Chapter 4 fit within the compressed sensing paradigm. While generating function approaches bypass costly matrix exponentiation and simulation-based techniques by exploiting mathematical properties in the branching structure, we've demonstrated that these techniques are now made scalable by additionally harnessing available sparsity structure. We show that when sparsity is present in the set of transition probabilities, computational cost can be reduced up to a logarithmic factor over existing methods. Note that the presence of

sparsity is the *only* additional assumption necessary to apply our CSGF method—no prior knowledge about the structure of the sparsity or support regions of the transition probabilities is necessary. Many real-world applications of branching process modeling exhibit such sparsity, and we have seen that CSGF achieves accurate results with significant efficiency gains in two such examples with realistic parameter settings from the scientific literature. Finally, we note that other statistically relevant quantities such as expectations of particle dwell times and restricted moments can be computed using similar generating function techniques [Minin and Suchard, 2008], and the CSGF framework applies analogously when sparsity is present.

Chapter 6

STATISTICAL INFERENCE IN PARTIALLY OBSERVED STOCHASTIC COMPARTMENTAL MODELS WITH APPLICATION TO CELL LINEAGE TRACKING OF *IN VIVO* HEMATOPOIESIS

Single-cell lineage tracking strategies enabled by recent experimental technologies have produced significant insights into cell fate decisions, but lack the quantitative framework necessary for rigorous statistical tasks such as parameter estimation. In this chapter, we develop such a framework with corresponding moment-based parameter estimation techniques for continuous-time stochastic compartmental models. We apply this method to *hematopoiesis*, the complex mechanism of blood cell production. Using branching process techniques, we derive closed-form expressions for higher moments in a general class of such models, enabling efficient rate estimation in a much richer, multi-compartment model of hematopoiesis than previous statistical studies. The method provides the first rate inference procedure to our knowledge for fitting such models to time series data generated from cellular barcoding experiments. After testing the methodology in simulation studies, we apply our estimator to hematopoietic lineage tracking data from rhesus macaques. Results provide a more complete understanding of cell fate decisions during hematopoiesis in non-human primates, which may be more relevant to human biology and clinical strategies than previous findings in murine studies. The methodology is transferrable to compartmental models and multi-type branching models more broadly, commonly used in studies of cancer progression, epidemiology, and many other fields.

6.1 Introduction

This chapter develops inferential tools for a class of hidden stochastic population processes. In particular, we present a correlation matching loss function based on method of moments for rate estimation in multi-type branching process models of *hematopoiesis*, the process of blood cell production. During hematopoiesis, self-renewing hematopoietic stem cells (HSCs) specialize or *differentiate* via a series of intermediate progenitor cell stages to produce mature blood cells [Weissman, 2000]. Understanding the details of this system is a fundamental problem in biology, and progress in this area will also help shed light on other areas of basic biology. For example, further advances in hematopoiesis research will yield insights into mechanisms of cellular interactions, cell lineage programming, and characterization of cellular phenotypes during cell differentiation [Orkin and Zon, 2008]. Moreover, understanding hematopoiesis is clinically important: all blood cell diseases, including leukemias, myeloproliferative disorders and myelodysplasia are caused by malfunctions in some part of the hematopoiesis process, and hematopoietic stem cell (HSC) transplantation has become a mainstay for gene therapy and cancer treatments [Whichard et al., 2010].

Hematopoiesis research was one of the earliest successes of mathematical modeling in cell biology [Becker et al., 1963, Siminovitch et al., 1963]. *Stochastic compartmental models* form one popular class of models used to study hematopoiesis, in which cells are assumed to self replicate and differentiate according to a *Markov branching process* [Kimmel and Axelrod, 2002]. While much is known about production of blood cells by progenitor cells, uncovering details of HSC and progenitor cell replication/differentiation dynamics has proven to be more difficult, requiring further modeling simplifications. Notably, experimental techniques developed to study feline hematopoiesis using X-chromosome inactivation markers have produced a series of statistical studies leading to a two-compartmental stochastic model of hematopoiesis [Abkowitz et al., 1990, Newton et al., 1995, Golinelli et al., 2006, Fong et al., 2009, Catlin et al., 2011]. However, this simple two-type representation cannot distinguish between stages of differentiation beyond the HSC, and results obtained from analyzing this

model have not resolved long standing questions about patterns and sizes of the clones descended from individual HSC lineages. It should be noted that even these simplified models capturing the clonal dynamics descended from an HSC have posed significant statistical and computational challenges. More complex multi-compartmental models have been studied mathematically under additional assumptions—for instance, the regulatory behavior of several multi-stage models have been studied by Colijn and Mackey [2005] and Marciniak-Czochra et al. [2009]. However, efforts in analyzing these more precisely specified structures have relied on deterministic modeling of the overall population with continuous-valued state variables. To study dynamics in detail at the single-cell level, continuous approximations are not suitable as HSC counts of a single lineage are very low and near zero, and while deterministic modeling may be appropriate for steady-state population level behavior, they are not suitable to model fate decisions at the cell level which are much more sensitive to stochastic events. Indeed, studies suggest that hematopoietic dynamics are stochastic in nature [Ogawa, 1993, Kimmel, 2014]. Additionally, as one cannot completely specify all details of such a complex system in a mathematical model, a stochastic modeling approach that quantifies uncertainty provides a natural safeguard against model selection to some extent.

Recently emergent experimental techniques now allow researchers to track the dynamics of distinct cell lineages descended from one ancestral progenitor or HSC cell. Collecting such high resolution data is made possible by lentiviral genetic barcoding coupled with modern high-throughput sequencing technologies [Gerrits et al., 2010, Lu et al., 2011, Wu et al., 2014]. Data collected from individual cell lineages, rather than from a population containing a mixture of cell lineages, comprise independent and identically distributed time series, potentially allowing for investigation of much more realistic models of hematopoiesis. More importantly, the ability to analyze individual lineage trajectories can be very useful in characterizing patterns of cell differentiation, shedding light on the larger tree structure of the differentiation process. While these data are certainly more informative than those from previous experiments, statistical methods capable of analyzing such data are only beginning to emerge. Perié et al. [2014] model genetic barcoding data in a murine study collected at

the end of the mice's lifespan, but do not account for the longitudinal aspect of the data nor read count information, instead working with a binarized simplification of the data. Goyal et al. [2015] present a neutral model of steady-state hematopoiesis applied to vector site integration data, but cannot infer crucial process parameters such as the rate of stem cell self-renewal.

Wu et al. [2014] provide a preliminary analysis of their cellular barcoding data using off-the-shelf techniques such as hierarchical clustering. While this already reveals important scientific insights, it lacks the ability to perform statistical tasks such as parameter estimation and model fitting/selection. This paper attempts to fill this methodological gap, developing new statistical techniques for studying the barcoded hematopoietic cell lineages from the rhesus macaque data. As we will discuss in more detail, the difficulty lies in the partially observed nature of a complex process with a massive hidden state space. Statistical challenges arise from several facets of the experimental design so that standard techniques for hidden Markov models and CTMC inference cannot be readily applied, instead requiring careful modeling that at once captures the complexity of the data yet allows feasible algorithms for inference. We propose a fully generative stochastic modeling framework and an efficient method of parameter estimation that allows much richer hematopoietic structures to be statistically analyzed than previously possible, allowing for many-compartmental models that consider HSC, progenitor, and mature cell stages. The following section details the experimental design and dataset we consider, and provides an overview of the stochastic model. Next, we motivate the approach by statistically formulating our inferential goal, provide a rigorous characterization of each component of our model, and derive the necessary mathematical expressions in Section 6.3. We then thoroughly validate these methods via several simulation studies, and fit the models to the rhesus macaque barcoding data. Finally, we close with a discussion of these results, their implications, and avenues for future work.

6.2 Data and model

We will analyze single cell lineage tracking data generated by the cellular barcoding experiments in [Wu et al., 2014]. Briefly, Wu et al. [2014] start by extracting HSCs and progenitors from rhesus macaques and labeling the extracted sample. Specifically, lentiviral vectors are created using high diversity oligonucleotides with known DNA sequences — these vector sequences each correspond to a unique ID, collectively forming a genetic *barcode library*. Next, autologous HSCs and early progenitor cells are extracted from the monkey and marked or transduced with these lentiviruses, and transduced cells are then infused back into the irradiated monkeys. Irradiation depletes the residual blood cells, so that reconstitution is supported by these extracted cells following the retransplantation.

During hematopoietic reconstitution, the animal is monitored indirectly by taking samples of the blood cells at observation intervals ranging from several weeks to months. All clones descended from a marked cell inherit its unique barcode ID, enabling lineage tracking. At each observation time, the blood sample is sorted into monocyte (Mono), granulocyte (Gr), T, B, and natural killer (NK) cell types. Next, polymerase chain reaction (PCR) is performed on purified DNA samples from each sorted cell population, and barcodes are retrieved from the PCR product using Illumina sequencing. Sequences are filtered in such a way that only barcode IDs with numbers of reads (barcode sequences) exceeding a specified minimum read threshold remain in the dataset, reducing the effect of nonlinearities and noise in the PCR procedure in the pool of sequences we work with. Thus, at each observation time, the experimental protocol yields a read count corresponding to each barcode ID present in each cell type sample. Together, read count data for a given barcode ID constitute an independent time series that inform us of contributions to different cell types over time across lineages. Having restricted our attention to only barcode lineages exceeding a threshold of at least 1000 read counts at some observation time as in [Wu et al., 2014], we arrive at the dataset consisting of over 110 million read counts across 9635 unique barcode lineages, observed at irregular time intervals over a total period of 30 months. An illustration of the

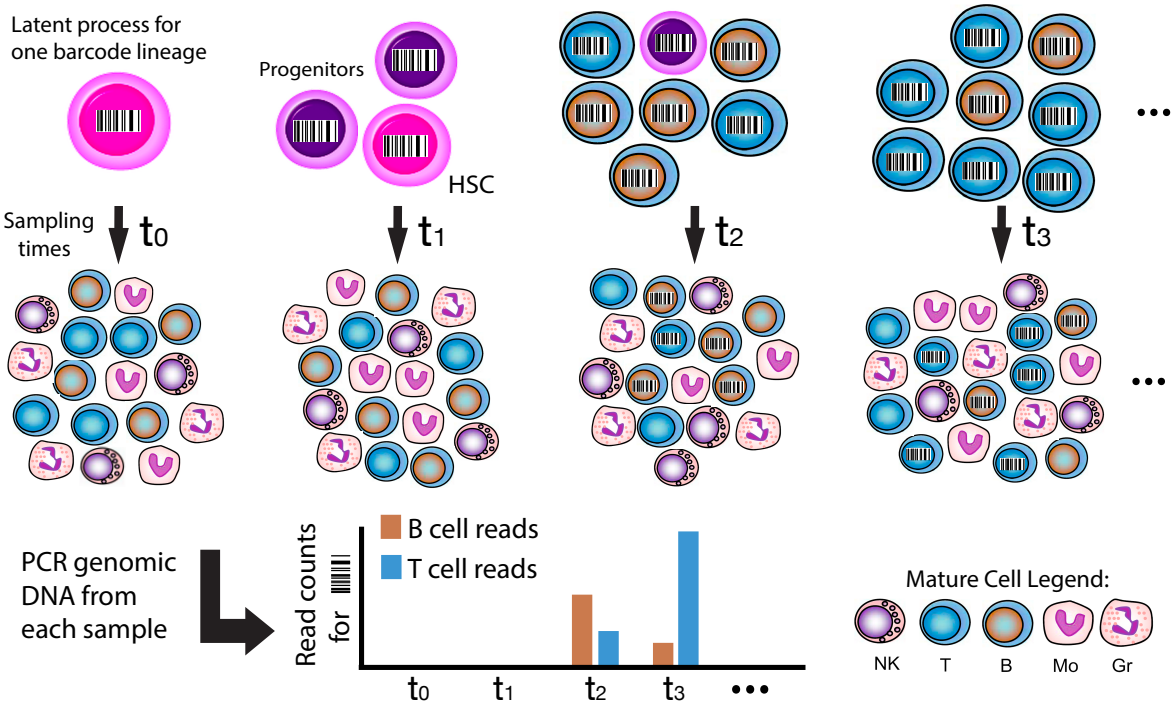


Figure 6.1: Illustration of the protocol for a fixed barcode lineage. The top panel represents the latent process starting with single HSC at several snapshots in time $t_0, \ldots, t_3$. Because HSCs and early progenitors reside in the marrow and are unobservable in blood, note that only mature cells descended from the given HSC, which are present after t_2, appear in the samples (second panel). Read counts corresponding to the given barcode after PCR and sequencing then reflect the number of cells sharing that barcode in the sample, which in turn reflect the barcoded population in the latent process.

process after transplantation corresponding to one lineage is provided in Figure 6.1.

In terms of the observed dataset, the collection of read counts for each mature cell type m can be represented as an $P \times J$ matrix $\mathbf{Y}_m = (\mathbf{Y}_m^1, \mathbf{Y}_m^2, \ldots, \mathbf{Y}_m^J)$ whose columns correspond to observation times $\mathbf{t} = (t_1, \ldots, t_J)$. Each pth row encodes the read count time series corresponding to a unique barcode ID $p \in 1, \ldots, P$ among the cell type population associated with this matrix.

To analyze the data, we first assume that the hematopoietic process evolves according to a continuous-time Markov branching process. The choice of a branching process model

is natural, as canonical differentiation trees that have been posited in the scientific literature follow such a structure, and equivalent stochastic models have been established and successfully studied in the statistical hematopoiesis literature [Kimmel and Axelrod, 2002, Catlin et al., 2011]. However, additional nontrivial challenges to modeling and inference arise from experimental protocol and incompleteness of data. Not only is the underlying process discretely observed at sampling times, but the available data at these times are only partially informative, as read counts are not directly relatable to true cell populations of each type at that time. Therefore, the model formally falls into the class of continuous-time hidden Markov models (CT-HMMs), as the latent process $\mathbf{X}(t)$ is Markov and the observed counts are independent across time points conditional on the latent process. Modeling must account for this additional uncertainty in the read data via an observation model or emission distribution of the CT-HMM. The latent process model, observation model, and estimation procedure are detailed in the following section.

6.3 Methods

When feasible, likelihood methods for CTMC model-based inference are preferable as they are most statistically efficient. However, the likelihood in our setting is doubly intractable: the observed data likelihood of the latent branching process is already computationally unwieldy—recent numerical techniques to compute this likelihood for multi-type branching processes and efforts to scale these techniques [Xu et al., 2015, Xu and Minin, 2015] fall short in our application due to the potential sizes of barcoded mature cell populations, which reach hundreds of thousands per type within a single barcode lineage. Further, marginalizing over all possible configurations of unobserved compartments and underlying cell populations that are consistent with observed reads requires an additional integration step over an enormous hidden state space. Without an expression to analytically integrate out the hidden variables, alternatives such as data augmentation are notoriously difficult when the hidden space is large. Although HMMs have been extensively studied, likelihood-based inference for HMMs is generally intractable when the state space of the hidden Markov process is

infinite or finite but massive [Cappé et al., 2006]. On the other hand, populations of HSCs and early progenitors in an individual lineage are likely to be very low and near zero, rendering approximations such as diffusion process and other continuous-space representations ill-suited.

In lieu of feasible likelihood methods, we consider inference based on the generalized method of moments, a computationally simpler alternative to maximum likelihood estimation that yields consistent estimators. This method relies on deriving equations relating a set of population moments to the target model parameters to be estimated. Next, the discrepancy between the population and sample moments are minimized to get parameters of interest. Although moment-based estimators are known to be less statistically efficient than MLEs, the choice is well-motivated for our dataset consisting of thousands of barcode lineages, each acting as an independent, identically distributed realization from the model. Recently, similar approaches have found success in application to stochastic kinetic models [Lakatos et al., 2015] and toward developing quasi- and pseudo-likelihood estimators [Chen and Hyrien, 2011].

Our estimator seeks to match pairwise empirical read count correlations across barcode lineages with their corresponding model-based population correlations. We derive explicit analytic forms for the first and second moments of a general class of branching models for hematopoiesis, allowing for the computation of marginal correlations between any two mature types. The advantage of working with correlations in the data is twofold: first, the observed correlation profiles between types are more time-varying and thus more informative than the mean and variance curves of read counts. Second, because correlations are scale invariant, we do not need to additionally model and estimate the effect of PCR amplification and fluctuations of absolute cell numbers on read counts in an already complex model. This robustness comes with a caveat — we may not expect all branching process rates to be identifiable with a scale free approach, instead requiring some parameters be fixed to provide scale information. This will be further discussed in Section 6.4.1.

With closed form moment expressions, model-based correlations can be computed very

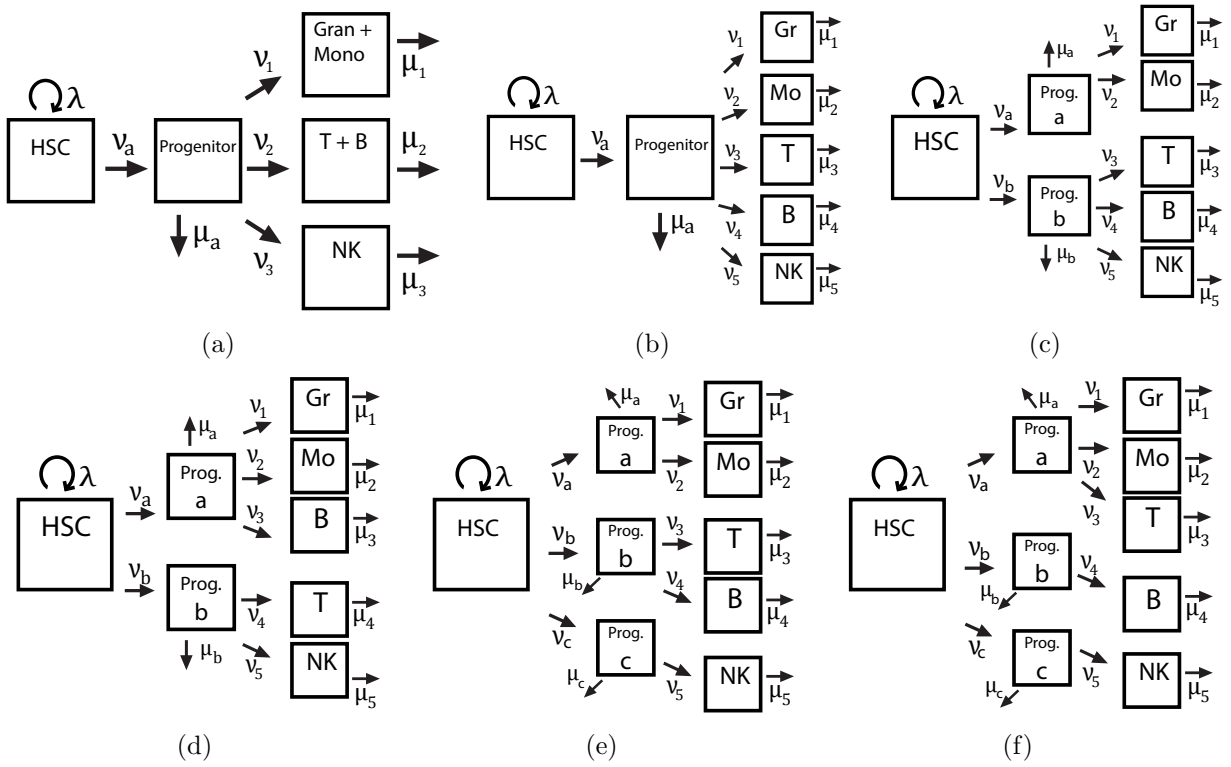


Figure 6.2: Differentiation trees to be considered in simulation study and real data analysis. In the first two models, mature cells are descended from one common multipotent progenitor: (a) groups mature cells in a model with three total mature cell compartments, and (b) assigns each mature cell type its own compartment. Note that previous statistical studies by Catlin et al. [2001], Golinelli et al. [2006], Fong et al. [2009] have modeled only the first two compartments. Models (c)—(f) include several biologically plausible topologies featuring two or three oligopotent progenitors, each specializing to produce only particular mature cells.

efficiently given any parameter setting, enabling the use of generic optimization methods to minimize a *loss function* relating the model-based correlations to observed correlations in the read count data. Throughout this section, we will often refer to the five-type branching process model of hematopoiesis depicted in Figure 6.2 (a) for clarity and ease of notation. Under the depicted model, HSCs can either self replicate or produce a more specialized progenitor cell that in turn can produce three types of mature cells separated by compartment. We will also fit richer models displayed in Figure 6.2 with additional nodes representing different types of early progenitor specialization, or more possible intermediate stages and mature types. The following derivations apply to models with arbitrarily many compartments at the progenitor level and mature cell level, enabling us to investigate arbitrary groupings of cell types and candidate branching pathways.

6.3.1 Correlation loss function

To estimate the parameters $\boldsymbol{\theta}$, we seek to match model-based correlations closely to the empirical correlations between observed read counts in the collection of distinct barcode lineages, which form many independent, identically distributed realizations of the stochastic process. This is achieved by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{t_j} \sum_{m} \sum_{n \neq m} \left[\psi^j_{mn}(\boldsymbol{\theta}; \mathbf{Y}) - \hat{\psi}^j_{mn}(\mathbf{Y}) \right]^2, \tag{6.1}$$

where ψ^j_{mn} represents model-based correlation between reads of type m and n cells at time t_j

$$\psi^j_{mn}(\boldsymbol{\theta}; \mathbf{Y}) = \rho(Y_m(t_j), Y_n(t_j)) = \frac{\mathrm{Cov}[Y_m(t_j), Y_n(t_j)]}{\sigma(Y_m(t_j))\sigma(Y_n(t_j))},$$

and $\hat{\psi}^j_{mn}$ denotes corresponding sample correlations across realizations p at time t_j

$$\hat{\psi}^j_{mn}(\mathbf{Y}) := \hat{\rho}(Y_m(t_j), Y_n(t_j) = \frac{\sum_{p=1}^{N} (Y^p_m(t_j) - \bar{Y}_m(t_j))(Y^p_n(t_j) - \bar{Y}_n(t_j))}{\sqrt{\sum_{p=1}^{N} (Y^p_m(t_j) - \bar{Y}_m(t_j))^2} \sqrt{\sum_{p=1}^{N} (Y^p_n(t_j) - \bar{Y}_n(t_j))^2}}.$$

Underlying this loss function is the system of moment equations $\left\{\psi^j_{mn}(\boldsymbol{\theta};\mathbf{Y}) = \hat{\psi}^j_{mn}\right\}$ equating theoretical normalized moments with their sample analogs at each time t_j. Because the dataset contains more observation times than parameters to be estimated, the loss function is motivated by minimizing the residuals as a nonlinear least squares objective. The problem of estimating hematopoietic events rates now translates to seeking $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta};\mathbf{Y})$. Our method is akin to an estimating equations approach; if

$$\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y}) := \boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}) - \hat{\boldsymbol{\psi}}(\mathbf{Y}),$$

where $\boldsymbol{\psi}(\boldsymbol{\theta};\mathbf{Y}), \hat{\boldsymbol{\psi}}(\mathbf{Y})$ are vectors containing all pairwise model-based and empirical correlations at each time point, respectively, then $\text{E}[\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})] \to 0$ for all $\boldsymbol{\theta}$ as $N \to \infty$. For this approach to be formally an estimating equation, we would require $\text{E}[\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})] = 0$ for all sample sizes N, a condition not satisfied by the sample correlation coefficient which exhibits finite-sample bias. The estimator $\hat{\boldsymbol{\theta}}$ is usually obtained by solving for roots of the estimating equation—in the overdetermined case, we seek the solution $\hat{\boldsymbol{\theta}}$ that minimizes $\|\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})\|_2^2$.

If $\boldsymbol{\theta}$ represents the true model parameters, the law of large numbers ensures that the sample correlations converge to the theoretical correlations as the sample size $N \to \infty$, so that $\mathcal{L}(\boldsymbol{\theta};\mathbf{Y})$ approaches 0 asymptotically. Thus, assuming identifiability of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ is a consistent estimator by continuity of $\boldsymbol{\psi}$—refer to [Van der Vaart, 2000, Chapter 2] for a detailed proof.

We further note the relationship of our estimator to the *generalized* method of moments (GMM), established by replacing the ℓ^2 norm in our loss function $\|\cdot\|_2$ by a family of norms $\|\cdot\|_W$ induced by a positive definite weight matrix $\mathbf{W}$. In the general setting, the estimator is now given by

$$\hat{\boldsymbol{\theta}}_W = \underset{\boldsymbol{\theta}}{\text{argmin}} \|\mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})\|_W^2 := \underset{\boldsymbol{\theta}}{\text{argmin}}\, \mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y})^T\, \hat{\mathbf{W}}\, \mathbf{G}_N(\boldsymbol{\theta};\mathbf{Y});$$

notice minimization of $\mathcal{L}(\boldsymbol{\theta};\mathbf{Y})$ is the special case of $\hat{\mathbf{W}} = \mathbf{I}$. The norm induced by $\mathbf{W}$ allows

different moment equations to have unequal contributions to the objective function, and its estimate $\hat{\mathbf{W}}$ from the data intuitively assigns less weight to components which have higher variance and thereby provide less information. While the method of estimating equations enjoys consistency and asymptotic normality (see Chapter 2 of [Wakefield, 2013] for details), GMM estimators $\hat{\boldsymbol{\theta}}_W$ are also asymptotically efficient under optimal choice of $\hat{\mathbf{W}}$ [Hansen, 1982]. While many algorithms exist for estimating the weight matrix $\hat{\mathbf{W}}$, the task is nontrivial [Hansen et al., 1996]. Again, introducing a weight matrix in our loss function is akin to GMM, but the moment conditions are not completely satisfied due to bias in the sample correlation: instead, each moment equation becomes unbiased in the limit as N grows. Additionally, because we have a large dataset so that finite-sample efficiency is of lesser concern and we do not expect particular time points or correlation pairs to be much more informative than others, we opt for the simpler case with $\hat{\mathbf{W}} = \mathbf{I}$, avoiding the inclusion of many additional entries of the weight matrix as parameters to be estimated.

Having established the framework for estimation here, the following sections detail the model specification and derivation of terms appearing in $\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y})$. We begin with the mathematical formulation of the stochastic compartmental model, and derive the second moments of the latent process $\mathbf{X}(t)$ using branching process techniques. While this enables us to compute model-based correlations of the branching process, we must then relate these quantities to those in the observed process $\mathbf{Y}$: we do so by next presenting an observation model that closely reflects the experimental sampling protocol using a multivariate hypergeometric distribution. Finally, the necessary quantities in Equation (6.1) are computable by connecting these model components via properties of conditional expectation, and rate estimation reduces to solving a nonlinear least squares optimization problem.

6.3.2 Stochastic branching model formulation

A branching process is a Markov process in which a collection of *independently acting* particles (cells) can reproduce and die according to a probability distribution. Here we consider a continuous-time, multi-type branching process taking values over a discrete state space of

cell counts. In this setting, each particle type has a distinct mean lifespan and reproductive probabilities, and can give rise to particles of its own type as well as other types at its time of death.

For concreteness, we introduce notation for the branching process corresponding to Figure 6.2 (a). The process is a stochastic vector $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_5(t))$ taking values in state space $\Omega = \mathbb{N}^5$, where $X_i(t)$ denotes the number of type i cells at time $t \geq 0$. Each type i cell produces j type 1 particles, l type 2 particles, m type 3 particles, n type 4 particles, and m type 5 particles at *instantaneous rates* $a_i(j,k,l,m,n)$ upon completion of its lifespan. The rate of no event occurring, beginning with one type 1 particle, is defined as $\alpha_1 := a_1(1,0,0,0,0) = -\sum_{(j,k,l,m,n)\neq(1,0,0,0,0)} a_1(j,k,l,m,n)$, with α_i defined analogously, so that $\sum_{j,k,l,m,n} a_i(j,k,l,m,n) = 0$ for $i = 1, \ldots, 5$.

Particle independence implies that the process is *linear*: overall rates are multiplicative in the number of particles. For example, in such a process, the infinitesimal probability of jumping to $\mathbf{X}(h) = (j,k,l,m,n)$ beginning with K type 1 particles over a short interval of time h is

$$\begin{aligned}\Pr_{(K,0,0,0,0),(j,k,l,m,n)}(h) &:= \Pr\{\mathbf{X}(h) = (j,k,l,m,n) | \mathbf{X}(0) = (K,0,0,0,0)\} \\ &= K \cdot a_1(j,k,l,m,n) \cdot h + o(h).\end{aligned}$$

Subsequently, offspring of each particle evolve according to the same set of instantaneous rates, and these rates $a_i(j,k,l,m,n)$ do not depend on t so that the process is *time-homogeneous*. Together these assumptions imply that each type i particle has exponentially distributed lifespan with rate $-\alpha_i$, and $\mathbf{X}(t)$ evolves over time as a CTMC [Guttorp, 1995, Chapter 3].

As depicted in Figure 6.2 (a), the parameters $\lambda, \nu_a, \mu_a, \nu_1, \nu_2, \nu_3, \mu_1, \mu_2$, and μ_3 define the infinitesimal rates of the process. The rates denoted ν_i correspond to differentiation, while μ_i denotes cell death/exhaustion; λ denotes HSC self-renewal. Specifying such a process classically using the infinitesimal generator or CTMC rate matrix is mathematically unwieldy, as this is an infinite matrix with no simplifying structure. However, in terms of branching

process rates, these event rates can now be equivalently and compactly expressed as

$$a_1(2,0,0,0,0) = \lambda, \quad a_1(0,1,0,0,0) = \nu_a, \quad a_1(1,0,0,0,0) = -(\lambda+\nu_a), \quad a_2(0,0,0,0,0) = \mu_a,$$

$$a_2(0,1,1,0,0) = \nu_1, \quad a_2 = (0,1,0,1,0) = \nu_2, \quad a_2(0,1,0,0,1) = \nu_3, \quad a_2(0,1,0,0,0) = -(\mu_a+\nu_1+\nu_2),$$

$$a_3(0,0,0,0,0) = \mu_1, \quad a_3(0,0,1,0,0) = -\mu_1, \quad a_4(0,0,0,0,0) = \mu_2,$$

$$a_4(0,0,0,1,0) = -\mu_2, \quad a_5(0,0,0,0,0) = \mu_3, \quad a_5(0,0,0,0,1) = -\mu_5,$$

with all other rates zero. Thus, the process is completely characterized by parameters in a ten-dimensional vector $\boldsymbol{\theta} = (\lambda, \nu_a, \mu_a, \nu_1, \nu_2, \nu_3, \mu_1, \mu_2, \mu_3, \pi_a)$ containing the rates and initial distribution parameter π_a representing the probability that the lineage is originally descended from a progenitor. In more general models with more than one progenitor compartment, the initial distribution is parametrized by a vector $\boldsymbol{\pi} = (\pi_a, \pi_b, \ldots)$.

6.3.3 Moments of the compartmental process

Here we derive analytic expressions for the first and second moments of the latent branching process, enabling efficient computation of model-based correlations $\psi^j(\boldsymbol{\theta}, \mathbf{Y})$ appearing in the loss function. Our approach is similar to the random variable technique introduced by Bailey [1964], but we derive expressions by way of probability generating functions rather than appealing to the cumulants. The derivation applies to the general class of models consisting of an HSC stage, progenitor stage, and mature cell stage, with arbitrary number of progenitor compartments and mature cell compartments, including all structures depicted in Figure 6.2. In this class, each mature cell type m is descended from only one progenitor compartment, so that its corresponding differentiation rate ν_m is unique and well-defined. The subscript 0 indicates rates relating to HSCs, and we use indices $a \in \mathcal{A}$ to denote progenitors, with mature cell types denoted by $m \in \mathcal{M}$. All intermediate progenitors are descended from the HSC compartment, and we use the notation $\{a \to m\}$ if progenitor a gives rise to type m mature cells, thus completely specifying a given branching model. The

total number of compartments or cell types is denoted by N, and we use the notation $\mathbf{e}_i$ to represent the vector of length N whose ith entry equals 1 and is 0 elsewhere.

From applying the process rates to the Kolmogorov backward equations, we can write *pseudo-generating functions* defined as

$$u_i(\mathbf{s}) = \sum_{k_1} \sum_{k_2} \cdots \sum_{k_N} a_i(k_1, \ldots, k_N) s_1^{k_1} s_2^{k_2} \cdot s_N^{k_N}, \tag{6.2}$$

where $\mathbf{s}$ is a vector of dummy variables. For our class of models, these are given by

$$\begin{aligned}
u_0(\mathbf{s}) &= \lambda s_0^2 + \sum_{a \in \mathcal{A}} \nu_a s_a - (\lambda + \sum_{a \in \mathcal{A}} \nu_a) s_0, \\
u_a(\mathbf{s}) &= \sum_{m \in \mathcal{M}} \nu_m s_a s_m \mathbf{1}_{\{a \to m\}} + \mu_a - \left(\mu_a + \sum_{m \in \mathcal{M}} \nu_m \mathbf{1}_{\{a \to m\}} \right) s_a \qquad \forall a \in \mathcal{A}, \\
u_m(\mathbf{s}) &= u_m(s_m) = \mu_m - \mu_m s_m, \qquad \forall m \in \mathcal{M}.
\end{aligned}$$

Next, we can write the probability generating function (PGF) of the process, beginning with one type 1 (HSC) particle, via a relation to the pseudo-generating function u_1 as follows:

$$\begin{aligned}
\phi_1(t; \mathbf{s}) &= \mathrm{E}\left[\prod_{j=1}^{N} s_j^{X_j(t)} | \mathbf{X}(0) = \mathbf{e}_1 \right] = \sum_{k_1=0}^{\infty} \cdots \sum_{k_N=0}^{\infty} \mathrm{Pr}_{\mathbf{e}_1,(k_1,k_2,\ldots k_N)}(t) s_1^{k_1} s_2^{k_2} \cdots s_N^{k_N} \\
&= \sum_{k_1=0}^{\infty} \cdots \sum_{k_N=0}^{\infty} \left[\mathbf{1}_{\{k_1=1, k_2=\ldots=k_N=0\}} + a_1(k_1, \ldots, k_N) t + o(t) \right] s_1^{k_1} s_2^{k_2} \cdots s_N^{k_N} \\
&= s_1 + u_1(\mathbf{s}) t + o(t).
\end{aligned} \tag{6.3}$$

Analogously defining ϕ_i for processes beginning with one type i particle for each $i = 1, \ldots, N$, Equation (6.3) yields the relation

$$\frac{\partial}{\partial t} \phi_i(t, \mathbf{s}) = u_i(\phi_1(t, \mathbf{s}), \ldots, \phi_N(t, \mathbf{s})).$$

Only expressions conditioning on one initial particle are required throughout, since each

latent process represents a lineage sharing one genetic barcode, which is always descended from a single marked cell. Now, let $M_{l|k}(t)$ denote the expected number of type l cells at time t, given one initial type k cell. From definition of ϕ_i, we see that we can relate the probability generating functions to these first moments via partial differentiation:

$$M_{l|k}(t) = \frac{\partial}{\partial s_l}\phi_k(t, \mathbf{s})|_{s_1=s_2=\ldots=s_N=1}.$$

Similarly, we may further differentiate the PGF to derive second moments used toward variance and covariance calculations. Define

$$U_{kl|1}(t) = \mathrm{E}\left[X_k(X_l - \mathbf{1}_{\{k=l\}})|\mathbf{X}(0) = \mathbf{e}_1\right],$$

with $U_{kl|i}(t)$ defined analogously beginning with one type i particle. Then $U_{kl|j}(t) = \left.\frac{\partial^2\phi_j}{\partial s_k \partial s_l}\right|_{\mathbf{s}=\mathbf{1}}$.

This relationship via partial differentiation enables us to write a system of differential equations governing the moments. Applying the multivariate chain rule and the Faà di Bruno formula,

$$\frac{\partial}{\partial t}M_{j|i}(t) = \frac{\partial^2\phi_i}{\partial t \partial s_j}\Big|_{\mathbf{s}=\mathbf{1}} = \sum_k \frac{\partial u_i}{\partial s_k}\frac{\partial \phi_k}{\partial s_j}\Big|_{\mathbf{s}=\mathbf{1}}, \tag{6.4}$$

$$\frac{\partial}{\partial t}U_{jk|i}(t) = \frac{\partial^3\phi_i}{\partial t \partial s_j \partial s_k}\Big|_{\mathbf{s}=\mathbf{1}} = \sum_{m=1}\left(\frac{\partial u_i}{\partial \phi_m}\frac{\partial^2\phi_m}{\partial s_j \partial s_k}\right) + \sum_{m,n=1}\left(\frac{\partial^2 u_i}{\partial \phi_m \partial \phi_n}\frac{\partial \phi_m}{\partial s_j}\frac{\partial \phi_k}{\partial s_k}\right)\Big|_{\mathbf{s}=\mathbf{1}}. \tag{6.5}$$

Notice equation (6.4) defines a system of ordinary differential equations (ODEs) determining the mean behavior, whose solutions can be plugged in to solve the second system of equations (6.5) governing second moments. These systems are subject to the initial conditions $M_{j|i}(0) = \mathbf{1}_{\{i=j\}}$, $U_{jk|i}(0) = 0$ for all i, j, k. For simplicity we introduce the notation $\kappa_{ij} = \frac{\partial u_i}{\partial s_j}\Big|_{\mathbf{s}=\mathbf{1}}$: for instance,

$$\kappa_{00} = \lambda - \sum_{a\in\mathcal{A}}\nu_a, \quad \kappa_{aa} = -\mu_a, \quad \kappa_{mm} = -\mu_m, \quad \kappa_{0a} = \nu_a, \quad \kappa_{am} = \nu_m \mathbf{1}_{\{a\to m\}} \qquad \forall a \in \mathcal{A}, m \in \mathcal{M}.$$

The system for first moments is relatively straightforward: first, the means $M_{m|m}(t)$

where $m \in \mathcal{M}$ are simply solutions to pure death equations, so that

$$M_{m|m}(t) = e^{\kappa_{mm}t} = e^{-\mu_m t}.$$

These solutions can now be substituted into simple first moment equations conditional on beginning with a marked progenitor: from (6.4), these equations are given by

$$\frac{\partial}{\partial t} M_{m|a}(t) = \kappa_{aa} M_{m|a}(t) + \mathbf{1}_{\{a \to m\}} \kappa_{am} M_{m|m}(t),$$

and upon rearrangement are of the general form

$$\frac{d}{dt} M_{m|a}(t) + P(t) M_{m|a}(t) = Q(t). \tag{6.6}$$

Such a differential equation can be solved using the integrating factor method, multiplying both sides $e^{\int P(t)dt}$ and rearranging for $M_{m|a}(t)$. Solving, we obtain

$$M_{m|a}(t) = \mathbf{1}_{\{a \to m\}} \frac{\kappa_{am}}{\kappa_{aa} - \kappa_{mm}} \left(e^{\kappa_{aa}t} - e^{\kappa_{mm}t} \right) \quad = \quad \mathbf{1}_{\{a \to m\}} \frac{\nu_m}{\mu_m - \mu_a} \left(e^{-\mu_a t} - e^{-\mu_m t} \right).$$

Next, (6.4) again gives us mean equations conditional on beginning with one marked HSC:

$$\frac{\partial}{\partial t} M_{m|0}(t) = \kappa_{00} M_{m|0}(t) + \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \kappa_{0a} M_{m|a}(t),$$

which clearly is also of the form (6.6). Thus, we can plug in the solutions we've obtained for $M_{m|a}(t)$ and solve the system using the same technique, yielding

$$\begin{aligned} M_{m|0}(t) &= e^{\kappa_{00}t} \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \frac{\kappa_{0a}\kappa_{am}}{\kappa_{aa} - \kappa_{mm}} \left(\frac{e^{(\kappa_{aa} - \kappa_{00})t} - 1}{\kappa_{aa} - \kappa_{00}} - \frac{e^{(\kappa_{mm} - \kappa_{00})t} - 1}{\kappa_{mm} - \kappa_{00}} \right) \\ &= e^{(\lambda - \sum_a \nu_a)t} \sum_{a \in \mathcal{A}} \mathbf{1}_{\{a \to m\}} \frac{\nu_a \nu_m}{\mu_m - \mu_a} \left(\frac{e^{((\sum_a \nu_a) - \mu_a - \lambda)t} - 1}{(\sum_a \nu_a) - \mu_a - \lambda} - \frac{e^{((\sum_a \nu_a) - \mu_m - \lambda)t} - 1}{(\sum_a \nu_a) - \mu_m - \lambda} \right). \end{aligned}$$

These expressions characterize the mean behavior of the system, and furthermore may now

be used toward solving for the second moments. We introduce for simplicity the additional notation $\kappa_{i,jk} := \frac{\partial^2 u_i}{\partial s_j \partial s_k}\Big|_{\mathbf{s}=\mathbf{1}}$; for instance, $\kappa_{0,00} = 2\lambda$. Further, the equations $U_{mm|m}(t) = \kappa_{mm} U_{mm|m}(t)$, and together with the initial condition are only satisfied by the trivial solution $U_{mm|m}(t) = 0$ for all final types $m \in \mathcal{M}$. Now, many terms in equation (6.5) have zero contribution, and the remaining equations in the system can be simplified to yield

$$\begin{aligned}\frac{d}{dt}U_{mn|a}(t) &= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\Big(\frac{\partial u_a}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m \partial s_n} + \frac{\partial^2 u_a}{\partial s_a \partial s_m}\frac{\partial \phi_a}{\partial s_n}\frac{\partial \phi_m}{\partial s_m} + \frac{\partial^2 u_a}{\partial s_a \partial s_n}\frac{\partial \phi_a}{\partial s_m}\frac{\partial \phi_n}{\partial s_n}\Big) \\ &= \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\left(\kappa_{aa}U_{mn|a} + \kappa_{a,am}M_{n|a}M_{m|m} + \kappa_{a,an}M_{m|a}M_{n|n}\right) \qquad \forall a \in \mathcal{A}, m \neq n \in \mathcal{M},\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}U_{mn|0}(t) &= \left(\frac{\partial u_0}{\partial s_0}\frac{\partial^2\phi_0}{\partial s_m s_n} + 2\frac{\partial^2 u_0}{\partial s_0^2}\frac{\partial \phi_0}{\partial s_m}\frac{\partial \phi_0}{\partial s_n} + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\frac{\partial u_0}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m s_n}\right)\Big|_{\mathbf{s}=\mathbf{1}} \\ &= \kappa_{00}U_{mn|0} + 2\kappa_{0,00}M_{m|0}M_{n|0} + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}}\kappa_{0a}U_{mn|a} \qquad \forall m \neq n \in \mathcal{M}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{d}{dt}U_{mm|a}(t) &= \mathbf{1}_{\{a\to m\}}\Big(\frac{\partial u_a}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m^2} + 2\frac{\partial^2 u_a}{\partial s_a \partial s_m}\frac{\partial \phi_a}{\partial s_m}\frac{\partial \phi_m}{\partial s_m} + 0\Big) \\ &= \mathbf{1}_{\{a\to m\}}\left(\kappa_{aa}U_{mm|a} + 2\kappa_{a,am}M_{m|a}M_{m|m}\right) \qquad \forall a \in \mathcal{A}, m \in \mathcal{M},\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}U_{mm|0}(t) &= \left[\frac{\partial u_0}{\partial s_0}\frac{\partial^2\phi_0}{\partial s_m^2} + \frac{\partial^2 u_0}{\partial s_0^2}\left(\frac{\partial \phi_0}{\partial s_m}\right)^2 + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\frac{\partial u_0}{\partial s_a}\frac{\partial^2\phi_a}{\partial s_m^2}\right]\Big|_{\mathbf{s}=\mathbf{1}} \\ &= \kappa_{00}U_{mm|0} + \kappa_{0,00}M_{m|0}^2 + \sum_{a\in\mathcal{A}}\mathbf{1}_{\{a\to m\}}\kappa_{0a}U_{mm|a} \qquad \forall m \in \mathcal{M}.\end{aligned}$$

Since we already have expressions for the means $M_{\cdot|\cdot}$, these equations $U_{\cdot|a}(t)$ each become a first order linear ODE and can now each be solved individually. Indeed, they again take the

form (6.6), and we find

$$U_{mm|a}(t) = \mathbf{1}_{\{a\to m\}} e^{\kappa_{aa}t} \int_0^t 2 \cdot e^{-\kappa_{aa}x} \kappa_{a,am} M_{m|a}(x) M_{m|m}(x)\, dx,$$
$$U_{mn|a}(t) = \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}} e^{\kappa_{aa}t} \int_0^t e^{-\kappa_{aa}x} \left(\kappa_{a,am} M_{n|a}(x) M_{m|m}(x) + \kappa_{a,an} M_{m|a}(x) M_{n|n}(x)\right)\, dx.$$

Replacing $\kappa_{\cdot}$ with model-based rates, we integrate and simplify these expressions to obtain

$$U_{mm|a}(t) = \mathbf{1}_{\{a\to m\}} \frac{2\nu_m^2}{\mu_m - \mu_a} e^{-\mu_a t} \left[\frac{\mu_a - \mu_m}{\mu_m(\mu_a - 2\mu_m)} - \frac{e^{-\mu_m t}}{\mu_m} - \frac{e^{(\mu_a - 2\mu_m)t}}{\mu_a - 2\mu_m}\right]$$
$$U_{mn|a}(t) = \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}} \left\{ \frac{\nu_m\nu_n}{\mu_n - \mu_a} e^{-\mu_a t} \left[\frac{\mu_a - \mu_n}{\mu_m(\mu_a - \mu_m - \mu_n)} - \frac{e^{-\mu_m t}}{\mu_m} - \frac{e^{(\mu_a - \mu_m - \mu_n)t}}{\mu_a - \mu_m - \mu_n}\right] \right.$$
$$\left. + \frac{\nu_m\nu_n}{\mu_m - \mu_a} e^{-\mu_a t} \left[\frac{\mu_a - \mu_m}{\mu_n(\mu_a - \mu_m - \mu_n)} - \frac{e^{-\mu_n t}}{\mu_n} - \frac{e^{(\mu_a - \mu_m - \mu_n)t}}{\mu_a - \mu_m - \mu_n}\right] \right\}.$$

Finally, we plug in these solutions into the differential equations beginning with an HSC governing $U_{\cdot|0}(t)$, which now take on the same general form and again can be solved by the integrating factor method:

$$U_{mn|0}(t) = e^{\kappa_{00}t} \int_0^t e^{-\kappa_{00}x} \left(\kappa_{0,00} M_{n|0}(x) M_{m|0}(x) + \sum_{a\in\mathcal{A}} \mathbf{1}_{\{a\to m\}}\mathbf{1}_{\{a\to n\}} \kappa_{0a} U_{mn|a}(x)\right)\, dx,$$
$$U_{mm|0}(t) = e^{\kappa_{00}t} \int_0^t e^{-\kappa_{00}x} \left(\kappa_{0,00} M_{m|0}^2(x) + \sum_{a\in\mathcal{A}} \mathbf{1}_{\{a\to m\}} \kappa_{0a} U_{mm|a}(x)\right)\, dx.$$

At this stage, we see that these integrals have closed form solutions as well, since their integrands only differ from the previous set of equations by including additional sums of exponentials from the $U_{\cdot|a}(t)$ expressions. We omit the integrated forms in the general case for brevity, but remark that while they appear lengthy, they are comprised of simple terms and can be very efficiently evaluated, enabling use within iterative algorithms. For completeness, we include the explicit solutions to the simplest model in the Appendix.

With closed form moment expressions in hand, we can readily recover variance and co-

variance expressions and thus calculate model-based correlations. For instance,

$$\mathrm{Cov}\left[X_4(t), X_5(t)|\mathbf{X}(0) = \mathbf{e}_1\right] = U_{45|1}(t) - M_{4|1}(t)M_{5|1}(t).$$

Because the initial state is uncertain, unconditional variances and covariances between mature types can be computed by marginalizing over the initial distribution vector $\boldsymbol{\pi}$, with details in the Appendix. We thus arrive at the marginal expressions by applying the law of total (co)variance:

$$\begin{aligned}\mathrm{Var}[X_i(t)] &= \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{i|k}^2] - \sum_{k=1}^{K} \pi_k^2 (\mathrm{E}[X_{i|k}])^2) - 2\sum_{j\neq k} \pi_j \pi_k \mathrm{E}[X_{i|j}]\mathrm{E}[X_{i|k}] \\ &= \sum_{k=1}^{K} \pi_k [U_{ii|k}(t) + M_{i|k}(t)] - \pi_k^2 M_{i|k}(t)^2 - 2\sum_{j\neq k} \pi_j \pi_k M_{i|k}(t) M_{i|j}(t). \qquad (6.7)\end{aligned}$$

$$\mathrm{Cov}[X_i(t), X_j(t)] = \sum_{k=1}^{K} \pi_k^2 (\mathrm{E}[X_{i|k}X_{j|k}] - \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|k}]) + \pi_k(1-\pi_k)\mathrm{E}[X_{i|k}X_{j|k}] - \sum_{k\neq l} \pi_k \pi_l \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|l}]. \qquad (6.8)$$

6.3.4 Observation model

Analytic expressions for the covariances and variances of the latent branching process enable calculation of pairwise correlations between mature cell type populations, but it remains to relate these expressions to the correlations between read counts, $\psi^j(\boldsymbol{\theta}; \mathbf{Y})$, appearing in our loss function. To complete the data generating model, it is necessary to specify the probability distribution of the barcode read counts conditional on the state of the branching process $\{\mathbf{X}(t)\}$. Read counts are observed between mature blood cells, and we denote these counts for cell type j corresponding to barcode p at time t by $Y_j^p(t)$. Read counts are assumed proportional to the number of blood cells with barcode p, $\widetilde{Y}_j^p(t)$, *sampled* at time t, so that $Y_j^p(t) = c_j(t) \times \widetilde{Y}_j(t)^p$ where constants $c_j(t)$ reflect the results of PCR amplification at time

t. Such a linear representation of PCR amplification is not unreasonable — applying minimum read count thresholds already ameliorate noise and nonlinearities in the amplification process. However, this has not accounted for uncertainty due to sampling: recall that at each observation time point, a fixed number of cells of each type is obtained from the blood sample. Within the purified DNA samples, a random number of barcodes is present, sampled in proportion to their prevalence in the cell population. Therefore, the distribution of sampled cells can be well-modeled by a multivariate hypergeometric distribution

$$\widetilde{\mathbf{Y}}_j(t) \mid \mathbf{X}(t) \sim \text{mvhypergeom}(N_j, \mathbf{X}_j(t), n_j), \tag{6.9}$$

where n_j is the known number of sampled type j cells, N_j is the total number of barcoded cells of type j in the animal, and $\mathbf{X}_j(t)$ again represents the underlying branching process, whose pth components contain the numbers of type j cells with barcode p. Note that n_j, N_j are known based on the experimental protocol, while $\mathbf{X}_j(t)$ is unknown. The pth component of the probability mass vector $Y_j^p(t)$ can be interpreted as the probability of drawing Y_j^p balls of color p out of an urn containing N_j total balls, $X_j^p(t)$ of which are of color p, in a sample of size n_j. In this setting, each color corresponds to a barcode lineage ID, so the mechanistic resemblance to the experimental sampling itself motivates the distributional choice.

Incorporating this observation model, computing the correlation $\psi^j(\boldsymbol{\theta}; \mathbf{Y})$ requires applying the laws of total variance and covariance to the covariance expressions obtained for the latent branching process. Conditioning the previously derived expressions moment expressions on the multivariate hypergeometric sampling distribution, we obtain the following expressions comprising $\psi^j_{mn}(\boldsymbol{\theta}; \mathbf{Y})$:

$$\begin{aligned}
\text{Cov}(Y_j, Y_k) &= \frac{n_j n_k}{N_j N_k} \text{Cov}(X_j, X_k) \\
\text{Var}(Y_j) &= \frac{n_j(N_j - n_j)}{N_j(N_j - 1)} \text{E}(X_j) - \frac{n_j(N_j - n_j)}{N_j^2(N_j - 1)} \text{E}(X_j^2) + \frac{n_j^2}{N_j^2} \text{Var}(X_j).
\end{aligned}$$

6.4 Results

6.4.1 Simulation study

To assess our methods, we examine the performance of our loss function estimator on simulated data generated from several hematopoietic tree structures in our branching process framework. Specifically, we consider models with three or five mature types with varying progenitor structures displayed in Figure 6.2. For each model, we simulate 400 independent datasets, each consisting of $20,000$ realizations representing barcoded lineages, from the continuous-time branching process model. True rates for simulating these processes were chosen such that summing over the $20,000$ barcode lineages, the total populations of each mature cell type are relatively constant after time $t = 2$, since true cell populations should be fairly constant for scientific realism. Note that while total populations are stable, individual barcode trajectories display a range of heterogeneous behaviors, with many trajectories becoming extinct and others reaching very high counts. This reflects the behavior we see in the real dataset.

From each of these synthetic datasets, we then produce an *observed dataset* by drawing samples of fixed size from the complete data according to the multivariate hypergeometric distribution, mimicking experimental sampling noise. Observations are recorded at irregular times over a two year period similar to the span and frequency of the experimental sampling schedule. Parameter estimation is then performed on these observed datasets.

To minimize the loss function objective, we use the general optimization implementation in package `nlminb`. Optimization is performed over 250 random restarts per observed dataset. We constrain rates to be non-negative, and include a simple log-barrier constraint to enforce that the overall growth of the HSC reserve is non-negative. In models with more than one progenitor cell, the initial distribution vector is constrained to a probability simplex. Rather than specifying additional hard constraints in the optimization problem, we use a multinomial logistic reparametrization so that each initial distribution parameter varies freely in $\mathbb{R}$; see Appendix for details. Finally, we remark that optimization over all free parameters

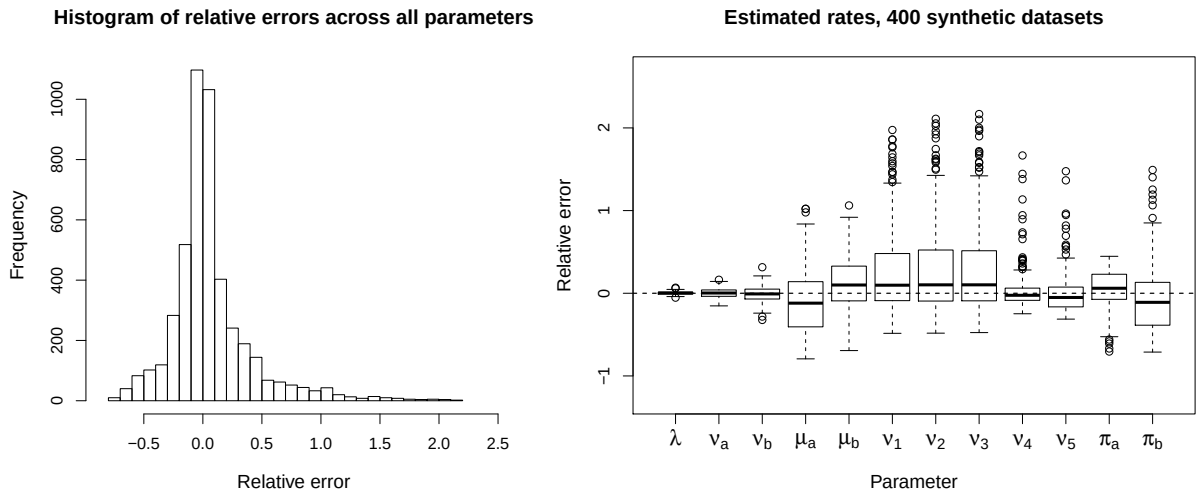


Figure 6.3: Performance of loss function estimator on synthetic data from model with five mature types and two progenitor compartments, i.e. model (c) or (d). While we see outlier influence, median estimates are accurate despite the parameter rich setting. Medians, median absolute deviations, and standard errors corresponding to the plotted estimates are included in the Appendix for completeness.

leads to mild identifiability problems—in particular, pairs of mature differentiation rates and death rates are often only identifiable up to a ratio. This is unsurprising: the correlations comprising the objective function are invariant to scale, so we would expect parameters to be distinguishable only up to a multiplicative constant. To remedy this, we choose to fix the death rates μ_i at their true value, supplying information that provides a sense of scale to infer all other parameters. Indeed, this is also justifiable in practice: mature cell types are observable in the bloodstream, and information about their behavior, i.e. average lifespans, is available in the scientific literature.

Correlation profiles from estimated parameters corresponding to the results in the tables above are displayed in Figure 6.4. Visually, we see the curves are very close to those corresponding to true parameters. We also note clear qualitative differences between models, with the two-progenitor model exhibiting two clear groupings of correlation profiles and exhibiting low and negative correlations.

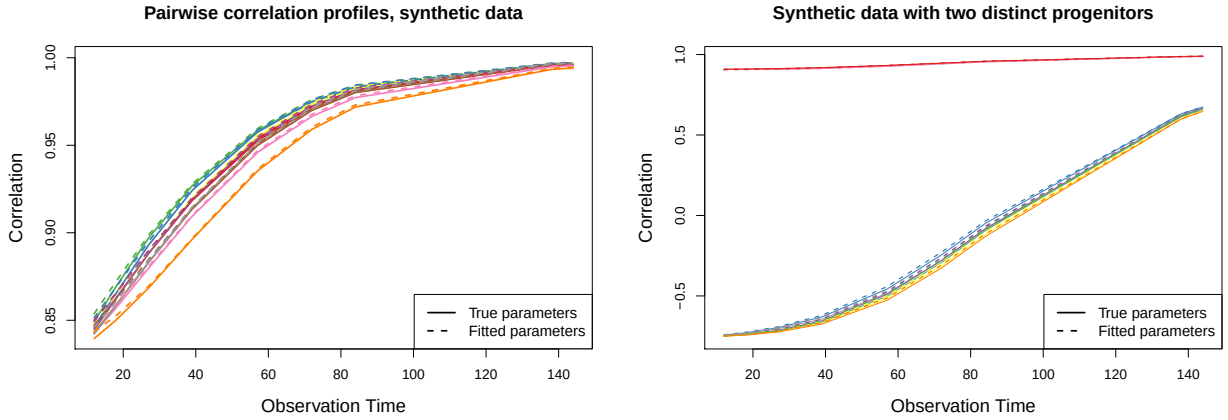


Figure 6.4: Pairwise correlation curves between five mature cell compartments descended from one common progenitor (left) or two distinct progenitors (right) calculated based on our point estimates. Solution curves from best fitting parameter estimates are almost indistinguishable from those corresponding to true parameters in both cases. Note that in the two-progenitor model, pairwise correlations among mature cell types display two distinct clusters of behavior, and that negative correlations are possible.

Model misspecification In the following simulation experiments, we examine the performance of the estimator in under- and over-specified models. We do so by fitting incorrect models, assuming the data are generated from a model with one common progenitor or with three intermediate progenitors, to the data simulated from the two-progenitor model which we have fitted in the previous section. Recall that in the true model, mature types 1,2, and 3 are descended from progenitor a, while the others are from progenitor b. Estimates reported in Figure 6.3 have near zero median relative error, and we note the median objective function value at convergence was 2.78×10^{-4}, with median absolute deviation 1.31×10^{-4} and standard deviation 2.47×10^{-4}.

The fitted correlation curves in under- and over-specified progenitor structures are displayed in Figure 6.5, with detailed tables containing estimates again included in the Appendix. We also examine the behavior when fitting a model with fewer compartments by "lumping" similar mature types together. To this end, we consider grouping mature types

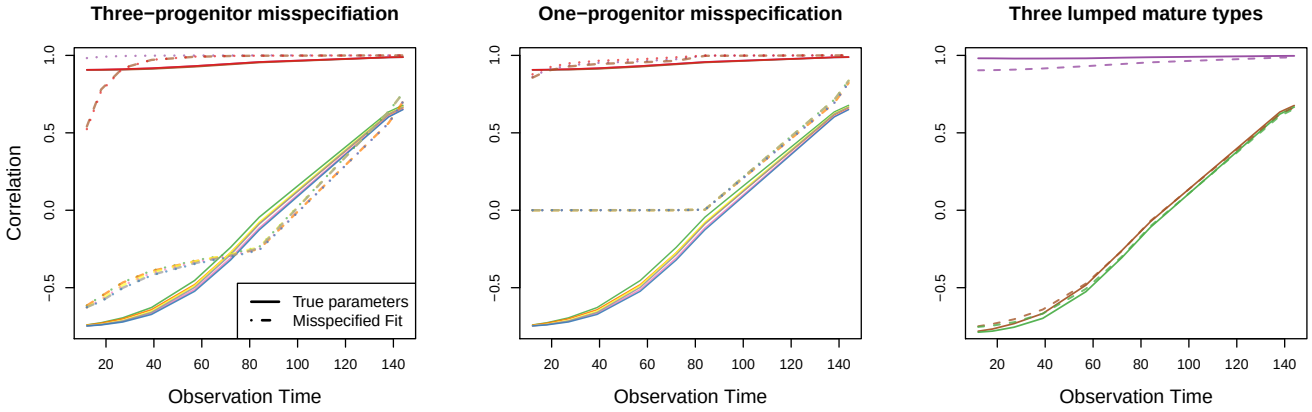


Figure 6.5: Fitted correlation curves corresponding to misspecified model estimates. Data are generated from a true model with two distinct progenitors and the true correlation profiles are the same as those displayed in the right panel of Figure 6.4. While we see a generic lack of fit in the three-progenitor model, notice that specifying one common progenitor fails to exhibit negative correlations necessary to explain the data. On the other hand, "lumping" mature compartments but properly specifying progenitor structure results in reasonable performance, as evident in the rightmost panel.

2 and 3 together, and types 4 and 5 together, thus fitting a model with three total mature cell compartments, but with a progenitor structure consistent with the true model. Results in Figure 6.5 suggest it is reasonable to group cells with shared lineages together, resulting in a much milder effect on model fit than progenitor structure misspecification. Note the objective value in Table C-7 is orders of magnitude lower than the five-type models with misspecified progenitor structures, suggesting that lumping mature types is a justifiable model simplification compared to the tradeoff of specifying a richer model with false assumptions on the intermediate structure. Such a grouping strategy can be important toward avoiding overfitting a model to real data when some degree of model misspecification is inevitable, and will be advantageous in settings where limited data requires aggregation to fit a simpler model with fewer parameters.

6.4.2 Cell lineage barcoding in rhesus macaques

Having validated our method on simulated data from the model, we are ready to analyze the data from the lineage barcoding experiments from [Wu et al., 2014]. We consider barcoding data collected from a rhesus macaque *zh33* over a 30 month period following transplantation. We consider only sampling times at which uncontaminated read data for each of the five cell types (granulocyte, monocyte, T, B, and Natural Killer) are available, and as in the original study, apply a filter so that we consider only barcode lineages exceeding a threshold of at least 1000 read counts at any time point. After restricting by these criteria, our dataset consists of 9635 unique barcode lineages with read data available at eleven unevenly spaced sampling times.

As inputs to the loss function estimator, we fix death rates, reported below, at biologically realistic parameters based on previous studies [Hellerstein et al., 1999, Zhang et al., 2007, Kaur et al., 2008]. Parameters of the multivariate hypergeometric sampling distribution are informed by circulating blood cell (CBC) data recorded at sampling times, detailed in the Appendix. These include $N_m(t)$, the total population of type m cells in circulation at time t across all barcodes, and K_m, the constant number of type m cells in the sample at each observation time. Finally, the initial barcoding level for HSCs π_1 is informed by levels of green fluorescent protein (GFP) positivity, which stabilize after 3 months. Because only HSCs have long-term regenerative capacity, the stable GFP marking level suggests the proportion of barcoded cells that were marked at the HSC stage as opposed to a later progenitor stage. While the GFP levels are observable and available to us, we will also infer π_1 independently of the GFP data in model (a) as additional validation.

We estimate the remaining rate parameters and initial barcoding distribution using the loss function estimator in all models displayed in Figure 6.2. Fitted pairwise correlation curves from estimates obtained loss function optimization with 2000 random restarts in models with one multipotent progenitor compartment are displayed in Figure 6.6: there are three such curves in the model with three mature compartments, with ten possible pairs

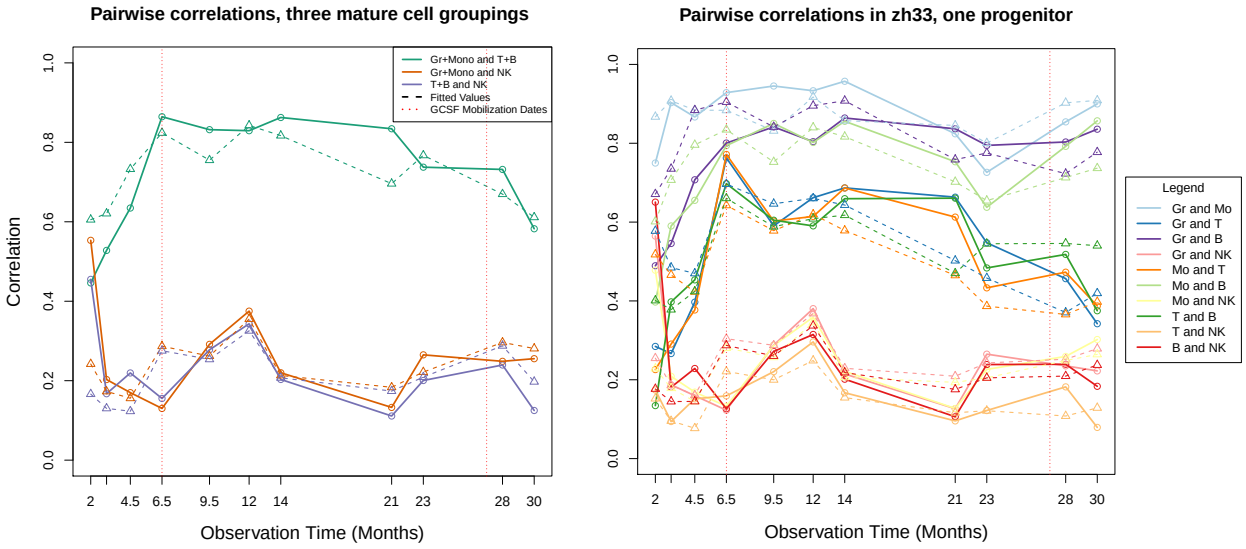


Figure 6.6: Dashed lines depict fitted correlations to read data in models (a) and (b) assuming one early progenitor compartment. GCSF mobilization dates are marked by vertical red lines. Solid lines connect the empirical correlations. Right figure does not feature complete legend for visibility; bottom cluster of four curves correspond to pairwise correlations between NK cells and other mature cells.

among the model consisting of all five mature types in the plot on the right. The raw data correlations are also displayed as solid lines, and we comment that at a qualitative level, there is visible separation into three clusters of correlation profiles among the five mature cell groups, consistent with the choice of three lumped compartments in the simpler model (a). Notably, empirical correlations between NK cells and any other cell type are significantly lower than all other pairwise correlations. This supports the main result in the pilot clustering-based analysis in the original study [Wu et al., 2014], reporting on distinctive NK lineage behavior, from a new perspective. In both plots, fitted curves successfully follow the shape of observed correlations over time, and we observe that the largest error occurs at the 6.5 month sample, coinciding with the application of granulocyte-colony stimulating factor (GCSF), a technical intervention that perturbs normal hematopoiesis in the animal. The corresponding plots for models with multiple progenitors are included in the Appendix.

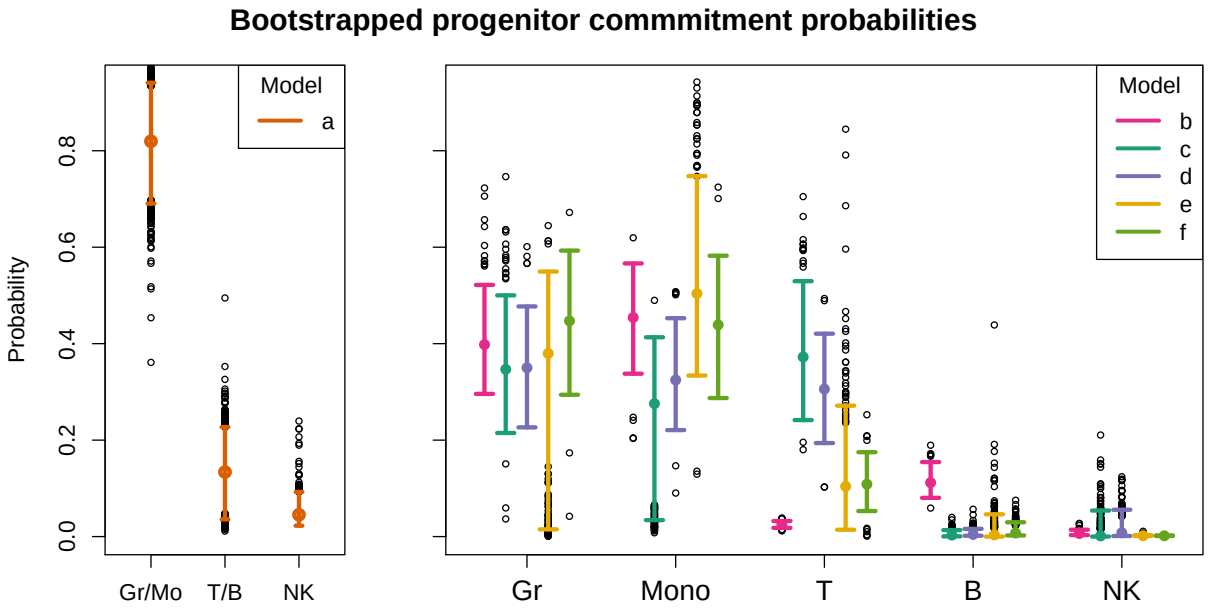


Figure 6.7: Comparison of fitted intermediate differentiation rates parametrized as fate decision probabilities. Displayed are the bootstrap estimates of normalized commitment rates to each mature compartment i, $\frac{\hat{\nu}_i}{\sum_j \hat{\nu}_j}$, in each model displayed in Figure 6.2 (a)-(f) fitted to rhesus macaque data.

Next, we display a visual comparison of intermediate differentiation rates normalized as fate decision probabilities in Figure 6.7 and fitted self-renewal rates in Figure 6.8 across models. The complete set of parameter estimates (used to generate fitted curves in Figure 6.6) and their corresponding confidence intervals are reported in the Appendix. Confidence intervals are produced via 2500 bootstrap replicate datasets. Nonparametric bootstrap resampling was performed over barcode IDs as well as over read count sampling, to account for variation across stochastic realizations and from sampling noise.

Rate estimates are parametrized as number of events per five days: for instance, the fixed death rates $\boldsymbol{\mu} = (0.4, 0.04, 0.3)$ in the lumped model correspond to half-lives of about eight days among granulocytes and monocytes, three months for T and B cells, and two weeks in NK cells. In all models with five mature compartments, we fix death rates at $\boldsymbol{\mu} = (0.8, 0.3, 0.04, 0.08, 0.4)$.

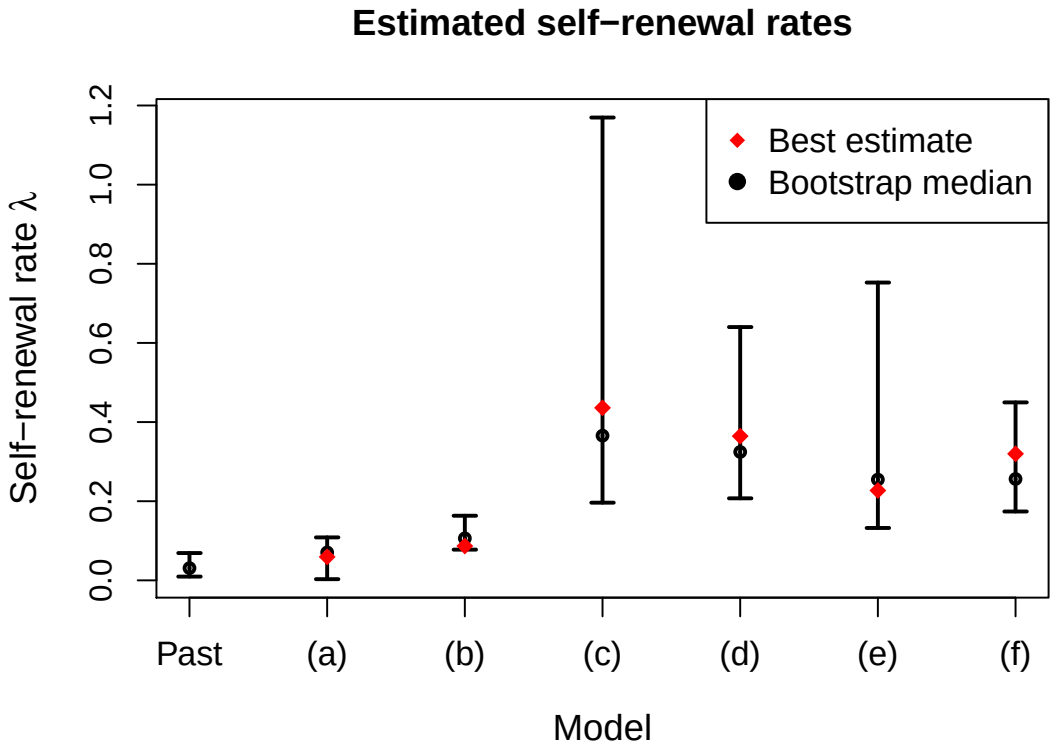


Figure 6.8: Comparison of fitted self-renewal rates $\hat{\lambda}$ and 95% confidence intervals across all models displayed in Figure 6.2 (a)-(f). Point estimates with lowest objective value (best estimates) are marked by red diamonds, while bootstrap confidence intervals and medians are plotted in black. The confidence interval around $\hat{\lambda}$ from model (a) overlaps with the interval obtained in previous telomere analyses focusing on HSC behavior in primates [Shepherd et al., 2007], while the interval from model (b) is very close and in reasonable range. The other models, which do not feature a multipotent common progenitor, result in less plausible estimated values.

Previous studies of HSC dynamics in nonhuman primates based on telomere analysis [Shepherd et al., 2007] estimate the HSC self-renewal rate at once every 23 weeks, with 11-75 week range, corresponding to an estimate of $\widetilde{\lambda} = 0.0310$, with interval $(0.0095, 0.0649)$ when translated to our parametrization. As we see in Figure 6.8, these findings coincide with our estimates and confidence intervals for $\hat{\lambda}$ in models with one multipotent progenitor compartment. While other rates pertaining to intermediate cell stages and initial barcoding level are quantities that have not been previously estimated, our results suggest that granulocytes and monocytes are produced much more rapidly than T, B and NK cells, and that individual progenitor cells are long-lived and can each produce thousands of these mature cells per day—biologically reasonable results that are newly supported from a statistical modeling perspective. Finally, we remark that the GFP data stabilize at around 13%. This level indicates the proportion of marked cells with long-term proliferative potential, suggesting the remaining 87% of barcoded cells are marked downstream at a progenitor stage. Holding out this information in Model 2(a), we estimate the initial progenitor marking level 86.1%, consistent with the GFP data as additional model validation.

In models (c)—(f) with multiple specialized, oligopotent progenitors compartments, we utilize the GFP data to fix the total progenitor marking level at 87% and estimate the proportion marked in each progenitor compartment. However, Table 9.0.7 shows that best estimates in these models lie on the boundary of the probability simplex. Along with wider confidence intervals, higher objective values, and less biologically plausible parameters, these results indicate a poorer model fit, reminiscent of the results discussed in the model misspecification experiments in Section 6.4.1.

While it may initially seem intuitive that a richer model with more compartments should result in a better fit, the models with multiple progenitors additionally assume the loss of lineage potential by restricting the types of mature cells that can be produced by each distinct progenitor. This may be a source of model misspecification. Indeed, recent studies dispute traditional assumptions about hematopoietic structures prescribing restricted differentiation pathways. For instance, Kawamoto et al. [2010] challenge the classical notion of a specialized

myeloid progenitor, showing that lymphocyte progenitors (i.e. T, B, NK) can also give rise to myeloid cells (Gr and Mono). Recent *in vitro* studies of human hematopoiesis find patterns that suggest multipotence of progenitors [Notta et al., 2015] in mature systems, and argue that oligopotent behavior is only present in early stages of development. Such oligopotent behavior in the specialization of progenitor cells is investigated in models (c)—(f), which support these recent findings in their lack of fit.

We emphasize that given fixed death rates for scale information, our method enables estimation of all remaining process rates, including those relating to unobservable intermediate progenitor stages, jointly with the initial barcoding level. The models we can consider are much more detailed and parameter-rich than those in previous statistical studies of hematopoiesis. We also note that while estimates are biologically plausible, they are obtained with an inevitable level of model misspecification, and rigorous approaches to model selection and to goodness of fit will be crucial to having more confidence in the validity of such model-based inference attempts.

6.5 Discussion

Our estimation procedure is the first method to our knowledge that enables parameter estimation in stochastic models including HSC, progenitor, and mature cell stages for time series data from hematopoietic lineage tracking experiments. Further, we show via simulation that the generalized loss function approach is very accurate when applied to data simulated from this class of models. Results from fitting experimental data have scientific bearing, newly estimating parameters such intermediate differentiation rates and initial marking levels in a multistage stochastic model. Our analysis also suggests a distinct NK cell lineage consistent with preliminary findings in [Wu et al., 2014]. While we do not provide a rigorous approach to model selection in this paper, our exploration of several models suggests that a single multipoint progenitor compartment provides a better fit to the data than models featuring loss of lineage potential, in agreement with recent *in vitro* studies of human hematopoiesis [Notta et al., 2015].

The class of models we consider and corresponding moment expressions are general in that an arbitrary number of intermediate progenitors and mature compartments can be specified, but have several limitations. First, we have three stages of cell development in our model, and future work may extend this to multiple stages. Second, the solution system only allows for each mature cell to be descended from one progenitor compartment, which makes it more difficult to investigate fully connected and nested models. Nonetheless, we are now able to perform parameter estimation in a much more detailed model than previous statistical studies, while accounting for missing information and experimental noise. Such models commonly arise in related fields such as chemical kinetics, oncology, population ecology, and epidemiology, and our methodology contributes broadly to the statistical toolbox for inference in partially observed stochastic processes, a rich area of research that still faces significant challenges.

We note that there are several additional limitations inherent to modeling hematopoiesis with a Markov branching process model. The assumptions of linearity and rate homogeneity imply a possibility of unlimited growth, and extending analysis to allow nonlinear effects such as feedback loops modulating the regulatory behavior as the system grows near a carrying capacity is merited. Similarly, the Markov assumption can be relaxed to include arbitrary lifespan distributions—age-dependent processes are one example falling under this model relaxation, and have been applied to analyzing stress erythropoiesis in recent studies [Hyrien et al., 2015]. Further details such as immigration or emigration in a random environment can also be explored in future studies: it is known that the cells we study in the peripheral bloodstream move in and out of tissue, for instance. While such extensions are mathematically difficult, they are trivial modifications to implement in simulation, and various forward simulation approaches or approximate methods such as approximate Bayesian computation (ABC) [Marjoram et al., 2003, Toni et al., 2009] may provide a promising alternative. Indeed, a Bayesian framework would allow existing prior information available from previous studies about average lifespans of mature blood cells to be incorporated without fixing these parameters.

As we have mentioned earlier, our correlation matching estimator can be generalized by including a weight matrix in the loss function allowing for different correlation profiles and time points to have unequal contributions to the objective function. Such a generalized method of moments approach is asymptotically efficient with optimal weight matrix, and in some settings, may outperform its counterpart with equal weights on the moment equations. Furthermore, the methods are extensible to related approaches such as quasi-likelihood estimation, which additionally can account for overdispersion in the mean-variance relationship [Wakefield, 2013].

Our fully generative framework and accompanying method of inference enable simulation studies and sensitivity analyses, and can be extended to develop tools for model selection. The larger scientific problem of inferring the most likely lineage pathway structure directly translates to the statistical problem of model selection. Many model selection approaches essentially build on parameter estimation techniques, balancing model complexity and goodness of fit by penalizing the number of model parameters via regularization. While model selection is generally difficult to perform in a loss function minimization framework, future work can investigate various penalization strategies applied to this class of models [Tibshirani, 1996, Fan and Li, 2001], or with shrinkage priors in a Bayesian setting [Park and Casella, 2008, Griffin et al., 2013]. Model selection using ABC is an active and rapidly developing area of research; see for instance [Toni et al., 2009, Liepe et al., 2014, Pudlo et al., 2014].

Attempts at modeling with more parameter-rich models enabling more pathways or including asymmetric division [Fong et al., 2009] should expect to be met with challenges involving overparametrization and identifiability, as well as added computational and mathematical complexity. However, such efforts and corresponding tools for model selection will be crucial in further progress toward understanding the structure of hematopoiesis. Finally, it should be noted that recent studies have challenged canonical multi-stage models of hematopoiesis [Kawamoto et al., 2010, Perié et al., 2014, Notta et al., 2015], and exploring a general class of models with model selection tools will allow a quantitative approach to inferring likely structures without using potentially misleading prior knowledge.

Chapter 7

LIKELIHOOD-BASED METHODS USING A TWO-TYPE BRANCHING APPROXIMATION TO THE GENERAL STOCHASTIC EPIDEMIC MODEL

7.1 Introduction

Mathematical and statistical analysis of infectious disease models have played a critical role in understanding the mechanisms influencing transmission, allowing scientists to better understand, predict, and make public health policy decisions about their spread. Stochastic models allow us to infer parameters relevant to the epidemic to quantify uncertainty in their variability, an advantage over deterministic models which largely provide information only about mean behaviors. Most prevalent are continuous-time, mechanistic models, including the widely used SIR model and several of its variants. Also known as the *general stochastic epidemic model*, the stochastic SIR model considers a closed population of size n, consisting of Susceptible, Infectious, and Recovered (or Removed) populations $S(t)+I(t)+R(t) = N$ that evolve over time as a Markov process. Individuals who become infected recover with a rate α, and individuals come in "close contact" with rate β— new infections occur when close contact occurs between susceptible and infected individuals. Recovered persons remain immune for the duration of the outbreak, and infected or recovered individuals cannot become susceptible again in this model. Thus, new infections cannot occur after the susceptible population is depleted, and the epidemic stops when the infected population reaches zero.

Many elaborations of this simple case have been studied— it is possible, for instance, to drop the constant population assumption, model infectious periods non-exponentially, accommodate heterogeneity in the population, and allow reinfection of recovered individuals. However, even in simple models, statistical inference is challenging: unlike the branching

processes models we have already discussed, this system does *not* satisfy particle independence and is thus nonlinear, as infection events depend on interactions between individuals of different types. Such an interacting system is much less amenable to stochastic analysis. Not only do the nonlinear dynamics inherently pose a challenge, but such epidemiological processes are almost never fully observed, so that times and exact counts of infection events are unavailable. Instead, partial counts of the observations and surrogate data such as onset of symptoms are available, requiring statistical methods to impute the missing information.

Due to the mathematical difficulties arising from nonlinear interaction between populations, mechanistic analysis of the stochastic SIR model is difficult, and the lack of an expression for transition probabilities has been a bottleneck for statistical inference. Renshaw [2011] remarks that while one can write out the Kolmogorov forward equation for the system, the "associated mathematical manipulations required to generate solutions can only be described as heroic." As we have seen in other applications, the likelihood of partially observed continuous-time Markov processes is rarely analytically available and often very difficult or infeasible to compute numerically. Instead, much work has been dedicated to likelihood-free methods including forward simulation techniques, as well as MCMC techniques relying on data augmentation [Gibson and Renshaw, 1998, ONeill and Roberts, 1999, Ionides et al., 2006]. Simulation-based approaches have the advantage of being highly flexible and can thus be used to study complex models, but are limited by the size of the augmented data, and can be too time-intensive for applications with large populations. For instance, the stochastic SIR model can be analyzed using ABC [McKinley et al., 2009], but we have already mentioned limitations of this approach. Particle filter methods can be used to analyze SIR models within maximum likelihood [Ionides et al., 2006, 2015] and Bayesian frameworks [Andrieu et al., 2010, Dukic et al., 2012, Koepke et al., 2016], but these methods are computationally very demanding and often suffer from convergence problems. Recent work [Ho et al., 2016] derives expressions for SIR transition probabilities using a continued fraction expansion approach, the state of the art for likelihood-based inference, but the method becomes computationally limited for studying large epidemics, or within data augmentation proce-

dures over a large hidden space. Furthermore, the derivation is delicate and may be difficult to extend to variations of the simple SIR model that include additional latent compartments, such as a SEIR (Susceptible-Exposed-Infected-Removed) model.

A number of approximations to the stochastic SIR model have been studied. Discrete-time simplifications have been proposed— the time-series SIR (TSIR) model is a well-known example [Finkenstädt and Grenfell, 2000], as are models by Knorr-Held and Richardson [2003] modeling the infection dynamics with an autoregressive Poisson rate. However, these simplifications also have their shortcomings, relying on the relatively strong assumption that populations are constant over each interval between observation times, that observations are regular and evenly spaced, and that observation periods are equal to the disease generation times. When examining large epidemics, it is reasonable to apply a continuous approximation to the large populations, modeling as a diffusion process with tractable solutions [Cauchemez and Ferguson, 2008]. However, such an approach is a poor proxy for the SIR model when observed counts are low.

In this chapter, we introduce a two-type branching process that enables mainstays of likelihood-based inference to fit SIR models to discretely and partially observed epidemic time-series data. We have seen in our treatment of the birth-death-shift process that a complex stochastic model can be closely approximated by a two-type branching process. The branching approximation reflects the intended dynamics of the model much more faithfully than previous analyses, which imposed the stringent assumption that at most one event can occur between observations. The rigid constraints in approximations such as the TSIR model are reminiscent of these kinds of assumptions, and motivate a similar approach using a multi-type branching process to more closely mimic SIR dynamics.

The idea of using a branching process to model epidemics is not a new one— birth-death processes have been applied to model the early stages of an epidemic, when the infective population is small compared to the number of susceptibles, justifying a single-type model of the $I(t)$ individuals [Becker, 1976, Farrington et al., 2003]. This approach models new infections as births and recoveries as deaths, but ignores the larger mechanism of exposure and

infection, and becomes ill-fit as the epidemic progresses. The birth-death approximation also allows for techniques based on coupling the epidemics with a limiting single-type branching process. This construction allows us to use branching process theory to study the extinction probabilities via criticality or sub-criticality of the branching process, and additionally gives us a rigorous definition of $R_0 := \lambda/\mu$ as the mean of its progeny distribution. This can even be used to derive a distribution of outbreak sizes in the case when $R_0 > 1$—see [Ball, 1983, Britton, 2010] for more details. Multivariate birth-death processes have been similarly applied to model early epidemics with several strains of infections [Griffiths, 1973], but essentially evolve as several copies of the birth-death approximation for each strain type—they do not utilize the additional types to more closely capture the mechanistic SIR dynamics.

7.2 SIR model and branching approximation

Our two-type branching process seeks to approximates the dynamics of the SIR model over any finite time interval by mimicking the dynamics of exposure and infection. If transition probabilities of such a process are tractable, we may use them as a proxy for the transition density of the original SIR process, allowing for approximate likelihood-based inference. Before describing the approximation, we review and define the stochastic SIR model.

The *deterministic* SIR model [McKendrick, 1926] compartmentalizes a closed population through which an infectious disease spreads into three categories: susceptible persons (S), infectious persons (I) and removed persons (R). Since the population is closed, the total population $N = S(t) + I(t) + R(t)$ at all times t. The deterministic dynamics of these three subpopulations follow a system of nonlinear ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt}(t) &= -\beta S(t)I(t), \\
\frac{dI}{dt}(t) &= \beta S(t)I(t) - \alpha I(t), \qquad (7.1) \\
\frac{dR}{dt}(t) &= \alpha I(t), \qquad (7.2)
\end{aligned}$$

where $\alpha > 0$ is the removal rate and $\beta > 0$ is the infection rate of the disease. Although this system does not have an analytic solution, it can easily be solved numerically, i.e. using an implicit Euler's method [Earn, 2008]:

$$\begin{aligned} S(t+\Delta t) &= S(t) - \beta S(t) I(t) \Delta t, \text{ and} \\ I(t+\Delta t) &= I(t) + \beta S(t) I(t) \Delta t - \alpha I(t) \Delta t, \end{aligned} \tag{7.3}$$

for sufficiently small step-size Δt. Here, the basic reproductive number is $R_0 = \beta N/\alpha$, and an epidemic occurs if $R_0 > 1$.

The deterministic model has several drawbacks: it is not suitable when the community is small [Britton, 2010]. Furthermore, it is undesirable that the model yields a threshold at which an epidemic will or will not occur: rather, it is much more realistic to discuss whether an outbreak will take off in terms of *probabilities*. A stochastic version of the model is preferable as it provides a safeguard to oversimplifying the many intricate dynamics in the spread of a disease by specifying a deterministic model. In addition, such models allow for a principled approach to uncertainty quantification.

The general stochastic epidemic model with infection rate β and recovery rate α extends the deterministic model by replacing the instantaneous rates of change by infinitesimal transition rates. The ordinary differential equations are thus replaced by the following nonzero event probabilities: as $h \to 0$,

$$\Pr(S(t+h) = x_h, I(t+h) = y_h | S(t) = x_t, I(t) = y_t)$$

$$= \begin{cases} \beta x_t y_t h + o(h) & \text{if } (x_h, y_h) = (x_t - 1, y_t + 1), \\ \alpha y_t h + o(h) & \text{if } (x_h, y_h) = (x_t, y_t - 1), \\ 1 - (\beta x_t y_t + \alpha y_t) h + o(h) & \text{if } (x_h, y_h) = (x_t, y_t). \end{cases}$$

We explicitly see the nonlinearity arising due to the interaction effect between susceptible and infected populations, evident in the xy product appearing in the probabilities above.

Because new infections occur at rate $\beta x_t y_t$ and recoveries occur with rate αy_t, it suffices to restrict attention to the $S(t), I(t)$ populations—recall that $S(t) + I(t) + R(t) = N$ so that $R(t)$ is recoverable given the susceptible and infectious populations.

7.2.1 A two-type branching process approximation

While branching processes fundamentally rely on particle *independence*, we can nonetheless make a very good approximation by mimicking the interaction effect over short time intervals. To this end, we propose a two-type branching process $\mathbf{X}(t)$ where $X_1(t)$ will represent the susceptible population and $X_2(t)$ denotes the infected population. Over any time interval $[t_0, t_1)$, we use the initial population X_{2,t_0} as a constant scaling the instantaneous rates. The only nonzero rates specifying the proposed model, in the notation introduced in Chapter 1, are

$$a_1(0,1) = \beta X_{2,t_0}, \qquad a_1(1,0) = -\beta X_{2,t_0}, \qquad a_2(0,1) = -\alpha, \qquad a_2(0,0) = \alpha. \tag{7.4}$$

This simple branching process is has instantaneous infection rate $\beta X_{2,t_0} X_1(t)$ and recovery rate $\alpha X_2(t)$ for all $t \in [t_0, t_1)$, closely resembling the true model rates, with the exception of fixing X_{2,t_0} in place of $X_2(t)$ in the rate of infection. This constant initial population fixes a piecewise homogeneous per-particle birth rate to satisfy particle independence while mimicking interactions, but notice that *both* populations are allowed to change over the time interval, offering much more flexibility than models such as TSIR and diffusion methods that assume constant populations and rates between discrete observations.

When such a model is appropriate, its simplicity and flexibility provide attractive mathematical properties toward many statistical methods. In particular, higher-order partial derivatives of the probability generating function (PGF) of this process surprisingly have closed form solutions that can be evaluated quickly and accurately. This enables us to directly compute transition probabilities given any fixed endpoints, so that numerical PGF evaluations and more advanced spectral techniques are not necessary.

7.2.2 Transition probabilities of the branching approximation

The probability generating functions of the branching process are derived via Kolmogorov equations: recall that $\phi_{mn}(t) = \phi_1(t)^m \phi_2(t)^n$, by particle independence, and the PGFs satisfy the backward equations for this system

$$\begin{aligned} \frac{d}{dt}\phi_1(t, s_1, s_2) &= \beta I_0\big(\phi_2(t, s_1, s_2) - \phi_1(t, s_1, s_2)\big) \text{ and} \\ \frac{d}{dt}\phi_2(t, s_1, s_2) &= \alpha - \alpha\phi_2(t, s_1, s_2). \end{aligned} \tag{7.5}$$

The ϕ_2 differential equation corresponds to that of a pure death process and is immediately solvable; suppressing the arguments of ϕ_2 for notational convenience, we obtain

$$\begin{aligned} \frac{d}{dt}\phi_2 &= \alpha - \alpha\phi_2 \\ \frac{d}{dt}\phi_2\left(\frac{1}{1-\phi_2}\right) &= \alpha \\ \ln(1-\phi_2) &= -\alpha t + C \\ \phi_2 &= 1 - \exp(-\alpha t + C). \end{aligned} \tag{7.6}$$

Plugging in the initial condition $\phi_2(0, s_1, s_2) = s_2$, we obtain $C = \ln(1 - s_2)$, and arrive at

$$\phi_2(t, s_1, s_2) = 1 + (s_2 - 1)\exp(-\alpha t). \tag{7.7}$$

Substituting this solution into the first differential equation and applying the integrating factor method provides

$$\begin{aligned} \phi_1 e^{\beta I_0 t} &= \int \beta I_0 e^{\beta I_0 t}\left(1 + \frac{s_2 - 1}{e^{\alpha t}}\right) dt = e^{\beta I_0 t} + \beta I_0(s_2 - 1)\int e^{(\beta I_0 - \alpha)t}\, dt \\ &= e^{\beta I_0 t} + \beta I_0(s_2 - 1)\frac{e^{(\beta I_0 - \alpha)t}}{\beta I_0 - \alpha} + C. \end{aligned} \tag{7.8}$$

Plugging in the initial condition $\phi_1(0, s_1, s_2) = s_1$ and rearranging yields

$$\phi_1 = 1 + \frac{\beta I_0(s_2 - 1)}{\beta I_0 - \alpha} e^{-\alpha t} + e^{-\beta I_0 t}\left(s_1 - 1 - \frac{\beta I_0(s_2 - 1)}{\beta I_0 - \alpha}\right). \tag{7.9}$$

Recall that transition probabilities are now related to the PGF via repeated partial differentiation; note that

$$\begin{aligned} P_{mn,kl}(t) &= \frac{1}{k!}\frac{1}{l!}\frac{\partial^k}{\partial s_1^k}\frac{\partial^l}{\partial s_2^l}\phi_{mn}(t, s_1, s_2)\Big|_{s_1=s_2=0} \\ &= \frac{1}{k!}\frac{1}{l!}\frac{\partial^k}{\partial s_1^k}\frac{\partial^l}{\partial s_2^l}\phi_1^m(t, s_1, s_2)\phi_2^n(t, s_1, s_2)\Big|_{s_1=s_2=0} \\ &= \frac{\partial^l}{\partial s_2^l}\sum_{i=0}^{k}\binom{k}{i}\frac{\partial^{k-i}}{\partial s_1^{k-i}}\phi_1^m(t, s_1, s_2)\frac{\partial^i}{\partial s_1^i}\phi_2^n(t, s_1, s_2)\Big|_{s_1=s_2=0}. \end{aligned} \tag{7.10}$$

This expression is generally unwieldy, which necessitates spectral techniques introduced in Chapter 4 in most settings. However, notice here that $\frac{\partial^i}{\partial s_1^i}\phi_2^n(t, s_1, s_2)\Big|_{s_1=0} = 0$ for all $i > 0$ in our model. Remarkably, this allows us to further simplify and ultimately arrive at closed-form expressions. Continuing, we see

$$\begin{aligned} P_{mn,kl}(t) &= \frac{\partial^l}{\partial s_2^l}\left[\binom{k}{0}\phi_2^n(t, s_1, s_2)\frac{\partial^k}{\partial s_1^k}\phi_1^m(t, s_1, s_2)\right]\Big|_{s_1=s_2=0} \\ &= \frac{\partial^l}{\partial s_2^l}\left\{\phi_2^n(t, s_1, s_2)\cdot\frac{m!}{(m-k)!}e^{-k\beta I_0 t}\left[1 + \frac{\beta I_0(s_2-1)}{\beta I_0 - \alpha}e^{-\alpha t}\right.\right. \\ &\qquad\left.\left. - e^{-\beta I_0 t}\left(1 + \frac{\beta I_0(s_2-1)}{\beta I_0 - \alpha}\right)\right]^{m-k}\right\}\Big|_{s_1=s_2=0} \\ &:= \frac{\partial^l}{\partial s_2^l}\left[\phi_2^n(t, s_1, s_2)\cdot h(t, s_1, s_2)\right]\Big|_{s_1=s_2=0} \\ &= \sum_{i=0}^{l}\binom{l}{i}\frac{\partial^{l-i}}{\partial s_2^{l-i}}h(t, s_1, s_2)\frac{\partial^i}{\partial s_2^i}\phi_2^n(t, s_1, s_2) \\ &:= \sum_{i=0}^{l}\binom{l}{i}A(l-i)B(i). \end{aligned} \tag{7.11}$$

Taking partial derivatives of $h(t, s_1, s_2)$ and of $\phi_2^n(t, s_1, s_2)$, we arrive at the following closed-

form solutions to the transition probabilities of the branching model:

Proposition 7.2.1 *The transition probabilities of the two-type branching approximation to the SIR model defined by* (7.4) *over any time interval of length* t *are given by*

$$Pr\{\mathbf{X}(t+\tau) = (k,l) | \mathbf{X}(\tau) = (m,n)\} := P_{mn,kl}(t) = \sum_{i=0}^{l} \binom{l}{i} A(l-i)B(i), \tag{7.12}$$

where

$$\begin{aligned} &B(i) = 0 \text{ for all } i \geq n, \text{ otherwise,} \\ &B(i) = \frac{n!}{(n-i)!}(1-e^{-\alpha t})^{n-i} e^{-i\alpha t} \end{aligned} \tag{7.13}$$

and

$$\begin{aligned} &A(l-i) = 0 \text{ for all } (l-i) \geq (m-k), \text{ otherwise,} \\ &A(l-i) = \frac{m!}{(m-k-(l-i))!} e^{-k\beta n t} \left[1 - \frac{\beta n}{\beta n - \alpha} e^{-\alpha t} - \left(1 - \frac{\beta n}{\beta n - \alpha}\right) e^{-\beta n t}\right]^{m-k-(l-i)} \\ &\qquad \times \left[\frac{\beta n}{\beta n - \alpha}(e^{-\alpha t} - e^{-\beta n t})\right]^{l-i}. \end{aligned} \tag{7.14}$$

This formula involving products of expressions (7.13) and (7.14) in equation (7.12) may still look unwieldy, but this sum is computed extremely quickly with a vectorized implementation, and with high degrees of numerical stability.

7.2.3 Empirical comparison to true transitions

We introduced the branching process approximation in [Ho et al., 2016] alongside the continued fraction expansion method in the same paper. The continued fraction method is slower to compute and perhaps less extensible to model variations, but provides exact expressions for SIR transition probabilities. Therefore, we will use it as a baseline to compare the approximate branching process transition probabilities.

As a brief summary, the method introduced by Ho et al. [2016] works in the Laplace domain of the forward Kolmogorov equation governing the SIR process. A recursive relationship for the Laplace transform of transition probabilities

$$f_{ab}(s) = \mathcal{L}[P_{(a_0 b_0, ab)}(t)](s) = \int_0^\infty e^{-st} P_{(a_0 b_0, ab)}(t) dt$$

can be solved so that $f_{ab}(s)$ is expressible as a continued fraction, which can be truncated to approximate $f_{ab}(s)$ with finitely many terms. Then, the transition probability can be recovered with a stable implementation of the numerical inverse Laplace transform,

$$\mathcal{L}^{-1}\left[f_{ab}(s)\right](t) \approx \frac{e^{H/2}}{2t}\mathcal{R}\left[f_{ab}\left(\frac{H}{2t}\right)\right] + \frac{e^{H/2}}{t}\sum_{k=1}^{\infty}(-1)^k \mathcal{R}\left[f_{ab}\left(\frac{H + 2k\pi i}{2t}\right)\right]. \tag{7.15}$$

We note that the following figures also appear in our work with Ho et al. [2016]. Figure 7.1 provides a comparison between methods of computing transition probabilities. Included are transition probabilities corresponding to the nine pairs of system states $\{(m,n),(k,l)\}_j$, $j = 1, \ldots, 9$, such that $P_{mn,kl}(0.5)$ is largest. Fixing these indices, we plot the set of probabilities $\{P_{mn,kl}(t)\}$ while varying t between 0.1 and 1.0. We see that transition probabilities computed using the continued fraction method under the death/birth-death model very closely match those computed empirically via simulation from the model, taken to be the ground truth. Almost all such probabilities in Figure 7.1 fall within the 95% confidence interval, while the branching process transitions follow a similar shape over time, but fall outside of the confidence intervals for many observation intervals. Figure 7.2 provides a heatmap visualization comparing the support of transition probabilities, and shows that the branching approximation is accurate with similar support to the empirical transition probabilities for a shorter time interval of length $t = 0.5$, but becomes noticeably further from the truth when we increase the observation length to $t = 1.0$. The transmission parameters for these comparisons were chosen based on rates inferred from the plague dataset in the following section, and $t = 0.5$ corresponds to an observation interval length of half a month. In many modern

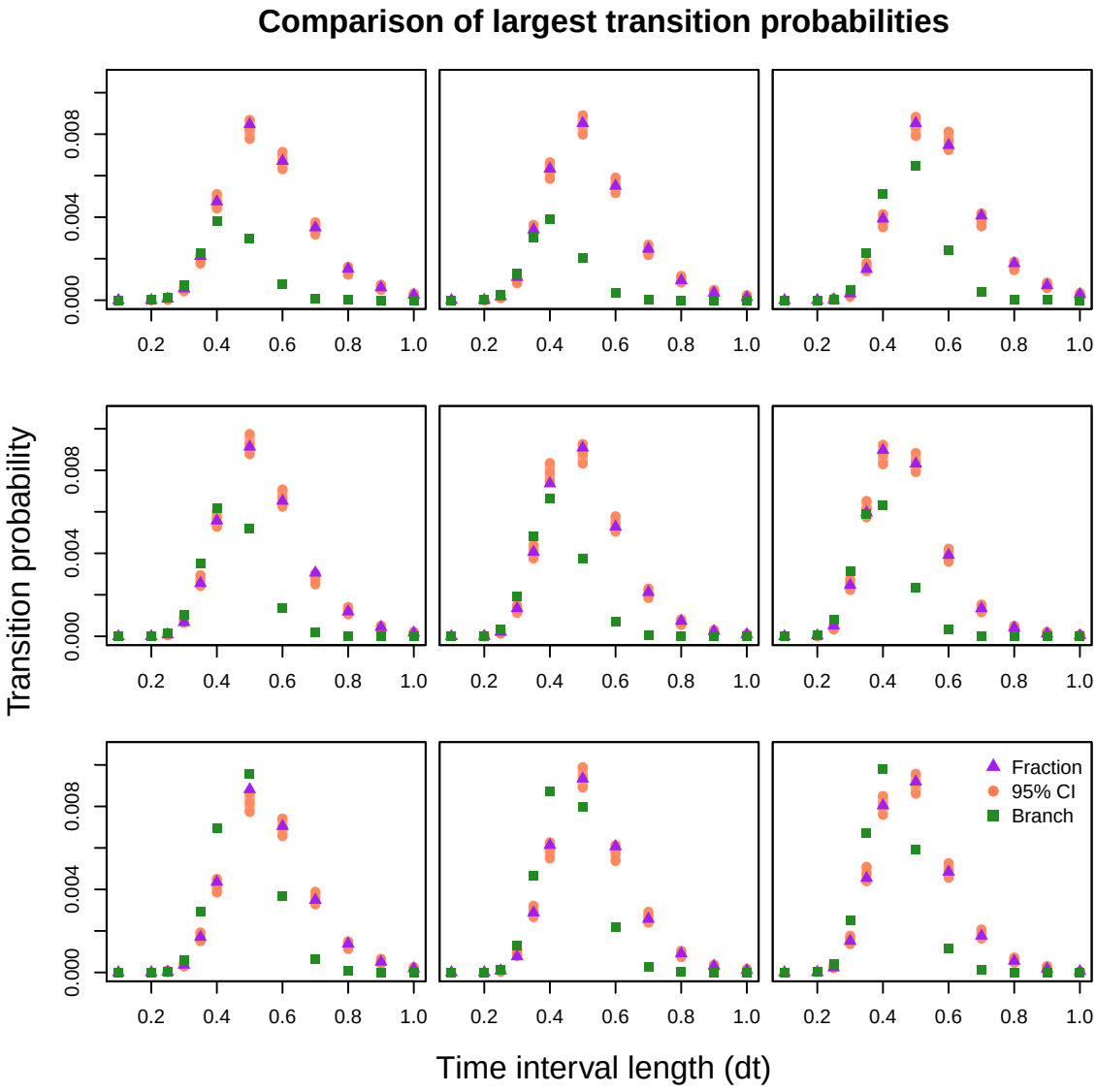


Figure 7.1: The plot above provides a comparison between methods to compute transition probabilities of the SIR model. Values of the nine pairs of states with largest transition probabilities when $t = 0.5$ are displayed, and we also plot transition probabilities as t varies from $0.1, \ldots, 1.0$, fixing these pairs of states. True parameters used to generate data are initialized at $I_0 = 15, S_0 = 110, \alpha = 3.2, \beta = 0.025$. Empirical Monte Carlo 95% confidence intervals over $150,000$ simulations from the true model are depicted in orange. Probabilities computed using the continued fraction expansion are depicted by purple triangles, while probabilities computed under the branching approximation are denoted by green squares.

datasets, data are available weekly or even daily due to improved disease surveillance. In these settings, we would expect the branching process model to offer a close approximation to the true SIR model, and thus its transition probabilities comprise a likelihood that serves as a good proxy for the intractable SIR likelihood.

7.3 Inference: the Great Plague in Eyam

We revisit the outbreak of plague in Eyam, a village in the Derbyshire Dales district, England, over the period from June 18th to October 20th, 1666. This plague outbreak is widely accepted to originate from the Great Plague of London, that killed about 15% of London's population at that time. We summarize data recording the spread of the disease [Raggett, 1982] in Table 7.1. As mentioned by Raggett [1982], this data are obtained by counting the number of deaths and estimating the infective population from the list of future deaths assuming a fixed length of illness prior to death. The susceptible population can then be computed easily assuming a closed population, an appropriate assumption as the town is isolated.

	Time (months)							
	0	0.5	1	1.5	2	2.5	3	4
Susceptible population	254	235	201	153	121	110	97	83
Infective population	7	14	22	29	20	8	8	0

Table 7.1: Susceptible and infectious population size in Eyam from June 18th to October 20th, 1666.

Raggett [1982] analyzes these data using the stochastic SIR model. The author uses a simple approximation method for the forward differential equation to obtain a point estimate $(\hat{\alpha}, \hat{\beta}) = (3.39, 0.0212)$. We revisit the data using a fully Bayesian approach.

With n observations $\{(s_k, i_k)\}_{k=1}^n$ at time $\{t_k\}_{k=1}^n$, the log discretely observed likelihood

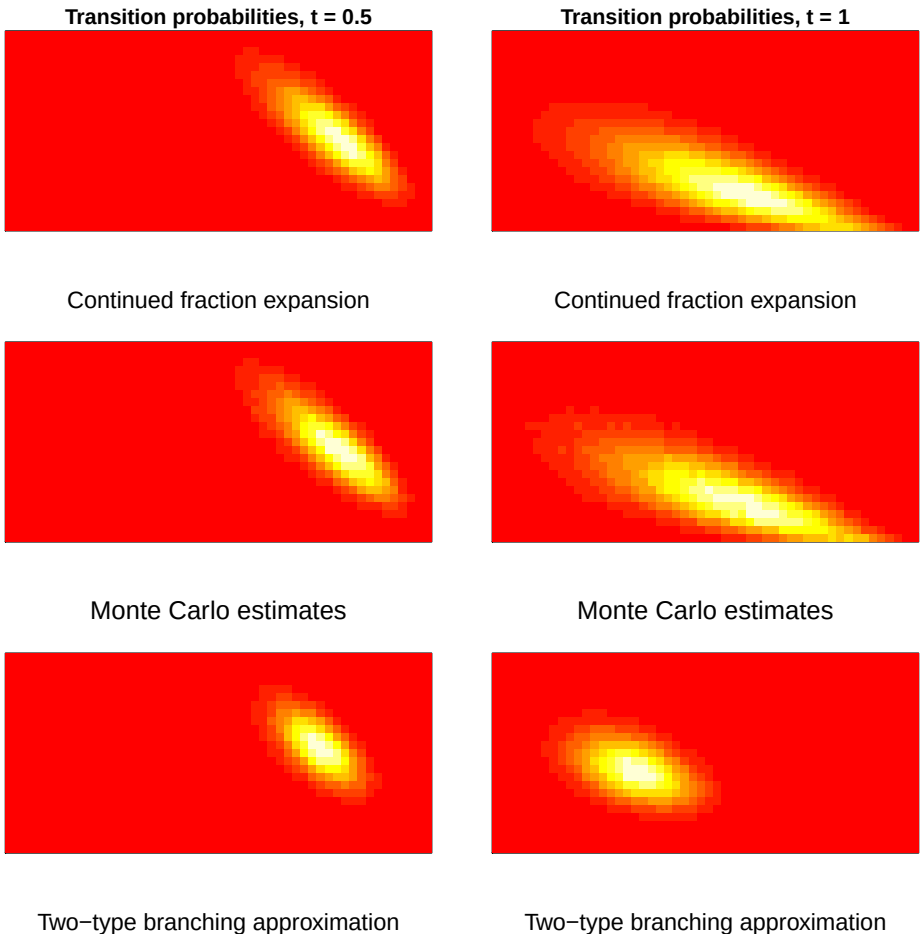
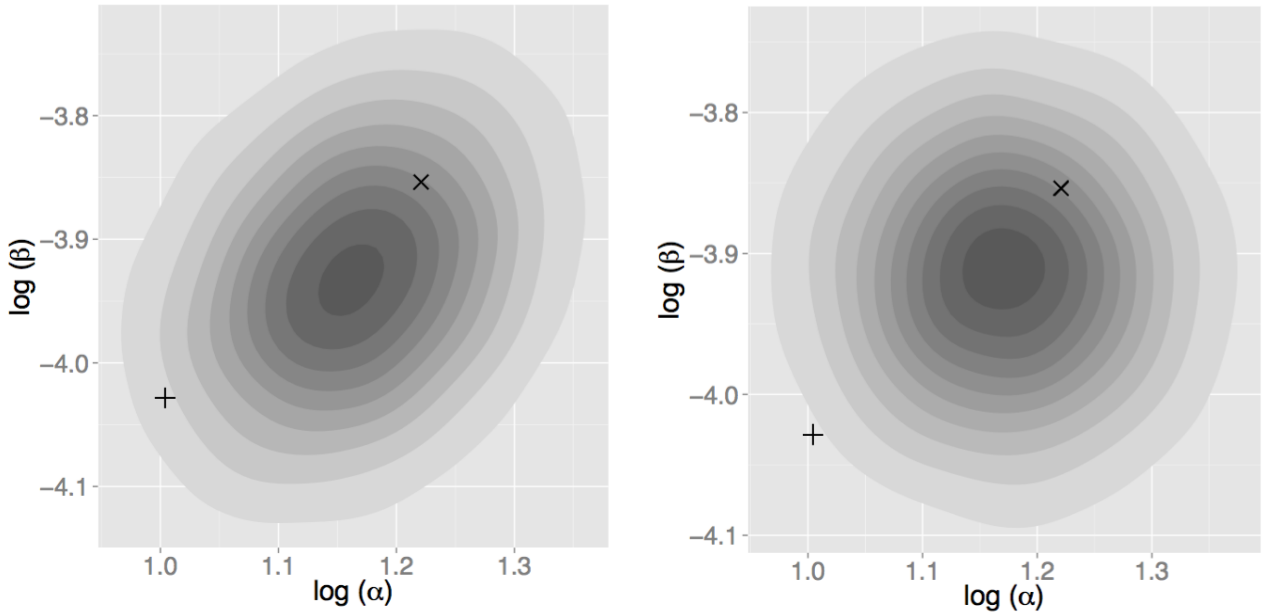


Figure 7.2: Heatmap visualizations of transition probabilities near the region of support across methods for $t = 0.5, 1$. We see that the branching approximation is noticeably different from the Monte Carlo ground truth when we increase t to 1, while the continued fraction approach remains accurate.



(a) Continued fraction method

(b) Branching approximation method

Figure 7.3: Posterior distributions (log scale) of the death rate α and the infection rate β during the plague of Eyam in 1666. The "+" symbol represents the estimate from Brauer [2008] using the deterministic SIR model, and the "×" symbol represents the Raggett's point estimate.

function is:

$$\ell_o\left(\alpha, \beta \middle| \{(s_k, i_k)\}_{k=1}^n\right) = \sum_{k=1}^{n-1} \log P_{(s_k,i_k),(s_{k+1},i_{k+1})}(t_{k+1} - t_k). \tag{7.16}$$

Because $\{S(t), I(t)\}$ is a death/birth-death process, the individual transition probabilities can be computed using the continued fraction expansion method of [Ho et al., 2016]. We compare the branching approximation to this method by making use of the approximate likelihood ℓ_b. In both cases, we proceed with inference via a Metropolis-Hastings sampler. Assuming *a priori* that $\log \alpha \sim \mathcal{N}(\mu = 0, \sigma = 100)$ and $\log \beta \sim \mathcal{N}(\mu = 0, \sigma = 100)$, we explore the posterior distribution of $(\log \alpha, \log \beta)$ using a random-walk Metropolis algorithm using `MCMCmetrop1R` from package `MCMCpack` [Martin et al., 2011]. We start the chain from Raggett's estimated value $(\log(3.39), \log(0.0212))$ and run it for $100,000$ iterations, discard-

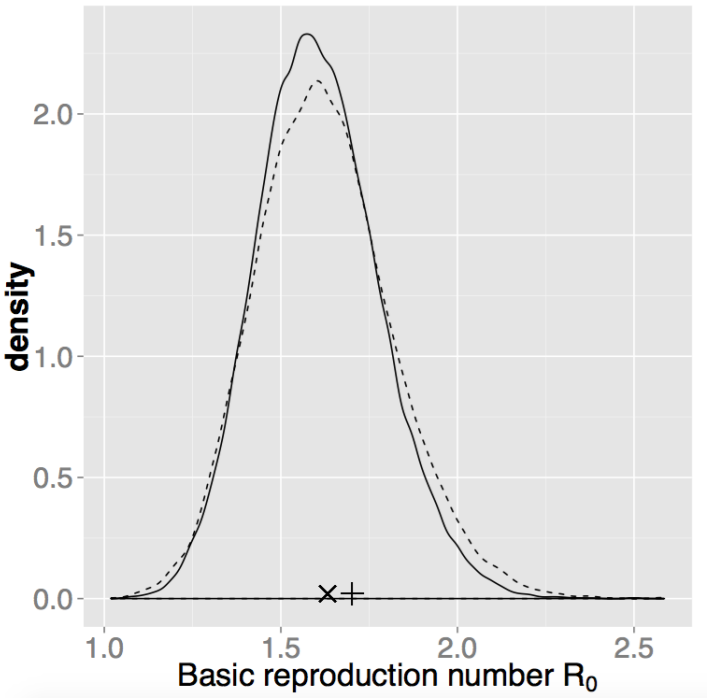


Figure 7.4: Posterior distribution of the basic reproduction number R_0 (solid line: continued fraction method, dashed line: branching approximation method). The "+", and the "×" symbols represent the estimate of R_0 from Brauer [2008], and from Raggett [1982] respectively.

ing $20,000$ iterations for burn-in. We illustrate the density of this posterior distribution in Figure 7.3(a). The posterior mean of α is 3.22 and the 95% Bayesian credible interval for α lies in $(2.69, 3.82)$. Those corresponding quantities for β are 0.0197 and $(0.0164, 0.0234)$. Under the branching approximation, the posterior mean of α is 3.237 and the 95% Bayesian credible interval for α is $(2.70, 3.84)$, while those quantities for β are 0.020 and $(0.017, 0.023)$. Although the posterior means and the 95% Bayesian credible intervals are similar to ones from the continued fraction method, we see in Figure 7.3(b) that this method fails to fully capture the posterior correlation structure between α and β. In both cases, credible intervals include the point estimate $(\hat{\alpha}, \hat{\beta}) = (2.73, 0.0178)$ from Brauer [2008] using the deterministic SIR model and Raggett's point estimate $(\hat{\alpha}, \hat{\beta}) = (3.39, 0.0212)$. A comparison of the poste-

rior distribution of the basic reproduction number R_0 between both methods is provided in Figure 7.4. The posterior mean of R_0 from the continued fraction method is 1.61 and from the branching approximation method is 1.62, while R_0 is estimated to be 1.70 in [Brauer, 2008], and 1.63 by Raggett [1982]. We see estimates are similar across methods, but remark in particular that the branching approximation estimate is very close to that under the continued fraction method which we take as a ground truth, offering an extremely efficient way to provide reasonable estimates of quantities such as R_0 despite being less accurate than the continued fraction approach.

7.4 Toward partially observed datasets and stratified populations

The two-type branching model allows for computation of an approximate discrete-data likelihood for the SIR model, enabling Bayesian inference as above as well as other likelihood-based methods similar to those introduced in previous chapters. Realistically, however, it is rare that we would fully observe the S, I, and R populations at each observation time—we may only observe removal data (deaths and recoveries) or *incidence data*, the number of new infections per period. Furthermore, the dynamics of an epidemic may vary across demographics, and we may want to consider stratified populations, for instance separating the data by gender, sex, or age groups.

This section presents ideas that are the topics of current and future work, making use of the two-type branching approximation toward developing efficient proposal densities and data augmentation schemes in these more challenging data settings. In particular, we will focus on the case of incidence data, which falls under the partially observed setting since recovery dynamics are unobserved. While terms such as transition probabilities are marginalized over all possible paths between observations (S_k, I_k) and (S_{k+1}, I_{k+1}), we now require an additional marginalization over all possible hidden populations consistent with the observed component at each time. This problem is notoriously difficult when the hidden state space is discrete and very large or high-dimensional.

Denote the incidence data $\mathbf{U} := \{U_k\}_{k=0,\ldots,K}$ with U_k the number of new infections/cases

in time period $[t_k, t_{k+1})$. Note that when we assume a closed total population size, U_k is equivalent to a decrease in susceptibles since $U_k = S_{k-1} - S_k$, so we may equivalently think of the incidence data as the case where the S_k population is discretely observed, but I_k is hidden. Unfortunately, there are many possible values of I_k that are consistent with each U_k: for a fixed period k, U_k does not correspond to a fixed initial infective population I_k, nor does it imply the value of I_{k+1} without knowing the number of recoveries occurring in $[t_k, t_{k+1})$.

As discussed in Chapter 2, data-augmented MCMC approaches provide a possible path toward inference when the latent space is too large for dynamic programming techniques for HMMs such as the Viterbi and Baum-Welch algorithms to be feasible [Rabiner, 1989]. In this case, the partial SIR likelihood given incidence data $L(\boldsymbol{\theta}; \mathbf{U}) = \pi(\mathbf{U}|\boldsymbol{\theta})$ is intractable, but we may obtain tractable forms by introducing latent or auxiliary variables $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$ so that a likelihood $L(\boldsymbol{\theta}; \mathbf{U}, \mathbf{z}) = \pi(\mathbf{U}, \mathbf{z}|\boldsymbol{\theta})$ *is* computable. For instance, if we augment by possible values of infectives at sampling times $\mathbf{z} = \mathbf{I}$, we return to the discretely-observed case. The discrete-data SIR likelihood $L(\boldsymbol{\theta}; \mathbf{U}, \mathbf{I})$ can therefore be approximated for any setting of $\mathbf{I}$ using the two-type branching process likelihood. It is also possible to augment by the complete event history, i.e. $\mathbf{z} = \{\eta_i\}, \{r_j\}$, the total set of exact infection and recovery times, respectively. In this case, the augmented likelihood $L(\boldsymbol{\theta}; \mathbf{U}, \boldsymbol{\eta}, \mathbf{r})$ is the complete-data SIR likelihood. As is always the case with CTMCs, this has a convenient and easily computable closed form

$$L(\mathbf{z}; \beta, \alpha, S_0, I_0) = \prod_{i=1}^{n} \alpha I(r_i) \prod_{j=1}^{m-1} \beta S(\eta_j) I(\eta_j) \exp\left(- \int_0^T \beta S(u) I(u) + \alpha I(u) du \right). \quad (7.17)$$

Not only is this function even faster to evaluate than the discretely-observed likelihood of the branching process approximation, but it has an exponential family form that leads to desirable properties such as conjugacy relationships allowing Gibbs parameter updates that reduce to computing sufficient statistics. Furthermore, augmenting by the complete data targets the *exact* posterior under the SIR model, rather than the approximate posterior

under the branching approximation.

In either approach, the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{U})$ is then explored via an MCMC algorithm, alternatively updating parameters and latent variables. Given a sample from the posterior, marginalization over the latent space $\boldsymbol{\mathcal{Z}}$ can be accomplished trivially by simply discarding or ignoring the $\mathbf{z}$ component of each sample.

While this data-augmentation framework is conceptually sound, it often becomes impractical due to a number of difficulties concerning the mechanism for proposing new values of latent variables. Data-augmented MCMC and psuedomarginal approaches considered for the SIR model generally rely on forward simulation , which quickly becomes computationally prohibitive in large epidemics where the number of events can be enormous [McKinley et al., 2009, Ionides et al., 2015]. Furthermore, forward simulating an epidemic is often unlikely to produce an augmented dataset compatible or consistent with the observed data, leading to poor acceptance rates and slow mixing, and in turn slow convergence. Together with a costly proposal mechanism, it becomes paramount to generate intelligent candidate values of $\mathbf{z}$.

In addition to yielding a efficient transition probability computations and a tractable discrete-data likelihood, the branching process approximation can be used to yield efficient proposal densities for data augmentation. The idea is to sample the times of infection and recovery, i.e. propose $\mathbf{z}$, based on the observed incidence data, rather than via forward simulation over the entire course of the epidemic, by exploiting a feature of the branching process model reminiscent of the well-known property of Poisson process event times. For a Poisson process with any rate, if we know that k events occured by some time T and condition on this information, then the event times are distributed as the order statistics of a uniform distribution over $[0, T)$.

Under the SIR model, the distribution of event times given the number of events that have occurred is not at all obvious. However, we find that a similar result holds for the two-type branching SIR model, relying on the following property of pure death processes due to Neuts and Resnick [1971]:

Theorem 7.4.1 *Let $\eta_1, \ldots, \eta_j, \ldots$ be the times of death in a linear pure death process with per-particle death rate μ. Given that U_k total deaths occur over a time interval of length t, the random variables $1 - e^{-\mu\eta_1}, \ldots, 1 - e^{-\mu\eta_{U_k}}$ are distributed as the order statistics of k independent uniform random variables over $(0, 1 - e^{-\mu t})$.*

Under the two-type branching model, the $X_1(t)$ component approximating the susceptible population follows a linear pure death process, and the per-particle "death rate" μ of the process is given by $X_{2,t_k}\beta$ for all $t \in [t_k, t_{k+1})$, where recall X_2 approximates the infective population. Conveniently and somewhat surprisingly, this distribution *does not* depend on the susceptible population size at the beginning of the interval.

Therefore, we can very efficiently augment the incidence data with the times of infection η_j by simply transforming a vector of uniform random variables. Not only is this update orders of magnitude faster than sampling η_j sequentially via forward simulation, but it is *guaranteed* to agree with the observed incidence data. Furthermore we expect proposed values of η_j to be reasonable since the branching model is mechanistically close to the true SIR model. On the other hand, the discrepancies in distribution of event times is corrected by the Metropolis-Hastings ratio, so that the exact SIR model posterior is still the target of inference.

Given these infection times η_j, the corresponding recovery times r_j may be proposed by adding an exponentially distributed time until recovery to each infection time, which coincides with the recovery dynamics specified in the SIR model. Together, these ingredients lead to an efficient data augmentation scheme guided mechanistically by the branching approximation, outlined below:

1. Begin with S_0, I_0, incidence data $\mathbf{U}$, and assume closed population

2. Sample infection $\eta_j = \frac{\log(1-u_j)}{-\beta I_{t_k}}$, $j = 1, \ldots, U_k$, for uniform random variables u_j drawn between $[0, 1 - e^{-\beta I_{t_k} dt})$.

3. Sample Exp(α) variables representing infectious periods, and add these values to η_j to

create recovery times $r_j, j = 1, \ldots U_k$ corresponding to each new infection.

4. Subtract the number of r_j such that $t_k < r_j < t_{k+1}$ from U_k to obtain the net infections occurring in this interval: this gives us the number of infected I_{k+1} at the beginning of next interval. Of course we also have $S_{k+1} = S_k - U_k$.

5. Repeat this process for the next interval starting with S_{k+1}, I_{k+1}.

This simple process produces augmented data $\mathbf{z}$ compatible with observations except in the case of recoveries occurring so that the epidemic ends prematurely while future incidence is still observed; this can also be conditioned at the cost of a more complicated procedure, but it is often easier to simply reject these rare cases.

Having applied these steps, we can use either $\{S_k, I_k\}$ and base inference on the discretely-observed branching likelihood, or $\{\eta_j, r_j\}$ and proceed using the complete-data SIR likelihood, embedding this augmentation step within a Metropolis-Hastings algorithm in either case. Which version outperforms in practice is not obvious and may vary depending on the given dataset.

7.5 Stratified populations

An important area of future work is the extension of these methods to SIR models whose populations are divided into multiple strata. Instead of only tracking S and I populations, we will consider $S^1, \ldots, S^m, I^1, \ldots, I^m$ populations across m strata, which may for instance represent different age groups that we expect to have different disease dynamics. Again, in a closed population, we can simply subtract from each of the closed population sizes to recover the $R^1, \ldots, R^m$ populations. In the stratified case, we still allow close contact to occur between distinct strata, so infected populations from any stratum I^i can infect any susceptible population S^j, with i not necessarily equal to j. While β represented the infection rates at close contact, we now have a set of $\{\beta_{ij}\}$ representing the transmission rate between susceptibles in stratum j and infectives from stratum i in the most general setting. Infected

persons in each stratum i then have their own recovery rate of α_i per individual. Depending on the number of strata, it is also possible to tie parameters to reduce the number of rates to be estimated. For instance setting a common $\beta_{ij} = \beta_j$ for all i, which would represent transmission parameters that vary by susceptibility across strata S^j but are independent of the characteristics of the individual transmitting the disease.

In the stratified SIR model described above, all nonzero instantaneous event probabilities for a time interval $h \to 0$, for each stratum i, are given by

$$\Pr(S(t+h) = \mathbf{x}_h, I(t+h) = \mathbf{y}_h | S(t) = \mathbf{x}_t, I(t) = \mathbf{y}_t)$$

$$= \begin{cases} (\beta_{i1} x_t^i y_t^1 + \ldots + \beta_{im} x_t^i y_t^m)h + o(h) & \text{if } (x_h^i, y_h^i) = (x_t^i - 1, y_t^i + 1), \mathbf{x}_h^{-i} = \mathbf{x}_t^{-i}, \mathbf{y}_h^{-i} = \mathbf{y}_t^{-i}, \\ \alpha_i y_t^i h + o(h) & \text{if } y_h^i = y_t^i - 1, \mathbf{x}_h = \mathbf{x}_t, \mathbf{y}_h^{-i} = \mathbf{y}_t^{-i}, \\ 1 - \sum_{i=1}^m \left(\alpha_i y_t^i + \sum_{j=1}^m \beta_{ij} x_t^i y_t^j \right) h + o(h) & \text{if } \mathbf{x}_h = \mathbf{x}_t, \mathbf{y}_h = \mathbf{y}_t. \end{cases}$$

where $\mathbf{x}_t^{-i}$ denotes all components of $\mathbf{x}_t$ except for the ith entry. Quite analogously to the case of the simple SIR model, there are still only two nonzero instantaneous event rates per stratum, not including the stratum's contribution to the remaining rate of no event occurring: a susceptible individual in stratum i becomes infected, or an infected individual in stratum i recovers. Interaction between strata only enters the dynamics via the *total force of transmission* to individuals in stratum i from *all* infectives across all other strata, yielding the simple additive transmission rate above consisting of the sums of nonlinear interaction terms that appear analogously in the simple SIR model. Importantly, however, there is no event rate where $\mathbf{x}$ changes in the j component, while $\mathbf{y}$ changes in the i component, for $i \neq j$.

Because the two-type branching approximation to the SIR model decouples precisely these interactions appearing in the transmission rate, derivations for its transition probabilities extend straightforwardly to a $2m$-type branching approximation in the stratified case. This process can be denoted $\mathbf{X}(t) = (X_1^1(t), \ldots, X_1^m(t), X_2^1(t), \ldots, X_2^m(t))$ where again

X_1^j mimics the susceptible populations and X_2^j represents the infectives. We replace each nonlinear term $\beta_{ij}S^i(t)I^j(t)$ appearing in the total transmission rate of the stratified SIR model by a linear counterpart $\beta_{ij}X_1^i(t)X_{2,t_0}^j$ for all $t \in [t_0, t_1)$. Again, X_{2,t_0}^j is the initial population I_0^j of infectives in stratum j at the beginning of the time interval t_0, which decouples nonlinearities by replacing the X_2^j component by a constant, and in turn yielding mathematical tractability via particle independence. Furthermore, because nonlinearity only enters through the overall transmission rate, but the stratified populations do not interact otherwise (i.e. there is no instantaneous event where S^i decreases but I^j increases for $i \neq j$), transition probabilities factor across strata, and can be computed as a product over known expressions in each group. That is, for two states of the process $\mathbf{x} = (x_1^1, \ldots, x_1^m, x_2^1, \ldots, x_2^m)$ and $\mathbf{y} = (y_1^1, \ldots, y_1^m, y_2^1, \ldots, y_2^m)$, we have

$$P_{\mathbf{x},\mathbf{y}}(t) = \prod_{i=1}^{m} P_{x_1^i, x_2^i \to y_1^i, y_2^i}(t),$$

where the per-stratum transition probabilities $P_{x_1^i, x_2^i \to y_1^i, y_2^i}(t)$ are derived analogously to the unstratified case, replacing βI_0 by $\sum_j \beta_{ij} I_0^j$ and α by α_i in the backward system in equation (7.5).

Given that the branching model and transition probabilities extend quite naturally to the stratified case, the methodologies described for the SIR model also extend in theory. Whether complicated data augmentation strategies will perform well in practice within higher-dimensional stratified models will be an important future direction, with much relevance to epidemiological studies.

Chapter 8

DISCUSSION AND FUTURE DIRECTIONS

This thesis has provided several novel contributions toward statistical inference of partially observed Markov branching processes. In Chapter 4, we presented spectral techniques to compute transition probabilities and restricted moments of multi-type branching processes, newly enabling likelihood-based inference via an EM algorithm or direct maximization of the discrete-data likelihood. The methodology was introduced in the context of a two-type branching process approximation to the birth-death-shift process, a model for transposon evolution with a complex state space whose discrete-data likelihood function was previously considered intractable. However, its utility applies generally to processes with an arbitrary number of types; the following chapters were therefore dedicated to computational considerations in settings with very large populations or numbers of particle types.

Chapter 5 demonstrated that the computational time required for these spectral generating function methods can be reduced logarithmically when transition probabilities are sparse, by recasting the problem in a compressed sensing framework. The resulting convex optimization problem is significantly easier to solve using standard algorithms; we demonstrate orders of magnitude improvements in runtime using an implementation of proximal gradient descent. Notably, we showed that transition probabilities of a two-type stochastic compartmental model of hematopoiesis are now tractable, while past studies have relied on simulation, estimating equations, or computationally intensive methods such as reversible jump MCMC to bypass likelihood computations.

While compressed sensing techniques help to overcome computational obstacles in systems with high population counts, they are not enough to make feasible models which additionally feature many particle types and datasets with a large collection of independent

realizations of the process. This motivates the moment-based approach in Chapter 6; although such methods are less statistically efficient than likelihood-based inference, the large number of processes becomes advantageous rather than computationally burdensome from this perspective. We introduced a loss function estimator based on second moments that successfully achieves accurate estimates in parameter-rich, multi-compartmental models, with application to fitting single-cell lineage data from recent hematopoiesis experiments.

Chapter 7 built upon the likelihood-based methods presented in Chapter 4 from a complimentary angle to the previous chapters, which target large and complex models, instead exploring more efficient ways to obtain the quantities we compute in Chapter 4 for *simpler* processes. We demonstrated that a two-type branching process can closely approximate the SIR model while admitting explicit closed form solutions for its transition probabilities, which enable much more efficient likelihood computations than a recent technique relying on continued fraction expansion. Parameter estimates from discretely observed epidemic data using the branching process approximation are very comparable to those obtained under the exact SIR model.

The methodological advances presented in this thesis have bearing on a broad range of scientific disciplines, as branching processes have found use in a plethora of fields. We have explored a glimpse of this here, producing new insights in several scientific applications. We provided a covariate-specific analysis of within-host IS*6110* transposon evolution, revealing an order of magnitude difference in the estimated shift rate than previous studies that impose restrictive assumptions. The loss function estimator was driven by time-series data generated by recent single-cell lineage barcoding experiments, whose high resolution suggests the ability to fit more complex, richer models than previous studies, enabling us to answer long-standing questions about the structure of differentiation patterns in advanced stages beyond the stem cell level. Our corresponding statistical methodology provides the first estimator to our knowledge for rate inference in stochastic compartmental models to such time-series data. Findings are consistent with past statistical studies and other scientific studies of hematopoiesis, and newly provide estimates for previously uninferred intermediate rates in a

general class of branching models. In addition to accurately inferring important epidemiological quantities such as transmission parameters and the basic reproduction number, efficient methods in the final chapter suggest the potential for likelihood-based inference in very large epidemics that are usually studied using a number of strict model simplifications.

A number of open questions and avenues for future directions follow naturally from these contributions; we mention some of them here.

Theoretical guarantees for branching approximations Both the branching approximation to the birth-death-shift process in Chapter 4 and the two-type SIR approximation in Chapter 7 are validated empirically, but lack theoretical results guaranteeing the fidelity of approximation. Work toward bounding a notion of distance between the true process and branching approximations is thus an important avenue for future work. We note that coupling bounds have been derived for the total variation distance between birth-death processes and approximations obtained by truncating the infinite-dimensional generator [Crawford et al., 2016]; these ideas may be useful toward multivariate settings. Coupling arguments have also been applied to study univariate branching process approximations for early stages of stochastic epidemics, showing that they agree up to some random point—see Ball [1983] for details. Such bounds are not only informative of when approximations break down, but may motivate alternate constructions to study various model properties in these cases: for instance, the Sellke construction [Sellke, 1983] enables us to study the final size of an epidemic that grows beyond the size well-approximable by a univariate branching process, and similar approaches may extend to the multi-type case.

Inhomogeneity and nonlinearity The mathematical developments in this thesis fundamentally rely on the assumption that rates parametrizing the branching process are time-homogeneous, and on the particle independence property of branching processes, which implies rate linearity in the size of each population. Inhomogeneous processes can be more appropriate in some cases, for instance when age-dependence of particles should be considered,

but pose significant difficulties both toward mathematical treatment and exact simulation. There has been some success in developing methods for these cases, such as pseudo- and quasi-likelihood estimators [Chen and Hyrien, 2011], but working toward likelihood-based inference requires further development. We note that a mathematical treatment of general birth-death processes whose rates need not be linear [Crawford and Suchard, 2014] has found success in a number of applications to data, and similar results may be obtained in the multi-type case. Moving beyond the realm of branching processes would enable the consideration of system-level regulatory behaviors such as feedback loops, which violate the particle independence assumption but may offer a more realistic model in some applications including hematopoiesis.

Simulation-based methods Monte Carlo simulation is typically used to compute the quantities we address in this thesis for systems that fall outside the class of processes to which our methodology applies. The ideas introduced here can potentially improve such simulation approaches. For instance, compressed sensing techniques introduced in Chapter 5 may be combined with τ-leaping [Cao et al., 2006], a method that discretizes a continuous-time process to skip computations over small intervals when simulation of large systems can be prohibitively slow. Compressed sensing methods can enable us to sequentially compute densities over time intervals short enough to induce sparsity in the support of transition probabilities, yet much longer than those necessary for τ-leaping. Samples can then be drawn from these densities repeatedly at negligible computational cost at much fewer discrete grid points, with the potential to directly improve methods such as Sequential Monte Carlo (SMC), which simulate and iteratively reweight a set of particles to empirically estimate the posterior density [Doucet et al., 2000]. Broadly, mechanistic information from transition probabilities can be used toward designing intelligent prior distributions or proposal densities toward improving MCMC methods, as we've discussed at the end of Chapter 7 for data augmented MCMC in large, partially observed epidemics. The loss function estimator introduced in Chapter 6 offers a fast way to estimate parameters that can then be

used to initialize a more complicated sampler, and is useful toward implementing approximate Bayesian computation (ABC) methods [Marjoram et al., 2003, Toni et al., 2009]. Akin to method of moments estimators from a Bayesian perspective, ABC methods base inference on low-dimensional summaries of the data—for instance the second moments used in our loss function estimator—but only rely on forward simulation from the model, allowing more complex models to be considered.

Model selection Within a likelihood-based framework, standard criteria such as AIC, BIC, and Bayes Factors apply for quantitatively choosing an optimal model while penalizing the number of parameters [Burnham and Anderson, 2002, Kass and Raftery, 1995], but do not apply to moment-based methods such as our loss function estimator in Chapter 6. Developing rigorous model selection techniques for such methods is an important future direction—indeed, in the application to hematopoiesis, the larger scientific problem of lineage pathway inference translates a model selection problem. Lasso-type penalties [Tibshirani, 1996] or more aggressive penalization methods such as Smoothly Clipped Absolute Deviation (SCAD) penalization [Fan and Li, 2001] are possible strategies to choose a rich enough model while avoiding overfitting. For Bayesian approaches, shrinkage priors offer an analogous approach toward penalty-based model selection strategies [Polson and Scott, 2010, Griffin et al., 2013, Bhattacharya et al., 2015, Carvalho et al., 2010]. Model selection for ABC is an active area of research, as Bayes Factors can be misleading and successful summary statistics can be distinct from those useful toward inference [Robert et al., 2011]; recent ideas borrowing from methods in machine learning provide a promising alternative [Pudlo et al., 2014].

BIBLIOGRAPHY

JL Abkowitz, ML Linenberger, MA Newton, GH Shelton, RL Ott, and P Guttorp. Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. Proceedings of the National Academy of Sciences, 87(22):9062–9066, 1990.

G Alsmeyer. On the Galton-Watson predator-prey process. The Annals of Applied Probability, pages 198–211, 1993.

C Andrieu, A Doucet, and R Holenstein. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(3):269–342, 2010.

NTJ Bailey. The Elements of Stochastic Processes; with Applications to the Natural Sciences. New York: Wiley, 1964.

F Ball. The threshold behaviour of epidemic models. Journal of Applied Probability, pages 227–241, 1983.

A Beck and M Teboulle. Gradient-based algorithms with applications to signal recovery. Convex Optimization in Signal Processing and Communications, 2009a.

A Beck and M Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009b.

AJ Becker, EA McCulloch, and JE Till. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature, 197:452–454, 1963.

N Becker. Estimation for an epidemic model. Biometrics, pages 769–777, 1976.

A Bhattacharya, D Pati, NS Pillai, and DB Dunson. Dirichlet-Laplace priors for optimal shrinkage. Journal of the American Statistical Association, page in press, 2015.

C Biémont. A brief history of the status of transposable elements: From junk DNA to major players in evolution. Genetics, 186(4):1085–1093, 2010.

F Brauer. Compartmental models in epidemiology. In Mathematical epidemiology, pages 19–79. Springer, 2008.

T Britton. Stochastic epidemic models: a survey. Mathematical Biosciences, 225(1):24–35, 2010.

KP Burnham and DR Anderson. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media, 2002.

JC Butcher. The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods. Wiley-Interscience, 1987.

EJ Candès. Compressive sampling. In Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures, pages 1433–1452, 2006.

EJ Candès. The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique, 346(9):589–592, 2008.

EJ Candès and T Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 51(12):4203–4215, 2005.

Y Cao, DT Gillespie, and LR Petzold. Efficient step size selection for the tau-leaping simulation method. The Journal of chemical physics, 124(4):044109, 2006.

O Cappé, E Moulines, and T Rydén. Inference in hidden Markov models. Springer, New York, USA, 2006.

CM Carvalho, NG Polson, and JG Scott. The horseshoe estimator for sparse signals. Biometrika, 97:465–480, 2010.

SN Catlin, JL Abkowitz, and P Guttorp. Statistical inference in a two-compartment model for hematopoiesis. Biometrics, 57(2):546–553, 2001.

SN Catlin, L Busque, RE Gale, P Guttorp, and JL Abkowitz. The replication rate of human hematopoietic stem cells in vivo. Blood, 117(17), 2011.

A Cattamanchi, PC Hopewell, LC Gonzalez, DH Osmond, L Masae Kawamura, CL Daley, and RM Jasmer. A 13-year molecular epidemiological analysis of tuberculosis in San Francisco. The International Journal of Tuberculosis and Lung Disease, 10(3):297–304, 2006.

S Cauchemez and NM Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. Journal of the Royal Society Interface, 5(25):885–897, 2008.

R Chen and O Hyrien. Quasi- and pseudo-maximum likelihood estimators for discretely observed continuous-time Markov branching processes. Journal of statistical planning and inference, 141(7):2209–2227, 2011.

C Colijn and MC Mackey. A mathematical model of hematopoiesisi. periodic chronic myelogenous leukemia. Journal of Theoretical Biology, 237(2):117–132, 2005.

FW Crawford and MA Suchard. Transition probabilities for general birth–death processes with applications in ecology, genetics, and evolution. Journal of Mathematical Biology, 65 (3):553–580, 2012.

FW Crawford and MA Suchard. Birth-death processes. arXiv preprint arXiv:1301.1305v2, 2014.

FW Crawford, VN Minin, and MA Suchard. Estimation for general birth-death processes. Journal of the American Statistical Association, 109(506):730–747, 2014.

FW Crawford, TC Stutz, and K Lange. Coupling bounds for approximating birth–death processes by truncation. Statistics & probability letters, 109:30–38, 2016.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39 (1):1–38, 1977.

DL Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52(4):1289–1306, 2006.

KS Dorman, JS Sinsheimer, and K Lange. In the garden of branching processes. SIAM review, 46(2):202–229, 2004.

CR Doss, Ma Suchard, I Holmes, MM Kato-Maeda, and VN Minin. Fitting birth–death processes to panel data with applications to bacterial DNA fingerprinting. The Annals of Applied Statistics, 7(4):2315–2335, 2013.

A Doucet, S Godsill, and C Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and computing, 10(3):197–208, 2000.

V Dukic, H F Lopes, and NG Polson. Tracking epidemics with Google flu trends data and a state-space SEIR model. Journal of the American Statistical Association, 107(500): 1410–1426, 2012.

DJD Earn. A light introduction to modeling recurrent epidemics. In Mathematical epidemiology, pages 3–17. Springer, 2008.

J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360, 2001.

CP Farrington, MN Kanaan, and NJ Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics, 4(2):279–295, 2003.

BF Finkenstädt and BT Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. Journal of the Royal Statistical Society: Series C (Applied Statistics), 49(2):187–205, 2000.

Y Fong, P Guttorp, and J Abkowitz. Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. The Annals of Applied Statistics, 3(4):1695–1709, 12 2009.

S Gagneux, K DeRiemer, T Van, M Kato-Maeda, BC de Jong, S Narayanan, M Nicol, S Niemann, K Kremer, MC Gutierrez, et al. Variable host–pathogen compatibility in Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America, 103(8):2869–2873, 2006.

A Gerrits, B Dykstra, OJ Kalmykowa, K Klauke, E Verovskaya, MJC Broekhuis, G de Haan, and LV Bystrykh. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood, 115(13):2610–2618, 2010.

GJ Gibson and E Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology, 15(1):19–40, 1998.

D Golinelli. Bayesian inference in hidden stochastic population processes. PhD thesis, University of Washington, 2000.

D Golinelli, P Guttorp, and JA Abkowitz. Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. Mathematical Medicine and Biology, 23(3): 153–172, 2006.

S Goyal, S Kim, ISY Chen, and T Chou. Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC biology, 13(1):85, 2015.

WK Grassmann. Transient solutions in Markovian queueing systems. Computers & Operations Research, 4(1):47–53, 1977.

JE Griffin, PJ Brown, et al. Some priors for sparse regression modelling. Bayesian Analysis, 8(3):691–702, 2013.

DA Griffiths. Multivariate birth-and-death processes as approximations to epidemic processes. Journal of Applied Probability, 10(1):15–26, 1973.

P Guttorp. Statistical inference for branching processes, volume 122. Wiley-Interscience, 1991.

P Guttorp. Stochastic modeling of scientific data. CRC Press, 1995.

P Guttorp, K Albertsen, JF Steffensen, and E Kristensen. Three papers on the history of branching processes. International statistical review, 63(2):233–245, 1995.

M Hajiaghayi, B Kirkpatrick, L Wang, and A Bouchard-Côté. Efficient continuous-time Markov chain estimation. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 638–646, 2014.

LP Hansen. Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, pages 1029–1054, 1982.

LPeter Hansen, J Heaton, and A Yaron. Finite-sample properties of some alternative GMM estimators. Journal of Business & Economic Statistics, 14(3):262–280, 1996.

TE Harris. The Theory of Branching Processes. Prentice-Hall Inc, 1963.

WK Hastings. Monte Carlo sampling methods using markov chains and their applications. Biometrika, 57(1):97–109, 1970.

M Hellerstein, MB Hanley, D Cesar, S Siler, C Papageorgopoulos, E Wieder, D Schmidt, R Hoh, R Neese, D Macallan, et al. Directly measured kinetics of circulating t lymphocytes in normal and hiv-1-infected humans. Nature medicine, 5(1):83–89, 1999.

P Henrici. Fast Fourier methods in computational complex analysis. Siam Review, 21(4): 481–527, 1979.

CC Heyde and E Seneta. The simple branching process, a turning point test and a fundamental inequality: A historical note on ij bienaymé. Biometrika, 59(3):680–683, 1972.

LST Ho, J Xu, FW Crawford, VN Minin, and MA Suchard. Birth (death)/birth-death processes and their computable transition probabilities with statistical applications. arXiv preprint arXiv:1603.03819, 2016.

A Hobolth and EA Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. The Annals of Applied Statistics, 3(3):1204, 2009.

M Huber. Spatial birth–death swap chains. Bernoulli, 18(3):1031–1041, 2012.

O Hyrien, SA Peslak, NM Yanev, and J Palis. Stochastic modeling of stress erythropoiesis using a two-type age-dependent branching process with immigration. Journal of mathematical biology, 70(7):1485–1521, 2015.

J Illian, A Penttinen, H Stoyan, and D Stoyan. Statistical analysis and modelling of spatial point patterns, volume 70. John Wiley & Sons, 2008.

EL Ionides, C Bretó, and AA King. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103(49):18438–18443, 2006.

EL Ionides, D Nguyen, Y Atchadé, S Stoev, and AA King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. Proceedings of the National Academy of Sciences, USA, 112(3):719–724, 2015.

CH Jackson. Multi-state models for panel data: the msm package for R. Journal of Statistical Software, 38(8):1–29, 2011.

P Jagers. Branching processes with biological applications. London [etc.]: Wiley, 1975.

JD Kalbfleisch and JF Lawless. The analysis of panel data under a Markov assumption. Journal of the American Statistical Association, 80(392):863–871, 1985.

S Karlin and J McGregor. Linear growth birth and death processes. J. Math. Mech., (7): 643–662, 1958.

RE Kass and AE Raftery. Bayes factors. Journal of the American Statistical Association, 90(430):773–795, 1995.

M Kato-Maeda, JZ Metcalfe, and L Flores. Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. Future Microbiology, 6(2):203–216, 2011.

A Kaur, M Di Mascio, A Barabasz, M Rosenzweig, HM McClure, AS Perelson, RM Ribeiro, and RP Johnson. Dynamics of t-and b-lymphocyte turnover in a natural host of simian immunodeficiency virus. Journal of virology, 82(3):1084–1093, 2008.

H Kawamoto, H Wada, and Y Katsura. A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. International immunology, 22(2):65–70, 2010.

N Keiding. Maximum likelihood estimation in the birth-and-death process. The Annals of Statistics, 3(2):363–372, 1975.

DG Kendall. On the generalized birth-and-death process. The Annals of Mathematical Statistics, 19(1):1–15, 1948.

M Kimmel. Stochasticity and determinism in models of hematopoiesis. In A Systems Biology Approach to Blood, pages 119–152. Springer, 2014.

M Kimmel and DE Axelrod. Branching Processes in Biology. Springer, New York, 2002.

L Knorr-Held and S Richardson. A hierarchical model for space–time surveillance data on meningococcal disease incidence. Journal of the Royal Statistical Society: Series C (Applied Statistics), 52(2):169–183, 2003.

A Koepke, IM Longini, Jr., ME Halloran, J Wakefield, and VN Minin. Predictive modeling of cholera outbreaks in Bangladesh. The Annals of Applied Statistics, 10(2):575–595, 2016.

E Lakatos, A Ale, PDW Kirk, and MPH Stumpf. Multivariate moment closure techniques for stochastic kinetic models. The Journal of Chemical Physics, 143(9), 2015.

JM Lange and VN Minin. Fitting and interpreting continuous-time latent Markov models for panel data. Statistics in Medicine, 32(26):4581–4595, 2013.

K Lange. Calculation of the equilibrium distribution for a deleterious gene by the finite Fourier transform. Biometrics, 38(1):79–86, 1982.

K Lange. A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society. Series B. Methodological, 57(2):425–437, 1995.

J Liepe, P Kirk, S Filippi, T Toni, CP Barnes, and MPH Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nature Protocols, 9(2):439–456, 2014.

R Lu, NF Neff, SR Quake, and IL Weissman. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature Biotechnology, 29(10):928–933, 2011.

A Marciniak-Czochra, T Stiehl, AD Ho, W Jäger, and W Wagner. Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. Stem cells and development, 18(3):377–386, 2009.

P Marjoram, J Molitor, V Plagnol, and S Tavaré. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, USA, 100(26):15324–15328, 2003.

AD Martin, KM Quinn, and JH Park. MCMCpack: Markov chain Monte Carlo in R. Journal of Statistical Software, 42(9):22, 2011. URL `http://www.jstatsoft.org/v42/i09/`.

CRE McEvoy, NCG van Pittius, TC Victor, PD van Helden, and RM Warren. The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. Tuberculosis, 87(5):393–404, 2007.

A.G. McKendrick. Applications of mathematics to medical problems. Proceedings of the Edinburgh Mathematics Society, 44:98–130, 1926.

T McKinley, AR Cook, and R Deardon. Inference in epidemic models without likelihoods. The International Journal of Biostatistics, 5(1):1557–4679, 2009.

VN Minin and MA Suchard. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology, 56(3):391–412, 2008.

C. Moler and C.V. Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review, 45:3–49, 2003.

JA Murphy and MR O'Donohoe. Some properties of continued fractions with applications in Markov processes. IMA Journal of Applied Mathematics, 16(1):57–71, 1975.

JA Nelder and R Mead. A simplex method for function minimization. The Computer Journal, 7(4):308–313, 1965.

Y Nesterov. A method of solving a convex programming problem with convergence rate O(1/k2). In Soviet Mathematics Doklady, volume 27, pages 372–376, 1983.

MF Neuts. Algorithmic Probability: A Collection of Problems, volume 3. CRC Press, 1995.

MF Neuts and SI Resnick. On the times of births in a linear birth process. Journal of the Australian Mathematical Society, 12(04):473–475, 1971.

MA Newton, P Guttorp, S Catlin, R Assunção, and JL Abkowitz. Stochastic modeling of early hematopoiesis. Journal of the American Statistical Association, 90(432):1146–1155, 1995.

J Neyman, T Park, and EL Scott. Struggle for existence. the Tribolium model: Biological and statistical aspects. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volume 4, pages 41–79. Univ of California Press, 1956.

F Notta, S Zandi, N Takayama, S Dobson, OI Gan, G Wilson, KB Kaufmann, J McLeod, E Laurenti, CF Dunant, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science, page aab2116, 2015.

M Ogawa. Differentiation and proliferation of hematopoietic stem cells. Blood, 81:2844–2844, 1993.

SH Orkin and LI Zon. Hematopoiesis: An evolving paradigm for stem cell biology. Cell, 132 (4):631–644, 2008.

PD ONeill and GO Roberts. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162(1):121–129, 1999.

T Park and G Casella. The Bayesian lasso. Journal of the American Statistical Association, 103(482):681–686, 2008.

L Perié, pd Hodgkin, SH Naik, TN Schumacher, RJ de Boer, and KR Duffy. Determining lineage pathways from cellular barcoding experiments. Cell Reports, 6(4):617 – 624, 2014.

NG Polson and JG Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. Bayesian Statistics, 9:501–538, 2010.

P Pudlo, JM Marin, A Estoup, JM Cornuet, M Gautier, and CP Robert. ABC model choice via random forests. ArXiv e-prints, 1406.6288, 2014.

PS Puri. Interconnected birth and death processes. Journal of Applied Probability, pages 334–349, 1968.

LR Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

GF Raggett. A stochastic model of the Eyam plague. Journal of Applied Statistics, 9(2): 212–225, 1982.

VA Rao and YW Teh. Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks. In Proceedings of the 27th International Conference on Uncertainty in Artificial Intelligence. 2011.

E Renshaw. Stochastic Population Processes: Analysis, Approximations, Simulations. Oxford University Press Oxford, UK, 2011.

CP Robert, JM Cornuet, JM Marin, and NS Pillai. Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences, USA, 108(37):15112–15117, 2011.

NA Rosenberg, AG Tsolaki, and MM Tanaka. Estimating change rates of genetic markers using serial samples: applications to the transposon IS*6110* in *Mycobacterium tuberculosis*. Theoretical Population Biology, 63(4):347–363, 2003.

RY Rubinstein and DP Kroese. Simulation and the Monte Carlo method, volume 707. John Wiley & Sons, 2011.

G Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 1978.

T Sellke. On the asymptotic distribution of the size of a stochastic epidemic. Journal of Applied Probability, pages 390–394, 1983.

CE Shannon. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1):3–55, 2001.

BE Shepherd, H Kiem, PM Lansdorp, CE Dunbar, G Aubert, A LaRochelle, R Seggewiss, P Guttorp, and JL Abkowitz. Hematopoietic stem-cell behavior in nonhuman primates. Blood, 110(6):1806–1813, 2007.

L Siminovitch, EA McCulloch, and JE Till. The distribution of colony-forming cells among spleen colonies. Journal of Cellular and Comparative Physiology, 62(3):327–336, 1963.

G Suwanpimolkul, LG Jarlsberg, JA Grinsdale, D Osmond, LM Kawamura, PC Hopewell, and M Kato-Maeda. Molecular epidemiology of tuberculosis in foreign-born persons living in San Francisco. American Journal of Respiratory and Critical Care Medicine, 187(9): 998–1006, 2013.

MM Tanaka and NA Rosenberg. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Statistics in Medicine, 20(16):2409–2420, 2001.

MA Tanner and WH Wong. The calculation of posterior distributions by data augmentation. Journal of the American statistical Association, 82(398):528–540, 1987.

R Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

T Toni, D Welch, N Strelkowa, A Ipsen, and MPH Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface, 6(31):187–202, 2009.

AW Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000.

JD Van Embden, MD Cave, JT Crawford, JW Dale, KD Eisenach, B Gicquel, P Hermans, C Martin, R McAdam, and TM Shinnick. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. Journal of Clinical Microbiology, 31(2):406–409, 1993.

J Wakefield. Bayesian and Frequentist Regression Methods. Springer Science & Business Media, 2013.

IL Weissman. Stem cells: Units of development, units of regeneration, and units in evolution. Cell, 100(1):157–168, 2000.

ZL Whichard, CA Sarkar, M Kimmel, and SJ Corey. Hematopoiesis and its disorders: a systems biology approach. Blood, 115(12):2339–2347, 2010.

C Wu, B Li, R Lu, SJ. Koelle, Y Yang, A Jares, AE Krouse, M Metzger, F Liang, K Loré, CO Wu, RE. Donahue, ISY Chen, I Weissman, and CE Dunbar. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. Cell Stem Cell, 14(4):486–499, 2014.

J Xu and VN Minin. bdsem: R package for MLE and EM inference in discretely observed birth-death-shift processes, 2014. URL `https://github.com/jasonxu90/bdsem`.

J Xu and VN Minin. Efficient transition probability computation for continuous-time branching processes via compressed sensing. Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pages 952–961, 2015.

J Xu, P Guttorp, MM Kato-Maeda, and VN Minin. Likelihood-based inference for discretely observed birth-death-shift processes, with applications to evolution of mobile genetic elements. Biometrics, 71(4):1009–1021, 2015. ISSN 1541-0420. doi: 10.1111/biom.12352. URL `http://dx.doi.org/10.1111/biom.12352`.

Y Zhang, DL Wallace, CM De Lara, H Ghattas, B Asquith, A Worth, GE Griffin, GP Taylor, DF Tough, PCL Beverley, et al. In vivo kinetics of human natural killer cells: the effects of ageing and acute and chronic viral infection. Immunology, 121(2):258–265, 2007.

Chapter 9

APPENDICES

Appendix A

Here we derive and solve the Kolmogorov backward equations of the two-type branching process representation of the BDS model in Chapter 4. These equations govern the generating functions whose coefficients yield transition probabilities. Our two-type branching process is represented by a vector $(X_1(t), X_2(t))$ that denotes the numbers of particles of two types at time t. Recall the quantities $a_1(k,l)$, the rates of producing k type 1 particles and l type 2 particles, starting with one type 1 particle, and $a_2(k,l)$, analogously defined but beginning with one type 2 particle. Then we may introduce respective pseudo-generating functions $u_i(s_1,s_2) = \sum_k \sum_l a_i(k,l) s_1^k s_2^l$ for $i = 1, 2$, and the probability generating functions can be expressed

$$\begin{aligned}\phi_{10}(t,s_1,s_2) &= E\left[s_1^{X_1(t)} s_2^{X_2(t)} \mid X_1(0)=1, X_2(0)=0\right] = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} P_{(1,0),(k,l)}(t) s_1^k s_2^l \\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \left[\mathbf{1}_{k=1,l=0} + a_1(k,l)t + o(t)\right] s_1^k s_2^l = s_1 + u_1(s_1,s_2)t + o(t). \qquad \text{(A-1)}\end{aligned}$$

An analogous expression for $\phi_{01}(t,s_1,s_2)$ is obtained similarly. For short, we write $\phi_{10} := \phi_1, \phi_{01} := \phi_2$, and thus we have the following relations between ϕ and u

$$\left.\frac{d\phi_1(t,s_1,s_2)}{dt}\right|_{t=0} = u_1(s_1,s_2), \qquad \left.\frac{d\phi_2(t,s_1,s_2)}{dt}\right|_{t=0} = u_2(s_1,s_2).$$

By particle independence, $\phi_{i,j} = \phi_1^i \phi_2^j$, so it suffices to work with only ϕ_1, ϕ_2. We now derive the backward equations for ϕ_1 and ϕ_2. Chapman-Kolmogorov equations yield the symmetric

relations

$$\phi_1(t+h, s_1, s_2) = \phi_1(t, \phi_1(h, s_1, s_2), \phi_2(h, s_1, s_2)) \tag{A-2}$$

$$= \phi_1(h, \phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2)). \tag{A-3}$$

To derive the backward equations, we begin by expanding $\phi_1(t+h, s_1, s_2)$ around t and applying (A-3):

$$\begin{aligned}
\phi_1(t+h, s_1, s_2) &= \phi_1(t, s_1, s_2) + \frac{d\phi_1(t+h, s_1, s_2)}{dh}\Big|_{h=0} h + o(h) \\
&= \phi_1(t, s_1, s_2) + \frac{d\phi_1(h, \phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2))}{dh}\Big|_{h=0} h + o(h) \\
&= \phi_1(t, s_1, s_2) + u_1(\phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2))h + o(h).
\end{aligned}$$

Since an analogous argument applies for ϕ_2, we arrive at the system

$$\begin{cases}
\frac{d}{dt}\phi_1(t, s_1, s_2) = u_1(\phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2)), \\
\frac{d}{dt}\phi_2(t, s_1, s_2) = u_2(\phi_1(t, s_1, s_2), \phi_2(t, s_1, s_2)),
\end{cases}$$

subject to initial conditions $\phi_1(0, s_1, s_2) = s_1, \phi_2(0, s_1, s_2) = s_2$.

We now substitute the rates specific to our birth-shift-death model into this general form: recall the rates defining the two-type branching process formulation presented in Section 2.4 of the main paper are

$$\begin{aligned}
&a_1(1,1) = \lambda, \qquad a_1(0,1) = \nu, \qquad a_1(0,0) = \mu \qquad a_1(1,0) = -(\lambda+\nu+\mu), \\
&a_2(0,2) = \lambda, \qquad a_2(0,1) = -(\lambda+\mu), \qquad a_2(0,0) = \mu,
\end{aligned} \tag{A-4}$$

so that the pseudo-generating functions and backward equations are

$$\begin{cases} u_1(s_1,s_2) = \lambda s_1 s_2 + \nu s_2 + \mu - (\lambda+\nu+\mu)s_1, & \frac{d}{dt}\phi_1 = \lambda\phi_1\phi_2 + \nu\phi_2 + \mu - (\lambda+\nu+\mu)s_1, \\ u_2(s_1,s_2) = \lambda s_2^2 - (\lambda+\mu)s_2 + \mu, & \frac{d}{dt}\phi_2 = \lambda\phi_2^2 - (\lambda+\mu)\phi_2 + \mu. \end{cases} \tag{A-5}$$

Upon rearranging, the expression for ϕ_2 becomes a Ricatti equation

$$\phi_2' - \lambda\phi_2^2 + (\lambda+\mu)\phi_2 = \mu,$$

and the constant solutions $\phi_2 = 1, \mu/\lambda$ are both particular solutions. Using the simpler root $\phi_2 = 1$, we can reduce the above Ricatti equation to a linear ODE by making a substitution $z = \frac{1}{\phi_2 - 1}$, so that $\phi_2 = 1 + \frac{1}{z}$:

$$\begin{aligned} \phi_2' = -\frac{z'}{z^2} = \mu - (\lambda+\mu)(\frac{1}{z}+1) + \lambda(1+\frac{1}{z})^2 \quad &= \mu - \frac{\lambda+\mu}{z} - (\lambda+\mu) + \lambda(\frac{1}{z^2}+\frac{2}{z}+1) \\ = -\frac{\mu-\lambda}{z} + \frac{\lambda}{z^2}. & \end{aligned}$$

Multiplying through by $-z^2$ and rearranging, we arrive at a linear equation that is easily solved via the integrating factor method:

$$z' + (\lambda-\mu)z = -\lambda \Rightarrow z = -\frac{\lambda}{\lambda-\mu} + Ce^{-(\lambda-\mu)t}.$$

Substituting ϕ_2 back into the expression, we obtain

$$\phi_2 = 1 + \frac{1}{\frac{\lambda}{\mu-\lambda} + Ce^{(\mu-\lambda)t}},$$

and plugging in the initial condition $\phi_2(0,s_1,s_2) = s_2$, we see $C = \frac{1}{s_2-1} + \frac{\lambda}{\lambda-\mu}$. Thus, we arrive at the closed form solution

$$\phi_2(t,s_1,s_2) = 1 + \left[\frac{\lambda}{\mu-\lambda} + (\frac{1}{s_2-1} + \frac{\lambda}{\lambda-\mu})e^{(\mu-\lambda)t}\right]^{-1} := g(t,s_1,s_2) \tag{A-6}$$

We can now plug this solution into the ODE for ϕ_1 to obtain

$$\frac{d}{dt}\phi_1 + (\lambda + \nu + \mu - \lambda g)\phi_1 = \nu g + \mu. \tag{A-7}$$

Closed form solution for ϕ_1

Equation (A-7) is linear with variable coefficients, and can again be solved by multiplying by an integrating factor. If we define the integrating factor $\psi := \exp\left[\int (\lambda + \nu + \mu - \lambda g)dt\right]$, then

$$\frac{d}{dt}(\phi_1\psi) = \psi(\nu g + \mu),$$

and after integration and rearranging,

$$\phi_1 = \psi^{-1}\left[\int \psi(\nu g + \mu)dt + C\right]. \tag{A-8}$$

After further simplification, we may write

$$\psi = e^{(\nu+\mu)t}(\lambda s_2 - \mu) + \lambda e^{(\lambda+\nu)t}(1 - s_2),$$

and the integrand becomes

$$\psi(\nu g + \mu) = (\nu + \mu)\psi + \frac{\nu\psi}{\frac{\lambda}{\mu-\lambda} + (\frac{1}{s_2-1} + \frac{\lambda}{\lambda-\mu})e^{(\mu-\lambda)t}}. \tag{A-9}$$

Integrating (A-9) and plugging into (A-8) with initial condition $\phi_1(0, s_1, s_2) = s_1$, we ultimately obtain a closed form expression

$$\begin{aligned}\phi_1(t,s_1,s_2) &= \left[e^{(\nu+\mu)t}(\lambda s_2-\mu)+\lambda e^{(\lambda+\nu)t}(1-s_2)\right]^{-1}\\ &\cdot\Bigg\{\nu(\mu-\lambda)e^{\nu t}\Bigg[\frac{e^{\mu t}(\lambda s_2-\mu)\,{}_2F_1\Big(1,\frac{\mu+\nu}{\mu-\lambda},\frac{\lambda-2\mu-\nu}{\lambda-\mu},\frac{e^{(\mu-\lambda)t}(\lambda s_2-\mu)}{\lambda(s_2-1)}\Big)}{\lambda(\mu+\nu)}\\ &+\frac{e^{\lambda t}(1-s_2)\,{}_2F_1\Big(1,\frac{\lambda+\nu}{\mu-\lambda},\frac{\mu+\nu}{\mu-\lambda},\frac{e^{(\mu-\lambda)t}(\lambda s_2-\mu)}{\lambda(s_2-1)}\Big)}{\lambda+\nu}\Bigg]+(\lambda s_2-\mu)e^{(\mu+\nu)t}\\ &+\frac{\lambda(\nu+\mu)(1-s_2)e^{(\lambda+\nu)t}}{\lambda+\nu}+\mu+s_1(\lambda-\mu)-\lambda s_2+\frac{\lambda(s_2-1)(\nu+\mu)}{\lambda+\nu}\\ &+\nu(\lambda-\mu)\Bigg[\frac{\lambda s_2-\mu}{\lambda(\mu+\nu)}\,{}_2F_1\Big(1,\frac{\mu+\nu}{\mu-\lambda},\frac{\lambda-2\mu-\nu}{\lambda-\mu},\frac{\lambda s_2-\mu}{\lambda(s_2-1)}\Big)\\ &+\frac{1-s_2}{\lambda+\nu}\,{}_2F_1\Big(1,\frac{\lambda+\nu}{\mu-\lambda},\frac{\mu+\nu}{\mu-\lambda},\frac{\lambda s_2-\mu}{\lambda(s_2-1)}\Big)\Bigg]\Bigg\}, \qquad \text{(A-10)}\end{aligned}$$

where ${}_2F_1$ indicates the hypergeometric function. In practice, we solve for ϕ_1 numerically rather than using this closed form solution: evaluating (A-7) via Runge-Kutta methods proves more stable than numerical evaluation of the hypergeometric functions arising in (A-10).

Appendix B

Here we derive the equations appearing in the main theorem of Chapter 4. The statement is repeated below:

Theorem 9.0.1 *Let $\{X_t\}$ be a two-type branching defined by the rates in equation* (5.15)*. Denote particle time and the number of births, shifts, and deaths over the interval $[0,t)$ by R_t, b_t, f_p, and d_t respectively. Define the generating functions corresponding to births as*

$$H_1^+(r,s_1,s_2,t) = E\left[r^{b_t}s_1^{X_1(t)}s_2^{X_2(t)} \mid \mathbf{X}(0)=(1,0)\right] \text{ and}$$
$$H_2^+(r,s_1,s_2,t) = E\left[r^{b_t}s_1^{X_1(t)}s_2^{X_2(t)} \mid \mathbf{X}(0)=(0,1)\right].$$

Then

$$H_2^+ = y_b + \left[\frac{-\lambda r}{2\lambda r y_b - \lambda - \mu} + \left(\frac{1}{s_2 - y_b} + \frac{\lambda r}{2\lambda r y_b - \lambda - \mu}\right) e^{-(2y_b\lambda r - \lambda - \mu)t}\right]^{-1},$$

where $y_b = (\lambda + \mu + \sqrt{\lambda^2 + 2\lambda\mu + \mu^2 - 4\lambda\mu r})/(2\lambda r)$, *and* H_1^+ *satisfies the following differential equation:*

$$\frac{d}{dt}H_1^+(t, s_1, s_2, r) = \lambda r H_1^+ H_2^+ + \nu H_2^+ + \mu - (\lambda + \mu + \nu)H_1^+, \tag{B-1}$$

subject to initial condition $H_1(r, s_1, s_2, 0) = s_1$.

The analogous generating functions for shifts, deaths, and particle time satisfy the following equations:

$$H_2^-(t, s_1, s_2, r) = y_d + \left[\frac{-\lambda}{2\lambda y_d - \lambda - \mu} + \left(\frac{1}{s_2 - y_d} + \frac{\lambda}{2\lambda y_d - \lambda - \mu}\right) e^{-(2y_d\lambda - \lambda - \mu)t}\right]^{-1},$$
$$H_2^{\rightarrow}(t, s_1, s_2, r) = 1 + \left[\frac{\lambda}{\mu - \lambda} + (\frac{1}{s_2 - 1} + \frac{\lambda}{\lambda - \mu})e^{(\mu - \lambda)t}\right]^{-1},$$
$$H_2^*(t, s_1, s_2, r) = y_* + \left[\frac{-\lambda}{2\lambda y_* - \lambda - \mu - r} + \left(\frac{1}{s_2 - y_*} + \frac{\lambda}{2\lambda y_* - \lambda - \mu - r}\right) e^{-(2y_*\lambda - \lambda - \mu - r)t}\right]^{-1},$$

$$\frac{d}{dt}H_1^-(t, s_1, s_2, r) = \lambda H_1^- H_2^- + \nu H_2^- + \mu r - (\lambda + \mu + \nu)H_1^-,$$
$$\frac{d}{dt}H_1^{\rightarrow}(t, s_1, s_2, r) = \lambda H_1^{\rightarrow} H_2^{\rightarrow} + \nu r H_2^{\rightarrow} + \mu - (\lambda + \mu + \nu)H_1^{\rightarrow},$$
$$\frac{d}{dt}H_1^*(t, s_1, s_2, r) = \lambda H_1^* H_2^* + \nu H_2^* + \mu - (\lambda + \mu + \nu + r)H_1^*,$$

where $y_d = (\lambda + \mu + \sqrt{\lambda^2 + 2\lambda\mu + \mu^2 - 4\lambda\mu r})/(2\lambda)$, $y_* = (\lambda + \mu + r + \sqrt{(\lambda + \mu + r)^2 - 4\lambda\mu})/(2\lambda)$ *and* $H_1^-(r, s_1, s_2, 0) = H_1^{\rightarrow}(r, s_1, s_2, 0) = H_1^*(r, s_1, s_2, 0) = s_1$.

Proof Begin by expanding

$$H_{10}^{+}(t,r,s_1,s_2) = \sum_n \sum_k \sum_l Pr(b_t = n, x_t = (k,l)|x_0 = (1,0)) s_1^k s_2^l r^n.$$

Recall the jump rates of the process in equation (5.15): a_1 correspond to the process beginning with 1 type one particle, and a_2 are jump rates starting with 1 type two particle. We can express the probability terms in H_{10}^{+} using the same type of first-order decomposition as in equation (B-14); for instance, in the event of a birth,

$$Pr(b_t = 1, x_t = (1,1)|x_0 = (1,0)) = a_1(1,1) + o(t) = \lambda + o(t)$$

and for other values of $n > 1$,

$$Pr(b_t = n, x_t = (1,1)|x_0 = (1,0)) = o(t).$$

In the case of a shift,

$$Pr(b_t = 0, x_t = (0,1)|x_0 = (1,0)) = a_1(0,1) + o(t) = \nu + o(t)$$

and for other values of $n \neq 0$,

$$Pr(b_t = n, x_t = (0,1)|x_0 = (1,0)) = o(t).$$

We see that the r^n term in the series H_{10}^{+} is either $r^1 = r$ if exactly one birth occurs, or $r^0 = 1$ as other powers correspond to more than one event and are absorbed into the $o(t)$ term. Thus,

$$\begin{aligned} H_{10}^{+}(t,r,s_1,s_2) &= \sum_k \sum_l g_{10,kl}(r,t) s_1^k s_2^l = \sum_n \sum_k \sum_l Pr(b_t = n, x_t = (k,l)|x_0 = (1,0)) s_1^k s_2^l r^n \\ &= s_1 + \lambda s_1 s_2 r + \nu s_2 + \mu - (\lambda + \nu + \mu) s_1 + o(t) := s_1 + u_1^b(s_1,s_2)t + o(t) \end{aligned}$$

with u_1^b denoting the pseudo-generating function, similarly to (B-14). With an analogous derivation for u_2^b, we arrive at the system

$$\begin{cases} u_1^b(s_1, s_2) = \lambda r s_1 s_2 + \nu s_2 + \mu - (\lambda + \nu + \mu) s_1 \\ u_2^b(s_1, s_2) = \lambda r s_2^2 - (\lambda + \mu) s_2 + \mu, \end{cases} \tag{B-2}$$

and since

$$\frac{dH_{10}^+(t, r, s_1, s_2)}{dt}\Big|_{t=0} = u_1^b(s1, s2, r), \qquad \frac{dH_{01}^+(t, r, s_1, s_2)}{dt}\Big|_{t=0} = u_2^b(s1, s2, r),$$

we obtain the backward equations system

$$\begin{cases} \frac{d}{dt} H_{10}^+(t, s_1, s_2, r) = u_1^b(H_{10}^+(t, s_1, s_2, r), H_{01}^+(t, s_1, s_2, r)), \\ \frac{d}{dt} H_{01}^+(t, s_1, s_2, r) = u_2^b(H_{10}^+(t, s_1, s_2, r), H_{01}^+(t, s_1, s_2, r)) \end{cases} \tag{B-3}$$

by the same Chapman-Kolmogorov argument used for transition probabilities, subject to initial conditions $H_{10}(t = 0, s_1, s_2, r) = s_1$ and $H_{01}(t = 0, s_1, s_2, r) = s_2$. The systems for deaths and shifts are derived analogously beginning with this first-order expansion technique, and are respectively given by

$$\begin{cases} u_1^d(s_1, s_2) = \lambda s_1 s_2 + \nu s_2 + r\mu - (\lambda + \nu + \mu) s_1 \\ u_2^d(s_1, s_2) = \lambda s_2^2 - (\lambda + \mu) s_2 + r\mu, \end{cases} \tag{B-4}$$

$$\begin{cases} u_1^{\rightarrow}(s_1, s_2) = \lambda s_1 s_2 + r\nu s_2 + \mu - (\lambda + \nu + \mu) s_1, \\ u_2^{\rightarrow}(s_1, s_2) = \lambda s_2^2 - (\lambda + \mu) s_2 + \mu. \end{cases} \tag{B-5}$$

To derive the system governing the particle time generating function, recall the quantity $q_{ij,kl}^*(x; t) := Pr(R_t \leq x, X(t) = (k, l) | X(0) = (i, j))$, and consider its Laplace-Stieltjes

transform

$$V_{ij,kl}(r;t) = \int_0^\infty e^{-rx} \mathrm{d}q^*_{ij,kl}(x;t). \tag{B-6}$$

The Laplace-Stieltjes transform of such a probability distribution corresponding to a *reward function*, where a_{ij} is the reward accrued per unit time spent in state (i,j), satisfies the forward equation

$$\frac{d}{dt}V_{ij,kl}(r;t) = -a_{ij}rV_{ij,kl}(r;t) + \sum_{m=1}^{K}\sum_{n=1}^{K} Q_{ij,mn}V_{ij,kl}(r;t), \tag{B-7}$$

where $\mathbf{Q}$ is the infinitesimal generator of the Markov chain, with finite or countable number of rows and columns and entries $Q_{ij,kl}$ the instantaneous rates of transitioning from state (i,j) to (k,l), and $Q_{ij,ij} = -\sum_{m,n\neq i,j} Q_{ij,mn}$. Following Neuts [1995], we derive the following integral equation:

$$q^*_{ij,kl}(x,t) = \mathbf{1}_{\{ij=kl\}}\mathbf{1}_{\{x\geq a_{ij}t\}}e^{Q_{ij,ij}t} + \sum_{m,n\neq i,j}\int_0^t e^{Q_{ij,ij}u}Q_{ij,mn}q^*_{mn.kl}(x-a_{ij}u, t-u)\mathrm{d}u.$$

Taking the Laplace transform of both sides and denoting $\widetilde{V}_{ij,kl}(r;t) = \int_0^\infty e^{-rx}q^*_{ij,kl}(x;t)\mathrm{d}x$, we obtain

$$\widetilde{V}_{ij,kl}(r,t) = \mathbf{1}_{\{ij=kl\}}r^{-1}\exp\left[(Q_{ij,ij}-a_{ij})t\right] + \sum_{m,n\neq i,j}\int_0^t e^{Q_{ij,ij}u}Q_{ij,mn}\mathrm{d}u\int_{a_{ij}u}^\infty e^{-rx}q^*_{mn,kl}(x-a_{ij}u;t-u)\mathrm{d}x.$$

Making a change of variables $y = x - a_{ij}u$ in the rightmost integral and multiplying both sides by $\exp\left[-(Q_{ij,ij}-a_{ij})t\right]$ yields

$$\exp\left[-(Q_{ij,ij}-a_{ij})t\right]\widetilde{V}_{ij,kl}(r,t) = \frac{1}{r} + \sum_{m,n\neq i,j}\int_0^t \exp\left[-(Q_{ij,ij}-a_{ij})(t-u)\right]Q_{ij,mn}\widetilde{V}_{mn,kl}(r;t-u).$$

Next, make another substitution $v = t - u$ and simplify after differentiating the above

equation with respect to t: we arrive at

$$\frac{\partial}{\partial t}\widetilde{V}_{ij,kl}(r;t) = -a_{ij}r\widetilde{V}_{ij,kl}(r;t) + \sum_{m=1}^{K}\sum_{n=1}^{K} Q_{ij,mn}\widetilde{V}_{mn,kl}(r;t).$$

Equation (B-7) then follows from $V_{ij,kl}(r;t) = s\widetilde{V}_{ij,kl}(r;t)$, with $V_{ij,kl}(t)(r;0) = \mathbf{1}_{\{ij=kl\}}$.

The matrix $\mathbf{V}(r;t) := \{V_{ij,kl}(r;t)\}$ can therefore be written as a matrix exponential

$$\mathbf{V}(r;t) := \exp[\mathbf{Q} - \mathrm{diag}(\mathbf{a})r)t] := \exp(\widetilde{\mathbf{Q}}t), \tag{B-8}$$

where $\mathrm{diag}(\mathbf{a})$ is the diagonal matrix with diagonal entries a_{ij}. In our case, $a_{ij} = 1$, since we are interested in particle time and the "reward" that accumulates per unit of time is that quantity of time itself. Strictly speaking we don't need infinite dimensional matrix manipulations here, but we use it to simplify our notation.

Note the similarity of equation (B-8) to the matrix exponential corresponding to transition probabilities $\mathbf{P}(t) = \exp(\mathbf{Q}t)$: thus, the system of backward equations for $V_{ij,kl}$ are almost identical to those for transition probabilities $p_{ij,kl}$. The generators $\widetilde{\mathbf{Q}} \neq \mathbf{Q}$ differ only in diagonal entries: instantaneous rates of no event occurring are augmented by an extra r term $\widetilde{Q}_{ij,ij} = -\sum_{m,n\neq i,j} Q_{ij,mn} - r$. The system of backward equations is thus given by

$$\begin{cases} u_1^*(s_1,s_2) = \lambda s_1 s_2 + \nu s_2 + \mu - (\lambda+\nu+\mu+r)s_1, \\ u_2^*(s_1,s_2) = \lambda s_2^2 + \mu - (\lambda+\mu+r)s_2, \end{cases} \tag{B-9}$$

and as we have seen in the derivation for expected births in equation (B-3), this implies that the generating function

$$H_{10}^*(r,s_1,s_2,t) = \sum_k\sum_l \int_0^\infty e^{-rx}\mathrm{d}q_{ij,kl}^*(x;t) = \sum_k\sum_l V_{10,kl}(r,t)s_1^k s_2^l$$

also satisfies the same system.

Reducing the systems

Each of the four systems for births, shifts, deaths, and particle time can be reduced to a single ODE by first solving the second equation analytically. We demonstrate this in the case of the birth equations (B-3), and abbreviate $H_{10}^{+} := H_1, H_{01}^{+} := H_2$. Plugging (B-2) into (B-3),

$$\begin{cases} \frac{d}{dt}H_1(t, s_1, s_2, r) = \lambda r H_1 H_2 + \nu H_2 + \mu - (\lambda + \nu + \mu)H_1, \\ \frac{d}{dt}H_2(t, s_1, s_2, r) = \lambda r H_2^2 - (\lambda + \mu)H_2 + \mu. \end{cases}$$

The second equation is a Ricatti equation. To solve it, we first identify a constant solution

$$y_b = \frac{\lambda + \mu + \sqrt{\lambda^2 + 2\lambda\mu + \mu^2 - 4\lambda\mu r}}{2\lambda r}$$

obtained by setting

$$\frac{d}{dt}H_2 = 0 = \lambda r H_2^2 - (\lambda + \mu)H_2 + \mu.$$

Next, perform a change of variables $z = \frac{1}{H_2 - y_b}$ so that $H_2 = y_1 + \frac{1}{z}$, and thus

$$\frac{dz}{dt} + (2y_b\lambda r - \lambda - \mu)z = -\lambda r$$

Using the multiplier method with multiplier $\exp\{(2y_b\lambda r - \lambda - \mu)t\}$, we obtain

$$z = e^{-(2y_b\lambda r - \lambda - \mu)t}\left[\int -\lambda r e^{(2\lambda r y_b - \lambda - \mu)t}dt + C\right] = \frac{-\lambda r}{2\lambda r y_b - \lambda - \mu} + Ce^{-(2y_b\lambda r - \lambda - \mu)t}.$$

Thus,

$$H_2 = y_b + \frac{1}{z} = y_b + \left[\frac{-\lambda r}{2\lambda r y_b - \lambda - \mu} + Ce^{-(2y_b\lambda r - \lambda - \mu)t}\right]^{-1}$$

and from $H_2(0, r, s_1, s_2) = s_2$, we see $C = \frac{1}{s_2 - y_b} + \frac{\lambda r}{2\lambda r y_b - \lambda - \mu}$. Finally, we arrive at the full solution to the second ODE

$$H_2 := g^b(t, s_1, s_2, r) = y_b + \left[\frac{-\lambda r}{2\lambda r y_b - \lambda - \mu} + \left(\frac{1}{s_2 - y_b} + \frac{\lambda r}{2\lambda r y_b - \lambda - \mu}\right)e^{-(2y_b\lambda r - \lambda - \mu)t}\right]^{-1}.$$

Plugging this solution into the equation for H_1, we have a single ODE that is numerically solvable:

$$\frac{d}{dt}H_1^+(t, s_1, s_2, r) = \lambda r H_1^+ g^b + \nu g^b + \mu - (\lambda + \mu + \nu)H_1^+.$$

An analogous solution beginning with Equations (B-4), (B-5), and (B-9) instead of (B-3) and solving the second Ricatti equation is used to simplify the other equation systems, yielding the results presented in Theorem 9.0.1.

Appendix C

Here we include additional figures that support, but are not crucial to, illustrating our simulation results from Chapter 4 in more detail.

Figure C-1 displays the transition probabilities $p_{(10,0),(ij)}$ for 25 randomly sampled (i, j) pairs with $0 \leq i, j \leq 32$, calculated by our generating function approach alongside their Monte Carlo estimates and confidence intervals. Monte Carlo estimates are based on 5000 realizations beginning with an initial count of 10 with $dt = 1.0, \lambda = .5, \mu = .45$ and ν ranging from 0.3 to 2.0.

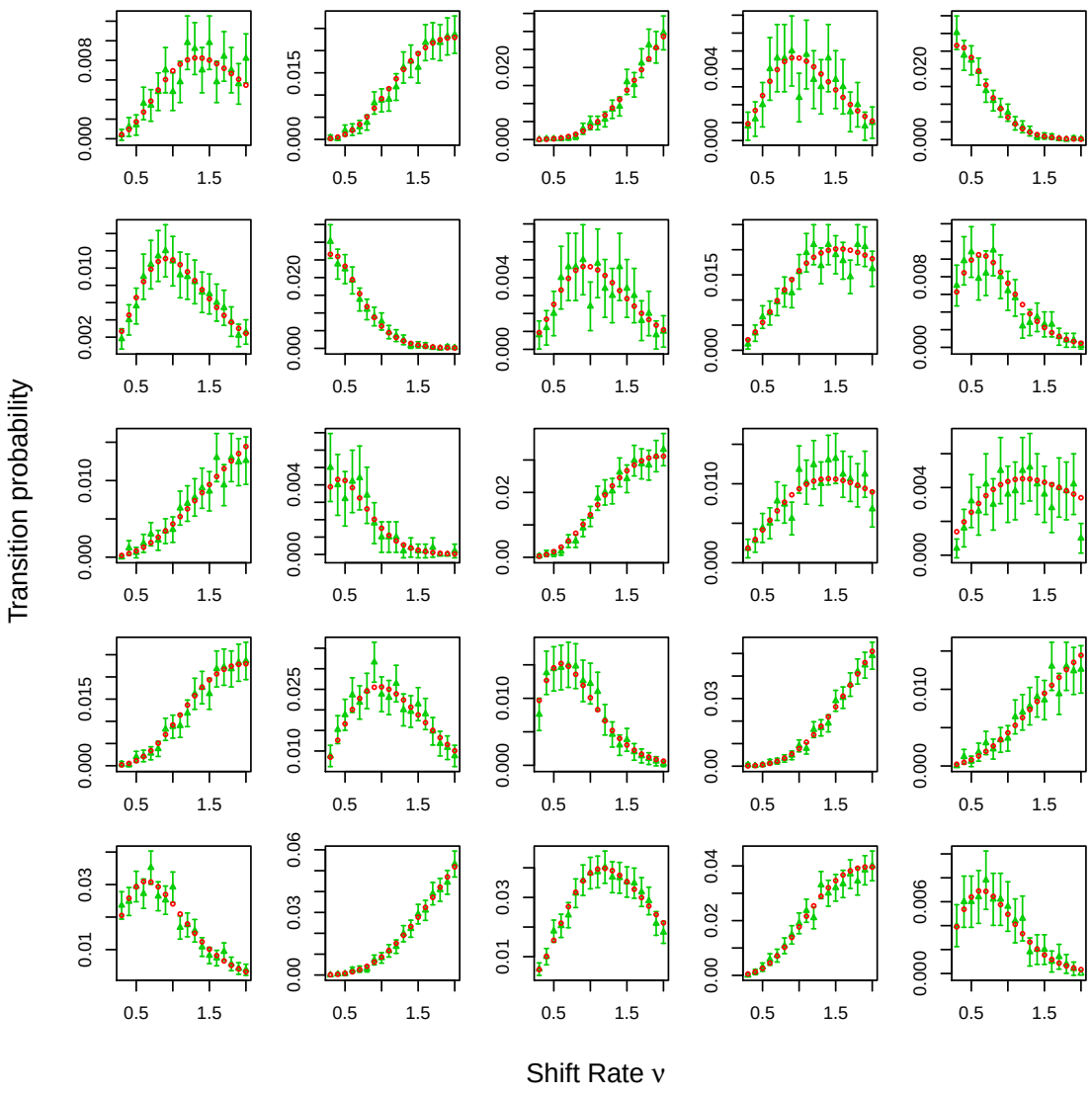


Figure C-1: Transition probabilities remain accurate when increasing all rates of process, presented over a wide range of ν values. Green points and intervals correspond to Monte Carlo estimates of transition probabilities and corresponding 95% confidence intervals. The red points denote probabilities computed with our generating function method.

Figure C-2 shows that restricted moment calculations performed during the E-step are indeed accurate: the following figure corresponds to simulations with 3 times the rates in the Rosenberg-Tanaka paper: $(\lambda, \nu, \mu) = 3 \cdot (.0188, .0026, .0147)$, with 10 initial particles and

varying time lengths.

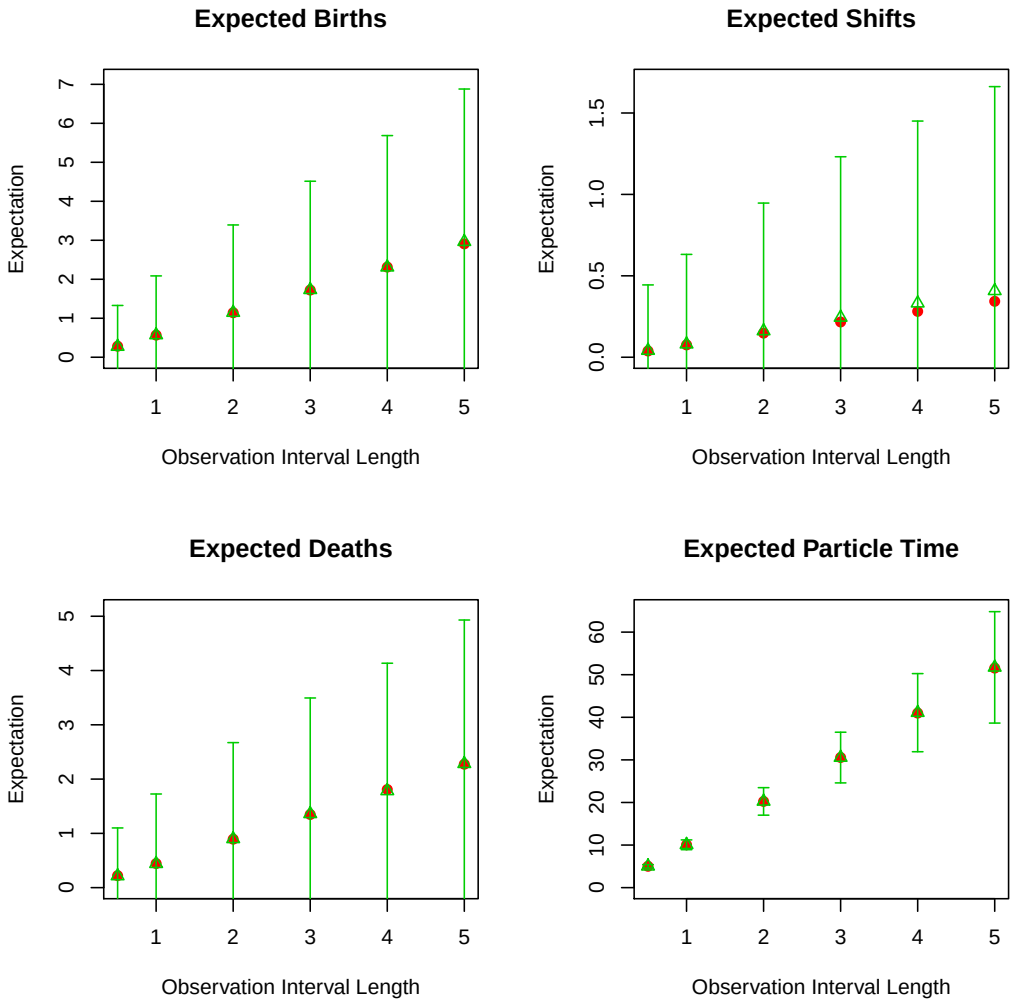


Figure C-2: Restricted moments calculated by our method (red) compared to approximation over 5000 Monte Carlo simulations and corresponding 95% confidence intervals (green).

Appendix D

Here we provide details related to the compressed sensing techniques presented in Chapter 5, including the line-search procedure used in our implementation of proximal gradient descent, as well as a novel solution to the PGFs of the two-compartment hematopoiesis model viewed

equivalently as a branching process.

9.0.1 Discrete Fourier matrix

The N by N discrete Fourier transform matrix $\mathbf{F}_N$ has entries

$$\{\mathbf{F}_N\}_{j,k} = \frac{1}{\sqrt{N}}(\omega)^{jk}$$

with $j, k = 0, 1, \ldots, N-1$ and $\omega = e^{i2\pi/N}$, and as we mention in the main paper, the inverse Fourier transform matrix $\boldsymbol{\psi}$ is given by its conjugate transpose. The partial M by N IDFT matrices $\mathbf{A}$ necessary in Algorithm 1 is obtained by only computing and stacking a subset of M random rows from $\boldsymbol{\psi}$.

9.0.2 Line search subroutine

We select step sizes with a simple line search algorithm summarized in the pseudocode below that works by evaluating an easily computed upper bound $\hat{f}$ on the objective f:

$$\hat{f}_L(Z, Y) := f(Y) + \nabla f(Y)^T(Z - Y) + \frac{L}{2}||Z - Y||_2^2. \tag{B-10}$$

We follow Beck and Teboulle [2009], who provide further details. In implementation, we select $L = .000005$ and $c = .5$, and reuse the gradient computed in `line-search` for step 10 of Algorithm 1 in the main paper.

Algorithm 2 `line-search` procedure.

1: **Input:** initial step size L, shrinking factor c, matrices $Y_k, \nabla g(Y_k)$.
2: Set $Z = \text{soft}\text{th}(Y_k - L\nabla g(Y_k))$
3: **while** $g(Z) > \hat{f}_L(Z, Y_k)$ **do**
4: Update $L = cL$
5: **end while**
6: **return** $L_k = L$

9.0.3 Two-compartment hematopoiesis process PGF

Recall the rates defining the two-compartment hematopoiesis model are given by

$$a_1(2,0) = \rho \qquad a_1(0,1) = \nu \qquad a_1(1,0) = -(\rho+\nu)$$
$$a_2(0,0) = \mu \qquad a_2(0,1) = -\mu$$

Thus, the pseudo-generating functions are

$$u_1(s_1, s_2) = \rho s_1^2 + \nu s_2 - (\rho+\nu)s_1$$

$$u_2(s_1, s_2) = \mu - \mu s_2 = \mu(1 - s_2)$$

Plugging into the backward equations, we obtain

$$\frac{d}{dt}\phi_1(t, s_1, s_2) = \rho\phi_1^2(t, s_1, s_2) + \nu\phi_2(t, s_1, s_2) - (\rho+\nu)\phi_1(t, s_1, s_2)$$

and

$$\frac{d}{dt}\phi_2(t, s_1, s_2) = \mu - \mu\phi_2(t, s_1, s_2).$$

The ϕ_2 differential equation corresponds to a pure death process and is immediately solvable: suppressing the arguments of ϕ_2 for notational convenience, we obtain

$$\begin{aligned}
\frac{d}{dt}\phi_2 &= \mu - \mu\phi_2 \\
\frac{d}{dt}\phi_2\left(\frac{1}{1-\phi_2}\right) &= \mu \\
\ln(1-\phi_2) &= -\mu t + C \\
\phi_2 &= 1 - \exp(-\mu t + C)
\end{aligned}$$

Pluggin in $\phi_2(0, s_1, s_2) = s_2$, we obtain $C = \ln(1 - s_2)$, and we arrive at

$$\phi_2(t, s_1, s_2) = 1 + (s_2 - 1)\exp(-\mu t) \tag{B-11}$$

Plugging this solution into the other backward equation, we obtain

$$\frac{d}{dt}\phi_1(t, s_1, s_2) = \rho\phi_1^2(t, s_1, s_2) - (\rho + \nu)\phi_1(t, s_1, s_2) + \nu(1 + (s_2 - 1)\exp(-\mu t)) \tag{B-12}$$

This ordinary differential equation can be solved numerically given rates and values for the three arguments, allowing computation of $\phi_{i,j} = \phi_1^i\phi_2^j$ which holds by particle independence.

Appendix E

Here we derive the second moments of the simplest instance in our class of branching models of hematopoiesis considered in Chapter 6 explicitly. The derivation considers a four-type model with one progenitor and two mature types (Figure 6.2 (a) ignoring the third mature compartment). We also derive the marginalized moment expressions after incorporating the sampling distribution.

From applying the process rates to the Kolmogorov backward equations, we can write *pseudo-generating functions* defined

$$u_i(s_1, s_2, s_3, s_4) = \sum_j \sum_k \sum_l \sum_m a_i(j, k, l, m) s_1^j s_2^k s_3^l s_4^m. \tag{B-13}$$

For the model depicted in Figure 6.2 (b), these are given by

$$\begin{aligned} u_1(s_1, s_2) &= \lambda s_1^2 + \nu_0 s_2 - (\lambda + \nu_0)s_1; \\ u_2(s_2, s_3, s_4) &= \nu_1 s_2 s_3 + \nu_2 s_2 s_4 + \mu_0 - (\mu_0 + \nu_1 + \nu_2)s_2; \\ u_3(s_3) &= \mu_1 - \mu_1 s_3; \qquad u_4(s_4) = \mu_2 - \mu_2 s_4. \end{aligned}$$

Next, we can write the probability generating function (PGF) of the process, beginning with

one type 1 particle, which is related to the pseudo-generating function u_1 as follows:

$$\begin{aligned}
\phi_1(t; s_1, s_2, s_3, s_4) &= \mathrm{E}[\prod_{j=1}^{4} s_j^{X_j(t)} | \mathbf{X}(0) = (1,0,0,0)] \\
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathrm{Pr}_{(1,0,0,0),(k,l,m,n)} s_1^k s_2^l s_3^m s_4^n \\
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \left[\mathbf{1}_{\{k=1,l=m=n=0\}} + a_1(k,l,m,n)t + o(t)\right] s_1^k s_2^l s_3^m s_4^n \\
&= s_1 + u_1(s_1, s_2, s_3, s_4)t + o(t). \qquad \text{(B-14)}
\end{aligned}$$

We may analogously define ϕ_i for processes beginning with one type i particle, for each $i = 1, \ldots, 4$, we have from Equation (B-14) the relation

$$\frac{\partial}{\partial t}\phi_i(t, s_1, \ldots, s_4) = u_i(\phi_1(t, s_1, \ldots, s_4), \ldots, \phi_4(t, s_1, \ldots, s_4)).$$

Now, let $M_{l|k}(t)$ denote the expected number of type l cells at time t, given one initial type k cell. From definition of ϕ_i, we see that we can relate the probability generating functions to these first moments via partial differentiation:

$$M_{l|k}(t) = \frac{\partial}{\partial s_l}\phi_k(t, s_1, \ldots, s_4)|_{s_1=s_2=s_3=s_4=1}$$

Similarly, we may further differentiate the PGF to derive second moments used toward variance and covariance calculations. Define

$$U_{kl|1}(t) = E\left[X_k(X_l - \mathbf{1}_{\{k=l\}})|\mathbf{X}(0) = (1,0,0,0)\right]$$

with $U_{kl|i}(t)$ defined analogously beginning with one type i particle. Then $U_{kl|j}(t) = \left.\frac{\partial^2 \phi_j}{\partial s_k \partial s_l}\right|_{\mathbf{s}=1}$,

and by the Faà di Bruno formula,

$$\frac{\partial^3 \phi_i}{\partial t \partial s_j \partial s_k} = \sum_{m=1}^{4} \left(\frac{\partial u_i}{\partial \phi_m} \frac{\partial^2 \phi_m}{\partial s_j \partial s_k} \right) + \sum_{m,n=1}^{4} \left(\frac{\partial^2 u_i}{\partial \phi_m \partial \phi_n} \frac{\partial \phi_m}{\partial s_j} \frac{\partial \phi_k}{\partial s_k} \right).$$

This relation allows us to write a system of non-homogeneous, linear ordinary differential equations (ODEs) governing second order moments:

$$\begin{aligned}
\frac{\partial}{\partial t} U_{33|1}(t) &= (\lambda - \nu_0) U_{33|1}(t) + \nu_0 U_{33|2}(t) + (2\lambda) M_{3|1}^2(t) \\
\frac{\partial}{\partial t} U_{44|1}(t) &= (\lambda - \nu_0) U_{44|1}(t) + \nu_0 U_{44|2}(t) + (2\lambda) M_{4|1}^2(t) \\
\frac{\partial}{\partial t} U_{34|1}(t) &= (\lambda - \nu_0) U_{34|1}(t) + \nu_0 U_{34|2}(t) + (2\lambda) M_{3|1}(t) M_{4|1}(t) \\
\frac{\partial}{\partial t} U_{34|2}(t) &= -\mu_0 U_{34|2}(t) + \nu_1 M_{4|2}(t) M_{3|3}(t) + \nu_2 M_{3|2}(t) M_{4|4}(t) \\
\frac{\partial}{\partial t} U_{33|2}(t) &= -\mu_0 U_{33|2}(t) + \nu_1 U_{33|3}(t) + 2\nu_1 M_{3|2}(t) M_{3|3}(t) \\
\frac{\partial}{\partial t} U_{44|2}(t) &= -\mu_0 U_{44|2}(t) + \nu_2 U_{44|4}(t) + 2\nu_2 M_{4|2}(t) M_{4|4}(t) \\
\frac{\partial}{\partial t} U_{33|3}(t) &= -\mu_1 U_{33|3}(t) \qquad\qquad \frac{\partial}{\partial t} U_{44|4}(t) = -\mu_2 U_{44|4}(t)
\end{aligned}$$

all with initial conditions $(\cdot)_{k,l}(0) = 0$. We immediately see that $U_{33|3}(t) = U_{44|4}(t) = 0$, and upon a series of solutions and substitutions, we successively solve the system of ODEs, yielding rather lengthy but simple closed form solutions for all required moments. These solutions, details of the derivation, and forms for the general case are included in the Appendix. The availability of analytic solutions allow for fast computations of the moments, making iterative algorithms for minimizing an objective function feasible. Given expressions for these higher moments, we can readily recover variance and covariance expressions, and thus calculate model-based correlations. For instance,

$$\mathrm{Var}\left[X_4(t) | \mathbf{X}(0) = (1,0,0,0)\right] = U_{44|1}(t) + M_{4|1}(t) - M_{4|1}(t)^2;$$

$$\mathrm{Cov}\left[X_3(t), X_4(t)|\mathbf{X}(0) = (0,1,0,0)\right] = U_{34|2}(t) - M_{3|2}(t)M_{4|2}(t).$$

Because the initial state is uncertain, the variances and covariances of X_3, X_4 can now be computed by marginalizing over the initial barcoding state:

$$X_3(t) = \pi\left[X_3(t)|\mathbf{X}(0) = (1,0,0,0)\right] + (1-\pi)\left[X_3(t)|\mathbf{X}(0) = (0,1,0,0)\right].$$

The marginalized means follow trivially by linearity and the law of total expectation. Dropping the dependence on t for notational simplicity, we use the law of total variance and law of total covariance to obtain the marginalized variance expressions

$$\begin{aligned}
\mathrm{Cov}(X_3, X_4) &= \pi^2(U_{34} - M_{3|1}M_{4|1}) + (1-\pi)^2(V_{34} - M_{3|2}M_{4|2}) \\
&\quad + \pi(1-\pi)(U_{34|1} + U_{34|2} - M_{3|2}M_{4|1} - M_{3|1}M_{4|2}) \\
\mathrm{Var}(X_3) &= \pi(U_{33|1} + M_{3|1}) + (1-\pi)(U_{33|2} + M_{3|2}) \\
&\quad - \pi^2 M_{3|1}^2 - (1-\pi)^2 M_{3|2}^2 - 2\pi(1-\pi)M_{3|1}M_{3|2} \\
\mathrm{Var}(X_4) &= \pi(U_{44|1} + M_{4|1}) + (1-\pi)(U_{44|2} + M_{4|2}) \\
&\quad - \pi^2 M_{4|1}^2 - (1-\pi)^2 M_{4|2}^2 - 2\pi(1-\pi)M_{4|1}M_{4|2}.
\end{aligned} \tag{B-15}$$

The explicit solutions for the simple model:

$$
\begin{aligned}
U_{33|2}(t) &= 2\frac{\nu_1^2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-2\mu_2 t}}{\mu_0-2\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-2\mu_2)}\right]\\
U_{44|2}(t) &= 2\frac{\nu_2^2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-2\mu_2 t}}{\mu_0-2\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-2\mu_2)}\right]\\
U_{34|2}(t) &= \frac{\nu_1\nu_2}{(\mu_2-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_1)t}}{\mu_1}-\frac{e^{-(\mu_1+\mu_2)t}}{\mu_0-\mu_1-\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_1(\mu_0-\mu_1-\mu_2)}\right]\\
&\quad+\frac{\nu_1\nu_2}{(\mu_1-\mu_0)}\left[\frac{e^{-(\mu_0+\mu_2)t}}{\mu_2}-\frac{e^{-(\mu_1+\mu_2)t}}{\mu_0-\mu_1-\mu_2}+\frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_2(\mu_0-\mu_1-\mu_2)}\right]\\
U_{33|1}(t) &= e^{(\lambda-\nu_0)t}\Bigg\{2\frac{\nu_0\nu_1^2}{\mu_1-\mu_0}\left[\frac{(\mu_0-\mu_1)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_1(\mu_0-2\mu_1)(\nu_0-\lambda-\mu_0)}-\frac{e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{\mu_1(\nu_0-\lambda-\mu_0-\mu_1)}-\frac{e^{(\nu_0-\lambda-2\mu_1)t}}{(\mu_0-2\mu_1)(\nu_0-\lambda-2\mu_1)}\right.\\
&\qquad\left.+\frac{\mu_1-\mu_0}{\mu_1(\mu_0-2\mu_1)(\nu_0-\lambda-\mu_0)}+\frac{1}{\mu_1(\nu_0-\lambda-\mu_0-\mu_1)}+\frac{1}{(\mu_0-2\mu_1)(\nu_0-\lambda-2\mu_1)}\right]\\
&\quad+\frac{2\lambda\nu_0^2\nu_1^2}{(\mu_1-\mu_0)^2}\left[\frac{e^{(\nu_0-\lambda-2\mu_0)t)}}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)}-\frac{2e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)}\right.\\
&\quad+\frac{2(\mu_0-\mu_1)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0)^2}+\frac{e^{(\nu_0-\lambda-2\mu_1)t}}{(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-2\mu_1)}+\frac{2(\mu_1-\mu_0)e^{-\mu_1 t}}{\mu_1(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)}\\
&\quad+\frac{(\mu_1-\mu_0)^2e^{(\lambda-\nu_0)t}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)^2}-\frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)}\\
&\quad+\frac{2}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)}-\frac{2(\mu_0-\mu_1)}{\mu_0(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0)^2}\\
&\quad-\frac{1}{(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-2\mu_1)}-\frac{2(\mu_1-\mu_0)}{\mu_1(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)}\\
&\quad\left.-\frac{(\mu_1-\mu_0)^2}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_1)^2(\nu_0-\lambda-\mu_0)^2)}\right]\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
U_{44|1}(t) = e^{(\lambda-\nu_0)t}\Bigg\{ & 2\frac{\nu_0\nu_2^2}{\mu_2-\mu_0}\Bigg[\frac{(\mu_0-\mu_2)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_2(\mu_0-2\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{\mu_2(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{e^{(\nu_0-\lambda-2\mu_2)t}}{(\mu_0-2\mu_2)(\nu_0-\lambda-2\mu_2)} \\
& + \frac{\mu_2-\mu_0}{\mu_2(\mu_0-2\mu_2)(\nu_0-\lambda-\mu_0)} + \frac{1}{\mu_2(\nu_0-\lambda-\mu_0-\mu_2)} + \frac{1}{(\mu_0-2\mu_2)(\nu_0-\lambda-2\mu_2)}\Bigg] \\
& + \frac{2\lambda\nu_0^2\nu_2^2}{(\mu_2-\mu_0)^2}\Bigg[\frac{e^{(\nu_0-\lambda-2\mu_0)t)}}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} - \frac{2e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} \\
& + \frac{2(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)^2} + \frac{e^{(\nu_0-\lambda-2\mu_2)t}}{(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-2\mu_2)} + \frac{2(\mu_2-\mu_0)e^{-\mu_2 t}}{\mu_2(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)} \\
& + \frac{(\mu_2-\mu_0)^2 e^{(\lambda-\nu_0)t}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)^2} - \frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} \\
& + \frac{2}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{2(\mu_0-\mu_2)}{\mu_0(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)^2} \\
& - \frac{1}{(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-2\mu_2)} - \frac{2(\mu_2-\mu_0)}{\mu_2(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)} \\
& - \frac{(\mu_2-\mu_0)^2}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_2)^2(\nu_0-\lambda-\mu_0)^2)}\Bigg]\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
U_{34|1}(t) = e^{(\lambda-\nu_0)t}\Bigg\{ & \frac{\nu_0\nu_1\nu_2}{\mu_2-\mu_0}\cdot\Bigg[\frac{(\mu_0-\mu_2)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_1(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_0)t}}{\mu_1(\nu_0-\lambda-\mu_1-\mu_0)} \\
& - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{\mu_2-\mu_0}{\mu_1(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{1}{\mu_1(\nu_0-\lambda-\mu_1-\mu_0)} + \frac{1}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)}\Bigg] \\
& + \frac{\nu_0\nu_1\nu_2}{\mu_1-\mu_0}\Bigg[\frac{(\mu_0-\mu_1)e^{(\nu_0-\lambda-\mu_0)t}}{\mu_2(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{e^{(\nu_0-\lambda-\mu_2-\mu_0)t}}{\mu_2(\nu_0-\lambda-\mu_2-\mu_0)} \\
& - \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{\mu_1-\mu_0}{\mu_2(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{1}{\mu_2(\nu_0-\lambda-\mu_2-\mu_0)} + \frac{1}{(\mu_0-\mu_1-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)}\Bigg] \\
& + \frac{2\lambda\nu_0^2\nu_1\nu_2}{(\mu_1-\mu_0)(\mu_2-\mu_0)}\cdot\Bigg[\frac{e^{(\nu_0-\lambda-2\mu_0)t}}{(\nu_0-\lambda-2\mu_0)(\nu_0-\lambda-\mu_0)^2} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_2)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} \\
& + \frac{(\mu_0-\mu_2)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_2)} - \frac{e^{(\nu_0-\lambda-\mu_0-\mu_1)t}}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)} \\
& + \frac{e^{(\nu_0-\lambda-\mu_1-\mu_2)t}}{(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} + \frac{(\mu_2-\mu_0)e^{-\mu_1 t}}{\mu_1(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{(\mu_0-\mu_1)e^{-\mu_0 t}}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)} + \frac{(\mu_1-\mu_0)e^{-\mu_2 t}}{\mu_2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} \\
& + \frac{(\mu_1-\mu_0)(\mu_2-\mu_0)e^{(\lambda-\nu_0)t)}}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)} - \frac{1}{(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-2\mu_0)} \\
& + \frac{1}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0-\mu_2)} - \frac{\mu_0-\mu_2}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_2)} \\
& + \frac{1}{(\nu_0-\lambda-\mu_0)(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_0-\mu_1)} - \frac{1}{(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_1-\mu_2)} \\
& + \frac{\mu_0-\mu_2}{\mu_1(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} + \frac{\mu_1-\mu_0}{\mu_0(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)} \\
& + \frac{\mu_0-\mu_1}{\mu_2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)(\nu_0-\lambda-\mu_0)} - \frac{(\mu_1-\mu_0)(\mu_2-\mu_0)}{(\lambda-\nu_0)(\nu_0-\lambda-\mu_0)^2(\nu_0-\lambda-\mu_1)(\nu_0-\lambda-\mu_2)}\Bigg]\Bigg\}
\end{aligned}
$$

9.0.4 Marginalized variance and covariance derivation

We now include the details behind Equation (B-15) and derive the expressions in the general case with K progenitors. Applying the law of iterated variance, the total variance for a type i mature cell population is given by

$$
\mathrm{Var}X_i(t) = \underbrace{\mathrm{E}[\mathrm{Var}[X_i(t)|\mathbf{X}(0)]]}_{(1)} + \underbrace{\mathrm{Var}[\mathrm{E}[X_i(t)|\mathbf{X}(0)]]}_{(2)}
$$

We drop the dependence on t in intermediate steps for simplicity. The inner part of the first term (1) expands as

$$\mathrm{Var}[X_i|\mathbf{X}(0)] = \mathrm{E}[X_i|\mathbf{X}(0)]^2 - (\mathrm{E}[X_i|\mathbf{X}(0)])^2,$$

and taking the outer expectation over initial barcoding probability, (1) simplifies to

$$\begin{aligned}
\mathrm{E}[\mathrm{Var}[X_i|\mathbf{X}(0)]] &= \mathrm{E}\left[\pi_1 X_{i|1}^2 + \ldots + \pi_K X_{i|K}^2 - (\pi_1 X_{i|1} + \ldots + \pi_K X_{i|K})^2\right] \\
&= \sum_{k=1}^{K} \pi_k(1-\pi_k)\mathrm{E}[X_{i|k}^2] - 2\sum_{j\neq k} \pi_j\pi_k \mathrm{E}[X_{i|j}X_{i|k}] \\
&= \sum_{k=1}^{K} \pi_k(1-\pi_k)\mathrm{E}[X_{i|k}^2] - 2\sum_{j\neq k} \pi_j\pi_k \mathrm{E}[X_{i|j}]\mathrm{E}[X_{i|k}]
\end{aligned}$$

where the final term factors in the last line by independence. Next, it is straightforward to expand (2) as

$$\begin{aligned}
\mathrm{Var}[\mathrm{E}[X_i|\mathbf{X}(0)]] &= \mathrm{Var}[\pi_1 X_{i|1} + \ldots + \pi_K X_{i|k}] \\
&= \sum_{k=1}^{K} \pi_k^2 \mathrm{Var}[X_{i|k}] = \sum_{k=1}^{K} \pi_k^2(\mathrm{E}[X_{i|k}^2] - (\mathrm{E}[X_{i|k}])^2)
\end{aligned}$$

Combining these simplifications (1) + (2), we arrive at the total variance expression marginalized over initial state:

$$\mathrm{Var}X_i(t) = \sum_{k=1}^{K} \pi_k \mathrm{E}[X_{i|k}^2] - \sum_{k=1}^{K} \pi_k^2(\mathrm{E}[X_{i|k}])^2) - 2\sum_{j\neq k} \pi_j\pi_k \mathrm{E}[X_{i|j}]\mathrm{E}[X_{i|k}]. \tag{B-16}$$

As the four-type model example in the main text, this is directly related to the closed form expressions we obtain from solving the systems of moment differential equations: equiva-

lently, (B-16) becomes

$$\mathrm{Var}X_i(t) = \sum_{k=1}^{K} \pi_k[U_{ii|k}(t) + M_{i|k}(t)] - \pi_k^2 M_{i|k}(t)^2 - 2\sum_{j\neq k} \pi_j \pi_k M_{i|k}(t) M_{i|j}(t).$$

The marginal covariance expressions are obtained exactly analogously, applying the law of total covariance instead of the law of total variance. The covariances are given by

$$\mathrm{Cov}X_i, X_j = \sum_{k=1}^{K} \pi_k^2(\mathrm{E}[X_{i|k}X_{j|k}] - \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|k}]) + \pi_k(1-\pi_k)\mathrm{E}[X_{i|k}X_{j|k}] - \sum_{k\neq l} \pi_k \pi_l \mathrm{E}[X_{i|k}]\mathrm{E}[X_{j|l}] \tag{B-17}$$

Given the variance and covariance expressions, incorporating the hypergeometric sampling distribution to obtain covariance and variance between read data $\mathbf{Y}$ applies identically by the equations in the main paper.

Unconstrained parametrization of initial barcoding vector:

For models with multiple progenitor types, the initial barcoding probabilities must be represented as a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ where π_1 denotes the probability of starting as an HSC, and π_i denotes the probability of starting as a type i progenitor for $i = 2, \ldots, K$. These parameters π_i are naturally constrained to a probability simplex, but in practice we reparametrize by borrowing from a technique used in multinomial logistic regression by defining a set of variables $\gamma_i := \ln \frac{\pi_i}{\pi_K}$ for $i = 1, \ldots, K-1$. Then notice $\pi_i = \pi_K e^{\gamma_i}$ for all $i \leq K-1$, and letting $\pi_K = \frac{1}{1+\sum_{i=1}^{K-1} e^{\gamma_i}}$, we ensure the simplex constraint that $\sum_{i=1}^{K} \pi_i = 1$. This enables us to equivalently consider the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K-1})$ as parameters instead of $\boldsymbol{\pi}$, and because γ_i vary freely in $\mathbb{R}$, we no longer need to add a constraint to the optimization problem.

Appendix F

Here we include all model fits in the simulation study and application to rhesus macaque data.

9.0.5 Detailed simulation results

Here, we include detailed tables of true parameters used to initiate simulation as well as median estimates, median absolute deviations, and standard deviations corresponding to the simulation study design discussed in section 6.4.1 for all model structures depicted in Figure 6.2. Some models depicted in Figure 6.2 are identical in simulation study– for instance, models (c) and (d) have no difference when final types are arbitrary. We also note that estimates reported in Table C-3 correspond to the results plotted in Figure 6.3 in the main text.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	π_a
True	0.0280	0.0200	0.0080	36	15	7	0.9000
Median	0.0283	0.0194	0.0086	34.84	14.18	6.624	0.8959
MAD	0.0008	0.0009	0.0021	6.31	2.797	1.167	0.0201
SD	0.0008	0.0010	0.0021	10.33	4.623	1.993	0.0199

Table C-1: Results of estimation on synthetic data from a model with three mature types and one common progenitor compartment, i.e. Model (a) in Figure 6.2. With fixed death rates at $\mu_1 = 0.24, \mu_2 = 0.14, \mu_3 = 0.09$, estimates are very close to true parameters used to simulate the data. Recall π_a denotes the proportion barcoded as progenitors, while $\pi_1 = 1 - \pi_a$ is the proportion marked at the HSC stage.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	ν_4	ν_5	π_a
True	0.0285	0.0200	0.0080	36.00	15.00	10.00	20.00	7.000	0.9000
Median	0.0284	0.0200	0.0076	37.16	15.54	10.35	20.69	7.246	0.9021
MAD	0.0007	0.0011	0.0016	5.851	2.568	1.693	3.399	1.178	0.0153
SD	0.0025	0.0019	0.2800	11.84	3.568	2.504	4.574	1.994	0.0465

Table C-2: Model with five mature types and one common progenitor compartment, i.e. Model (b) in Figure 6.2. Death rates fixed at $\mu_1 = 0.26, \mu_2 = 0.13, \mu_3 = 0.11, \mu_4 = 0.16, \mu_5 = 0.09$.

	λ	ν_a	ν_b	μ_a	μ_b	ν_1	ν_2	ν_3	ν_4	ν_5	π_a	π_b
True	0.0285	0.0130	0.0070	0.0050	0.0040	36	15	10	20	7	0.60	0.30
Med.	0.0286	0.0130	0.0069	0.0045	0.0043	38.01	16.29	10.92	19.64	6.65	0.6333	0.2706
MAD	0.0005	0.0008	0.0006	0.0021	0.0013	13.35	5.826	3.894	2.240	1.277	0.1399	0.1194
SD	0.0006	0.0007	0.0007	0.0019	0.0012	17.61	7.828	5.241	5.347	1.925	0.1388	0.1255

Table C-3: Model with five mature types and two distinct progenitor compartments, i.e. Model (c) in Figure 6.2. In this model, progenitor a gives rise to type 1 and 2 mature cells, and b produces type $3, 4$, and 5 type cells. Estimates remain accurate in this parameter rich setting with multiple progenitor compartments. These correspond to estimates plotted in Figure 6.3.

9.0.6 Model misspecification experiments

Tables C-5 and C-6 display the estimates obtained under over specified and misspecified models, along with objective values of the loss function at converged estimates; these correspond to total ℓ_2 loss between fitted and observed correlations. Note that these estimates correspond to the correlation plots displayed in Figure 6.5 in the main text.

	λ	ν_a	ν_b	ν_c	μ_a	μ_b	μ_c	ν_1
True	0.0500	0.0280	0.0140	0.0070	0.0080	0.0060	0.0020	40.0000
Median	0.0539	0.0303	0.0150	0.0075	0.0091	0.0058	0.0034	40.7977
MAD	0.0081	0.0047	0.0032	0.0016	0.0038	0.0072	0.0041	11.8020
SD	0.0143	0.0080	0.0052	0.0024	0.0037	0.0060	0.0053	18.0492
	ν_2	ν_3	ν_4	ν_5	π_a	π_b	π_c	
True	18.0000	14.0000	20.0000	8.0000	0.5500	0.2000	0.1500	
Median	18.1527	17.7127	26.4716	10.6550	0.5595	0.2017	0.1578	
MAD	5.0599	7.8044	9.4919	5.7547	0.0412	0.0120	0.0106	
SD	7.0998	6.6657	8.8583	157.5674	0.0369	0.0159	0.0137	

Table C-4: Synthetic data from a model with five mature types and three oligopotent and unipotent progenitors, i.e. Model (f) in Figure 6.2. Death rates fixed at $\boldsymbol{\mu} = (.24, .13, .12, .18, .1)$. While the standard deviation reveals influence of extreme outliers on the estimate or ν_4, median estimates are again accurate in a parameter rich model, and reasonably stable in terms of MAD.

	λ	ν_a	ν_b	ν_c	μ_a	μ_b	μ_c	ν_1
Med.	0.19365	0.05938	0.05475	0.00002	0.01136	0.19085	0.00056	56.89686
MAD	0.06633	0.02942	0.02433	0.00003	0.01683	0.28294	0.00083	38.64162
SD	0.07195	0.03583	0.03360	0.00015	0.59207	0.95299	0.00226	238.53322
	ν_2	ν_3	ν_4	ν_5	π_a	π_b	π_c	Objective
Med.	22.10742	16.70475	33.70443	11.65951	0.00121	0.00038	0.83830	2.90319
MAD	14.61860	9.40493	19.63489	5.62197	0.00179	0.00057	0.07087	0.75504
SD	11.94989	7.67183	15.40404	5.04462	0.01796	0.00948	0.08250	1.08521

Table C-5: Model fit in overspecified case with three progenitors: note that the objective value is higher than the correct specification, and note that the estimates seem more spread apart than the correctly specified inference while representative of the overall shape of true correlation profiles.

	λ	ν_a	μ_a	ν_1	ν_2	ν_3	ν_4	ν_5	π_a	Objective
Med.	0.131	0.00468	0.0332	71.1	30.0	20.6	0.000	0.000	1.000	21.358
MAD	0.0091	0.0041	0.0123	29.9	12.7	7.87	0.00000	0.00000	0.00001	0.221
SD	0.0096	0.0078	0.0149	23.7	9.91	6.43	0.00000	0.00000	0.00001	0.227

Table C-6: Underspecified model fit. Interestingly, this model seems to correctly identify that types $1, 2, 3$ are linked from a common progenitor, but because one shared progenitor is not compatible with the observed correlations, and in particular cannot explain negative correlations between types from distinct lineages, the model assigns almost zero mass to rates ν_4, ν_5 of producing the other mature types. The solution seems to be strongly a boundary solution with all barcoded cells starting in the progenitor compartment, resulting in a very poor objective function value.

	λ	ν_a	ν_b	μ_a	μ_b	ν_1	ν_2	ν_3	π_2	π_3	Obj.
Med.	0.0286	0.0130	0.0080	0.0077	0.0014	31.43	21.32	46.57	0.533	0.358	9.093×10^{-5}
MAD	0.0007	0.0009	0.0011	0.0022	0.0020	6.595	5.097	29.47	0.1096	0.1009	4.629×10^{-5}
SD	0.0067	0.0043	0.0028	0.0026	0.0022	22.31	21.54	63.07	0.1791	0.1943	6.950×10^{-5}

Table C-7: Results corresponding to three grouped mature cell compartments with correctly specified progenitor structure are much more sensical, with noticeably lower objective values at convergence.

9.0.7 Tables of complete estimated parameters fitted to lineage barcoding data

Par	(a)	(b)	(c)	(d)	(e)	(f)
$\hat{\lambda}$	0.0593	0.0867	0.4360	0.3644	.2271	0.3198
$\hat{\nu}_a$	1.00e-6	1.80e-7	0.4090	0.3521	0.1446	0.0033
$\hat{\nu}_b$			0.0257	0.0121	0.0725	0.3131
$\hat{\nu}_c$					0.0101	0.0033
$\hat{\mu}_a$	7.95e-6	0.0367	1.150	4.096	0.7037	0.1449
$\hat{\mu}_b$			4.023	3.699	4.022	1.253e-3
$\hat{\mu}_c$					3.602	1.434
$\hat{\nu}_1$	2042.0	1486.3	1305.5	866.1	1896.8	1959.09
$\hat{\nu}_2$	434.7	1764.3	201.4	391.3	221.3	560.4
$\hat{\nu}_3$	147.4	74.0	113.6	264.4	112.3	127.5
$\hat{\nu}_4$		326.4	448.7	299.5	417.1	287.9
$\hat{\nu}_5$		17.9	17.0	54.1	79.3	104.2
$\hat{\pi}_a$	0.861	0.87^*	0.870	0.870	0.0	0.0
$\hat{\pi}_b$			0.0	0.0	0.0	0.870
$\hat{\pi}_c$					.870	0.0
Loss	0.4071	1.653	3.465	3.330	3.836	2.91

Table C-8: Best fitting parameters are estimated for all models displayed in Figure 6.2. Model (a) has fixed deaths $(.6, .04, .4)$. All other models have fixed death rates $(.8, .3, .04, .08, .4)$.

Par	(a)	(b)	(c)	(d)	(e)	(f)
$\hat{\lambda}$	$(0.003, 0.109)$	$(0.077, 0.163)$	$(0.196, 1.168)$	$(0.207, 0.640)$	$(0.132, 0.752)$	$(0.174, 0.449)$
$\hat{\nu}_a$	$(0.0, 0.004)$	$(0.0, 0.001)$	$(0.085, 1.148)$	$(0.131, 0.611)$	$(0.014, 0.617)$	$(0.074, 0.396)$
$\hat{\nu}_b$			$(0.006, 0.168)$	$(0.006, 0.112)$	$(0.015, 0.486)$	$(0.021, 0.154)$
$\hat{\nu}_c$					$(0.000, 0.019)$	$(0.000, 0.010)$
$\hat{\mu}_a$	$(0.0, 0.002)$	$(0.028, 0.046)$	$(0.000, 3.879)$	$(0.267, 3.603)$	$(0.0, 2.854)$	$(0.246, 2.651)$
$\hat{\mu}_b$			$(0.434, 4.022)$	$(0.437, 4.102)$	$(0.447, 4.023)$	$(0.811, 4.543)$
$\hat{\mu}_c$					$(0.102, 4.103)$	$(0.283, 4.100)$
$\hat{\nu}_1$	$(956.0, 2239.9)$	$(830.1, 1838.6)$	$(627.7, 1482.2)$	$(613.3, 1487.3)$	$(600.9, 1495.0)$	$(615.7, 1474.2)$
$\hat{\nu}_2$	$(52.0, 488.7)$	$(1021.9, 2055.3)$	$(131.8, 521.4)$	$(135.6, 344.6)$	$(187.9, 477.2)$	$(255.9, 448.7)$
$\hat{\nu}_3$	$(39.7, 148.2)$	$(60.7, 99.8)$	$(30.6, 294.6)$	$(134.5, 305.0)$	$(4.279, 291.3)$	$(126.3, 297.8)$
$\hat{\nu}_4$		$(275.2, 470.7)$	$(146.2, 558.7)$	$(126.4, 297.4)$	$(127.8, 321.7)$	$(128.2, 295.7)$
$\hat{\nu}_5$		$(10.1, 44.65)$	$(3.786, 9.559)$	$(1.137, 10.63)$	$(6.488, 84.9)$	$(26.4, 74.0)$
$\hat{\pi}_a$	$(0.017, 0.861)$	$(0.87, 0.87)$	$(0.0, .0.599)$	$(0.0, .598)$	$(0.0, 0.038)$	$(0.000, 0.001)$
$\hat{\pi}_b$			$(0.0, 0.999)$	$(0.0, 0.999)$	$(0.0, 1.0)$	$(0.999, 1.0)$
$\hat{\pi}_c$					$(0.0, 1.0)$	$(0.000, 0.000)$
Loss	$(0.352, 0.696)$	$(1.485, 2.341)$	$(2.591, 4.771)$	$(2.472, 4.489)$	$(3.092, 5.763)$	$(2.566, 4.777)$

Table C-9: Corresponding 95% confidence intervals produced via nonparametric bootstrap of 2500 replicate datasets. Recall sum of progenitor barcoding proportions fixed to be .87 for models (b)-(f).

9.0.8 Additional fitted correlation profiles fitted to lineage barcoding dataset

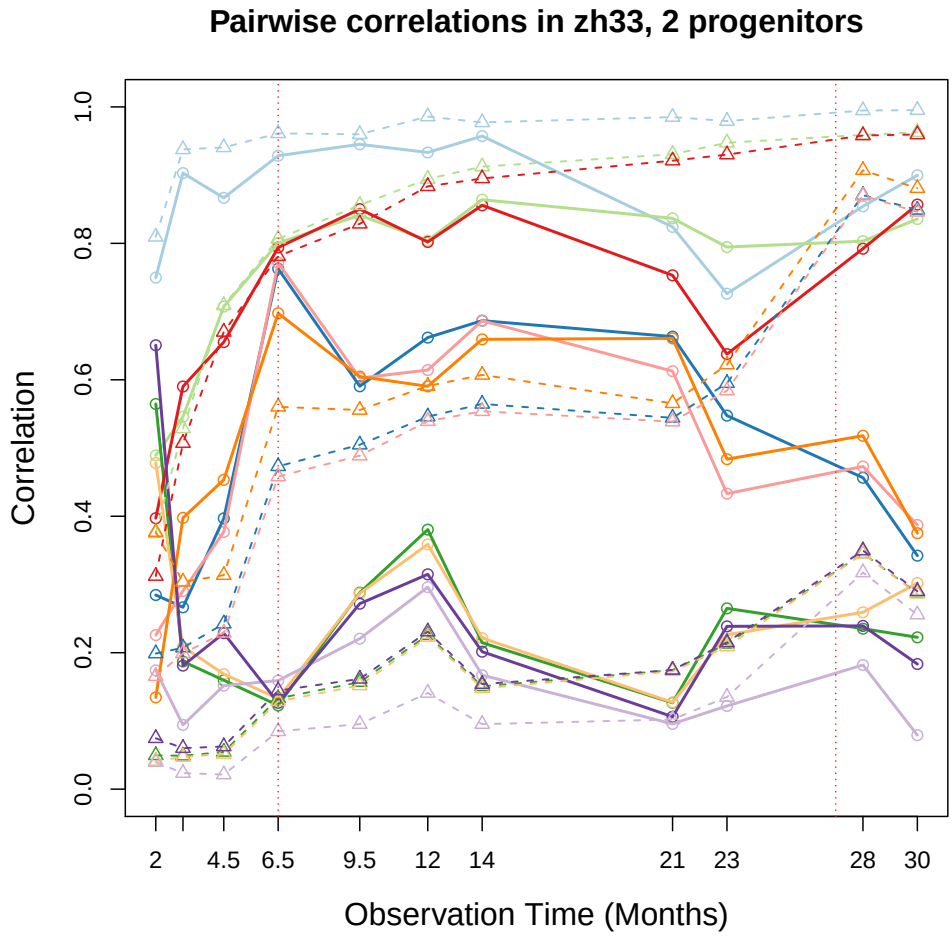


Figure C-3: Fitted curves for real data in models with two progenitors, corresponding to model (c) displayed in Figure 6.2. The "misgrouped" fitted curves apparent after 23 months visually suggest the misspecification in designating specialized oligopotent progenitors.

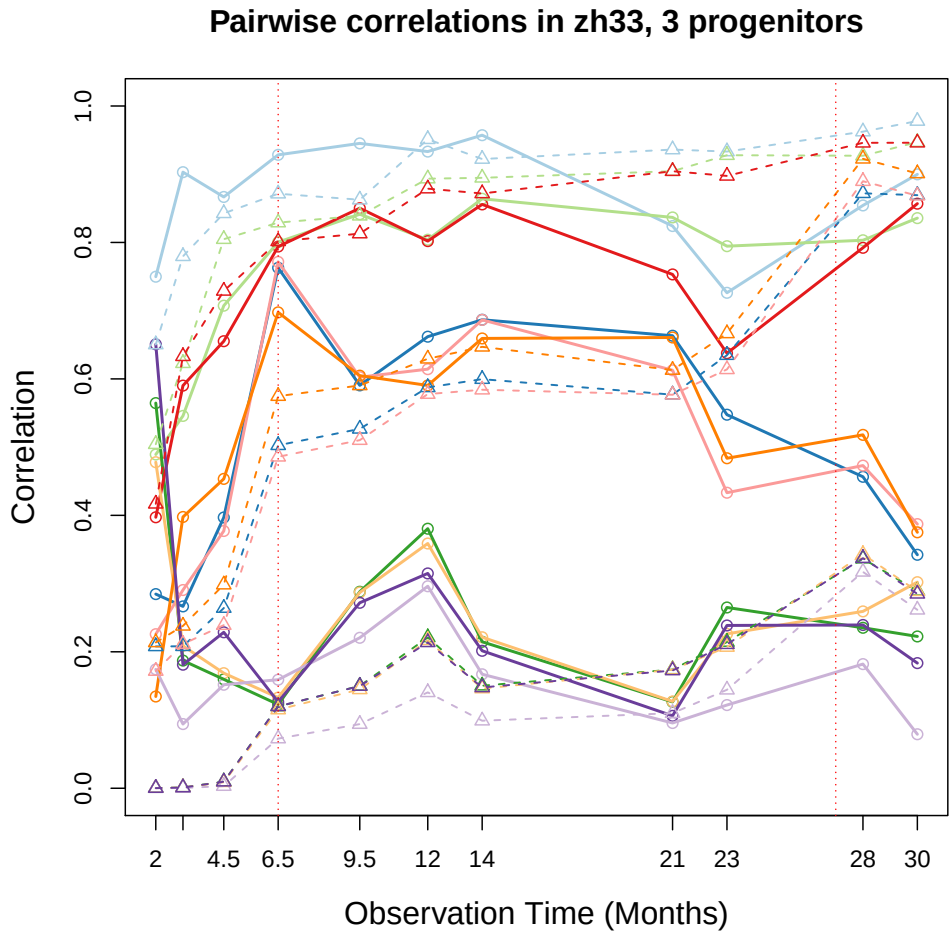


Figure C-4: Fitted curves for real data in models with two and three specialized progenitors, i.e. model (e) in Figure 6.2. Again, a misgrouping is visually apparent in fitted curves after 23 months