A Stochastic Dollo model for lateral transfer

Luke J. Kelly

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

St John's College

Trinity, 2016

I would like to dedicate this thesis to Pepper and Fiona.

Acknowledgements

I would like to acknowledge the support of my family throughout my education, and particularly that of Jessie, Gurjinder and Zhe during my doctorate. I would also like to acknowledge the contribution of the following people: Simon Greenhill, for providing the Polynesian language data set and feedback on our manuscript; Robin Ryder, for talking me through `TraitLab`; and my examiners, Jotun Hein and Alex Bouchard-Côté. I gratefully received financial support from St John's College and The Engineering and Physical Sciences Research Council (partnership award EP/J500495/1) and the Department of Statistics. Most of all, I would like to thank my supervisor, Geoff, for being a constant source of inspiration.

Abstract

Lateral transfer, a process whereby species exchange evolutionary traits through non-ancestral relationships, is a frequent source of model misspecification in phylogenetic inference. Lateral transfer obscures the phylogenetic signal in the data — the signal of the taxa ancestry — as the histories of affected traits are mosaics of the species phylogeny and may conflict with the underlying phylogeny. We control for the effect of lateral transfer in a Stochastic Dollo model and a Bayesian setting. We infer rooted phylogenetic trees. Our likelihood is highly intractable as its parameters are given by the solution of a sequence of systems of differential equations which represent the expected evolution of traits along a tree and grow exponentially in dimension with the number of taxa under consideration. We construct an accurate parameter approximation framework, and from this we derive an efficient exact-approximate inference scheme. We illustrate our method on data sets of lexical traits in Eastern Polynesian and Indo-European languages and obtain improved fits over the corresponding model without lateral transfer.

Table of contents

List of figures

List of tables

Introduction

Phylogenetics is the study of the evolutionary relationships between a set of *phyla*, or *taxa*, descended from a common ancestor. These evolutionary relationships define the *phylogeny* of the taxa. We describe some common types of phylogenies in Figure 1.1 of Chapter 1. A taxon is distinguished by the complex evolutionary traits shared by its members. These traits take many forms, such as characters in a DNA sequence or phenotypes displayed by individuals. We are interested in inferring the phylogeny of the taxa from these overlapping sets of traits.

Solely tree-based models of evolution assume that traits pass vertically from one generation to the next through ancestral relationships. *Lateral transfer* violates this assumption. Although lateral transfer is not a factor in every evolutionary process, *horizontal gene transfer* is a common feature of prokaryote evolution (Jain et al., 1999; Koonin et al., 2001), for example, where it facilitates the rapid spread of antibiotic resistance in bacterial populations (Barlow, 2009). Similarly, trait *borrowing* is a factor in the diversification of languages (Gray et al., 2010) and cultures (O'Brien et al., 2001; Tehrani et al., 2010). This thesis focuses on the problem of inferring phylogenies when traits may undergo both vertical and lateral transfer.

To address this problem, we build a detailed *ab initio* model of tree and trait dynamics which explicitly incorporates lateral transfer in the evolutionary process. We model trait presence and absence across taxa. Exact inference under our model is a difficult computational task and our efforts to reduce this burden form a major part of this thesis. In doing so, we do not compromise the model or the exactness

of our inference. Rather, we exploit properties of the model and techniques from numerical analysis to accurately approximate the likelihood parameters and build an efficient *exact-approximate* inference scheme. By exact-approximate here, we mean that the Markov chain Monte Carlo (MCMC) algorithm is asymptotically exact, but approximate in the sense that we do not use exact transition probabilities at every step of the chain.

We present three major results in this thesis.

- The first result is our novel Bayesian phylogenetic model which we successfully apply to a number of data sets where lateral transfer is a known problem and existing models are misspecified.

- The second result is our likelihood parameter calculation whereby we describe how to solve systems of differential equations on a tree.

These results are of particular interest to practitioners as lateral transfer is a frequent problem in many applications yet they lack fully model-based tools to account for it.

- The third major result is our exact-approximate inference scheme. We describe a framework to reduce the computational cost of evaluating the likelihood parameters on a tree with L leaves from $\mathcal{O}[L2^L C(L)]$ operations, where $C(L)$ grows linearly with L, to $\mathcal{O}[L2^L D(k)]$, where $D(k) = 2^{k+4} \ll C(L)$. The parameter k is chosen by the user and does not depend on L. The resulting decrease in sampling efficiency here compared to the corresponding exact inference scheme is more than offset by the decrease in computation time, thereby yielding a higher effective number of samples per unit time.

This result relates to computing the product of the exponential of a matrix $\mathbf{A}$ and a vector $\mathbf{x}$. For the model we describe in this thesis, the matrix $\mathbf{A}$ is a large, sparse matrix of Markov chain transition rates. Computing the exponential of a matrix is a difficult numerical problem (Moler and Van Loan, 1978, 2003). Furthermore, we cannot perform this operation explicitly for the size of problems we consider as $\exp(\mathbf{A})$

is dense and does not fit in computer memory. A straightforward approach which avoids this memory issue is to evaluate $e^{\mathbf{A}}\mathbf{x}$ as the solution of the differential equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. Typically, the number of matrix-vector multiplications required to compute the solution increases with the size of the problem. The accuracy of our approximation scheme does not depend on the size of the problem and it reduces the computational cost of the drawing an effectively independent sample from our model by a factor between 5 and 10 compared to the differential equation solver-based approach.

The structure of this thesis is as follows:

- In Chapter 1, we present an overview of the field of phylogenetics, discuss various types of phylogenies and the evolutionary processes giving rise to them, and illustrate the problem of lateral transfer and why it is important to control for it when performing inference.

- We describe our model in Chapter 2. We begin with a description of the data to motivate the generative model of the observation process which we then describe. We follow this with a description of the likelihood-parameter calculation and conclude with our Bayesian model.

- In Chapter 3, we illustrate how to perform exact inference under the model using Markov chain Monte Carlo methods. It is difficult to quantify how well a MCMC sampler explores its target distribution, particularly in a complex setting such as a posterior distribution on phylogenetic trees. To address this issue, we implement the Wang–Landau algorithm, an exact MCMC algorithm which forces the chain to explore its sample space. Finally, we test our implementation of the simulation algorithms by sampling from the prior distribution and comparing the output to the corresponding marginal prior distributions.

- We return to the likelihood parameter calculation in Chapter 4. We describe how to exploit symmetries in the sequence of initial value problems describing the evolution of the parameters along the tree, and from this we construct an efficient

approximation scheme with a significantly lower computational cost than the exact approach. We then describe how to construct an unbiased estimate of the likelihood and perform exact inference in a pseudo-marginal MCMC framework.

- In Chapter 5, we describe a number of tests to validate our model and software implementation by sampling from the posterior distribution of the model fit to synthetic data sets. We describe two general frameworks for assessing goodness-of-fit and performing model selection.

- In Chapter 6, we apply our exact and exact-approximate methods to a data set of lexical traits in Eastern Polynesian languages. Here, a taxon represents a language and a trait is a sound-meaning pair. We then illustrate our approximate inference scheme on a larger data set of lexical traits in Indo-European languages. Both language families are extensively studied, but these represent the first analyses to properly control for lateral transfer.

We conclude the thesis with some remarks about the model and possible directions for future work. The appendices contain a proof of a result in the main text, along with figures of secondary importance for the analyses in Chapters 5 and 6.

With the exception of Chapter 4, many of the results in this thesis appear in the following manuscript:

> L.J. Kelly and G.K. Nicholls. Lateral transfer in Stochastic Dollo models. Manuscript submitted for publication, ArXiv:1601.07931, 2016,

currently under review in *The Annals of Applied Statistics*.

Chapter 1

A brief introduction to phylogenetics

Chapter overview

We first present a broad overview of the field of phylogenetics then focus on the problem at the heart of this thesis, lateral trait transfer.

1.1 Types of phylogenies

Broadly speaking, we can divide phylogenies into *rooted* and *unrooted trees* and *networks*. In each case, the labelled external *leaf* nodes of a phylogeny represent the observed taxa. A tree describes the ancestry of taxa which evolve through speciation events only. In addition to ancestral relationships, networks depict evolutionary contact events between species. Internal nodes in trees represent ancestral states, but this is not always the case for a network. Similarly, the edges in a tree represent evolving species, but again this is not necessarily true for a network. In a rooted phylogeny, edges are directed forwards in time from the root to the leaves. As the name suggests, we do not know the root in an unrooted tree, but these trees may still contain a time

component. We illustrate each of these graph structures in Figure 1.1. Typically, the more information a phylogeny contains, the more difficult it is to infer accurately. For example, an unrooted tree is relatively simple structure to infer, whereas a rooted graph is much more difficult.

There is a huge variety of methods for reconstructing phylogenies of each type. Most techniques have their origins in the *comparative method*: they compare the differences and similarities in traits across taxa to infer the phylogeny (Felsenstein, 1985). We consider phylogeny reconstruction to be a statistical inference problem so we mainly focus on probabilistic, model-based methods in this chapter.

Model-based approaches to phylogenetic inference are often tailored to a specific type of data-generating process and, as such, rely on explicit, restrictive sets of assumptions. These methods attempt to find a combination of phylogeny and model parameters which best describe the data under the model, or combination thereof. These models are more difficult to fit than the non-model-based approaches we describe below.

Non-model-based methods do not propose a generative model for the data. They typically rely on a few basic assumptions and, as a result, may be applied to many data sets. For example, distance-based methods aim to return the phylogeny which minimises the distance between taxa according to some measure on the phylogeny and data. The space of trees is enormous — there are $(2L-3)!! = (2L-3)\cdot(2L-5)\cdots 3\cdot 1$ rooted and $(2L-5)!!$ unrooted bifurcating tree topologies on L taxa (Billera et al., 2001) — so a key element of any method, model-based or otherwise, is how we perform this search through tree space. Although fast algorithms exist for fitting many of these non-model-based methods, we only have heuristic tools such as bootstrapping to quantify the uncertainty in our inference. This is a major weakness of non-model-based approaches.

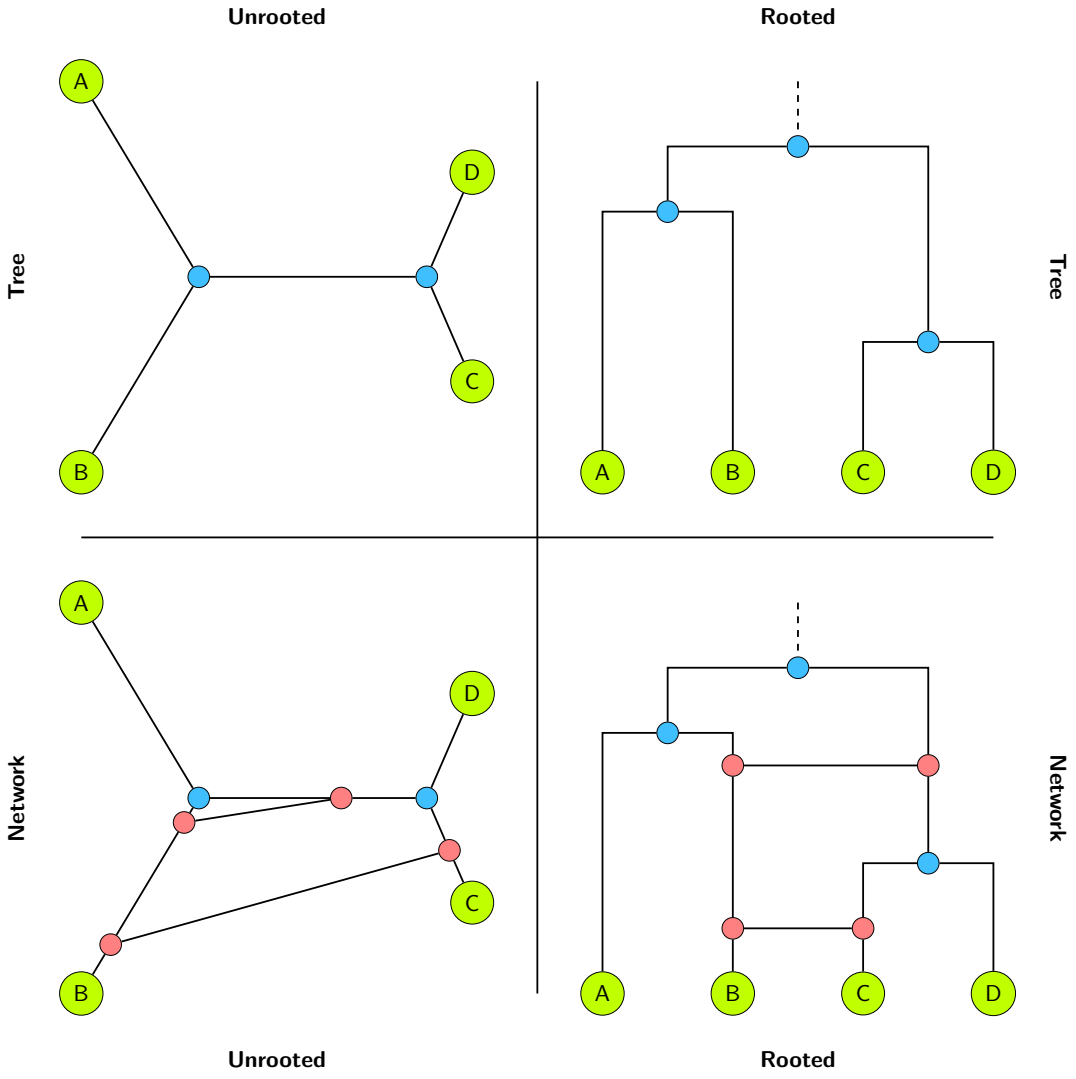


Figure 1.1: Rooted and unrooted phylogenetic trees and networks. Ancestral nodes are coloured blue and leaf nodes are in green. Red nodes and the edges between them represent contact events. Here, we represent contact events occurring at the same time on the respective branches; however, this may not be the case in practice.

1.2 Model-based phylogenetic methods

To reiterate a point in the previous section, model-based phylogenetic methods are specific to the types of processes they attempt to model. A recurring theme in this chapter will be that, typically, the more assumptions we make for our model, the better and more precise our inference will be, assuming that the assumptions we make are satisfied in practice. We begin this section with a description of vertical trait transfer, an evolutionary process whereby traits pass through ancestral relationships only, then discuss some methods to infer phylogenies in this setting. We then describe the problem posed by lateral transfer.

1.2.1 Vertical trait transfer

In a homogeneous population, traits evolve and spread through its members via breeding or other means. Traits which convey a selective advantage randomly occur and pass from one generation to the next through ancestral relationships at the expense of other, less advantageous traits. Eventually these traits spread throughout the population and may become a distinguishing feature of the species. A speciation event occurs when the population fragments into separate groups. These groups subsequently evolve in isolation, unable to interbreed, and eventually become distinct. If we consider a set of taxa descended from a common ancestor in this way, their shared ancestry represents a phylogenetic tree, such as the rooted and unrooted trees in Figure 1.1, for example. This tree structure, introduced by Darwin (1859), clearly describes the sequence of speciation events which lead to the diverse set of taxa we observe, and, in the case of trees with a time component, the elapsed evolutionary time between these branching events.

The challenge is to infer the tree from the trait data. When the individual trait relationships are unknown, we may also infer them as part of the inference (Redelings and Suchard, 2005; Hall and Klein, 2010). We are not concerned with this aspect of

phylogenetic inference in this thesis, however, and from here on we assume that all trait relationships are identifiable. We make the equivalent, simplifying assumption that *parallel* and *convergent evolution* do not occur; that is, the same trait does not evolve independently in different species.

Returning to our inference problem, to proceed, we specify a probabilistic model of tree and trait dynamics then take either a frequentist or Bayesian approach to inference. The trait process parameters typically parameterise a continuous-time stochastic process along the branches of the tree. In the absence of constraints on the branch lengths, parameters, or both, these models suffer from issues with rate-time identifiability; that is, we cannot uniquely identify the event times, equivalently the branch lengths, from the rate parameters. We return to this issue below. The assumption that different lineages evolve independently of each other is a significant one from a modelling perspective. The unobserved trait histories are latent variables in the model so this assumption means that on a tree with L leaves, we can easily integrate the ancestral states out of the model likelihood in $\mathcal{O}(L)$ operations with Felsenstein's pruning algorithm (Felsenstein, 1981).

Frequentist approaches typically seek to maximise the model likelihood function. Erdõs et al. (1999) describe conditions under which these models are identifiable — these results also apply to the Bayesian models we describe below for reasonable choices of prior distributions. We can test hypotheses about trees in a frequentist framework, and Steel et al. (2009) show that testing whether a candidate tree is the true tree is a relatively simple task compared to finding it in the first place. The shape of the likelihood function in phylogenetic models typically contains many peaks and troughs (Whidden and Matsen IV, 2016) and may be difficult to optimise even when all but one of the parameters are fixed (Dinh and Matsen IV, 2015). Annealed importance sampling (Neal, 2001a) provides one possible solution to this NP-hard optimisation problem (Roch, 2006). The primary difficulty with frequentist phylogenetic inference is quantifying the uncertainty in our inference. For example. building on work by

Billera et al. (2001) about the geometry of the space of trees, Willis (2016) describes a method to estimate frequentist confidence sets of phylogenetic trees from a set of samples. To do so, Willis first maps the sampled trees into Euclidean space, forms a confidence set there, then finds the trees which map into this set. However, the inverse of this map may only be computed approximately, and the mapping operation does not necessarily preserve all the properties of interest in the tree samples. Furthermore, it is non-trivial to account for uncertainty in the method used to construct the input trees themselves.

Bayesian approaches are a more recent development in phylogenetic inference, and we may easily quantify the uncertainty in our inference in this setting. Rannala and Yang (1996) perform conditional Bayesian inference. They first propose a birth-death process prior on a five-taxon tree. For each *labelled history* (a time-ordering of the internal nodes representing the relative timing of the speciation events), they compute maximum likelihood estimates of the process parameters. Finally, their method returns the *max a posteriori* labelled history. Rannala and Yang work with labelled histories in order to avoid the issue with rate-time identifiability we describe above. Furthermore, there are $L!(L-1)!/2^{L-1}$ labelled histories on a L-taxon tree (Edwards, 1970), so this approach is only feasible when there are sufficiently few taxa that we may enumerate the entire space of labelled histories.

In practice, many topologies are extremely unlikely and, under some distance metric in tree space, are often surrounded by similarly unlikely topologies. Following their earlier work, Yang and Rannala (1997) use a Markov chain Monte Carlo approach to perform a probabilistic integration over the space of labelled histories on a larger number of taxa. They sample each labelled history in proportion to its posterior probability. From a set of such samples, we can easily estimate probabilities and summary statistics and test hypotheses about our inference. We describe MCMC inference for phylogenies in greater detail in Chapter 3.

Notable advances in Bayesian phylogenetic inference since the breakthrough by Yang and Rannala include methods to jointly estimate the tree and rate parameters of the evolutionary process (Wilson and Balding, 1998; Drummond et al., 2002), and a technique to simultaneously estimate the alignment of molecular sequences and the phylogenetic tree relating them (Redelings and Suchard, 2005). More recent developments include sequential Monte Carlo methods to infer phylogenies (Bouchard-Côté et al., 2012; Wang et al., 2015), and methods which control for the correlated evolution of traits through latent variables (Cybis et al., 2015).

Of particular interest to us is the *Stochastic Dollo* (SD) model for unordered sets of traits (Nicholls and Gray, 2008). The SD model posits a birth-death process of traits along branches of the tree: new traits are born at rate λ; traits are copied into offspring lineages at a speciation event; and instances of traits die independently at rate μ. The basic process respects *Dollo parsimony* (Farris, 1977): each trait may be born exactly once on the tree, and if all the copies of a trait die, it remains extinct. The basic SD model describes trait presence/absence only. Alekseyenko et al. (2008) extend the SD model to multiple character states, and Ryder and Nicholls (2011) introduce rate heterogeneity across branches and missing data. Bouchard-Côté and Jordan (2013) describe the *Poisson Indel Process*, a sequence-valued counterpart to Stochastic Dollo. The binary SD model of Nicholls and Gray and its extension by Ryder and Nicholls form the basis of the method we describe in this thesis.

While undoubtedly important, the methods we describe so far in this chapter only apply when species evolve solely through vertical transfer. This is not always the case in practice, and we now describe a more general class of methods which allow for both vertical and lateral trait transfer.

1.2.2 Vertical and lateral trait transfer

Vertical transfer is not the only mechanism by which species diversify. Lateral trait transfer, such as *horizontal gene transfer* in biology or *borrowing* in linguistics, is a

process driving the evolution of populations whereby species acquire traits from contemporary species through non-ancestral relationships. The term *transfer* is a misnomer here as we assume that the recipient acquires a copy of the trait from the donor, rather than the trait itself.

When both vertical and lateral transfer occur, individual trait histories are a mosaic of the species phylogeny. As such, while each individual trait history is a tree, it may conflict with the overall phylogeny; that is, the order of speciation events in the trait tree do not reconcile with the species tree. This has the effect of obscuring the phylogenetic signal of the branching events in the data. Models which ignore lateral transfer are misspecified in this setting, often severely so. In our experience, this model error can lead to spurious levels of confidence in poorly fitting models. For example, for the class of Stochastic Dollo models we describe above, simulation studies show that topology estimates are typically only robust when the true topology is balanced (Greenhill et al., 2009), and estimates of the root time tend to be biased towards the present (Nicholls and Gray, 2008; Ryder and Nicholls, 2011).

Hybridisation is commonly understood to correspond to instantaneous bursts of lateral transfer between a pair of species, and corresponds to the rooted phylogenetic graph in Figure 1.1, for example, where the horizontal branches connecting non-ancestral nodes may correspond to hybridisation events. The graph is *hard wired* if a trait transfers along every contact edge it encounters, and *soft wired* otherwise (Huson and Scornavacca, 2011). We focus our attention on *hybridisation* models for now and then switch to more general models of lateral transfer.

Patterson et al. (2012) review a number of tests for *admixture*, such as hybridisation, in allele frequency data. Similarly, there are many methods to test for lateral transfer in sequence data which compare gene trees to an overall species tree inferred *a priori* (Daubin et al., 2002; Beiko and Hamilton, 2006; Abby et al., 2010). None of these methods are model based, and in the case of the latter rely heavily on the quality of the input species trees.

The unrooted phylogenetic graph in Figure 1.1 is an example of an *implicit* phylogenetic network. Internal nodes in these graphs accommodate incompatibilities in the data with the assumption of an underlying species tree, but do not necessarily represent the evolutionary history of the taxa (Huson and Bryant, 2006; Oldman et al., 2016). The `Neighbor-Net` algorithm (Bryant and Moulton, 2004) is an example of implicit phylogenetic network inference method. This method takes a distance matrix between data points as input and uses a neighbour-joining algorithm to return a phylogenetic network. Gray et al. (2010) perform a `Neighbor-Net` analysis of Polynesian languages, and produce a network which describes the complex relationships among languages due to vertical and lateral transfer. We return to this analysis in Chapter 6 when we analyse a subset of this data set.

We are not concerned with only testing for lateral transfer in a data set or building implicit phylogenetic networks. Rather, we are interested in inferring meaningful rooted phylogenies from data which has undergone lateral transfer so, from here on, we focus on model-based inference methods. Incidentally, the method we propose in this thesis allows us to test for lateral transfer and, with a certain parameterisation, may replicate a rooted, explicit phylogenetic network.

Lathrop (1982) and Pickrell and Pritchard (2012) propose frequentist inference methods for allele frequency data. Both methods attempt to infer the order of population splits and a finite number of instantaneous hybridisation events. Lathrop performs exact, likelihood-based inference, whereas Pickrell and Pritchard make an approximation so that their model is tractable. The respective authors describe frameworks to test the significance of the inferred hybridisation events under their models. However, similar to the frequentist methods in the previous section, we cannot adequately quantify the uncertainty in our inference.

The *coalescent* is a retrospective model in population genetics for gene genealogies (Kingman, 1982). The *multispecies* coalescent extends this framework to multiple gene trees within a species tree, a setting more suited for phylogenetic inference (Ran-

nala and Yang, 2003). Kubatko (2009) uses the multispecies coalescent model to perform model selection on a soft-wired *explicit* phylogenetic network with a fixed number of hybridisation events. Wen et al. (2016) perform Bayesian inference under the same model with a variable number of reticulation nodes. Both of these methods take gene trees inferred *a priori* as input and reconcile them with the network via the hybridisation events. The number of possible gene-tree/species-tree reconciliations increases rapidly with the number of hybridisation nodes, and the authors focus on data sets with few taxa where inference is tractable. For example, Wen et al. (2016) state that seven taxa represents the upper limit for their method on available hardware. In addition, as with the tests for lateral transfer described above, these methods are at the mercy of the input gene tree reconstructions, and the potential biases here are well studied (Pamilo and Nei, 1988; Hahn, 2007; Rasmussen and Kellis, 2007).

Returning to the point that histories of traits which underwent lateral transfer are a mosaic of the species tree, Suchard et al. (2003) fit a multiple changepoint model which adaptively partitions the data and infers different trees for each partition. Suchard (2005) proposes a model whereby each trait tree is formed by a random walk through tree space starting at the species tree. This approach requires a matrix exponential operation to compute the transition probabilities, thereby restricting its application to small data sets. Recent developments in approximating transition probabilities (Crawford et al., 2015; Ho et al., 2016), as well as the techniques described in Chapter 4 of this thesis, may allow this method to be extended to larger numbers of taxa and traits.

Nakhleh et al. (2005) describe a model whereby additional contact edges are added to an input tree inferred *a priori* to facilitate lateral transfer. This method is not model based, rather, it seeks to optimise a *weighted maximum compatibility* criterion. Warnow et al. (2006) describe but do not fit a probabilistic version of this model. Warnow et al. propose that when a trait encounters a contact edge, it transfers along it with some

probability. In the terminology above, both of these methods return networks which are explicit and soft wired.

Hybridisation differs from *recombination* in population genetics which is when two parent lineages merge to form a single offspring. Similar to the trait histories in phylogenetic models, the *ancestral recombination graph* (Hudson, 1983) is often treated as a nuisance parameter in population genetics as the aim is to model variability within populations rather than across them. Bloomquist and Suchard (2010) use the ancestral recombination graph as a basis for modelling lateral transfer. In their model, each trait history is a tree derived from the graph. In more detail, for each trait and recombination node, they include a variable to indicate which parent branch a trait tree derives from. The space of ancestral recombination graphs is enormous, as is the number of trait trees which may be derived from a graph, so inference under this model is difficult.

The approaches we describe so far are realistic when lateral transfer occurs in instantaneous bursts of hybridisation. Outside of this, they are feasible when the overall level of lateral transfer is so low that explicit reticulation nodes can capture most of the variability without introducing noticeable biases into the rest of the model.

One frequent approach in these settings is to simply discard known-transferred traits from the data and fit a tree-based model of vertical transfer to the remainder (Gray and Atkinson, 2003; Nicholls and Gray, 2008; Ryder and Nicholls, 2011; Bouckaert et al., 2012; and many others). This approach is problematic when we consider that recently transferred traits are more readily identified and removed from the data. While this action may improve the resolution of the branching events near the leaves, two issues arise: unidentified transfers remain in the data, so a vertical model is still misspecified here; and the discarded traits are not a random thinning of the data. We return to this source of model misspecification in Chapter 5.

As the number of contact events increase, all of the above methods struggle to adequately represent the data-generating process. We now turn to the more general

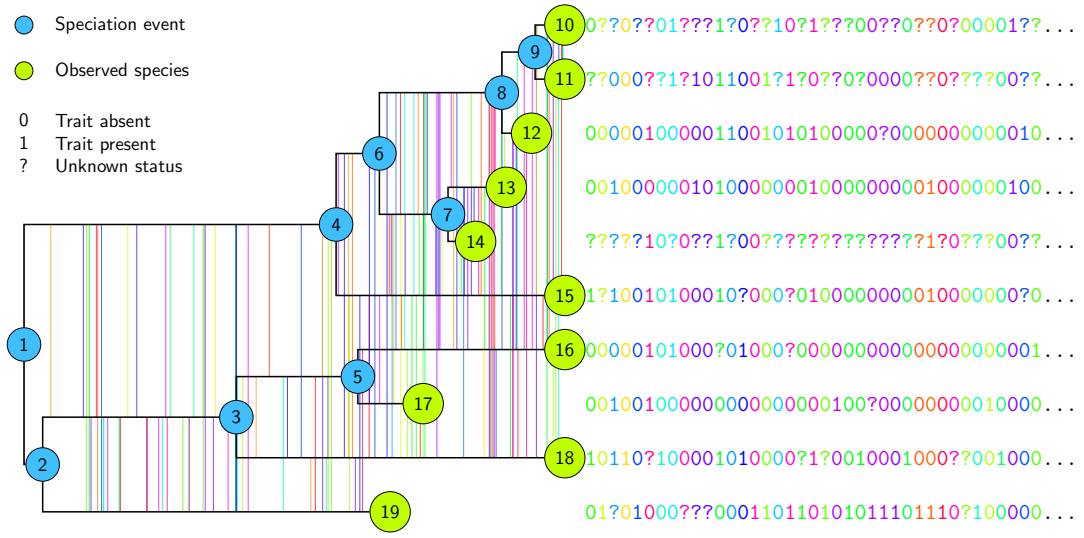


Figure 1.2: Phylogenetic tree and trait data with both vertical and horizontal trait transfer. Coloured lines represent lateral trait transfer events.

problem of traits transferring independently of each another. We illustrate this type of behaviour in Figure 1.2.

For a simple lateral transfer model whereby traits transfer into randomly chosen lineages at random times, such as in Figure 1.2, Roch and Snir (2013) show that distance-based methods may recover the true phylogeny up to relatively high transfer rates. This is similar to the model of trait dynamics that we consider in this thesis. Although their analysis is probabilistic, distance-based methods are not, so it is difficult to quantify the uncertainty in any estimates using this approach in practice.

Szöllősi et al. (2012) propose a discrete-time model for lateral transfer which may be considered a hybridisation model at its coarsest level and a general lateral transfer model in the style of Roch and Snir (2013) and Figure 1.2 at its finest level. Their method takes a set of gene trees as input and computes the probability under the model to reconcile each gene tree with an overall species tree. To achieve this, they

- Discretise time on the tree into subintervals
- Add an event node to each lineage on each subinterval

- Allow a trait to transfer to another lineage when it encounters an event node.

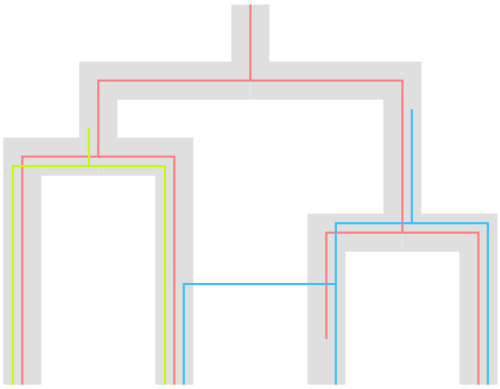
The authors describe systems of differential equations to compute the probability of a given gene-tree/species-tree reconciliation under their model. Their method seeks the species tree which maximises the likelihood given by the product of these reconciliation probabilities. The authors do not incorporate a molecular clock in their model, although this is possible, and, as a result, their method returns a labelled history. Sjöstrand et al. (2014) perform approximate Bayesian inference under a similar, time-discretised model and use MCMC to sample from the model posterior. The species tree is fixed in this case, and they estimate the gene trees from sequence data. Inference suffers if we do not jointly model the variability in the gene and species trees, particularly when lateral transfer occurs. As we have touched on previously, the number of possible gene-tree/species-tree reconciliations explodes as the number of event nodes increases. As such, the subintervals in these approaches must be sufficiently short that the model can accurately represent the data, and sufficiently long that inference is tractable.

In Figure 1.3, we illustrate how a selection of the methods in this section might represent the histories of three traits which undergo lateral transfer.

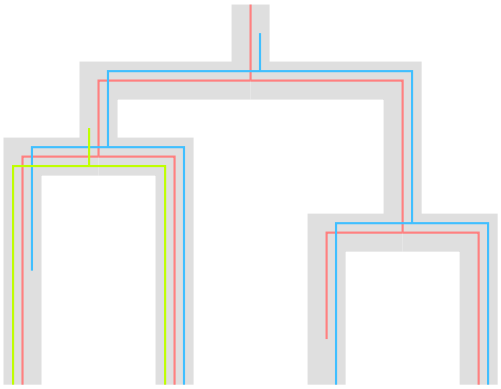
1.3 Discussion

To infer a phylogeny from trait presence/absence data which have undergone lateral transfer, we desire a continuous-time model of tree and trait dynamics which

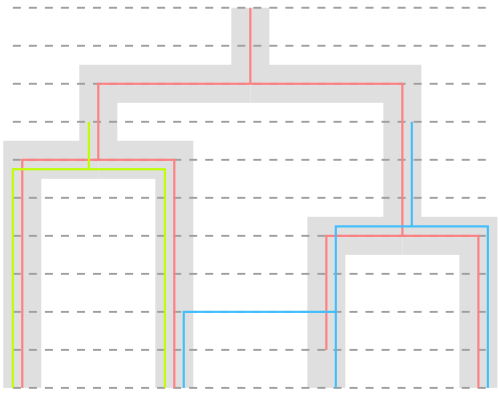
- Explicitly incorporates lateral transfer
- Can infer the timing of internal nodes
- Does not make approximations which introduce uncontrolled errors
- Allows us to quantify uncertainty in our inference in a principled manner



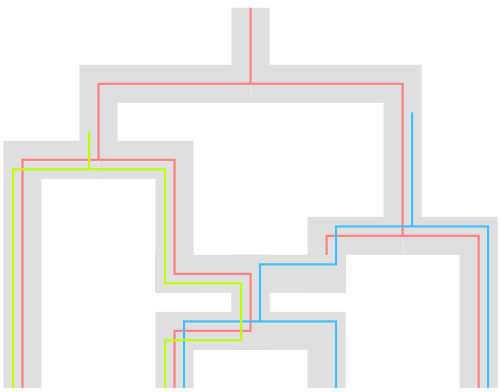
(a) A continuous-time model which allows for lateral transfer can correctly represent the trait histories (Suchard, 2005).



(b) If we ignore lateral transfer then the trait which transfers in reality is assumed arise above the most recent common ancestor of the leaves which display it (Gray and Atkinson, 2003; Bouckaert et al., 2012).



(c) In a time-discretised model, trait events may only occur at discrete time points (denoted here by grey, dashed lines) (Szöllősi et al., 2012; Sjöstrand et al., 2014).



(d) For low levels of lateral transfer, we can model the trait histories as being derived from a species-level ancestral recombination graph (Bloomquist and Suchard, 2010).

Figure 1.3: Four model-based approaches to the problem of lateral transfer on a four-taxon tree or network. The network is depicted in grey, and trait histories are offset to improve visibility.

- Returns a meaningful phylogeny as the number of lateral transfer events increase.

By this final point, we mean that the method does not return a phylogenetic graph with many more contact edges than ancestry edges.

Controlling for the effect of lateral transfer in phylogenetic inference is a difficult computational problem. This partially explains why none of the methods we discuss in the previous section satisfy all of the above conditions. As a result, these methods which attempt to control for lateral transfer are not fully model based or rely on approximations in one form or another. In this thesis, we do not attempt to solve the problem of lateral transfer for all classes of models and data sets. We do claim to make a breakthrough in the problem of lateral transfer for a particular class of problems, namely processes which generate the unordered sets of trait presence/absence data in our description of the Stochastic Dollo model.

Nicholls and Gray (2008) describe how to simulate a lateral transfer process in the basic Stochastic Dollo model. Similar to the process considered by Roch and Snir (2013), each species randomly acquires copies of traits possessed by other contemporary species. In Chapter 2, we describe how to perform exact likelihood-based inference satisfying all of the above conditions under this model. We do not first infer trait histories then reconcile them with a species tree in the manner of many of the methods we describe above. Instead, we integrate numerically over all possible trait histories under our model on a given tree. Equivalently, we integrate over all possible soft-wired phylogenetic graphs formed by adding contact edges between contemporaneous lineages, as in Figure 1.2, for example. This integration step comes with a massive computational cost — it is exponential in the number of taxa under consideration — and we address this issue in Chapter 4.

As a special case of our model, the SD model is an obvious baseline comparison for our results in Chapters 5 and 6. We shall demonstrate that when lateral transfer occurs, the effect of properly controlling for it in our inference outweighs the resulting increase in computational cost. We do not compare our method to the other model-

based approaches for genetic data sets that we describe earlier in this chapter as the data is gathered under a fundamentally different experimental design: we choose a trait and record whether it is present or absent in the taxa, and these patterns of trait presence and absence across the taxa are informative of the tree structure; genes are complex traits, so the characters they display are informative of the tree structure. The model which Szöllősi et al. (2012) describe is therefore closest to that which we develop in this thesis as it allows for gene loss. However, we cannot form realistic gene trees from our trait presence/absence data to feed into this models, so we cannot hope to obtain a valid comparison by fitting it here.

Chapter 2

Model description

Chapter overview

In this chapter, we first discuss the data encoding. This motivates the model which we then define. We describe how to compute the likelihood parameters as a sequence of initial value problems on a tree. This is a difficult computational hurdle and we return to this issue in Chapter 4

2.1 Homologous trait data

Homologous traits, or *cognates* in historical linguistics, for example, are derived from a common ancestral trait through a combination of vertical inheritance (*orthologous*) and lateral transfer (*xenologous*) events. We assign each set of homologous traits a unique common label from the set of trait labels, $\mathbb{Z}$; that is, if a pair of traits are homologous then they possess the same label. We record the status of trait h in taxon i as

$$d_i^h = \begin{cases} 0, & \text{trait } h \text{ is absent in taxon } i, \\ 1, & \text{trait } h \text{ is present in taxon } i, \\ ?, & \text{the status of trait } h \text{ in taxon } i \text{ is unknown.} \end{cases}$$

Language	Woman/female	Mother
Marquesan	1000	000001
Hawaiian	1000	011000
Rarotongan	1000	000100
Maori	1000	000010
Rurutuan	1000	011100
Tahitian	1000	011100
Penrhyn	1000	100000
Mangareva	1000	100000
Tuamotu	1000	011000
Rapanui	1110	001000
Manihiki	1001	011100

Table 2.1: Encoding of Eastern Polynesian lexical data. Among the languages in the data set, there are four unique traits which mean *woman* or *female*. For example, Tuamotu possesses one of these traits whereas Manihiki possesses two, one of which is not observed in the other languages.

We let $\mathbf{D}$ denote the array recording the status of each trait across the observed taxa. A column $\mathbf{d}^h$ of $\mathbf{D}$ is a *site-pattern* recording the status of trait h across the taxa. These patterns of trait presence and absence form the basis of the generative model we describe in the next section.

For the first application in Chapter 6, each trait is a word in one of 210 meaning categories and each taxon is an Eastern Polynesian language. For example, the Maori and Hawaiian words for *woman* and *wife*, both *wahine*, are derived from a common ancestral trait h, say, so $d^h_{\text{Maori}} = d^h_{\text{Hawaiian}} = 1$. On the other hand, the Maori word for *mother*, *whaea*, is not related to its Hawaiian counterpart, *makuahine*. As Maori lacks an instance of the trait from which the Hawaiian *makuahine* is derived, we record a 0 in the corresponding entry in the data array, and vice versa. We reproduce the columns corresponding to these words in the data set in Table 2.1, and we plot the entire data set in Chapter 6.

An important point regarding the data is that we assume each column is independent; that is, our method does not account for multiple traits being members of the

same meaning category, for example, or any interaction between traits. We return to this point in the concluding remarks of this thesis.

2.2 Generative model

A branching process on sets of traits determines the phylogeny of the observed taxa. Each set represents an evolving species. A branching event on the tree corresponds to a speciation event and a leaf represents an observed taxon. The set contents diversify according to a trait process. Finally, the status of each trait is recorded in the taxa to form the data array. We first define our model and inference method in terms of binary patterns of trait presence and absence in taxa which we observe simultaneously. We then extend the model to more complex scenarios. We illustrate the notation and evolutionary processes in this section in Figures 2.1 and 2.2.

A rooted phylogenetic tree $g = (V, E, T)$ on L leaves is a connected, acyclic graph with node set $V = \{0, 1, \ldots, 2L-1\}$, directed edge set E and node times $T \in \{-\infty\} \times \mathbb{R}^{2L-1}$. The node set V comprises: one *Adam* node labelled 0 of degree 1; the internal nodes $V_A = \{1, 2, \ldots, L-1\}$ of degree 3; and the leaf nodes $V_L = \{L, L+1, \ldots, 2L-1\}$ of degree 1. Node $i \in V$ arises at time $t_i \in T$ relative to the current time, 0. For convenience, we label the internal nodes in such a way that $t_1, \ldots, t_{L-1}$ is a strictly increasing sequence of node times.

Edges represent evolving species and are directed forwards in time. We label each edge by its offspring node, so if $\mathrm{pa}(i)$ denotes the parent of node $i \in V \setminus \{0\}$, edge $i \in E$ runs from node $\mathrm{pa}(i)$ at time $t_{\mathrm{pa}(i)}$ to i at time t_i. We assume that the Adam node arose at time $t_0 = -\infty$; therefore a branch of infinite length connects it to the *root* node 1 at time t_1. For the time being, we observe each of the leaves at time 0.

If we slice across the tree at time t, there are $L^{(t)}$ species labelled $\mathbf{k}^{(t)} = (i \in E : t_{\mathrm{pa}(i)} \leq t < t_i)$. We let $H_i(t) \subset \mathbb{Z}$ denote the set of traits possessed by species

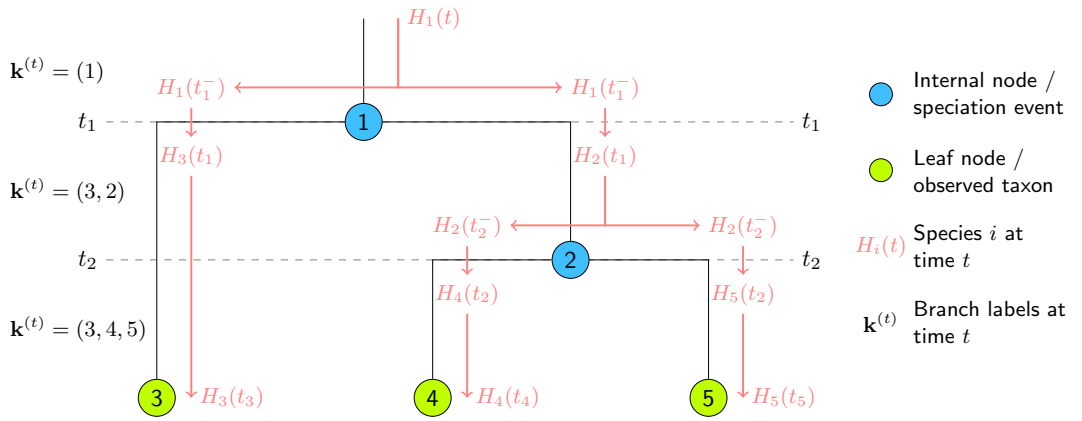


Figure 2.1: A branching process on sets of traits determines the phylogeny of the observed taxa. Branch lengths represent elapsed evolutionary time.

$i \in \mathbf{k}^{(t)}$ at time t. We now define four properties of the set-valued evolutionary process $H(t) = \{H_i(t) : i \in \mathbf{k}^{(t)}\}$.

Property T1 (Set branching event). At a speciation event, the traits present in the parent are copied into the offspring. Species $i \in \mathbf{k}^{(t_i^-)}$ branches at time t_i and is replaced by two identical offspring, j and $k \in \mathbf{k}^{(t_i)}$,

$$H_j(t_i) \leftarrow H_i(t_i^-),$$
$$H_k(t_i) \leftarrow H_i(t_i^-),$$

where t_i^- denotes the time just before the branching event.

We illustrate this concept in Figure 2.1.

Property T2 (Trait birth). New traits are born at rate λ over time in each extant species. Equivalently, new traits arise according to a Poisson process of rate λ along the branches of the tree. If trait $h \in \mathbb{Z}$ is born in species i at time t,

$$H_i(t) \leftarrow H_i(t^-) \cup \{h\}.$$

Property T3 (Trait death)**.** A species kills off each of the traits it possesses independently at rate μ. If trait $h \in H_i(t^-)$ in species i dies at time t,

$$H_i(t) \leftarrow H_i(t^-) \setminus \{h\}.$$

Property T4 (Lateral trait transfer)**.** A species acquires a copy of a trait by lateral transfer at rate β scaled by the fraction of species which already possess it. If species i acquires a copy of trait $h \in \mathcal{H}^{(t^-)} = \bigcup_{i \in \mathbf{k}^{(t^-)}} H_i(t^-)$ at time t,

$$H_i(t) \leftarrow H_i(t^-) \cup \{h\}.$$

Clearly if $h \in H_i(t^-)$ already then the transfer event has no effect.

We illustrate each of these events for a single trait in Figure 2.2.

Starting from a single set $H(-\infty) = \{\emptyset\}$, the process $H(t)$ evolves as a continuous-time Markov chain through a combination of branching (T1) and trait (T2–4) events to yield the diverse set of taxa $H(0) = \{H_i(0) : i \in V_L\}$ we observe at time 0. Figure 2.2 depicts a realisation of the process. When the lateral transfer rate $\beta = 0$, we recover the binary Stochastic Dollo process of Nicholls and Gray (2008). From here on, we shall drop all explicit dependence on time from our notation at time 0 when it is clear from the context.

We do not allow for the possibility of unsampled lineages interacting with the process on the phylogeny of the observed taxa. We return to this issue in the concluding remarks of this thesis.

2.3 Likelihood calculation

For a given trait history, we may calculate its likelihood under the process we describe above (T1–4) in terms of the joint distribution of jumps between states and the exponentially distributed time intervals between them (Norris, 1997). However, we are interested in the overall phylogeny so the unobserved trait histories are nuisance

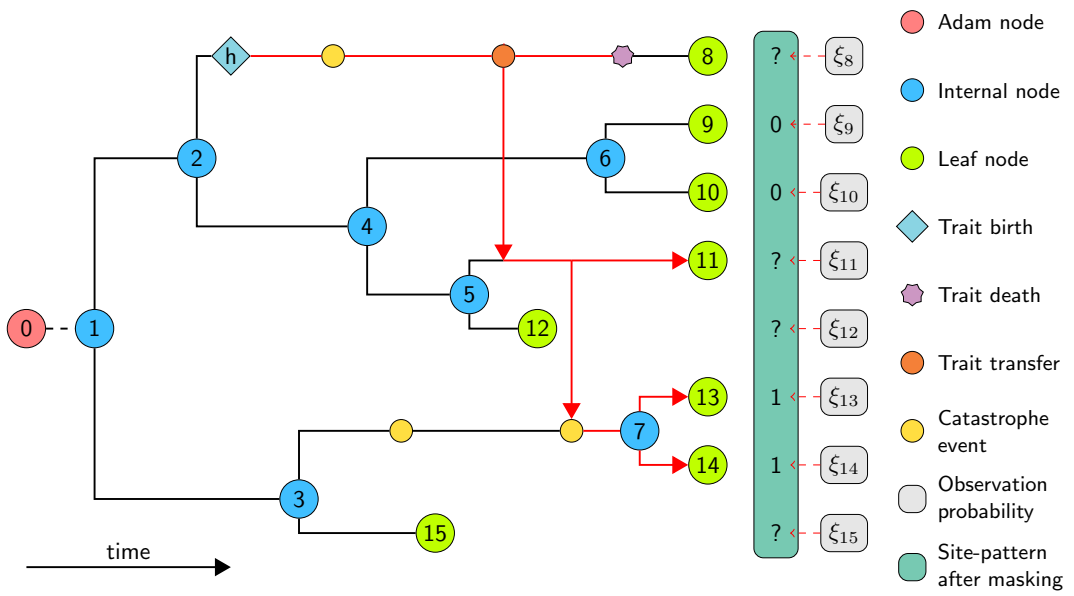


Figure 2.2: Illustration of the Stochastic Dollo with Lateral Transfer model. Catastrophes, missing data and offset leaves are introduced in Section 2.4.

parameters here, and therefore we must integrate them out of our inference. This is a simple operation under the Stochastic Dollo model as we can integrate out the history of each trait and account for extinct traits using simple recursions (Nicholls and Gray, 2008; Ryder and Nicholls, 2011). We now describe how to integrate over all possible trait histories under our model as a sequence of initial value problems. Unfortunately, this is a much more computationally expensive task than the equivalent integration step under the Stochastic Dollo model; we address this issue in detail in Chapter 4.

2.3.1 Pattern evolution

If we cut through the tree at time t, each trait $h \in \mathcal{H}^{(t)}$ displays a *pattern* of presence and absence across the $L^{(t)}$ extant species $\mathbf{k}^{(t)} = (k_i^{(t)} : i \in [L^{(t)}])$, where $[L^{(t)}] = \{1, \ldots, L^{(t)}\}$. These patterns of presence and absence evolve over time as new branches arise and instances of h die and transfer. The pattern displayed by trait h at time t is

the $L^{(t)}$-tuple $\mathbf{p}^h(t) = (p_i^h(t) : i \in [L^{(t)}])$ where

$$p_i^h(t) = \begin{cases} 1, & h \in H_{k_i^{(t)}}(t), \\ 0, & \text{otherwise}, \end{cases}$$

indicates the presence or absence of trait h on lineage $k_i^{(t)}$ at time t. Figure 2.3 illustrates how the pattern displayed by trait h in Figure 2.2 evolves along the tree.

The space of binary patterns of trait presence or absence across $L^{(t)}$ lineages is $\mathcal{P}^{(t)} = \{0,1\}^{L^{(t)}} \setminus \{\mathbf{0}\}$, where $\mathbf{0}$ is an $L^{(t)}$-tuple of zeros, and there are $N_{\mathbf{p}}(t) = |\{h \in \mathcal{H}^{(t)} : \mathbf{p}^h(t) = \mathbf{p}\}|$ traits displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time t. The *empty* pattern, $\mathbf{0}$, is the absorbing state for the pattern process. The dynamics of the pattern frequency process $\mathbf{N}(t) = (N_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ follow directly from Properties T1–4 of the trait process.

Patterns at branching events

At a branching event, patterns increase in length and the space of patterns expands to accommodate the new patterns which traits may display. The tuple $\mathbf{k}^{(t)}$ of branch labels is consistent across speciation events in the sense that when lineage $j = k_i^{(t_j^-)}$ branches at time t_j, the branch labels are

$$\mathbf{k}^{(t_j)} = \left(k_1^{(t_j^-)}, \ldots, k_{i-1}^{(t_j^-)}, k_i^{(t_j)}, k_{i+1}^{(t_j)}, k_{i+1}^{(t_j^-)}, \ldots, k_{L^{(t_j^-)}}^{(t_j^-)}\right),$$

where species $k_i^{(t_j)}$ and $k_{i+1}^{(t_j)}$ are the offspring of species $j = k_i^{(t_j^-)}$ (T1). For example, in Figure 2.1, $\mathbf{k}^{(t_2^-)} = (3,2) \to \mathbf{k}^{(t_2)} = (3,4,5)$ due to the branching event on lineage 2, the right-most branch in our representation, at time t_2. It follows that each trait $h \in \mathcal{H}^{(t_j)}$ displays a pattern $\mathbf{p}^h(t_j)$ with entries $p_i^h(t_j) = p_{i+1}^h(t_j) \leftarrow p_i^h(t_j^-)$. We illustrate this property in Figure 2.3.

A pattern $\mathbf{p} \in \mathcal{P}^{(t_j)}$ with entries $p_i = p_{i+1}$ is consistent with the branching event on lineage $k_i^{(t_j^-)}$ as it may be formed by duplicating the ith entries of a pattern in $\mathcal{P}^{(t_j^-)}$. On the other hand, the trait process cannot generate a pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ with $p_i \neq p_{i+1}$

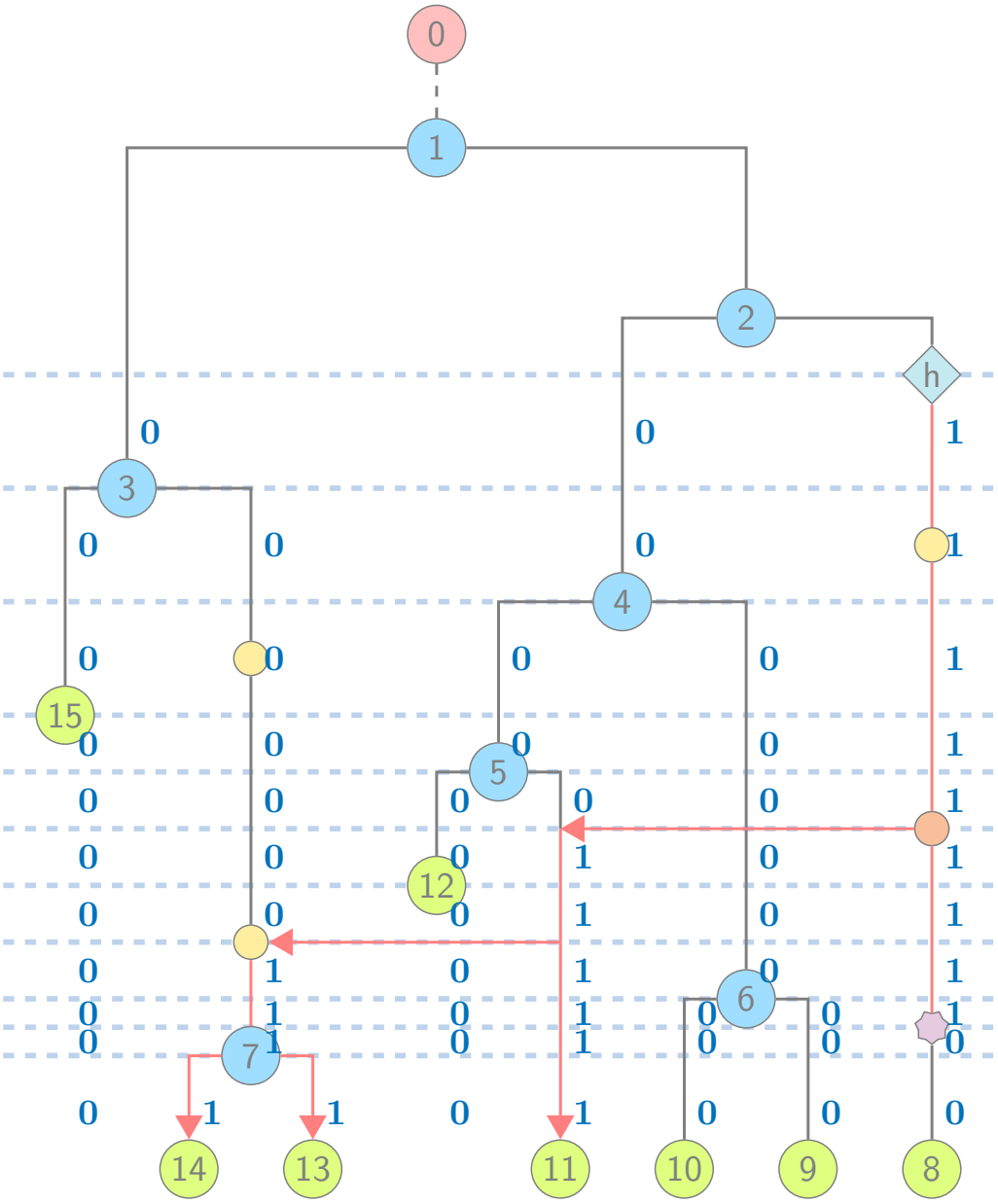


Figure 2.3: The patterns of presence and absence displayed by trait *h* as it evolves along the tree in Figure 2.2.

at time t_j by definition (T1). We denote by $\mathbf{T}^{(j)}$ the linear operation

$$\mathbf{N}(t_j) \leftarrow \mathbf{T}^{(j)}\mathbf{N}(t_j^-),$$

which initialises the pattern frequencies $\mathbf{N}(t_j)$ at the beginning of the jth interval between branching events. The vector $\mathbf{N}(t_j)$ comprises entries of $\mathbf{N}(t_j^-)$ for patterns consistent with the branching event, and zeros otherwise. For example, in Figure 2.3,

$$N_{0001}(t_j) \leftarrow [T^{(j)}\mathbf{N}(t_j^-)]_{0001} = N_{001}(t_j^-),$$

but $[\mathbf{N}(t_j)]_{0101} \leftarrow 0$ as the pattern $(0,1,0,1)$ is inconsistent with the branching event on lineage 3. We return to this initialisation operation when we describe how to evaluate the expected pattern frequencies in Section 2.3.2.

Patterns between branching events

In order to formally describe the Markovian evolution of the pattern frequencies $\mathbf{N}(t)$ between branching events, we first define how patterns are related. The Hamming distance between patterns $\mathbf{p}$ and $\mathbf{q} \in \mathcal{P}^{(t)}$ is $d(\mathbf{p},\mathbf{q}) = |\{i \in [L^{(t)}] : p_i \neq q_i\}|$ and $s(\mathbf{p}) = d(\mathbf{p},\mathbf{0})$ is the Hamming weight of $\mathbf{p}$. A trait displaying pattern $\mathbf{p}$ at time t *communicates* with patterns in the sets

$$S_{\mathbf{p}}^- = \{\mathbf{q} \in \mathcal{P}^{(t)} \cup \{\mathbf{0}\} : s(\mathbf{q}) = s(\mathbf{p}) - 1,\, d(\mathbf{p},\mathbf{q}) = 1\},$$
$$S_{\mathbf{p}}^+ = \{\mathbf{q} \in \mathcal{P}^{(t)} : s(\mathbf{q}) = s(\mathbf{p}) + 1,\, d(\mathbf{p},\mathbf{q}) = 1\},$$

the patterns which differ from $\mathbf{p}$ through a single death (T3) or transfer (T4) event. Figure 2.4 describes the transition rates between pattern states $\mathbf{p}$ and $\mathbf{q} \in S_{\mathbf{p}}^- \cup S_{\mathbf{p}}^+$. New traits displaying patterns of Hamming weight 1 arise on each branch through trait birth events (T2).

The pattern displayed by a trait at time t follows a random walk on the $L^{(t)}$-cube with vertex set $\mathcal{P}^{(t)}$ and edges between communicating patterns. The transition rates along edges follow from Figure 2.4, and we illustrate this representation in Figure 2.5

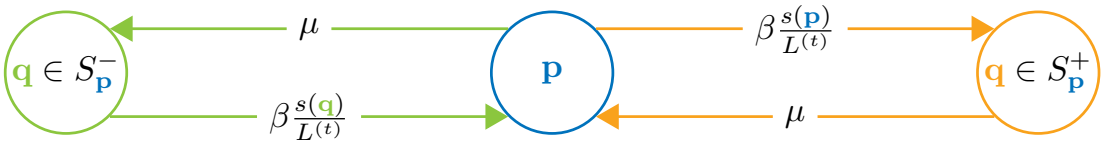


Figure 2.4: Transition rates between pattern states $\mathbf{p} \in \mathcal{P}^{(t)}$ and $\mathbf{q} \in S_{\mathbf{p}}^{-} \cup S_{\mathbf{p}}^{+}$.

for $L^{(t)} = 3$. We shall return to this representation of the process shortly. The proportion of branches where the trait is present evolves in a similar manner to the Moran model in population genetics (Moran, 1958) with the fundamental difference that here we look at trait presence or absence whereas Moran focuses on the expressed character state. In Chapter 4, we shall consider a representation of the trait process which allows us to analyse the number of branches displaying a given trait.

2.3.2 Expected pattern frequencies

Traits which display the same pattern evolve independently of each other and of other traits. If we sum over the rates in Figure 2.4 for each trait displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, then on a short interval of length $\mathrm{d}t$ between branching events, by a standard argument for Markov chains (Norris, 1997; Grimmett and Stirzaker, 2001),

$$\begin{aligned}
&\mathbb{P}[N_{\mathbf{p}}(t+\mathrm{d}t) - N_{\mathbf{p}}(t) = k | g, \lambda, \mu, \beta] \\
&\quad = \begin{cases} s(\mathbf{p}) N_{\mathbf{p}}(t) \left[\mu + \beta\left(1 - \frac{s(\mathbf{p})}{L^{(t)}}\right)\right] \mathrm{d}t + o(\mathrm{d}t), & k = -1, \\ \left[\lambda \mathbf{1}_{\{s(\mathbf{p})=1\}} + \beta \sum_{\mathbf{q} \in S_{\mathbf{p}}^{-}} \frac{s(\mathbf{q})}{L^{(t)}} N_{\mathbf{q}}(t) \right. & \\ \qquad \left. + \mu \sum_{\mathbf{q} \in S_{\mathbf{p}}^{+}} N_{\mathbf{q}}(t)\right] \mathrm{d}t + o(\mathrm{d}t), & k = 1, \end{cases}
\end{aligned} \tag{2.3.1}$$

where $\lim_{\mathrm{d}t \to 0} o(\mathrm{d}t)/\mathrm{d}t = 0$.

Let $x_{\mathbf{p}}(t; g, \lambda, \mu, \beta) = \mathbb{E}[N_{\mathbf{p}}(t) | g, \lambda, \mu, \beta]$, the expected number of traits in $\mathcal{H}^{(t)}$ displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time t. From Equation 2.3.1 we derive that $x_{\mathbf{p}}(t)$ evolves

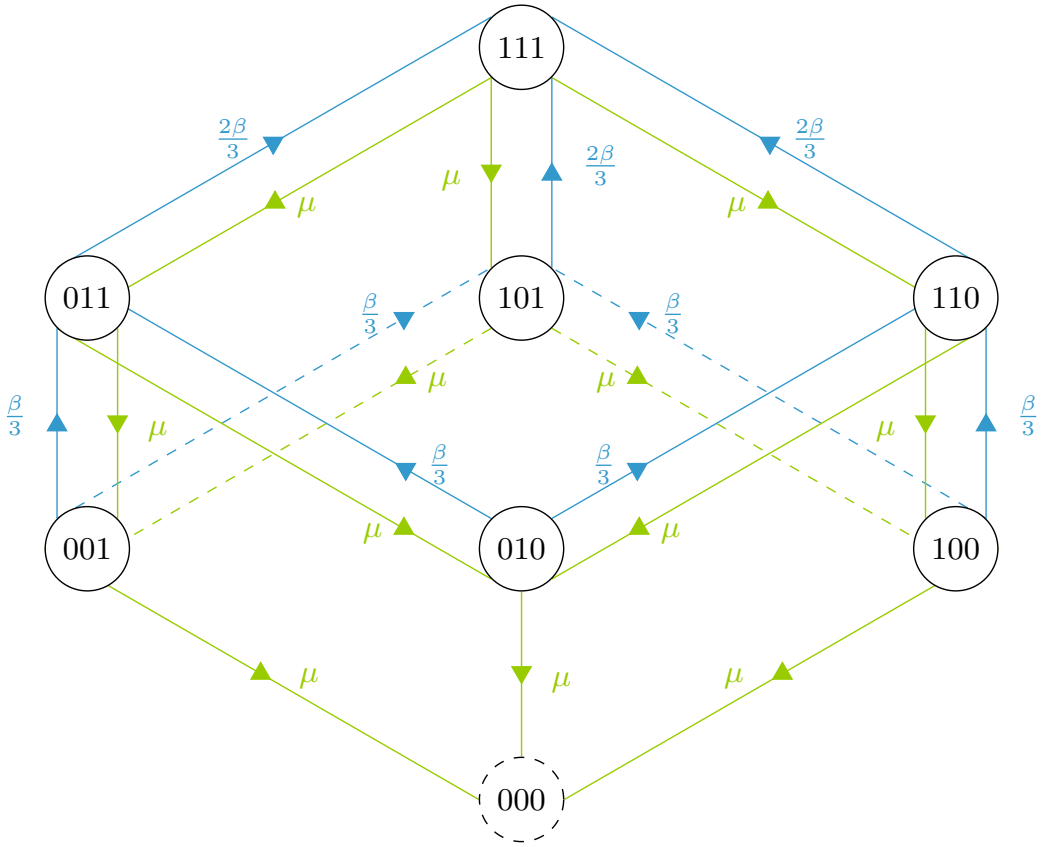


Figure 2.5: The pattern displayed by a trait at time t follows a random walk on the $L^{(t)}$-cube independently of patterns displayed by other traits. The state $\mathbf{0}$ is absorbing.

according to the following differential equation:

$$\begin{aligned}\dot{x}_{\mathbf{p}}(t) &= \lim_{\mathrm{d}t\to 0} \frac{\mathbb{E}\left[N_{\mathbf{p}}(t+\mathrm{d}t) - N_{\mathbf{p}}(t) | g, \lambda, \mu, \beta\right]}{\mathrm{d}t} \\ &= -s(\mathbf{p})x_{\mathbf{p}}(t)\left[\mu + \beta\left(1 - \tfrac{s(\mathbf{p})}{L^{(t)}}\right)\right] + \lambda \mathbf{1}_{\{s(\mathbf{p})=1\}} \\ &\quad + \beta \sum_{\mathbf{q}\in S_{\mathbf{p}}^{-}} \tfrac{s(\mathbf{q})}{L^{(t)}} x_{\mathbf{q}}(t) + \mu \sum_{\mathbf{q}\in S_{\mathbf{p}}^{+}} x_{\mathbf{q}}(t).\end{aligned} \tag{2.3.2}$$

There are $|\mathcal{P}^{(t)}| = 2^{L^{(t)}} - 1$ coupled differential equations describing the expected evolution of pattern frequencies (2.3.2). We may write these equations as $\dot{\mathbf{x}}(t) = \mathbf{A}^{(t)}\mathbf{x}(t) + \mathbf{b}^{(t)}$ where: $\mathbf{x}(t) = (x_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ is the vector of expected pattern

frequencies at time t; the sparse matrix $\mathbf{A}^{(t)} = (a_{\mathbf{p},\mathbf{q}} : \mathbf{p}, \mathbf{q} \in \mathcal{P}^{(t)})$ with entry

$$a_{\mathbf{p},\mathbf{q}} = \begin{cases} -s(\mathbf{p})\left[\mu + \beta\left(1 - \frac{s(\mathbf{p})}{L^{(t)}}\right)\right], & \mathbf{q} = \mathbf{p}, \\ \mu, & \mathbf{q} \in S_{\mathbf{p}}^{-}, \\ \beta\frac{s(\mathbf{q})}{L^{(t)}}, & \mathbf{q} \in S_{\mathbf{p}}^{-}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3.3}$$

describes the flow between patterns as a result of trait death and transfer events; and the sparse vector $\mathbf{b}^{(t)} = (b_{\mathbf{p}} : \mathbf{p} \in \mathcal{P}^{(t)})$ with entry

$$b_{\mathbf{p}} = \begin{cases} \lambda, & s(\mathbf{q}) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3.4}$$

describes the flow into patterns of Hamming weight 1 through trait birth events. We revisit the structure of the systems of differential equations in Chapter 4.

In Section 2.2, we state that a branch of infinite length connects the Adam and root nodes. As a result, the pattern frequency process $\mathbf{N}(t)$ is in equilibrium immediately before the first branching event, so the expected frequency $\mathbf{x}(t_1^-) = x_1(t_1^-) = \lambda/\mu$. This is the solution of the Equation 2.3.2 for one dimension. We can prove a stronger result here, that $N_1(t_1^-) \sim \text{Poisson}(\lambda/\mu)$, by solving the detailed balance equation

$$\lambda \mathbb{P}[N_1(t_1^-) = n] = \mu \mathbb{P}[N_1(t_1^-) = n+1], \quad n = 0, 1, \ldots$$

with the condition that $\sum_{n=0}^{\infty} \mathbb{P}[N_1(t_1^-) = n] = 1$.

Recall from Section 2.3.1 the linear, pattern frequency initialisation operator $\mathbf{T}^{(j)}$ which propagates the pattern frequencies $\mathbf{N}(t_j^-)$ across the jth branching event to form $\mathbf{N}(t_j)$. Expectation is also a linear operation so we can apply $\mathbf{T}^{(j)}$ here to obtain

$$\mathbf{x}(t_j) = \mathbb{E}[\mathbf{N}(t_j)|g, \lambda, \mu, \beta] = \mathbb{E}[T^{(j)}\mathbf{N}(t_j^-)|g, \lambda, \mu, \beta] = T^{(j)}\mathbf{x}(t_j^-), \quad j = 1, 2, \ldots$$

the expected pattern frequencies immediately after the branching event. We can write the expected pattern frequencies at the leaves, $\mathbf{x} = \mathbf{x}(0)$, recursively as the following sequence of initial value problems between branching events: for $i = 1, \ldots, L-1$,

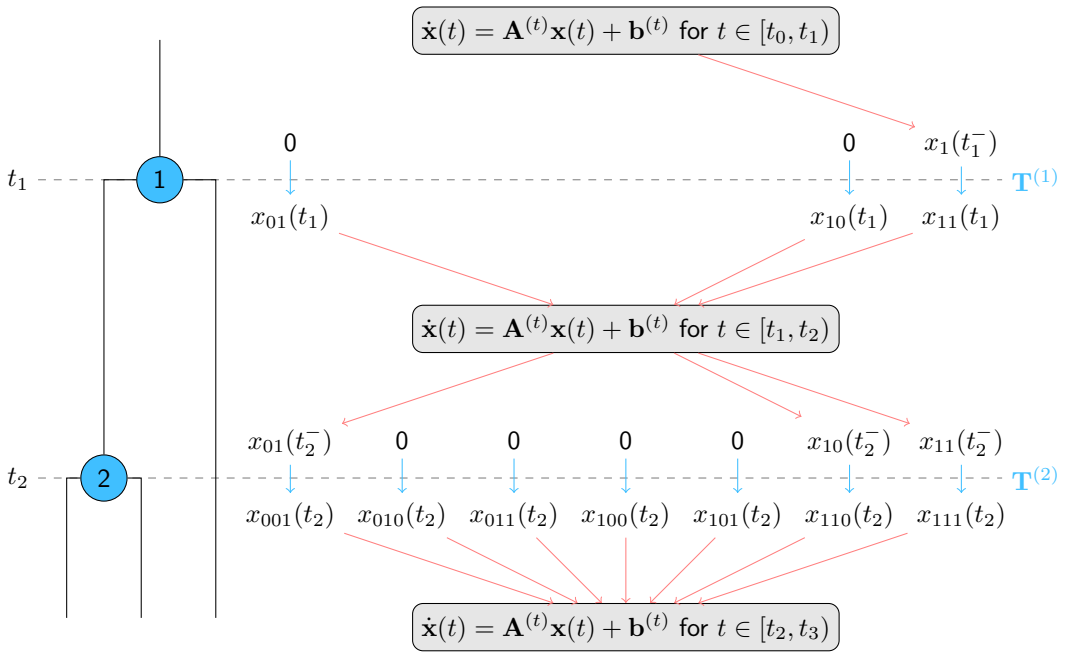


Figure 2.6: Calculating the expected pattern frequencies on a given tree as a sequence of initial value problems (2.3.5). The initialisation operation $\mathbf{T}^{(i)}\mathbf{x}(t_i^-) = \mathbf{x}(t_i)$ from Section 2.3.1 provides the initial condition at the start of the ith interval between branching events.

solve

$$\dot{\mathbf{x}}(t) = \mathbf{A}^{(t)}\mathbf{x}(t) + \mathbf{b}^{(t)} \quad \text{for} \quad t \in [t_i, t_{i+1}) \quad \text{where} \quad \mathbf{x}(t_i) = \mathbf{T}^{(i)}\mathbf{x}(t_i^-), \tag{2.3.5}$$

with the initial condition $\mathbf{x}(t_1^-) = \lambda/\mu$. We illustrate this procedure graphically in Figure 2.6.

2.3.3 Likelihood

We are now in a position to formally state the distribution of the pattern frequencies $\mathbf{N}(t)$ under our model. We note that

- New traits arise according to a Poisson process along the branches of the tree
- Instances of a trait evolve independently of
  - other instances of the same trait

  - instances of other traits

- The labels of traits which display the same patterns of presence and absence across the leaves are exchangeable,

so the frequencies of traits displaying the same pattern are independent Poisson random variables. Theorem 1 formalises this result.

Theorem 1 (Binary data distribution). *For each pattern* $\mathbf{p} \in \mathcal{P}^{(t)}$,

$$N_{\mathbf{p}}(t)|g, \lambda, \mu, \beta \sim \text{Poisson}\big(x_{\mathbf{p}}(t; g, \lambda, \mu, \beta)\big),$$

independently of the other patterns in $\mathcal{P}^{(t)}$, *where* $\mathbf{x}(t; g, \lambda, \mu, \beta)$ *is given by Equation 2.3.5.*

We prove Theorem 1 in Appendix A. In computing the likelihood, we must account for all the patterns which we could observe in the data; that is, the patterns we do record as well as those we do not. For a vector $\mathbf{N}(t)$, the likelihood is

$$\mathbb{P}[\mathbf{N}(t)|g, \lambda, \mu, \beta] = \prod_{\mathbf{p} \in \mathcal{P}^{(t)}} \frac{x_{\mathbf{p}}(t)^{N_{\mathbf{p}}(t)} e^{-x_{\mathbf{p}}(t)}}{N_{\mathbf{p}}(t)!}. \tag{2.3.6}$$

In this thesis, we do not analyse any data sets containing the *empty* pattern, $\mathbf{0}$, so we do not include it in the pattern sets $\mathcal{P}$ or $\mathcal{Q}$. If our experimental design allows us to observe such patterns then we can make a straightforward adjustment to the definition of our model to account them. The frequency of traits displaying the empty pattern $N_{\mathbf{0}}(t)$ at time t is not well defined when we have a branch of infinite length leading into the root node. Bouchard-Côté and Jordan (2013) address this problem by replacing such a branch with a Poisson(λ/μ) number of birth events at the root and setting $N_{\mathbf{0}}(t_1) = 0$. The empty pattern does not interact with traits evolving on the tree so we can calculate its expected frequency at the leaves by simply augmenting the initial value problems in Equation 2.3.5 or using the identity

$$\frac{\lambda}{\mu} + \lambda \sum_{i \in E \setminus \{1\}} (t_{\text{pa}(i)} - t_i) = x_{\mathbf{0}} + \sum_{\mathbf{p} \in \mathcal{P}} x_{\mathbf{p}},$$

as mass can only leave the system for the empty pattern. In Section 2.4.4, describe how to modify the likelihood in Equation 2.3.6 when we discard other patterns from the data.

2.4 Model extensions

We describe how to extend the model and likelihood calculation for rate variation, missing data, offset leaves and the systematic removal of patterns from the data.

2.4.1 Rate heterogeneity

To allow the trait event rates λ, μ and β to vary across lineages and time would require time- or state-dependent initial value problems (2.3.5) to compute the expected pattern frequencies. We instead introduce spikes of activity in the form of *catastrophes* as a surrogate for rate heterogeneity. We first describe the catastrophe process of Ryder and Nicholls (2011), then extend it for lateral transfer.

Catastrophes, such as those in Figure 2.2, occur at rate ρ along each branch of the tree. Let C denote the set of catastrophes on a tree. We parameterise each catastrophe $c \in C$ as $c = (b, t)$ where $b \in E$ is the catastrophe branch and $t \in \mathbb{R}$ is its location along branch b. We extend our definition of a tree to $g = (V, E, T, C)$ and update the sample space of the model accordingly.

When the trait process encounters a catastrophe $(k_i^{(t)}, t) \in C$ on branch $k_i^{(t)}$ at time t, each trait in $H_i(t^-)$ is killed with probability $\kappa \in (0, 1)$, the catastrophe *severity*, and a Poisson$(\lambda\kappa/\mu)$ number of new traits are born. The trait process is frozen on the other lineages during a catastrophe so each catastrophe is equivalent to artificially lengthening the branch where it occurs by $\delta = -\mu^{-1}\log(1-\kappa)$ units of time. This catastrophe definition is reversible, in the sense that if we observe the trait process as it passes through a catastrophe, we cannot distinguish the direction of time (Kelly, 1979). Each lineage in the tree evolves independently, so the effect of a catastrophe in

this setting is invariant to its location along a branch. As a result, we need only know the total number of catastrophes on a branch in the SD model.

To the catastrophe process of Ryder and Nicholls, we introduce that traits may transfer into the catastrophe branch during a catastrophe. In greater detail, the catastrophe duration is unchanged but the catastrophe branch may

- Acquire traits through birth and transfer events
- Lose traits through death events.

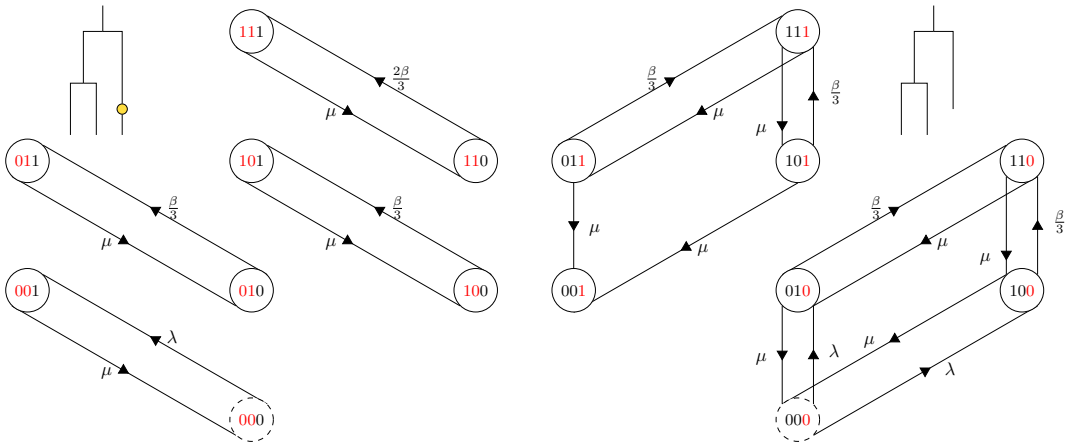
We illustrate this process in Figure 2.7a. The trait process at a catastrophe is equivalent to thinning the overall trait process from Section 2.2 to events on the catastrophe branch only, so we account for a catastrophe $(k_i^{(t)}, t)$ in the expected pattern frequency calculation (2.3.5) with the update:

$$\begin{aligned} x_{\mathbf{p}}(t) &\leftarrow e^{-\mu\delta} x_{\mathbf{p}}(t^-) + (1 - e^{-\mu\delta})\frac{\lambda}{\mu} && \mathbf{p} \in \mathcal{P}^{(t)},\ s(\mathbf{p}) = 1, \\ & && p_i = 1, \\ \begin{pmatrix} x_{\mathbf{q}}(t) \\ x_{\mathbf{r}}(t) \end{pmatrix} &\leftarrow \exp\left[\begin{pmatrix} -\beta\frac{s(\mathbf{q})}{L^{(t)}} & \mu \\ \beta\frac{s(\mathbf{q})}{L^{(t)}} & -\mu \end{pmatrix} \delta \right] \begin{pmatrix} x_{\mathbf{q}}(t^-) \\ x_{\mathbf{r}}(t^-) \end{pmatrix} && \mathbf{q}, \mathbf{r} \in \mathcal{P}^{(t)},\ d(\mathbf{q}, \mathbf{r}) = 1, \\ & && q_i = 0,\ r_i = 1, \end{aligned} \tag{2.4.1}$$

where we exploit the property, illustrated in Figure 2.7a, that each pattern communicates with only one other pattern during a catastrophe. We can prove the update in Equation 2.4.1 is correct by adapting the proof of Theorem 1; and our catastrophe process is consistent with Ryder and Nicholls when the lateral transfer rate $\beta = 0$.

The pairwise expected pattern frequency update $\mathbf{x}(t^-) \to \mathbf{x}(t)$ in Equation 2.4.1 has an analytic solution which gives our catastrophe process a physical interpretation. For patterns $\mathbf{q}$ and $\mathbf{r}$ in Equation 2.4.1, $s(\mathbf{q}) = s(\mathbf{r}) - 1$ and

$$\begin{aligned} x_{\mathbf{q}}(t) &\leftarrow \frac{\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}} e^{-\delta\left(\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}\right)}}{\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}} x_{\mathbf{q}}(t^-) + \frac{\mu - \mu e^{-\delta\left(\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}\right)}}{\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}} x_{\mathbf{r}}(t^-), \\ x_{\mathbf{r}}(t) &\leftarrow \frac{\beta\frac{s(\mathbf{q})}{L^{(t)}} - \beta\frac{s(\mathbf{q})}{L^{(t)}} e^{-\delta\left(\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}\right)}}{\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}} x_{\mathbf{q}}(t^-) + \frac{\beta\frac{s(\mathbf{q})}{L^{(t)}} + \mu e^{-\delta\left(\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}\right)}}{\mu + \beta\frac{s(\mathbf{q})}{L^{(t)}}} x_{\mathbf{r}}(t^-). \end{aligned}$$



(a) Pattern process during a catastrophe on the right-hand branch.

(b) Pattern process when the right-hand branch is offset.

Figure 2.7: (a) A catastrophe reduces the pattern process to a set of $2^{L^{(t)}-1}$ independent one- and two-dimensional systems; (b) n offset leaves reduce the pattern process to a set of 2^n independent $(2^{L^{(t)}-n}-1)$- and $2^{L^{(t)}-n}$-dimensional systems.

Now, traits which display either of the patterns undergo a Poisson$(\mu+\beta s(\mathbf{q})/L^{(t)})$ number of attempted transfer or death events during the catastrophe, a result of the superposition property of Poisson processes (Kingman, 1992). Each event is an attempted death event with probability $\mu/(\mu+\beta s(\mathbf{q})/L^{(t)})$, and an attempted transfer event otherwise. An attempted death event has no effect if the trait is not absent on the branch, and similarly, an attempted transfer event has no effect if the trait is already present on a branch. Therefore, only a fraction of these attempted events succeed, and only the final event matters. The probability that a trait h displays pattern $\mathbf{r}$ at the end of the catastrophe, given that it displayed pattern $\mathbf{q}$ at the beginning, is

$$\mathbb{P}\left[\mathbf{p}^h(t)=\mathbf{r}\middle|\,\mathbf{p}^h(t^-)=\mathbf{q},(k_i^{(t)},t)\in C\right]=\left[1-e^{-\delta\left(\mu+\beta\frac{s(\mathbf{q})}{L^{(t)}}\right)}\right]\frac{\beta\frac{s(\mathbf{q})}{L^{(t)}}}{\mu+\beta\frac{s(\mathbf{q})}{L^{(t)}}},$$

where the first term on the right-hand side is the probability that at least one event involving trait h occurs during the catastrophe, and the second term is the probability that the final event is a transfer event given an event occurs. Subtracting this from

1 gives the probability that a trait displays pattern **q** at the end of the catastrophe having also displayed **q** at the outset. The remaining terms follow by the same logic.

The model is now, in a sense, overparameterised and not uniquely identifiable as we can represent the overall trait process through multiple short catastrophes — this result follows from the statement above that a catastrophe is a restriction of the overall trait process to events on a single branch as we may equivalently write that the overall trait process is a superposition of catastrophe events. The complication in practice is that data is not distributed exactly according to the process we describe, so our inference scheme may attempt to compensate for this through catastrophes. We consider multiple short catastrophes to be a sign of model misspecification in the overall trait process; therefore we wish to restrict the possibility of this pathological behaviour in practice. One obvious solution here is to limit the number of catastrophes through either a hard bound on $|C|$ or by restricting the catastrophe rate $\rho < \mathcal{O}([\sum_{i \in E \setminus \{1\}} t_{\mathrm{pa}(i)} - t_i]^{-1})$, so that in either case the *a priori* expected number of catastrophes on the tree is relatively small. We do not want to rule this behaviour out completely as we want to know when our model is misspecified on a data set; instead, we enforce a minimum catastrophe severity $\kappa \geq 0.25$ so that catastrophes are identifiable relative to the overall trait process. This choice of threshold is somewhat arbitrary and a response to the behaviour of our model without this constraint when fit to the Eastern Polynesian data set in Chapter 6. We return to this issue in Chapter 3 in the context of our inference algorithm.

2.4.2 Missing data

In Section 2.1, we record a trait as missing in a taxon when we are unable to confirm whether it is present or absent. We illustrate this in Figure 2.2 where the true binary site-pattern, $(0,0,0,1,0,1,1,0)$, is recorded as $(?,0,0,?,?,1,1,?)$. Following Ryder and Nicholls (2011), we assume data is *missing at random*. The true binary state of trait h at taxon $i \in V_L$ is recorded with probability $\xi_i = \mathbb{P}(d_i^h \in \{0,1\})$ independently

of the other traits and taxa. Let $\Xi = (\xi_i : i \in V_L)$, the set of true-state observation probabilities.

The space of observable site-patterns across L taxa with missing data is $\mathcal{Q} = \{0, 1, ?\}^L \setminus \{\mathbf{0}\}$ and $u(\mathbf{q}) = \{\mathbf{p} \in \mathcal{P} : p_i = q_i \text{ if } q_i \neq ?, i \in [L]\}$ is the set of binary patterns consistent with $\mathbf{q} \in \mathcal{Q}$ before masking. The frequency of traits displaying pattern $\mathbf{q} \in \mathcal{Q}$ is a Poisson random variable with mean

$$\mathbf{x}_q(0; g, \lambda, \mu, \beta, \Xi) = \sum_{\mathbf{p} \in u(\mathbf{q})} x_{\mathbf{p}}(0; g, \lambda, \mu, \beta) \prod_{i=1}^{L} \xi_{k_i}^{\mathbf{1}_{\{q_i \in \{0,1\}\}}} (1 - \xi_{k_i})^{\mathbf{1}_{\{q_i = ?\}}}. \tag{2.4.2}$$

The first term on the right-hand side is the expected number of traits displaying binary patterns consistent with $\mathbf{q}$ before masking, and the second term is the probability of obscuring their entries to display $\mathbf{q}$.

The result in Equation 2.4.2 is a straightforward application of the theory of Poisson processes (Kingman, 1992). If $N_{\mathbf{q}|\mathbf{p}}$ denotes the frequency of traits whose true pattern is $\mathbf{p} \in \mathcal{P}$ but have entries obscured to display $\mathbf{q} \in \mathcal{Q}$, then by the restriction property of Poisson processes,

$$N_{\mathbf{q}|\mathbf{p}} | g, \lambda, \mu, \beta, \Xi \sim \text{Poisson}\left(x_{\mathbf{p}}(0; g, \lambda, \mu, \beta, \Xi) \prod_{i=1}^{L} \xi_{k_i}^{\mathbf{1}_{\{q_i \in \{0,1\}\}}} (1 - \xi_{k_i})^{\mathbf{1}_{\{q_i = ?\}}} \right),$$

and by the superposition property

$$N_{\mathbf{q}} = \sum_{\mathbf{p} \in \mathcal{P}} N_{\mathbf{q}|\mathbf{p}} \sim \text{Poisson}\left(\sum_{\mathbf{p} \in \mathcal{P}} x_{\mathbf{p}}(0; g, \lambda, \mu, \beta, \Xi) \prod_{i=1}^{L} \xi_{k_i}^{\mathbf{1}_{\{q_i \in \{0,1\}\}}} (1 - \xi_{k_i})^{\mathbf{1}_{\{q_i = ?\}}} \right).$$

We now briefly discuss three other approaches to missing data. The observation probabilities $\Xi = \{\xi_i : i \in V_L\}$ are typically strongly informed by the data, so we could estimate these parameters in advance with negligible loss of coverage for our inference. It would not be difficult to extend this masking process to depend on the status of the unobserved trait; that is, the trait observation probability in a taxon depends on whether the unobserved trait is possessed by the taxon, or not. Finally, traits are often missing in blocks across taxa in practice. Ryder (2009) investigates this

source of model error in the Stochastic Dollo model, and concludes that it has little effect on the inference.

2.4.3 Non-isochronous data

Data are not isochronous when the taxa are not sampled simultaneously, and the taxa appear as *offset* leaves in the phylogeny; nodes 12 and 15 in Figure 2.2, for example. Similar to catastrophes, the trait process is frozen on offset leaves, so a pattern may only communicate with those patterns identical to it on the extinct lineages and differing at a single entry on the extant lineages. We illustrate this for one offset leaf in Figure 2.7b. We now describe how to generalise the expected pattern frequency calculation (2.3.5) to account for offset leaves.

The $L^{(t)}$ extinct and evolving lineages at time t, of which $\hat{L}^{(t)}$ are extant, are labelled $\mathbf{k}^{(t)} = (i \in E : t_{\mathrm{pa}(i)} \leq t < t_i \mathbf{1}_{\{i \in V_A\}})$. The Hamming distance between patterns $\mathbf{p}$ and $\mathbf{q} \in \mathcal{P}^{(t)}$ across extant lineages only is $\hat{d}(\mathbf{p}, \mathbf{q}) = |\{i \in [L^{(t)}] : p_i \neq q_i,\, t < t_{k_i^{(t)}}\}|$ and the Hamming weight of $\mathbf{p}$ across extant lineages is $\hat{s}(\mathbf{p}) = \hat{d}(\mathbf{p}, \mathbf{0})$. Recalling $S_{\mathbf{p}}^-$ and $S_{\mathbf{p}}^+$ from Section 2.3.1, pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ communicates with patterns in the sets

$$\hat{S}_{\mathbf{p}}^- = \{\mathbf{q} \in S_{\mathbf{p}}^- : \hat{s}(\mathbf{q}) = \hat{s}(\mathbf{p}) - 1,\, \hat{d}(\mathbf{p}, \mathbf{q}) = 1\},$$
$$\hat{S}_{\mathbf{p}}^+ = \{\mathbf{q} \in S_{\mathbf{p}}^+ : \hat{s}(\mathbf{q}) = \hat{s}(\mathbf{p}) + 1,\, \hat{d}(\mathbf{p}, \mathbf{q}) = 1\},$$

and its expected frequency evolves

$$\begin{aligned}\dot{x}_{\mathbf{p}}(t) = &- \hat{s}(\mathbf{p})\left[\mu + \beta\left(1 - \tfrac{\hat{s}(\mathbf{p})}{\hat{L}^{(t)}}\right)\right] x_{\mathbf{p}}(t) + \lambda \mathbf{1}_{\{s(\mathbf{p}) = \hat{s}(\mathbf{p}) = 1\}} \\ &+ \beta \sum_{\mathbf{q} \in \hat{S}_{\mathbf{p}}^-} \tfrac{\hat{s}(\mathbf{q})}{\hat{L}^{(t)}} x_{\mathbf{q}}(t) + \mu \sum_{\mathbf{q} \in \hat{S}_{\mathbf{p}}^+} x_{\mathbf{q}}(t).\end{aligned}$$

As with rate heterogeneity in Section 2.4.1, we may prove that this equation correctly describes the evolution of patterns displaying pattern $\mathbf{p}$ using the techniques in the proof of Theorem 1.

Unregistered traits	Unregistered patterns $\mathcal{Q} \setminus R(\mathcal{Q})$
Observed in j taxa or fewer	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i = 1\}\rvert \leq j\}$
Observed in j or more taxa	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i = 1\}\rvert \geq j\}$
Potentially present in j or more taxa	$\{\mathbf{q} \in \mathcal{Q} : \lvert\{i \in [L] : q_i \neq 0\}\rvert \geq j\}$
Absent in taxon k_i	$\{\mathbf{q} \in \mathcal{Q} : q_i = 0\}$

Table 2.2: Registration rules proposed by Alekseyenko et al. (2008) and Ryder and Nicholls (2011).

2.4.4 Data registration

Patterns which are deemed uninformative or unreliable are typically removed from the data. For example, in a parsimony-based analysis, traits which are present in either a single taxon or all of the taxa are not informative of the tree structure and are discarded from the data. This is not the case for our model as many traits present in a single taxon are indicative of a long branch and a lack of lateral transfer. Similarly, many traits present in all of the taxa indicate short branches near the leaves or high levels of lateral transfer.

To proceed, we define a registration rule R, which may be a composition of other simpler rules such as those in Table 2.2, and discard the columns in the data array $\mathbf{D}$ not satisfying R, leaving the registered data $R(\mathbf{D})$. The consequence for the model likelihood in Theorem 1 is trivial as we simply restrict the product over patterns to $R(\mathcal{P})$ or $R(\mathcal{Q})$ where necessary. In all of the analyses in this thesis, we discard the traits not marked present in at least one taxon.

The likelihood in Equation 2.3.6 includes a term for the expected number of registered patterns in the data. If we do not remove any patterns from the data, then

$$\mathbb{E}\left[\sum_{\mathbf{q}\in\mathcal{Q}} N_{\mathbf{q}} \middle| g, \lambda, \mu, \beta, \Xi, \dots\right] = \sum_{\mathbf{q}\in\mathcal{Q}} x_{\mathbf{q}}(0; g, \lambda, \mu, \beta, \Xi, \dots) = \sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{p}}(0; g, \lambda, \mu, \beta, \dots).$$

If we discard the patterns corresponding to traits potentially absent in all of the taxa, then

$$\sum_{\mathbf{q}\in R(\mathcal{Q})} x_{\mathbf{q}}(0;g,\lambda,\mu,\beta,\Xi,\dots) = \sum_{\mathbf{p}\in\mathcal{P}}\left[1-\prod_{i=1}^{L}(1-\xi_{k_i})^{p_i}\right] x_{\mathbf{p}}(0;g,\lambda,\mu,\beta,\dots).$$

Similarly, if we discard the patterns which correspond to traits potentially present in all of the taxa, we have

$$\sum_{\mathbf{q}\in R(\mathcal{Q})} x_{\mathbf{q}}(0;g,\lambda,\mu,\beta,\Xi,\dots) = \sum_{\mathbf{p}\in\mathcal{P}}\left[1-\prod_{i=1}^{L}(1-\xi_{k_i})^{1-p_i}\right] x_{\mathbf{p}}(0;g,\lambda,\mu,\beta,\dots).$$

There is some overlap between the sets of patterns displayed by traits either potentially absent or potentially present in all of the taxa, in which case the total expected pattern frequency is

$$\begin{aligned}&\sum_{\mathbf{q}\in R(\mathcal{Q})} x_{\mathbf{q}}(0;g,\lambda,\mu,\beta,\Xi,\dots)\\ &\quad= \sum_{\mathbf{p}\in\mathcal{P}}\left[1-\prod_{i=1}^{L}(1-\xi_{k_i})^{1-p_i}\right]\left[1-\prod_{i=1}^{L}(1-\xi_{k_i})^{p_i}\right] x_{\mathbf{p}}(0;g,\lambda,\mu,\beta,\dots).\end{aligned}$$

2.5 Bayesian model

2.5.1 Prior distributions

In order to estimate both node times and rate parameters, we calibrate the space Γ of rooted phylogenetic trees on L taxa with *clade constraints*. The constraint $\Gamma^{(0)} = \{g \in \Gamma : \underline{t}_1 \leq t_1 < 0\}$ restricts the earliest admissible root time to $\underline{t}_1$. Each additional constraint $\Gamma^{(c)}$ places either time or ancestry constraints on the remaining nodes. We denote by $\Gamma^{\mathrm{C}} = \bigcap_c \Gamma^{(c)}$ the calibrated space of phylogenies satisfying the clade constraints.

Nicholls and Ryder (2011) describe a prior distribution on trees with the property that the marginal distribution of the root time t_1 is approximately uniform across an interval $[\underline{t}_1, \bar{t}_1] \subset \mathbb{R}$ specified in advance. For a given tree $g = (V, E, T, C)$, there are $Z(g)$ possible time orderings of the nodes amongst the admissible node times $T(g) =$

Parameter	Prior	Reasoning
Trait birth rate	$\lambda \sim 1/\lambda$	Improper, scale invariant
Trait death rate	$\mu \sim \Gamma(10^{-3}, 10^{-3})$	Approximately $1/\mu$
Trait transfer rate	$\beta \sim \Gamma(10^{-3}, 10^{-3})$	Approximately $1/\beta$
Catastrophe rate	$\rho \sim \Gamma(1.5, 5 \times 10^3)$	$\mathbb{E}[\rho^{-1}] = 10^4$ years
Catastrophe severity	$\kappa \sim \mathrm{U}[0.25, 1]$	$\mathbb{E}[\delta\|\mu] = \mu^{-1}[1 - \log(0.75)]$ years
Observation probabilities	$\Xi \sim \mathrm{U}[0,1]^L$	Independent, uniform

Table 2.3: Prior distributions on parameters.

$\{T' : (V, E, T', C) \in \Gamma^C\}$. For each node $i \in V$, $\underline{t}_i = \inf_{T \in T(g)} t_i$ and $\bar{t}_i = \sup_{T \in T(g)} t_i$ are respectively the earliest and most recent times that i may achieve in an admissible tree with topology (V, E). If $S(g) = \{i \in V : \underline{t}_i = \underline{t}_1\}$ denotes the set of *free* nodes with times bounded below by $\underline{t}_1$, we define the prior distribution on trees with density

$$f_G(g) \propto \frac{\mathbf{1}_{\{g \in \Gamma^C\}}}{Z(g)} \prod_{i \in S(g)} \frac{\underline{t}_1 - \bar{t}_i}{t_1 - \bar{t}_i},$$

Provided that $\underline{t}_1 \ll \min_{i \in V \setminus S} \underline{t}_i$, the corresponding marginal prior distributions on the topology (V, E) and root time t_1 are approximately uniform (Ryder and Nicholls, 2011). Uniform priors on offset leaf times completes our prior specification on g. Heled and Drummond (2012) describe an exact, computationally intensive method to compute uniform calibrated tree priors, but we do not pursue that approach here. Table 2.3 lists the prior distributions on the remaining parameters.

In Section 2.4.1, we state that catastrophes occur at rate ρ according to a Poisson process along the branches of the tree. For branch $i \in E \setminus \{1\}$, let $C^{(i)}$ denote the set of catastrophes on i and $\Delta^{(i)} = t_{\mathrm{pa}(i)} - t_i$ its length. We denote by $\Delta = \sum_{i \in V \setminus \{0,1\}} \Delta^{(i)}$ the total length of the tree below the root. The conditional prior on the catastrophe set C is

$$\pi_{C|R}(C|\rho) = \prod_{i \in E \setminus \{1\}} \frac{(\rho \Delta_i)^{|C^{(i)}|} e^{-\rho \Delta_i}}{|C^{(i)}|!} \frac{|C^{(i)}|!}{\Delta^{|C^{(i)}|}} = \rho^{|C|} e^{-\rho \Delta},$$

where we account for the fact that, conditional on their number on a branch, catastrophes are invariant to relabelling and *a priori* uniformly distributed across the branch. This is equivalent to a Poisson$(\rho\Delta)$ number of catastrophes uniformly distributed

across the entire tree. The prior on ρ in Table 2.3 is $\Gamma(a, b)$, where $a = 1.5$ and $b = 5 \times 10^3$, so the marginal prior on the catastrophe set C is

$$\begin{aligned}\pi_C(C) &= \int_0^\infty \pi_{C|R}(C|\rho)\pi_R(\rho)\mathrm{d}\rho \\ &= \frac{b^a}{\Gamma(a)} \frac{\Gamma(|C| + a)}{(\Delta + b)^{|C|+a}} \\ &= \frac{\Gamma(|C| + a)}{\Gamma(a)|C|!} \left(\frac{\Delta}{\Delta + b}\right)^{|C|} \left(\frac{b}{\Delta + b}\right)^a \frac{|C|!}{\Delta^{|C|}},\end{aligned} \tag{2.5.1}$$

a Negative Binomial$[a, b/(b + \Delta)]$ distribution on the total number of catastrophes. Conditional on their total number, catastrophes locations remain *a priori* uniformly distributed across the tree. The number of catastrophes on each branch is a Negative Multinomial random variable, so the marginal distribution on a single branch $i \in E \setminus \{1\}$ below the root is Negative Binomial with parameters a and $b/(b + t_{\mathrm{pa}(i)} - t_i)$.

The *relative transfer rate* β/μ is a more natural indicator of the level of lateral transfer in the data than the trait transfer rate β. The relative transfer rate is the expected number of times that a single instance of a trait is copied by other lineages before it dies. This result follows a similar argument to the physical interpretation of a catastrophe in Section 2.4.1. For an instance of a trait, we know from Properties T3 and T4 in Section 2.2 that the time until the trait dies is an Exp(μ) random variable, and the time until its next transfer event is a Exp(β) random variable. By the superposition and thinning properties of Poisson processes, the time until the next event involving the trait is an Exp$(\mu + \beta)$ random variable, and this is a death event with probability $\mu/(\mu + \beta)$ (Kingman, 1992). Focusing on the pure event process, the trait is killed after a Geometric$[\mu/(\mu + \beta)] - 1$ number of transfer events. This random variable has expectation $(\mu + \beta)/\mu - 1 = \beta/\mu$, the relative transfer rate. From the Gamma marginal prior distributions on μ and β in Table 2.3, we derive that the prior on the relative transfer rate is a Beta Distribution of the Second Kind with infinite mean and

variance.[1] On the basis of simulation studies, Nicholls and Gray (2008) and Greenhill et al. (2009) consider a relative transfer rate of 0.5 to be high as it is easily detected from summary statistics of the data. We revisit this quantity in Chapters 5 and 6.

2.5.2 Posterior distribution

As we remark earlier, new traits arise according to a Poisson process of rate λ and evolve independently along the tree. This has two implications for our inference:

- The expected pattern frequencies are a linear function of the birth rate λ; that is, $\mathbf{x}(t; g, \lambda, \ldots) = \lambda \mathbf{x}(t; g, 1, \ldots)$
- We may integrate λ out of the likelihood in Equation 2.3.6 analytically.

This first result follows from three properties of the expected pattern frequency calculation (2.3.5): the initial condition $\mathbf{x}(t_1^-) = \lambda/\mu$ at the root; the linearity of the initial value problems between branching events; and the linear expected pattern frequency initialisation operations $\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots$ at the branching events. To see this,

[1] The probability density transform for the change of variables $(\mu, \beta) \overset{g}{\mapsto} (x, y) = (\mu, \beta/\mu)$ is

$$\pi_{X,Y}(x, y) = \pi_{M,B}[g_1^{-1}(x, y), g_2^{-1}(x, y)]|J|^{-1} = \pi_M[g_1^{-1}(x, y)]\pi_B[g_2^{-1}(x, y)]|J|^{-1}$$

where $g_1^{-1}(x, y) = \mu = x$, $g_2^{-1}(x, y) = \beta = xy$. The Jacobian of this transformation is

$$J = \frac{\partial g(\mu, \beta)}{\partial(\mu, \beta)} = \begin{bmatrix} 1 & -\frac{\beta}{\mu^2} \\ 0 & \frac{1}{\mu} \end{bmatrix},$$

so $|J|^{-1} = \mu = x$. If π_M and π_B are both $\Gamma(a, b)$ densities then

$$\pi_{B/M}(y) = \int \pi_M(x)\pi_B(xy)x\mathrm{d}x = \frac{\Gamma(2a)}{\Gamma^2(a)} \frac{y^{a-1}}{(1+y)^{2a}},$$

a Beta Distribution of the Second Kind density with both parameters set to a. This Beta Distribution of the Second Kind has infinite mean when $a \leq 1$ and infinite variance when $a \leq 2$. From Table 2.3, we set $a = 10^{-3}$ for our analyses.

let $\mathbf{y}(t) = \mathbf{x}(t; g, 1, \dots)$ and observe that

$$\dot{\mathbf{x}}(t_1; g, \lambda, \dots) = \mathbf{A}^{(t_1)}\mathbf{x}(t_1; g, \lambda, \dots) + \mathbf{b}^{(t_1)} = \mathbf{A}^{(t_1)} \begin{bmatrix} 0 \\ 0 \\ \lambda/\mu \end{bmatrix} + \begin{bmatrix} \lambda \\ \lambda \\ 0 \end{bmatrix} = \lambda \dot{\mathbf{y}}(t_1),$$

by construction, so $\mathbf{x}(t_2^-) = \lambda \mathbf{y}(t_2^-)$. The initialisation operation $\mathbf{T}^{(2)}$ is linear, so $\mathbf{x}(t_2) = T^{(2)}\mathbf{x}(t_2^-) = \lambda T^{(2)}\mathbf{y}(t_2^-)$. We then repeat this argument to see that $\mathbf{x}(t; g, \lambda, \dots) = \lambda \mathbf{y}(t)$ for any time t. For $\mathbf{y} = \mathbf{y}(0; g, 1, \dots)$, from the Poisson likelihood in Theorem 1 and the improper prior on λ in Table 2.3, we have

$$\begin{aligned} \pi(\mathbf{D}|g, \mu, \beta, \dots) &= \int_0^\infty \pi(\mathbf{D}|g, \lambda, \dots)\pi_\Lambda(\lambda)\mathrm{d}\lambda \\ &\propto \int_0^\infty \frac{1}{\lambda} \prod_{\mathbf{p} \in R(\mathcal{Q})} (\lambda y_{\mathbf{p}})^{n_{\mathbf{p}}} e^{-\lambda y_{\mathbf{p}}} \mathrm{d}\lambda \\ &\propto \prod_{\mathbf{p} \in R(\mathcal{Q})} \left(\frac{y_{\mathbf{p}}}{\sum_{\mathbf{q} \in R(\mathcal{Q})} y_{\mathbf{q}}} \right)^{n_{\mathbf{p}}}, \end{aligned}$$

a multinomial likelihood whereby a pattern $\mathbf{p} \in R(\mathcal{Q})$ is observed with probability proportional to its expected frequency $x_{\mathbf{p}}$.

Let $n_{\mathbf{p}} = |\{h \in \mathcal{H}(0) : \mathbf{p} = \mathbf{d}^h \in R(\mathbf{D})\}|$ denote the frequency of traits in the registered data displaying pattern $\mathbf{p} \in R(\mathcal{Q})$. Putting everything together, the posterior distribution is

$$\pi(g, \mu, \beta, \kappa, \Xi | R(\mathbf{D})) \propto f_G(g) f_M(\mu) f_B(\beta) \prod_{\mathbf{p} \in R(\mathcal{Q})} \left(\frac{x_{\mathbf{p}}}{\sum_{\mathbf{q} \in R(\mathcal{Q})} x_{\mathbf{q}}} \right)^{n_{\mathbf{p}}}, \tag{2.5.2}$$

where the expected pattern frequencies $\mathbf{x} \equiv \mathbf{x}(0; g, 1, \mu, \beta, \kappa, \Xi)$ account for catastrophes, missing data and offset leaves, where necessary. This completes the specification of the Stochastic Dollo with Lateral Transfer (SDLT) model.

2.6 Discussion

In this chapter, we describe a novel Bayesian model of species diversification and extensions to settings such as rate heterogeneity. The model posterior (2.5.2) is intractable: we may only evaluate it numerically at a given point, and only up to an unknown normalising constant. In Chapter 3, we describe how to perform exact MCMC inference under the model using Markov chain Monte Carlo methods. In Chapter 4, we address the computational cost of our likelihood calculation due to the integration over all possible trait histories under the model and describe our exact-approximate inference scheme.

Chapter 3

Exact MCMC inference

Chapter overview

We now describe how to perform inference under the Stochastic Dollo with Lateral Transfer model using two exact Markov chain Monte Carlo methods in the sense that the limiting distribution of the samples we draw is the posterior distribution in Equation 2.5.2.

3.1 Metropolis–Hastings algorithm

3.1.1 Description

To perform inference under the SDLT model, we construct a Markov chain Monte Carlo (MCMC) algorithm whose invariant distribution is the SDLT model posterior in Equation 2.5.2. From these samples, we estimate statistics and test hypotheses. From an initial configuration, $X_0 = [g_0, \mu_0, \beta_0, \kappa_0, \Xi_0]$, the algorithm generates a dependent sequence $X_1, X_2, \ldots, X_T$ of states according to the following rule: for $t = 0, \ldots, T-1$, draw the candidate state X^* from the proposal distribution $Q(X_t, \cdot)$ at the current

Algorithm 1 The Metropolis–Hastings algorithm

1: **Initialise** the chain at state X_0
2: **For** t = 1 **to** T **do**
3: **Sample** X_t from the Metropolis–Hastings kernel $K(X_{t-1}, \cdot)$ (3.1.2) targeting the posterior $\pi(\cdot|\mathbf{D})$ (2.5.2)
4: **End For**
5: **Return** samples $X_1, \ldots, X_T$

state and with probability $\alpha(X_t, X^*)$, where

$$\alpha(X_t, X^*) = \min\left[1, \frac{\pi(X^*|\mathbf{D})}{\pi(X_t|\mathbf{D})} \frac{Q(X^*, X_t)}{Q(X_t, X^*)}\right], \tag{3.1.1}$$

we set $X_{t+1} \leftarrow X^*$, and otherwise we set $X_{t+1} \leftarrow X_t$ (Metropolis et al., 1953; Hastings, 1970). The transition kernel of this Markov chain is

$$K(X_t, X_{t+1}) = Q(X_t, X_{t+1})\alpha(X_t, X_{t+1}) + \mathbf{1}_{\{X_t = X_{t+1}\}}\left(1 - \int_{\mathcal{X}} Q(X_t, X^*)\alpha(X_t, X^*)\mathrm{d}X^*\right), \tag{3.1.2}$$

where the integral is taken over the entire sample space $\mathcal{X}$ of the SDLT model parameters. We summarise this Metropolis–Hastings sampling scheme in Algorithm 1.

3.1.2 Proposal distributions

The SDLT model parameters comprise a discrete topology (V, E) with continuous node times $T = \{t_i : i \in V\}$, a set C of catastrophes which change the dimension of the model, and $L + 3$ continuous parameters μ, β, κ and $\Xi = (\xi_i : i \in V_L)$. Drummond et al. (2002) describe a family of proposal distributions $Q(\cdot, \cdot)$ for phylogenetic trees and rate parameters which yield an ergodic, Harris-recurrent chain. They perturb one component of the current state at a time to form the proposed state, so the algorithm falls into the category of *random-walk Metropolis* algorithms. Nicholls and Gray (2008) and Ryder and Nicholls (2011) discuss these moves in the context of the Stochastic Dollo model with parameters $[(V, E, T), \mu]$ and $[(V, E, T, C), \mu, \rho, \kappa, \Xi]$ respectively. In order to perform inference under the SDLT model, we construct a new move for the lateral transfer rate β and modify existing moves to account for catastrophes locations

Parameter	Move
Topology (V, E)	Exchange a pair of nodes Exchange a pair of subtrees
Node times T	Scale some or all node times Vary one internal node time Vary leaf times
Catastrophes C	Add or delete a catastrophe on a branch Move a catastrophe to a neighbouring branch
Continuous parameters	Rescale the death rate μ, lateral transfer rate β, or catastrophe severity κ Rescale one or all of the observation probabilities Ξ

Table 3.1: Possible moves in our MCMC algorithm.

on the tree. Table 3.1 lists the possible moves in our algorithm. We now discuss the moves particular to the SDLT model in greater detail.

Many of the moves in Table 3.1 either change the dimension of the model or the unit of volume in the proposal distribution; that is, for a proposed move X_t to X^*, we have $\mathrm{d}X_t \neq \mathrm{d}X^*$. In order to construct such moves, we use the more general description of Equation 3.1.1 by Green (1995). At the current state X_t, we generate a random variable $\varrho \in \mathbb{R}$ from some distribution with density ϖ, and construct the proposed state X^* through the invertible, deterministic map $g : (X_t, \varrho) \mapsto (X^*, \varrho^*)$. We rewrite Equation 3.1.1 for this approach as

$$\alpha(X_t, X^*) = \min\left[1, \frac{\pi(X^*|\mathbf{D})}{\pi(X_t|\mathbf{D})} \frac{\varpi(\varrho^*)}{\varpi(\varrho)} \left|\frac{\partial g(X_t, \varrho)}{\partial(X_t, \varrho)}\right|\right]. \tag{3.1.3}$$

For a set $\mathrm{d}X^*$, the proposal distribution $Q(\cdot, \cdot)$ in Equation 3.1.1 is now

$$Q(X_t, \mathrm{d}X^*) = \int_{\{\varrho : g(X_t, \varrho) \in \mathrm{d}X^* \times \mathbb{R}\}} \varpi(\varrho)\mathrm{d}\varrho.$$

In the following discussion, the current state of the chain is $X = [(V, E, T, C), \mu, \beta, \kappa, \Xi]$ and the proposed state is X^*.

Lateral transfer rate β

We apply the same scaling update to the lateral transfer rate β as the death rate μ. We sample the auxiliary variable $\varrho \sim \mathrm{U}[\varkappa^{-1}, \varkappa]$, for some constant $\varkappa > 1$, and compute the proposed lateral transfer rate β^* in state X^* through the transformation $g : (\beta, \varrho) \mapsto (\beta^*, \varrho^*) = (\varrho\beta, \varrho^{-1})$. The relevant terms in the Metropolis–Hastings–Green acceptance ratio (3.1.3) are

$$\frac{\varpi(\varrho^*)}{\varpi(\varrho)} \left| \frac{\partial g(\beta, \varrho)}{\partial(\beta, \varrho)} \right| = \frac{\beta}{\beta^*} = \frac{1}{\varrho},$$

as $\varpi(\varrho) = \varpi(\varrho^{-1})$.

Catastrophe set C

In Section 2.4.1, we parameterise a catastrophe $c = (b, t)$, where b is the branch index and t is the catastrophe time. To simplify our discussion of some of the moves here, we introduce the reparameterisation $c = (b, u)$, where $u \in (0, 1)$ satisfies $t = t_b + u(t_{\mathrm{pa}(b)} - t_b)$ in the definition above. We say that u is the *relative location* of catastrophe c along branch b. The relative location of a catastrophe remains fixed throughout the duration that it is in the state of the Markov chain. This parameterisation is scale invariant, in the sense that if we scale all the node times by a constant ϱ, the catastrophe time is scaled accordingly. Of course, we must still account for the change in the unit of volume of the catastrophe sampling distribution.

Addition and deletion of catastrophes. The location for a new catastrophe $c^* = (b^*, u^*)$ is chosen uniformly at random across the branches of the tree to form the proposed state X^* with catastrophe set $C^* = C \cup \{c^*\}$. Catastrophes are chosen uniformly at random for deletion in the reverse move so the ratio of proposal distributions in Equation 3.1.1 is

$$\frac{Q(X^*, X)}{Q(X, X^*)} = \frac{p_{DC}}{p_{AC}} \frac{1}{|C| + 1} \sum_{i \in E \setminus \{1\}} (t_{\mathrm{pa}(i)} - t_i), \tag{3.1.4}$$

where p_{AC} and p_{DC} denote the probabilities of proposing to add and delete a catastrophe respectively.

Adding and deleting catastrophes changes the dimension of the state. Alternatively, we could sample the location ϱ uniformly across the branches of the tree below the root so $\varpi(\varrho) = 1/\sum_{i \in E \setminus \{1\}} (t_{\mathrm{pa}(i)} - t_i)$ and write the forward move as $g : (c_1, \ldots, c_{|C|}, \varrho) \mapsto (c_1, \ldots, c_{|C|}, c_{|C|+1}) = C^*$. The Jacobian of this transformation is 1. For the reverse move, we draw ϱ^* from the discrete uniform distribution $\mathrm{U}\{[|C| + 1]\}$ on the set $\{1, \ldots, |C| + 1\}$, which has probability mass function $\varpi^*(\varrho^*) = 1/|C^*| = 1/(|C| + 1)$, and the terms in Equation 3.1.3 corresponding to Equation 3.1.4 follow.

Move catastrophe branch. We chose catastrophe $c = (b, u)$ uniformly from the catastrophe set C to move to branch b^* chosen uniformly from the $\deg(b) + \deg[\mathrm{pa}(b)] - 2$ branches neighbouring branch b. If $c^* = (b^*, u^*) = (b^*, u)$ replaces c in the proposed state X^*, the ratio of proposal distributions in Equation 3.1.1 is

$$\frac{Q(X^*, X)}{Q(X, X^*)} = \frac{\deg(b) + \deg[\mathrm{pa}(b)] - 2}{\deg(b^*) + \deg[\mathrm{pa}(b^*)] - 2} \frac{t_{\mathrm{pa}(b^*)} - t_{b^*}}{t_{\mathrm{pa}(b)} - t_b}.$$

The first term on the right-hand side here is the ratio of $\pi(b|b^*)$ and $\pi(b^*|b)$, where $\pi(b^*|b) = 1/(\deg(b) + \deg[\mathrm{pa}(b)] - 2)$ is the probability mass function of choosing branch b^* when we elect to move a catastrophe currently on branch b. The second term on the right-hand side accounts for the change in catastrophe sampling densities due to the different lengths of branches b and b^*. Although we do not resample the relative location u of the catastrophe along the branch, our model is parameterised in terms of the catastrophe times, so we must account for this in the transition kernel (3.1.2) of the chain. Therefore, if $c = (b, t)$ and $c^* = (b^*, t^*)$, then $\pi(t|b) = 1/(t_{\mathrm{pa}(b)} - t_b)$ and $\pi(t^*|b^*) = 1/(t_{\mathrm{pa}(b^*)} - t_{b^*})$.

Tree moves. We describe two different approaches to account for catastrophes when we update the tree topology or node times. Both approaches lead to the same result.

Suppose that we propose to scale all the node times T in the current state X by a factor $\varrho \in [\varkappa^{-1}, \varkappa]$ to form the proposed state $X^* = [(V, E, \varrho T, C), \dots]$. We can

- Update the tree with the catastrophes *in situ*
- Remove the catastrophes then update the tree and replace the catastrophes in the same relative locations.

Suppose that $T^C = \{t : (b, t) \in C\}$, the set of catastrophe times in the current state. For the first approach, the proposed state is given by the transformation $g : (T, T^C, \varrho) \mapsto (T^*, T^{C^*}, \varrho^*) = (\varrho T, \varrho T^C, \varrho^{-1})$, and the Jacobian term in the acceptance ratio in Equation 3.1.3 is

$$\left| \frac{\partial g(T, T^C, \varrho)}{\partial (T, T^C, \varrho)} \right| = \varrho^{(2L-1)+|C|-2},$$

as there are $2L - 1$ nodes beneath the Adam node in the tree. For the second interpretation above, the probability of proposing to remove all of the catastrophes in either state is equal. The Jacobian term in Equation 3.1.3 for the map $g : (T, \varrho) \mapsto (\varrho T, \varrho^{-1})$ is $\varrho^{(2L-1)-2}$. Now, recall the notation from Section 2.5.1 whereby $C^{(i)}$ denotes the catastrophe set on branch $i \in E \setminus \{1\}$. The relative sampling densities for adding back the catastrophes after the reverse (numerator) and forward (denominator) scaling moves is

$$\frac{\prod_{i \in E \setminus \{1\}} \frac{|C^{(i)}|!}{(t_{\mathrm{pa}(i)} - t_i)^{|C^{(i)}|}}}{\prod_{i \in E^* \setminus \{1\}} \frac{|C^{(i)*}|!}{(t^*_{\mathrm{pa}(i)} - t^*_i)^{|C^{(i)^*}|}}} = \prod_{i \in E \setminus \{1\}} \frac{[\varrho(t^*_{\mathrm{pa}(i)} - t^*_i)]^{|C^{(i)^*}|}}{(t_{\mathrm{pa}(i)} - t_i)^{|C^{(i)}|}} = \varrho^{|C|},$$

where $\pi(C^{(i)}||C^{(i)}|) = |C^{(i)}|!/(t_{\mathrm{pa}(i)} - t_i)^{|C^{(i)}|}$, the *a priori* uniform sampling distribution of the individual catastrophe locations in set $C^{(i)}$ in state X, given their total number $|C^{(i)}|$ on the branch.

This discussion leads us to a more general statement. For every move in Table 3.1 which affects the branch lengths — the subtree-prune-and-regraft moves we

describe below, for example — or the locations of catastrophes, the Jacobian term in Equation 3.1.3 is

$$\left|\frac{\partial g(X,\varrho)}{\partial(X,\varrho)}\right| \propto \prod_{i\in E\setminus\{1\}} \frac{|C^{(i)}|!}{(t_{\mathrm{pa}(i)}-t_i)^{|C^{(i)}|}} \frac{(t^*_{\mathrm{pa}(i)}-t^*_i)^{|C^{*(i)}|}}{|C^{*(i)}|!},$$

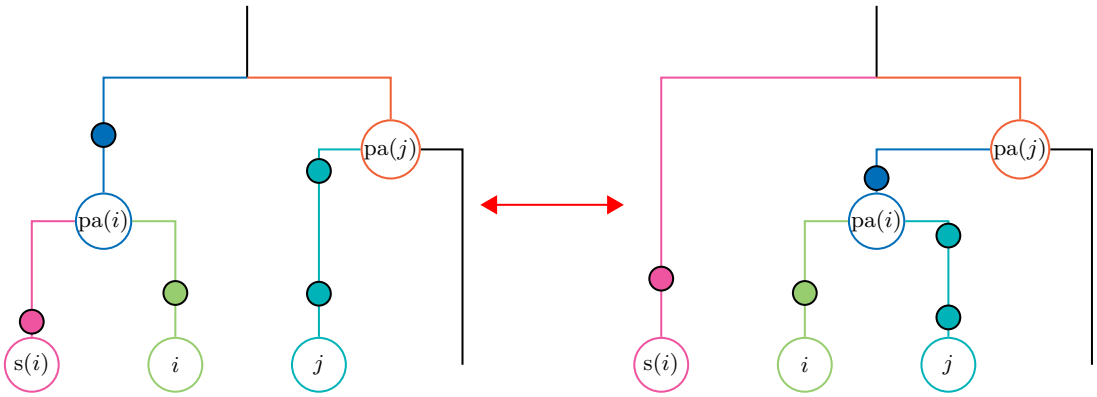
to account for the relative catastrophe sampling distributions in the current and proposed states.

We define subtree-prune-and-regraft moves on the tree in such a way that the total number of catastrophes on the tree remains constant and we do not need to combine tree moves with moves on catastrophes. Let $\langle \mathrm{pa}(i), i\rangle \in E$ denote a time-directed branch. From the current state X, we choose a node $i \in V \setminus \{0,1\}$ below the root, prune the subtree beneath its parent $\mathrm{pa}(i)$ and reattach it at a location chosen uniformly along a randomly chosen branch $\langle \mathrm{pa}(j), j\rangle \in E$ to create state X^*. Now, recall that catastrophes are indexed by the offspring node of the branch they lie on and suppose that vertices retain their labels in the move from state X to state X^*. There are three possible outcomes:

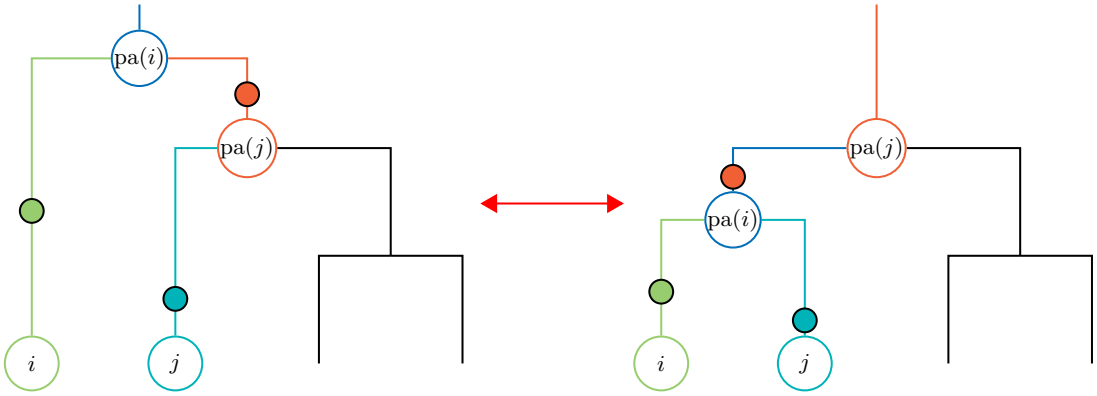
- If neither node i or $\mathrm{pa}(j)$ is the root in X, then catastrophes remain on their assigned branches in state X^*. In more detail, if a catastrophe $c = (b,u)$ is on branch $\langle \mathrm{pa}(b), b\rangle$ in state X then $c = (b,u)$ in state X^* also, with the possible difference that node b's parent may change, thereby affecting when the catastrophe occurs but not its relative location. We illustrate this point in Figure 3.1a.

- If $\mathrm{pa}(i)$ is the root in state X, then i's sibling $\mathrm{s}(i)$ is the root in state X^*. There are no catastrophes on $\langle 0, \mathrm{pa}(i)\rangle$ in X by definition so we move the catastrophes currently on $\langle \mathrm{pa}(i), \mathrm{s}(i)\rangle$ in X to $\langle \mathrm{pa}(j), \mathrm{pa}(i)\rangle$ in X^*, and the catastrophes on $\langle \mathrm{pa}(j), j\rangle$ in X to $\langle \mathrm{pa}(i), j\rangle$ in X^*. We illustrate this move in Figure 3.1b.

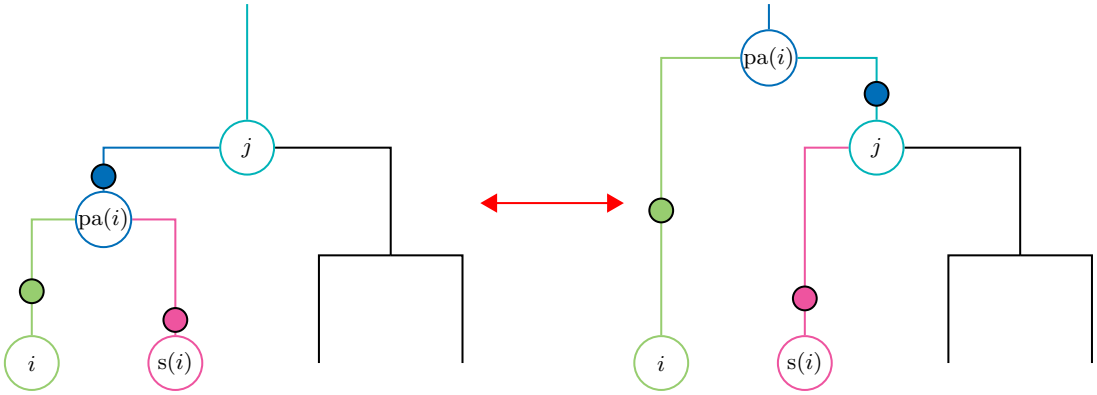
- Finally, if j is the root in X, then $\mathrm{pa}(i)$ becomes the root in X^* so we move the catastrophes on $\langle \mathrm{pa}(\mathrm{pa}(i)), \mathrm{pa}(i)\rangle$ in X to $\langle \mathrm{pa}(i), j\rangle$ in X^*. We illustrate this move, the reverse of the above, in Figure 3.1c.



(a) Neither node pa(i) nor pa(j) is the root in either state. Catastrophes remain on their assigned branches.



(b) Node pa(i) is the root in the left-hand state and node pa(j) is i's sibling. The catastrophes on branch $\langle \text{pa}(i), \text{pa}(j) \rangle$ in the left-hand state are moved to edge $\langle \text{pa}(j), \text{pa}(i) \rangle$ in the right-hand state when node j becomes the root.



(c) Node j is the root in the left-hand state. The catastrophes on branch $\langle j, \text{pa}(i) \rangle$ in the left-hand state are moved to edge $\langle \text{pa}(i), j \rangle$ in the right-hand state.

Figure 3.1: Subtree-prune-and-regraft moves do not affect the number of catastrophes on the tree.

3.1.3 Implementation

In addition to manually checking the code, we validate our implementation of the prior distributions and proposal moves through simulation. We now illustrate this with an example. We sample from the SDLT/SD prior using the Metropolis–Hastings algorithm to estimate the marginal distributions of the number of catastrophes on each branch. In Figure 3.2, we compare these distributions with samples from the corresponding Negative Binomial distributions in Section 2.5.1 on the same set of sampled trees. From this experiment, we conclude that our implementation of the prior distributions and Metropolis–Hastings algorithm is correct. In Chapter 5, we use synthetic data sets to validate our overall implementation of the SDLT model.

3.2 Convergence and mixing

We present a standard discussion of practical issues with MCMC algorithms, then focus on aspects unique to MCMC with catastrophes.

3.2.1 Efficiency

The Metropolis–Hastings algorithm in the previous section produces a dependent sequence of samples asymptotically distributed according to the posterior distribution in Equation 2.5.2. In practice, we monitor the trace and autocorrelation plots for each parameter to assess *convergence* (Geyer, 1992), and base our inference on the samples produced thereafter. Peskun (1973) proposes to rank sampling algorithms by the asymptotic variance of their Monte Carlo averages. We consider the effective number of independent samples from the posterior (2.5.2) produced by our algorithm. For T dependent samples, the *effective sample size* is

$$\text{ESS} = \frac{T}{1 + \sum_{k=1}^{\infty} \rho_k},$$

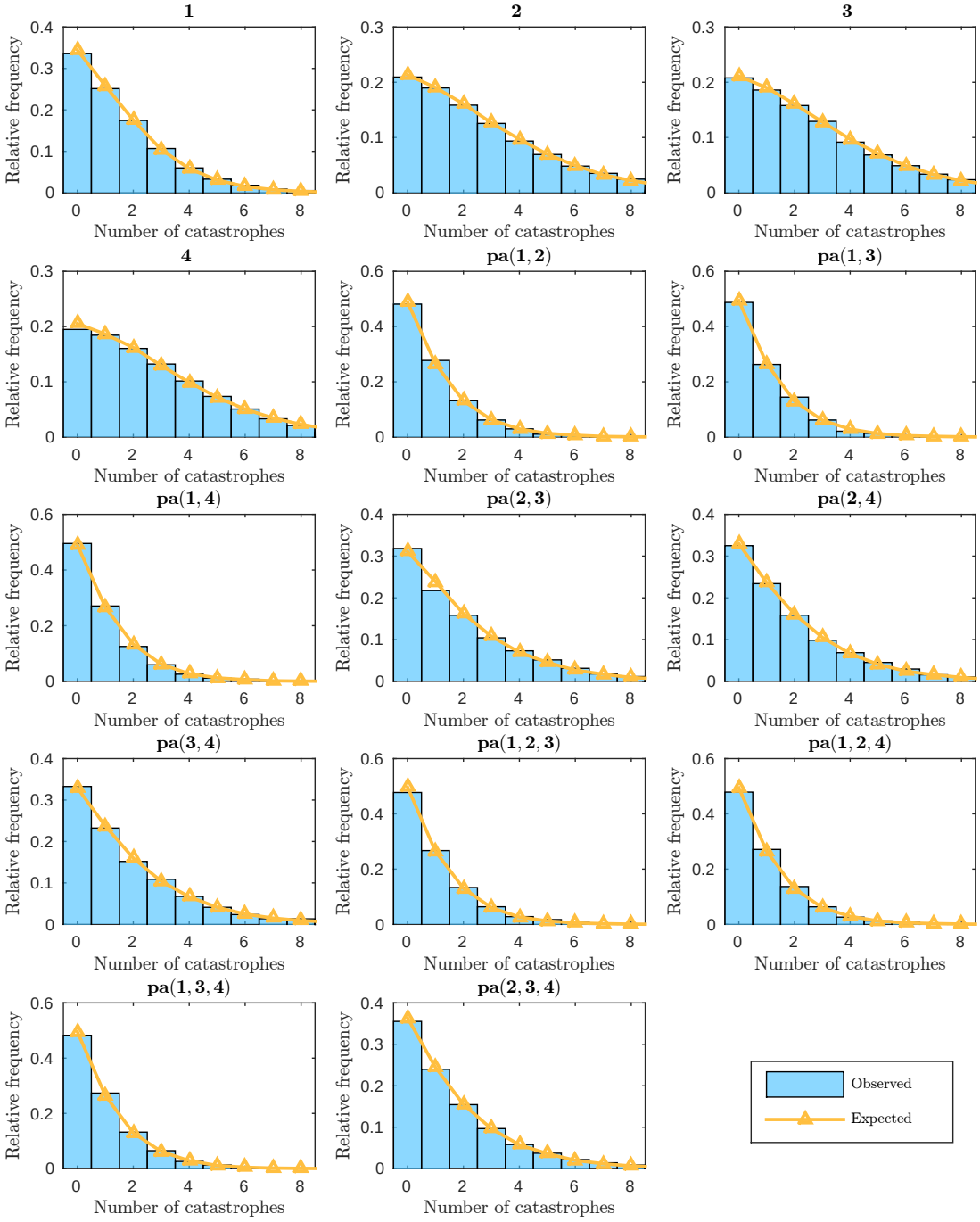


Figure 3.2: Marginal distributions of the number of catastrophes on branches of a four-taxon tree in samples the SDLT/SD prior. Not every branch appears in every sampled tree. We index each branch by its offspring node. We overload our notation here so $\mathrm{pa}(i, j)$ denotes the most recent common ancestor of leaves i and j.

where ρ_k is the lag-k autocorrelation between samples. Faster *mixing* chains have lower correlations between successive samples and therefore yield higher effective sample sizes. Geyer describes practical guidelines for computing the effective sample size when the autocorrelation terms must also be estimated from the MCMC output.

3.2.2 Mixing over catastrophes

Following our remarks in Section 2.4.1, our inference scheme is unable to distinguish between the effect of the overall trait process and multiple weak catastrophes. The likelihood (2.3.6) is extremely sensitive to the location and strength of catastrophes so, in spite of the prior on the number of catastrophes, our Metropolis–Hastings Markov chain may get stuck in any number of local modes. In addition to this mixing problem, a large number of catastrophes slows the algorithm as we must solve more initial value problems (2.3.5). This hints at a larger problem: unless the shape of the posterior distribution is known, the methods we describe above are unable to diagnose whether a chain is exploring the entire posterior or is merely stuck in a local mode. As we remark in Chapter 1, this is especially true with phylogenetic trees as the posterior distribution typically contains many peaks and troughs (Whidden and Matsen IV, 2016).

In a recent attempt to address the lack of convergence diagnostics for phylogenetic trees, Whidden and Matsen IV (2015) equip the space of tree topologies with the *subtree-prune-and-regraft* metric and compute both the mean time to transition back and forth between the most frequently sampled topologies, and a topological Gelman–Rubin-type convergence statistic for multiple chains (Gelman and Rubin, 1992). An alternative approach, popular in phylogenetic inference, is the *Metropolis-coupled Markov chain Monte Carlo* algorithm (Geyer and Thompson, 1995). This technique uses multiple tempered chains which exchange states: the *hottest* chain targets the prior distribution — so it is essentially free to move over the entire sample space — and the *coldest* chain targets the posterior. We take two steps to address our mixing

issue with catastrophes. First of all, as we describe in Section 2.4.1, we enforce a catastrophe severity $\kappa \geq 0.25$ so that catastrophes are identifiable relative to the overall trait process. Secondly, to ensure that our inference is not based on chains which are unable to escape local modes, we repeat our analyses with an exact sampling algorithm which actively forces the chain to explore its sample space.

3.3 Wang–Landau algorithm

The Wang–Landau is an exact sampling method which penalises the chain for spending time in a partition of space, thereby forcing it to explore its entire state space. We would like to force our MCMC samplers to move between states with varying numbers of catastrophes. As we only add or remove one catastrophe at a time in our MCMC algorithm, the chain may easily get stuck in a local mode. Rather than have many bins with a single number of catastrophes in each, we shall take an indirect approach and apply the Wang–Landau algorithm to the catastrophe severity κ instead. We thereby force the chains to explore states with severe catastrophes, which in turn pushes them towards areas of the sample space with fewer catastrophes.

We first give a general overview of the Wang–Landau algorithm and its properties, then describe our implementation. We note that we only use the Wang–Landau algorithm to search for modes in the posterior (2.5.2) in order to validate our Metropolis–Hastings-based inferences — it is not our primary inference tool.

3.3.1 Description

We follow the description by Jacob and Ryder (2014). For a target density π, we create a disjoint partition $\mathcal{X}_1, \ldots, \mathcal{X}_d$ of its sample space $\mathcal{X}$. We define the function J such that $J(x) = i$ when $x \in \mathcal{X}_i$. Let $\theta(i)$ denote the penalty on bin $\mathcal{X}_i$, and $\boldsymbol{\theta} = [\theta(1), \ldots, \theta(d)]$. We define the penalised density $\pi_{\boldsymbol{\theta}}$, where

$$\pi_{\boldsymbol{\theta}}(x) \propto \frac{\pi(x)}{\theta[J(x)]}, \quad x \in \mathcal{X}. \tag{3.3.1}$$

The Wang–Landau algorithm targets this penalised density, and simultaneously estimates the unknown bin penalties $\boldsymbol{\theta}$ (Wang and Landau, 2001a,b).

Samples $X_1, X_2, \ldots, X_T$ produced by the Wang–Landau algorithm possess the following properties:

- Let $\mathrm{d}\mathcal{X}_i$ denote a subset of $\mathcal{X}_i$. Samples in bin $\mathcal{X}_i$ are distributed according to the target distribution π restricted to $\mathcal{X}_i$:

$$\frac{\sum_{t=1}^{T} \mathbf{1}_{\{X_t \in \mathrm{d}\mathcal{X}_i\}}}{\sum_{t=1}^{T} \mathbf{1}_{\{X_t \in \mathcal{X}_i\}}} \xrightarrow{T\to\infty} \int_{\mathrm{d}\mathcal{X}_i} \pi_{\boldsymbol{\theta}}(X|X \in \mathcal{X}_i)\mathrm{d}X = \int_{\mathrm{d}\mathcal{X}_i} \pi(X|X \in \mathcal{X}_i)\mathrm{d}X.$$

- The proportion of samples in bin $\mathcal{X}_i$ tends towards a value ϕ_i chosen by the user:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{X_t \in \mathcal{X}_i\}} \xrightarrow{T\to\infty} \int_{\mathcal{X}_i} \pi_{\boldsymbol{\theta}}(X)\mathrm{d}X = \phi_i.$$

Let $\psi_i = \int_{\mathcal{X}_i} \pi(X)\mathrm{d}X$, the probability mass in bin $\mathcal{X}_i$ under the target distribution. From the above properties, we deduce that

$$\pi_{\boldsymbol{\theta}}(x) = \frac{\pi(x)}{\theta[J(x)]}\left[\sum_{j=1}^{d} \frac{\psi_j}{\theta(j)}\right]^{-1} = \phi_{J(x)}\frac{\pi(x)}{\psi_{J(x)}}. \tag{3.3.2}$$

We return to this normalised version of the penalised density (3.3.1) later.

Let $\boldsymbol{\theta}_t$ denote the estimate of the penalties $\boldsymbol{\theta}$ at iteration t of the Wang–Landau algorithm, where $\theta_0(i) \leftarrow 0$ for each $i \in [d]$. To proceed, we sample X_t from the Metropolis–Hastings transition kernel (3.1.2) targeting $\pi_{\boldsymbol{\theta}_t}$ and then update the estimated penalty vector by

$$\log\theta_{t+1}(i) \leftarrow \log\theta_t(i) + \gamma_k(\mathbf{1}_{\{X_{t+1}\in\mathcal{X}_i\}} - \phi_i), \quad i \in [d], \tag{3.3.3}$$

where $(\gamma_k)_{k=1}^{\infty}$ is a decreasing sequence of real numbers such that $\sum_{k=1}^{\infty}\gamma_k = \infty$ and $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$. The index k may be updated deterministically as $k = t$, for example, or stochastically. We implement the following stochastic schedule. Let $\nu_t(i) = \nu_{t-1}(i) + \mathbf{1}_{\{X_i\in\mathcal{X}_i\}}$ track the number of samples in bin $\mathcal{X}_i$, with $\nu_0(i)$ initialised at 0 for each bin $i \in [d]$. Let $c \in [0,1]$ denote a parameter chosen in advance by the

Algorithm 2 The Wang–Landau algorithm with stochastic schedule

Partition the sample space $\mathcal{X}$ into d disjoint bins $\mathcal{X}_1, \ldots, \mathcal{X}_d$, select the target weights $\phi_1, \ldots, \phi_d$ and sequence $\gamma_1, \gamma_2, \ldots$
Initialise the penalties $\boldsymbol{\theta}_0 = (\theta_1, \ldots, \theta_d)$, state $X_0 \in \mathcal{X}$, index $k = 1$ and counters $\nu_0(i), \ldots, \nu_0(d)$
For t = 1 **to** T **do**
    Update the estimates of the penalties to $\boldsymbol{\theta}_t$ as $\log \theta_t(i) \leftarrow \log \theta_{t-1}(i) + \gamma_k(\mathbf{1}_{\{X_{t-1} \in \mathcal{X}_i\}} - \phi_i)$ for each bin $i \in [d]$
    Sample X_t from the Metropolis–Hastings kernel $K_{\boldsymbol{\theta}_t}(X_{t-1}, \cdot)$ (3.1.2) targeting the penalised density $\pi_{\boldsymbol{\theta}_t}$
    Update the count $\nu_t(i)$ for each bin $i \in [d]$ as $\nu_t(i) \leftarrow \nu_{t-1}(i) + \mathbf{1}_{\{X_t \in \mathcal{X}_i\}}$
    If the flat histogram criterion (3.3.4) is satisfied **then**
        Reset $\nu_t(i) \leftarrow 0$ for each bin $i \in [d]$
        Increment $k \leftarrow k + 1$
    End If
End For
Return samples $X_1, \ldots, X_T$ and penalties $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$

user. We say that the *flat histogram criterion* is met at iteration t if

$$\left| \frac{\nu_t(i)}{\sum_{j \in [d]} \nu_t(j)} - \phi_i \right| < c\phi_i, \quad \text{for all } i \in [d]. \tag{3.3.4}$$

When this occurs, we reset the bin counts $\nu_t(1), \ldots, \nu_t(d)$ to zero and increment $k \to k + 1$. We describe this algorithm in pseudo-code in Algorithm 2.

For samples from the Wang–Landau algorithm to be asymptotically distributed according to the target distribution π, the algorithm must satisfy the *diminishing adaptation* condition of Roberts and Rosenthal (2007). This condition states that as the number of iterations increases either the probability that successive transition kernels vary tends to 0 or the variation in successive transition kernels tends to 0. Jacob and Ryder (2014) demonstrate that the flat histogram criterion (3.3.4) is reached in a finite number of steps when we update the penalty estimates according to Equation 3.3.3, so $\gamma_k \to 0$ asymptotically and this Wang–Landau algorithm is exact in a Monte Carlo sense.

The stochastic schedule for updating the sequence $(\gamma_k)_{\kappa=0}^{\infty}$ is a marked improvement on its deterministic counterpart (Jacob and Ryder, 2014). If the sequence de-

creases too quickly, the algorithm may fail to explore the entire sample space and, in doing so, estimate the penalty terms $\boldsymbol{\theta}$ with high variance. On the other hand, if the sequence decreases too slowly, the estimates of the bin penalties will be accurate but the chain will converge slowly. The behaviour of the algorithm thus depends on our strategy (3.3.3) for estimating the penalties $\boldsymbol{\theta} \propto (\psi_1/\phi_1, \ldots, \psi_d/\phi_d)$, which in turn depend on our partition $\mathcal{X}_1, \ldots, \mathcal{X}_d$ of the sample space, the unknown density π, and the proportions of time $\phi_1, \ldots, \phi_d$ that we force the chain to spend in each bin. We return to these practical issues below.

We may estimate integrals with respect to the target density π through *importance sampling*. For a function h and samples $X_1, \ldots, X_T$ from the Wang–Landau algorithm,

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t)\frac{\pi(X_t)}{\pi_{\boldsymbol{\theta}_t}(X_t)} \xrightarrow{T\to\infty} \int_{\mathcal{X}} h(X)\pi(X)dX = \mathbb{E}_\pi[h(X)]. \tag{3.3.5}$$

We now describe how to evaluate this estimator in practice. From Equation 3.3.2, we know that

$$\frac{\pi(x)}{\pi_{\boldsymbol{\theta}}(x)} = \theta[J(x)]\sum_{j=1}^{d}\frac{\psi_j}{\theta(j)}. \tag{3.3.6}$$

We cannot substitute this identity into Equation 3.3.5 as we do not know the ψ_j terms in the summation. However, we note that

$$\begin{aligned}
\mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\frac{\pi(X)}{\pi_{\boldsymbol{\theta}}(X)}\right] &= \left[\sum_{i=1}^{d}\theta(i)\int_{\mathcal{X}_i}\pi_\theta(X)\mathrm{d}X\right]\left[\sum_{j=1}^{d}\frac{\psi_j}{\theta(j)}\right] \\
&= \left[\sum_{i=1}^{d}\theta(i)\phi_i\right]\left[\sum_{j=1}^{d}\frac{\psi_j}{\theta(j)}\right] \\
&= 1,
\end{aligned}$$

so we can replace $\sum_{j=1}^{d}[\psi_j/\theta(j)]$ in Equation 3.3.6 by $[\sum_{i=1}^{d}\theta(i)\phi_i]^{-1}$. Substituting this identity into Equation 3.3.5, we obtain the unbiased estimator

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t)\frac{\theta_t[J(X_t)]}{\sum_{i=1}^{d}\theta_t(i)\phi_i} \xrightarrow{T\to\infty} \mathbb{E}_\pi[h(X)]. \tag{3.3.7}$$

Observing that $T^{-1}\sum_{t=1}^{T}[\pi(X_t)/\pi_{\boldsymbol{\theta}}(X_t)] \to 1$ as the number of samples $T \to \infty$, Casella and Robert (1998) claim that the biased estimator

$$\frac{\sum_{t=1}^{T} h(X_t)\frac{\theta_t[J(X_t)]}{\sum_{i=1}^{d}\theta_t(i)\phi_i}}{\sum_{t=1}^{T}\frac{\theta_t[J(X_t)]}{\sum_{i=1}^{d}\theta_t(i)\phi_i}}$$

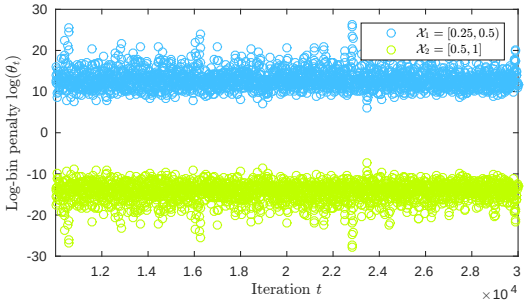
of $\mathbb{E}_\pi[h(X)]$ in Equation 3.3.5 may have lower variance than the estimator in Equation 3.3.7. We are concerned with performing unbiased inference so we do not consider this approach to be a viable alternative to the estimator in Equation 3.3.7.

3.3.2 Implementation

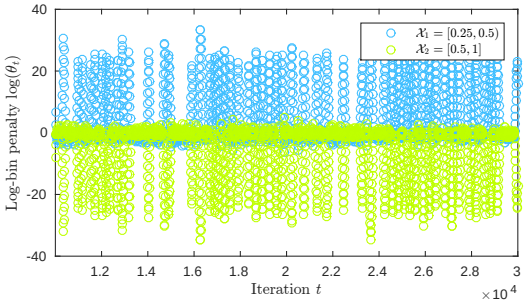
For our analyses, we bin the sample space of the catastrophe severity κ, $[0.25, 1]$ into two bins: $\mathcal{X}_1 = [0.25, 0.5)$ and $\mathcal{X}_2 = [0.5, 1]$. These bins are sufficient for our purposes because regardless of the value of κ in a given state, we can always transition from one bin to the other in a single step of our MCMC sampling algorithm. In addition, we set $\phi_1 = \phi_2 = 1/2$. We set $\gamma_k = k^{-2/3}$ on the basis of multiple trial runs. Atchadé and Liu (2010) recommend that the parameter c in the flat histogram criterion (3.3.4) is between 0.2 and 0.4. On the basis of further trial runs, we set $c = 0.3$.

The Wang–Landau algorithm is less efficient than the Metropolis–Hastings algorithm, assuming that the latter is mixing well, as we force the chain to spend more time in sub-optimal states. In Chapters 5 and 6, we use the Wang–Landau algorithm to verify that our inferences using the Metropolis–Hastings algorithm are not based on poorly mixing chains. We report trace plots of the estimated bin penalties $\boldsymbol{\theta}_t$ for each of these analyses in Figure 3.3. For the SDLT model, we are satisfied that our choice of bins and target proportions yields stable estimates of the bin penalties, and we do not attempt to further optimise this choice of parameters. The trace plots of the bin penalties are stable for the SD model fit to the data set without lateral transfer, `SIM-N`, and the data set with recently transferred traits removed, `SIM-T`. For the data

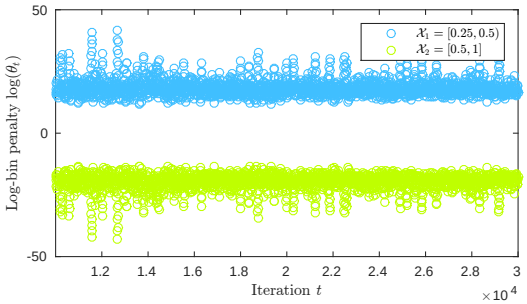
sets with lateral transfer, `SIM-B` and `POLY-0`, there are some instances where the chain takes many iterations to escape $\mathcal{X}_1$, with the net effect that $\theta_t(1)/[\theta_t(1) + \theta_t(2)] \to 1$ for these components of the importance sampling estimator in Equation 3.3.7. This is likely a result of the fact that we only update the catastrophe severity parameter κ in a small proportion of the moves, yet we update the estimated bin penalties $\boldsymbol{\theta}_t$ at every iteration.



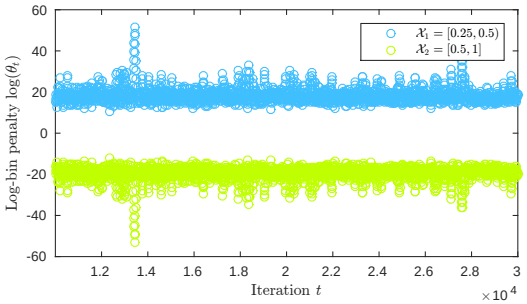
SDLT model on `SIM-B`.



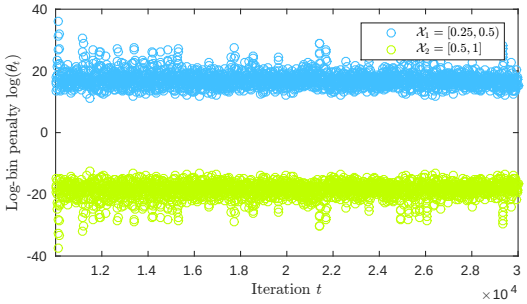
SD model on `SIM-B`.



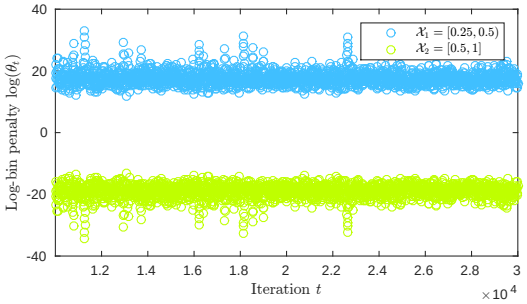
SDLT model on `SIM-N`.



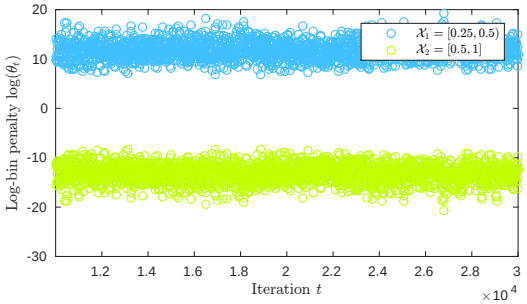
SD model on `SIM-N`.



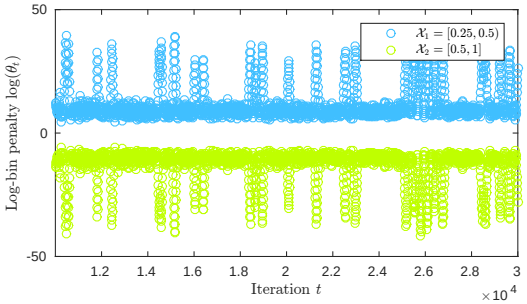
SDLT model on `SIM-T`.



SD model on `SIM-T`.



SDLT model on `POLY-0`.



SD model on `POLY-0`.

Figure 3.3: Bin penalty estimates in our Wang–Landau analyses of four data sets in Chapters 5 and 6.

Chapter 4

Exact-approximate MCMC inference

Chapter overview

We introduce the Stochastic Dollo with Lateral Transfer model in Chapter 2 and describe how to perform exact inference under the model in Chapter 3. The computational cost of calculating the expected pattern frequencies $\mathbf{x}(t) = (x_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ in Equation 2.3.5 grows exponentially with the number of leaves L. We now describe a scheme to accelerate our inference procedure. This chapter has three parts. We first discuss properties of the initial value problems (2.3.5) and the computational cost of this calculation. We then describe how to exploit symmetry in the pattern-based initial value problems to construct an efficient approximation scheme. We analyse the error in our approximation and show that it is negligible in practice. Finally, we construct an unbiased estimator of the likelihood and use it to perform exact-approximate MCMC inference.

4.1 Expected pattern frequencies calculation revisited

4.1.1 Computational cost

Although simple to state, the expected pattern frequency calculation (2.3.5) is a challenging computational problem for a number of reasons. A trait may display any valid binary pattern across the extant lineages during an interval and the transition rates between pattern states, in Figure 2.4, for example, depend on the Hamming weight of the current pattern state. Therefore, we must integrate over all possible configurations of trait presence and absence on the tree regardless of the observed site-patterns or the initial conditions at the start of intervals. We cannot decouple this integration across individual branches or subtrees like in a vertical model. Finally, the systems of differential equations describing the expected pattern evolution across an interval grow exponentially with the number of branches under consideration.

For the time being, we focus on a single time interval $[t, t+\Delta]$ with $L^{(t)}$ branches. The solution of Equation 2.3.5 across the interval is

$$\mathbf{x}(t+\Delta) = e^{\mathbf{A}^{(t)}\Delta}\mathbf{x}(t) + \left(\mathbf{I} - e^{\mathbf{A}^{(t)}\Delta}\right)\left(-\mathbf{A}^{(t)^{-1}}\mathbf{b}^{(t)}\right), \tag{4.1.1}$$

where $\mathbf{I}$ denotes the corresponding identity matrix. The matrix $\mathbf{A}^{(t)}$ is the generator matrix for a continuous time Markov chain with states $\mathcal{P}^{(t)} \cup \{\mathbf{0}\}$, transitions between patterns which differ by a single entry, and $\mathbf{0}$ is an absorbing state. The $(\mathbf{p}, \mathbf{q})$ entry of $\exp(\mathbf{A}^{(t)}\Delta)$ is the probability that a trait displaying pattern $\mathbf{q}$ at time t displays pattern $\mathbf{p}$ at time $t+\Delta$. The vector $\mathbf{x}(t+\Delta)$ is a weighted average of the initial condition $\mathbf{x}(t)$ and steady-state solution $-\mathbf{A}^{(t)^{-1}}\mathbf{b}^{(t)}$, which respectively account for the expected number of traits present at the start of the interval which survive to the end, and the expected number of traits born during the interval which survive until the end. In fact, we could obtain Equation 4.1.1 as $\mathbf{x}(t+\Delta) = \mathbf{y}(\Delta) + \mathbf{z}(\Delta)$, where $\mathbf{y}(\Delta)$ and $\mathbf{z}(\Delta)$ are the solutions of the following homogeneous and inhomogeneous initial value

problems: for $u \in [0, \Delta)$, solve

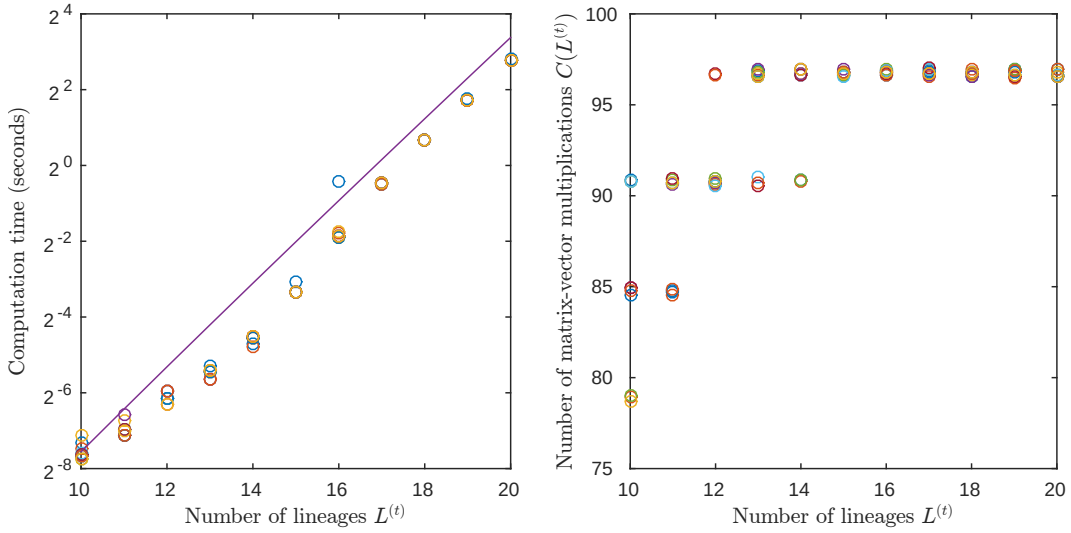
$$\begin{aligned} \dot{\mathbf{y}}(u) &= \mathbf{A}^{(t)}\mathbf{y}(u) \\ \dot{\mathbf{z}}(u) &= \mathbf{A}^{(t)}\mathbf{z}(u) + \mathbf{b}^{(t)} \end{aligned} \quad \text{where} \quad \begin{aligned} \mathbf{y}(0) &= \mathbf{x}(t), \\ \mathbf{z}(0) &= \mathbf{0}. \end{aligned} \tag{4.1.2}$$

We shall return to this representation at various points in this chapter.

Although the matrix $\mathbf{A}^{(t)}$ is sparse with only $L^{(t)}$ or $L^{(t)} + 1$ non-zero entries in each of its $2^{L^{(t)}} - 1$ rows, its exponent is dense as a trait may display any pattern in $\mathcal{P}^{(t)}$ on an interval provided there are no offset leaves, in which case the set of possible patterns is restricted. Consequently, so we cannot compute $\exp(\mathbf{A}^{(t)}\Delta)$ explicitly, let alone store it in memory for anything other than small $L^{(t)}$. For example, when $L^{(t)} = 16$, the matrix $\mathbf{A}^{(t)}$ in sparse format and double precision occupies roughly 18 megabytes of memory, whereas $\exp(\mathbf{A}^{(t)})$, on the other hand, requires approximately 34 gigabytes of memory. We instead turn to numerical differential equation solvers which propagate a solution vector across the interval.

The initial value problems in Equation 2.3.5 are linear and non-stiff. To estimate the solution in Equation 4.1.1 with an ordinary differential equation (ODE) solver requires $\mathcal{O}[L^{(t)}2^{L^{(t)}}C(L^{(t)})]$ operations, where $\mathcal{O}(L^{(t)}2^{L^{(t)}})$ is the cost of one matrix-vector multiplication $\mathbf{A}^{(t)}\mathbf{x}(t)$, and the solver cost $C(L^{(t)})$ accounts for the number of steps taken in order to maintain the solver error below a set tolerance. The `Matlab` solver `ODE45` is a fourth-order Runge–Kutta method. At each time step, `ODE45` computes fourth- and fifth-order estimates of the solution at the next time step. If the absolute and relative errors between the estimates are below their respective tolerances, typically 10^{-6} for the absolute error and 10^{-3} for the relative error, the method returns the fourth-order estimate of the solution. The solver uses the magnitude of these errors is used to adapt the step size.

We say that $\theta = \Delta \max(\mu, \beta)$ is the *effective interval length*. In Figure 4.1, we plot the time and number of matrix-vector multiplications taken by `ODE45` with its default settings to solve the initial value problems in Equation 2.3.5 for an effective interval



(a) Computation time.

(b) Cost function $C(L^{(t)})$.

Figure 4.1: Computational cost of solving the initial value problem in Equation 2.3.5 with the `Matlab` `ODE45` solver for effective interval length $\theta = 10^{-2}$ and various initial conditions. (a) The reference line increases by a factor $2(L^{(t)}+1)/L^{(t)}$ from $L^{(t)}-1$ lineages to $L^{(t)}$. (b) We jitter the vertical coordinates to improve visibility.

length $\theta = 10^{-2}$. In each case, we draw the initial expected pattern frequencies from an Exponential distribution with mean 10. We see that the ODE solver takes more steps as the number of lineages $L^{(t)}$ increases. When there are $L = 20$ leaves, the likelihood calculation on a catastrophe-free tree with isochronous leaves takes approximately four seconds on current hardware. As a result, our exact MCMC approach in Chapter 3 quickly becomes infeasible for $L > 20$ leaves.

If we repeat the analyses in Figure 4.1 with the initial and maximum step sizes that `ODE45` may take increased to $\theta = 10^{-2}$, for example, then the number of steps taken by the solver drops dramatically. However, if we compare the resulting expected pattern frequencies with those returned by `ODE45` with its more conservative default settings, then we observe absolute errors above 10^{-6}. To elaborate on a point made above, the reason for this behaviour is the fact that the absolute and relative errors which the solver uses are between its fourth- and fifth-order estimates of the solution, not the solution itself.

4.1.2 Pattern process symmetry

The solution of Equation 2.3.5 in Equation 4.1.1 is fully determined by the initial condition $\mathbf{x}(t)$, the matrix $\mathbf{A}^{(t)}\Delta$ of expected flow between patterns on the interval, and the vector $\mathbf{b}^{(t)}\Delta$ of expected pattern births. Returning to our interpretation of Equation 4.1.1, let

$$y_{\mathbf{p}}^{\mathbf{q}}(\Delta) = [\exp(\mathbf{A}^{(t)}\Delta)]_{\mathbf{p},\mathbf{q}}, \quad \mathbf{p}, \mathbf{q} \in \mathcal{P}^{(t)}$$

the probability that a trait which displays pattern $\mathbf{q}$ at time t displays pattern $\mathbf{p}$ at time $t + \Delta$. We could, for example, obtain these terms as solutions to the homogeneous initial value problem in Equation 4.1.2 for various initial conditions. For $u \in [0, \Delta)$, we first define the vector $\mathbf{y}^{\mathbf{q}}(u) = (y_{\mathbf{p}}^{\mathbf{q}}(u) : \mathbf{p} \in \mathcal{P}^{(t)})$ for each pattern $\mathbf{q} \in \mathcal{P}^{(t)}$, where $y_{\mathbf{p}}^{\mathbf{q}}(0) = \mathbf{1}_{\{\mathbf{p}=\mathbf{q}\}}$, then define the $|\mathcal{P}^{(t)}| = 2^{L^{(t)}} - 1$ homogeneous initial value problems: for $u \in [0, \Delta)$, solve

$$\dot{\mathbf{y}}^{\mathbf{q}}(u) = \mathbf{A}^{(t)}\mathbf{y}^{\mathbf{q}}(u) \quad \text{where} \quad \mathbf{y}^{\mathbf{q}}(0) = (y_{\mathbf{p}}^{\mathbf{q}}(0) : \mathbf{p} \in \mathcal{P}^{(t)}) \quad \text{for each} \quad \mathbf{q} \in \mathcal{P}^{(t)}. \tag{4.1.3}$$

Each initial value problem in Equation 4.1.3 starts with a unit impulse at a single pattern. We can then rewrite the expected pattern frequencies in Equation 4.1.1 as the linear combination

$$x_{\mathbf{p}}(t + \Delta) = \sum_{\mathbf{q} \in \mathcal{P}^{(t)}} y_{\mathbf{p}}^{\mathbf{q}}(\Delta) x_{\mathbf{q}}(t) + z_{\mathbf{p}}(\Delta), \quad \mathbf{p} \in \mathcal{P}^{(t)}, \tag{4.1.4}$$

where $\mathbf{z}(\Delta) = (z_{\mathbf{p}}(\Delta) : \mathbf{p} \in \mathcal{P}^{(t)})$ solves the inhomogeneous initial value problem in Equation 4.1.2. As with Equation 4.1.2, this example in Equation 4.1.4 is purely pedagogical as we would never solve $2^{L^{(t)}}$ separate systems of differential equations when, in practice, we can solve a single initial value problem of the same dimension instead. The point here is illustrate how we may write the expected pattern frequencies in Equation 4.1.1 as a linear combination of the solutions to Equation 4.1.3 which only depend on the length of the interval, the initial condition $\mathbf{x}(t)$ and $\mathbf{z}(\Delta)$, the expected increase in pattern frequencies from traits born during the interval. Many of the terms

in the solution to the initial value problems in Equation 4.1.3 are repeated. We shall exploit this repetition to instead evaluate the equivalence class expected frequencies as the solutions to a linear number of initial value problems whose dimensions are polynomial functions of $L^{(t)}$.

Returning to the $L^{(t)}$-cube representation of the pattern process in Section 2.3.1 on the interval $[t, t+\Delta]$, the cube structure and the transition rates between vertices, in Figure 2.5, for example, do not change across an interval. Furthermore, the branch ordering $\mathbf{k}^{(t)}$ which determines the displayed patterns on the interval is arbitrary; the only requirement is that the pattern frequency transfer operations $\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots$ are consistent across branching events.

Recall from Section 2.3.1 that for a pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, its Hamming $s(\mathbf{p})$ is the number of non-zero entries in $\mathbf{p}$, and its Hamming distance $d(\mathbf{p}, \mathbf{q})$ from pattern $\mathbf{q} \in \mathcal{P}^{(t)}$ is the number of entries where they differ. In Figure 4.2a, we colour the vertices of the cube in Figure 2.5 by their respective Hamming weights and Hamming distances from the *initial pattern* $\boldsymbol{\sigma} = (0, 0, 1)$. We note that

- The patterns in each colour group are invariant under permutation of the index sets $\{i \in [L^{(t)}] : \sigma_i = 0\}$ and $\{i \in [L^{(t)}] : \sigma_i = 1\}$
- All paths on the graph which start at $\boldsymbol{\sigma}$ and yield the same sequence of vertex colours are equally likely as the corresponding sequences of transition rates along edges are identical.

These results become clear when we plot patterns by their Hamming weight and Hamming distance from $\boldsymbol{\sigma}$ in Figure 4.2b. We repeat this exercise for the initial patterns $\boldsymbol{\sigma} = (0, 1, 1)$ and $\boldsymbol{\sigma}(1, 1, 1)$ in Figures 4.2c and 4.2d respectively and observe the same behaviour. Furthermore, to obtain the corresponding plot for pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, we permute the patterns in the plot for the initial pattern $\boldsymbol{\sigma}$ with $s(\boldsymbol{\sigma}) = s(\mathbf{p})$ until $\mathbf{p}$ is the initial pattern. This leads us to the important result in Theorem 2.

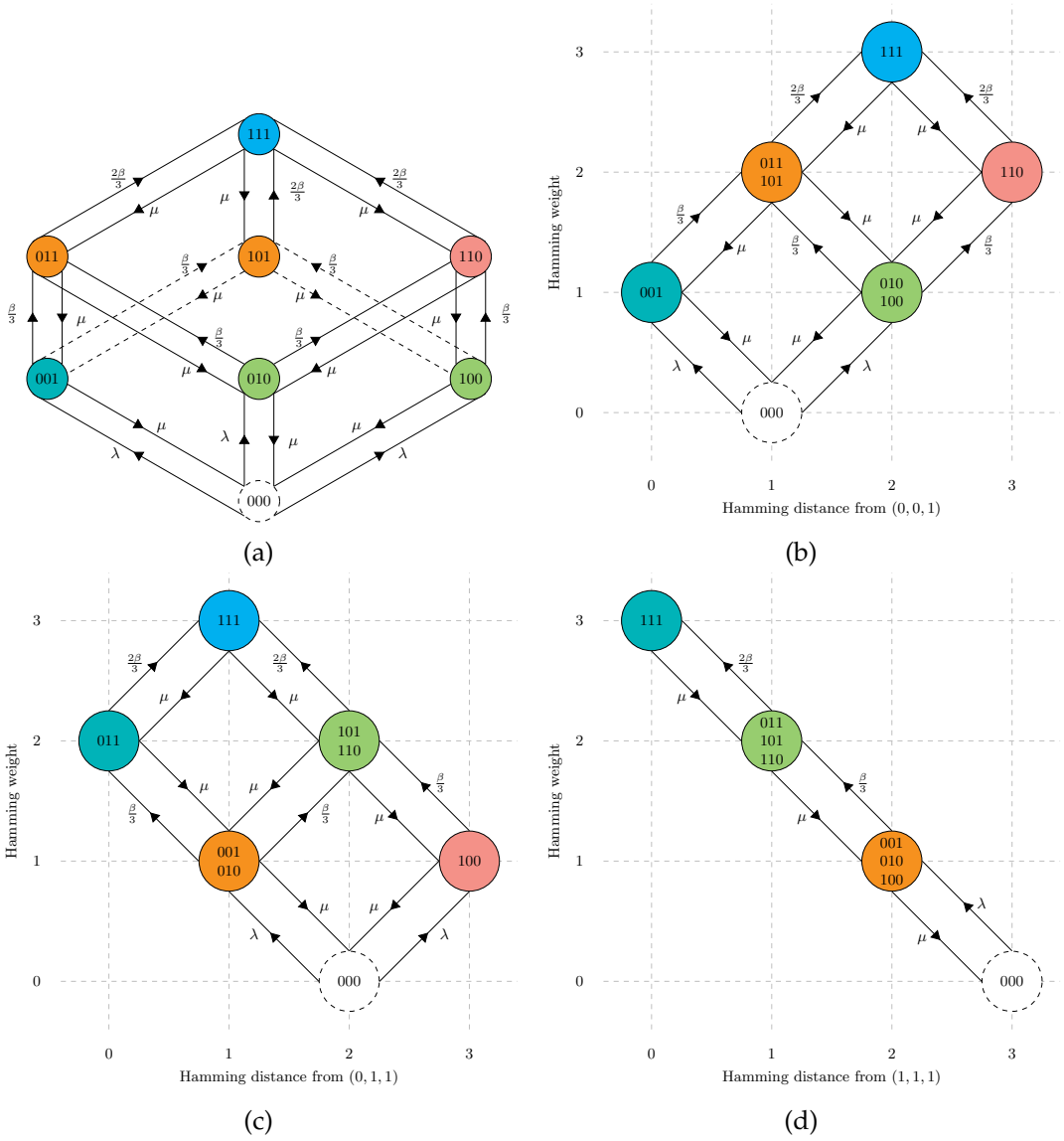


Figure 4.2: Symmetry among patterns when $L^{(t)} = 3$. (a) A trait displaying the pattern $\sigma = (0, 0, 1)$ at the start of the interval is equally like to display patterns of the same colour at the end of the interval. (b) We group these patterns by Hamming weight and distance from $(0, 0, 1)$ into equivalence classes. We repeat this grouping operation when the initial pattern σ is (c) $(0, 1, 1)$ and (d) $(1, 1, 1)$.

Theorem 2 (Pattern symmetry). *A trait which displays pattern $\boldsymbol{\sigma} \in \mathcal{P}^{(t)}$ at the start of an interval is equally likely to display any of the patterns in the set $\{\mathbf{p} \in \mathcal{P}^{(t)} : s(\mathbf{p}) = s,\, d(\mathbf{p}, \boldsymbol{\sigma}) = d\}$ at the end of the interval.*

As an example of the result in Theorem 2, suppose a trait $h \in \mathcal{H}^{(t)}$ currently displays pattern $\mathbf{p}^h(t) = \boldsymbol{\sigma} = (0,1,1)$. By our symmetry argument, it must be the case that for any interval length $\Delta \geq 0$,

$$\mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (0,0,1) \middle| \mathbf{p}^h(t) = (0,1,1)\right] = \mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (0,1,0) \middle| \mathbf{p}^h(t) = (0,1,1)\right],$$
$$\mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (1,0,1) \middle| \mathbf{p}^h(t) = (0,1,1)\right] = \mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (1,1,0) \middle| \mathbf{p}^h(t) = (0,1,1)\right],$$

as we can permute the second and third entries of each pattern without changing the result. Following this, if $\mathbf{p}^h(t) = (1,1,0)$ then we simply permute the entries in the above equation to obtain

$$\mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (1,0,1) \middle| \mathbf{p}^h(t) = (1,1,0)\right] = \mathbb{P}\left[\mathbf{p}^h(t+\Delta) = (0,1,1) \middle| \mathbf{p}^h(t) = (1,1,0)\right],$$

and so on.

4.1.3 Equivalence class process

We now describe how to exploit the symmetry result in the previous section. We can collapse the pattern-based initial value problems in Equation 2.3.5, which grow exponentially with $L^{(t)}$ in dimension, into multiple smaller problems of polynomial dimension and compute the solutions to the initial value problems in Equation 4.1.3. To do this, we collapse the initial value problems along the same lines as Figures 4.2b–d. More precisely, we solve $L^{(t)}$ homogeneous systems: one for each possible pattern Hamming weight at the start of an interval; and one inhomogeneous system for the contributions of traits born during the interval. Similar to Equation 4.1.4, the expected pattern frequencies at the end of the interval (4.1.1) are given by a linear combination of the initial expected pattern frequencies and the solutions to the homogeneous initial

value problems and we simply add on the contribution of traits born in the interval given by the solution of the inhomogeneous system.

For an initial pattern $\boldsymbol{\sigma} \in \mathcal{P}^{(t)}$ of Hamming weight $s(\sigma) = c$, we define the index set

$$S^c = \left\{(s,d) \in \mathbb{N}_0^2 : s + d = c + 2k,\, 0 \leq k \leq L^{(t)} - c\right\},$$

of valid equivalence class Hamming weights and Hamming distances relative to $\boldsymbol{\sigma}$. We define the set $\mathcal{C}^{\sigma} = \{C^{\sigma}_{s,d} : (s,d) \in S^c\}$ of *equivalence classes* with entry

$$C^{\sigma}_{s,d} = \{\mathbf{p} \in \mathcal{P}^{(t)} : s(\mathbf{p}) = s,\, d(\boldsymbol{\sigma}, \mathbf{p}) = d\},$$

the set of patterns in $\mathcal{P}^{(t)}$ of Hamming weight s and Hamming distance d from $\boldsymbol{\sigma}$, the arbitrary initial pattern. The $m(c) = (c+1)(L^{(t)} - c + 1)$ valid pairs in S^c do not include every point in the convex hull of $\{(0,c),\, (c,0),\, (L^{(t)} - c, L^{(t)}),\, (L^{(t)}, L^{(t)} - c)\}$ as we cannot change the Hamming weight of a pattern without changing its distance from $\boldsymbol{\sigma}$, and vice versa. This construction is evident in Figures 4.2b–d, and we illustrate S^c in general in Figure 4.3.

In our description of Figure 4.2 in the previous section, we state that we may permute the equivalence classes in the set $\mathcal{C}^{\sigma}$ to form the corresponding set for any pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ such that $s(\mathbf{p}) = s(\boldsymbol{\sigma})$. The pattern $\boldsymbol{\sigma}$ is then arbitrary, so we reparameterise the set $\mathcal{C}^{\sigma}$ of equivalence classes as $\mathcal{C}^{s(\sigma)}$. For example, when $L^{(t)} = 4$, we may choose

$$\boldsymbol{\sigma} = \begin{cases} (0,0,0,1), & c = 1, \\ (0,0,1,1), & c = 2, \\ (0,1,1,1), & c = 3, \\ (1,1,1,1), & c = 4, \end{cases}$$

to act as the initial patterns in the construction of the equivalence class sets $\mathcal{C}^1, \ldots, \mathcal{C}^4$.

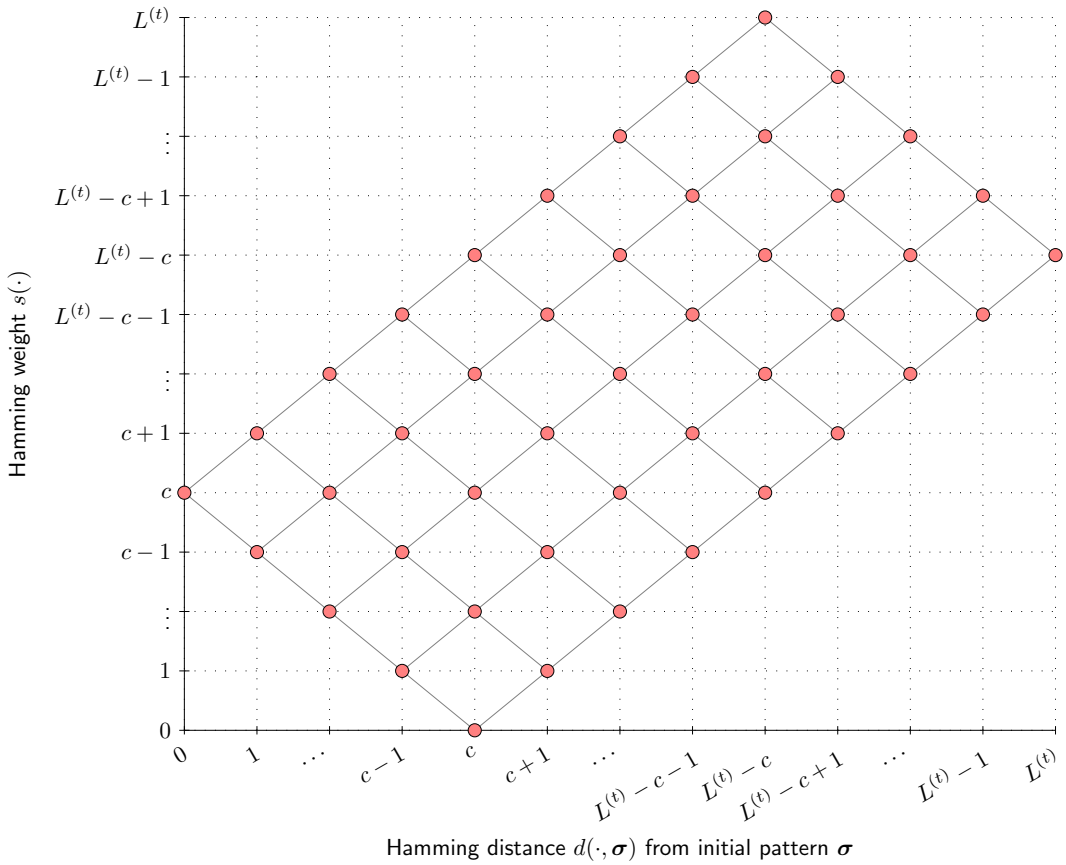


Figure 4.3: The index set $S^{s(\sigma)}$ of valid coordinate pairs (s,d) for entries in the set $\mathcal{C}^{\sigma}$ of equivalence classes.

For a given pattern, we require the number of patterns in neighbouring equivalence classes that it communicates with. A pattern $\mathbf{p} \in C^{c}_{s,d}$ communicates with:

$$
\begin{aligned}
m^{s-1,d-1}_{s,d} &= |\{i \in [L^{(t)}] : \sigma_i = 0, p_i = 1\}| = \frac{+d+s-c}{2} \text{ patterns in } C^{c}_{s-1,d-1},\\
m^{s-1,d+1}_{s,d} &= |\{i \in [L^{(t)}] : \sigma_i = 1, p_i = 1\}| = \frac{-d+s+c}{2} \text{ patterns in } C^{c}_{s-1,d+1},\\
m^{s+1,d-1}_{s,d} &= |\{i \in [L^{(t)}] : \sigma_i = 1, p_i = 0\}| = \frac{+d-s+c}{2} \text{ patterns in } C^{c}_{s+1,d-1},\\
m^{s+1,d+1}_{s,d} &= |\{i \in [L^{(t)}] : \sigma_i = 0, p_i = 0\}| = L^{(t)} - \frac{+d+s+c}{2} \text{ patterns in } C^{c}_{s+1,d+1},
\end{aligned}
$$

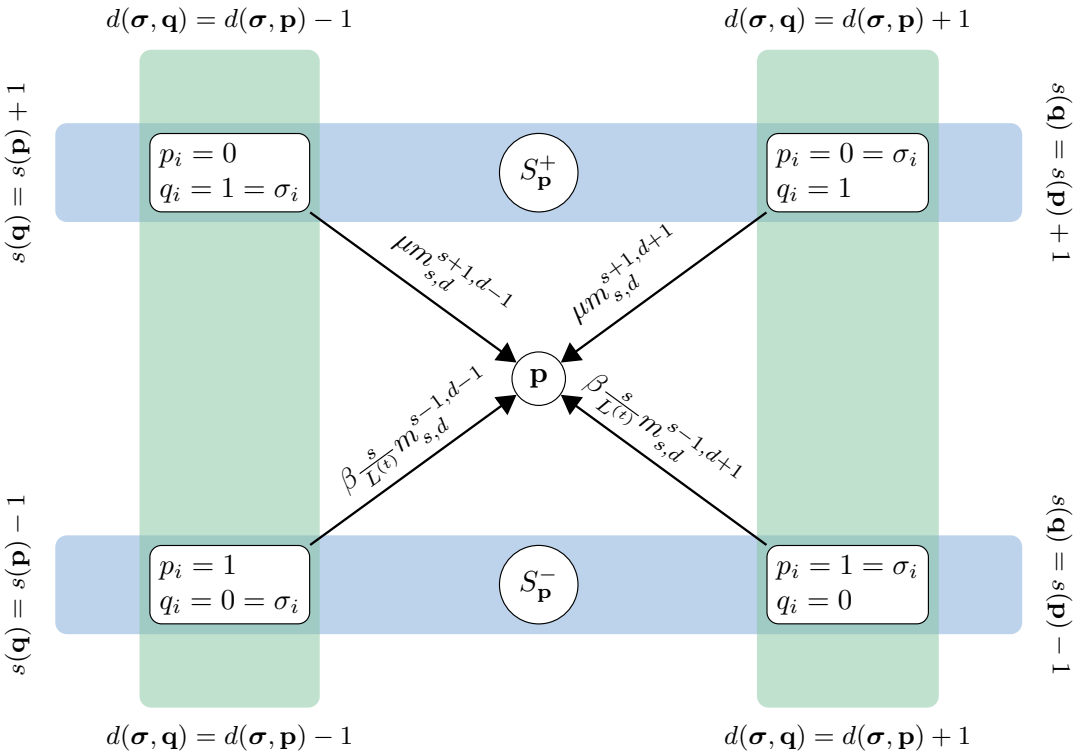


Figure 4.4: Structure of patterns in equivalence classes $C^{s(\sigma)}_{s\pm1,d\pm1}$ neighbouring pattern $\mathbf{p} \in C^{s(\sigma)}_{s,d}$.

where the multipliers $m^{s-1,d-1}_{s,d}$, $m^{s-1,d+1}_{s,d}$, $m^{s+1,d-1}_{s,d}$ and $m^{s+1,d+1}_{s,d}$ satisfy the following system of equations:

$$
\begin{aligned}
m^{s-1,d-1}_{s,d} &+ m^{s-1,d+1}_{s,d} &+ m^{s+1,d-1}_{s,d} &+ m^{s+1,d+1}_{s,d} &= L^{(t)} &= |S^-_{\mathbf{p}} \cup S^+_{\mathbf{p}}| = L^{(t)}, \\
m^{s-1,d-1}_{s,d} &+ m^{s-1,d+1}_{s,d} & & & = s &= |\{i \in [L^{(t)}] : p_i = 1\}|, \\
m^{s-1,d-1}_{s,d} & & + m^{s+1,d-1}_{s,d} & & = d &= |\{i \in [L^{(t)}] : p_i \neq \sigma_i\}|, \\
& m^{s-1,d+1}_{s,d} & + m^{s+1,d-1}_{s,d} & & = c &= |\{i \in [L^{(t)}] : \sigma_i = 1\}|.
\end{aligned}
$$

For example, there are $m^{s-1,d-1}_{s,d}$ patterns in $C^c_{s-1,d-1}$ which we can form from pattern $\mathbf{p}$ by taking one index $j \in \{i \in [L^{(t)}] : \sigma_i = 0, p_i = 1\}$ and setting $p_j \leftarrow 0$. Figure 4.4 illustrates this point.

We are now ready to redefine the pattern-based initial value problem in Equation 2.3.5 on the sets of equivalence classes $\mathcal{C}^1, \ldots, \mathcal{C}^{L^{(t)}}$.

4.1.4 Equivalence class frequencies

Similar to the terms in Equation (4.1.4), let $y^c_{s,d}(\Delta)$ denote the probability that a pattern in class $(c,0)$ at time t is in class (s,d) at time $t+\Delta$. Figure 4.4 describes the rates that traits displaying patterns in neighbouring classes transition to display a given pattern in $C^c_{s,d}$. Following a similar argument to the pattern-based expected pattern frequency calculation in Equations 2.3.1 and 2.3.2, this probability evolves across the interval $[t, t+\Delta]$ according to the following differential equation:

$$\begin{aligned}\dot{y}^c_{s,d}(u) = &-s\left[\mu+\beta\left(1-\frac{s}{L^{(t)}}\right)\right]y^c_{s,d}(u)\\ &+\mu\left[m^{s+1,d-1}_{s,d}y^c_{s+1,d-1}(u)+m^{s+1,d+1}_{s,d}y^c_{s+1,d+1}(u)\right]\\ &+\beta\frac{s}{L^{(t)}}\left[m^{s-1,d-1}_{s,d}y^c_{s-1,d-1}(u)+m^{s-1,d+1}_{s,d}y^c_{s-1,d+1}(u)\right],\end{aligned} \tag{4.1.5}$$

where $u \in [0,\Delta)$ and the initial condition is

$$y^c_{s,d}(0) = \begin{cases} 1, & s=c,\ d=0,\\ 0, & \text{otherwise.}\end{cases} \tag{4.1.6}$$

We may write the coupled system of equations (4.1.5) across equivalence classes as $\dot{\mathbf{y}}^c(u) = \mathbf{A}^c\mathbf{y}(u)$, where the $m(c)\times m(c)$ matrix $\mathbf{A}^c$ has entry

$$a^c_{(s,d),(s',d')} = \begin{cases} -s\left[\mu+\beta\left(1-\frac{s}{L^{(t)}}\right)\right], & (s',d')=(s,d),\\ \beta\frac{s}{L^{(t)}}m^{s',d'}_{s,d}, & (s',d')=(s-1,d\pm1),\\ \mu m^{s',d'}_{s,d}, & (s',d')=(s+1,d\pm1),\\ 0, & \text{otherwise,}\end{cases} \qquad (s,d),(s',d')\in S^c,$$

and the $m(c)$-vector of initial conditions $\mathbf{y}^c(0)$ is defined in Equation 4.1.6. We do this for each possible initial pattern Hamming weight $c = 1,\ldots,L^{(t)}$.

A constructive way to form the $m(c)\times m(c)$ rate matrix $\mathbf{A}^c$ across equivalence classes is to take the $(2^{L^{(t)}}-1)\times(2^{L^{(t)}}-1)$ rate matrix $\mathbf{A}^{(t)}$ across patterns in Equation 2.3.5, replace each column by the sum of the columns of patterns in the same equivalence class then delete the duplicate rows and columns.

Each trait born during the interval $[t, t+\Delta]$ starts in state **0**, so to speak. In Equation 4.1.4, the expected contribution of these traits to the expected pattern frequencies at the end of the interval is

$$\mathbf{z}(\Delta) = (z_{\mathbf{p}}(\Delta) : \mathbf{p} \in \mathcal{P}^{(t)}) = (\mathbf{I} - e^{\mathbf{A}^{(t)}\Delta})(\mathbf{A}^{(t)^{-1}}\mathbf{b}^{(t)}),$$

the solution of Equation 4.1.1 when the initial condition $\mathbf{x}(t)$ is a vector of zeros. By our symmetry argument, there are only $L^{(t)}$ possible equilibrium expected pattern frequencies: one for each possible Hamming weight. As a result, we need only solve one inhomogeneous system of equations across equivalence classes and, in doing so, we choose the initial pattern $\sigma = (1, 1, \ldots, 1)$ which yields the smallest system of differential equations. Equivalently, we could set $\sigma = \mathbf{0}$; but if we choose any other initial pattern σ then we end up with repetition in the corresponding equivalence class solutions. Let $\mathbf{b}^{L^{(t)}}$ denote the vector of birth rates collapsed into equivalence classes $\mathcal{C}^{L^{(t)}}$, and define the $(L^{(t)}+1)$-vector $\mathbf{z}(u) = (z_{s,d}(u) : (s,d) \in S^{L^{(t)}})$ where $u \in [0, \Delta)$ and $\mathbf{z}(0) = \mathbf{0}$; that is, $\mathbf{z}_{s,d}(0) = 0$ for $(s,d) \in S^{L^{(t)}}$. The entries of $S^{L^{(t)}}$ lie along the diagonal from $(s,d) = (L^{(t)}, 0)$ to $(0, L^{(t)})$ so only the Hamming weight is of interest here and we drop the Hamming distance index from our notation.

The complete set of initial value problems across equivalence classes $\mathcal{C}^1, \ldots, \mathcal{C}^{L^{(t)}}$ for the time interval $[t, t+\Delta]$ is as follows: for $u \in [0, \Delta)$, solve

$$\begin{aligned}
\dot{\mathbf{y}}^1(u) &= \mathbf{A}^1\mathbf{y}^1(u) & \mathbf{y}^1(0) &= \left(\mathbf{1}_{\{(s,d)=(1,0)\}} : (s,d) \in S^1\right), \\
&\vdots & &\vdots \\
\dot{\mathbf{y}}^c(u) &= \mathbf{A}^c\mathbf{y}^c(u) & \mathbf{y}^c(0) &= \left(\mathbf{1}_{\{(s,d)=(c,0)\}} : (s,d) \in S^c\right), \\
&\vdots \quad \text{where} & &\vdots \\
\dot{\mathbf{y}}^{L^{(t)}}(u) &= \mathbf{A}^{L^{(t)}}\mathbf{y}^{L^{(t)}}(u) & \mathbf{y}^{L^{(t)}}(0) &= \left(\mathbf{1}_{\{(s,d)=(L^{(t)},0)\}} : (s,d) \in S^{L^{(t)}}\right), \\
\dot{\mathbf{z}}(u) &= \mathbf{A}^{L^{(t)}}\mathbf{z}(u) + \mathbf{b}^{L^{(t)}} & \mathbf{z}(0) &= \mathbf{0}.
\end{aligned} \tag{4.1.7}$$

In contrast to the exponential in $L^{(t)}$ dimension of Equation 2.3.5, each of these initial value problems has a polynomial in $L^{(t)}$ number of terms. Equation 4.1.7 is the equivalence class representation of the initial value problem defined on patterns in Equation 2.3.5 on the interval $[t, t+\Delta]$.

Similar to Equation 4.1.4, we may write the expected frequency of pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time $t+\Delta$ in terms of the equivalence class solutions of Equation 4.1.7 as

$$x_{\mathbf{p}}(t+\Delta) = \sum_{\mathbf{q} \in \mathcal{P}^{(t)}} y^{s(\mathbf{q})}_{s(\mathbf{p}),d(\mathbf{p},\mathbf{q})}(\Delta) x_{\mathbf{q}}(t) + z_{s(\mathbf{p})}(\Delta). \tag{4.1.8}$$

The correspondence between the terms in Equations 4.1.1 and 4.1.8 is

$$\begin{aligned} \left[e^{\mathbf{A}^{(t)}\Delta}\right]_{\mathbf{p},\mathbf{q}} &= y^{s(\mathbf{q})}_{s(\mathbf{p}),d(\mathbf{p},\mathbf{q})}(\Delta), \qquad \mathbf{p},\mathbf{q} \in \mathcal{P}^{(t)}, \\ \left[\left(\mathbf{I} - e^{\mathbf{A}^{(t)}\Delta}\right)\left(-{\mathbf{A}^{(t)}}^{-1}\mathbf{b}^{(t)}\right)\right]_{\mathbf{p}} &= z_{s(\mathbf{p})}(\Delta), \qquad \mathbf{p} \in \mathcal{P}^{(t)}, \end{aligned}$$

which exploits the fact that many entries in $e^{\mathbf{A}^{(t)}\Delta}$ and $(\mathbf{I} - e^{\mathbf{A}^{(t)}\Delta})(-{\mathbf{A}^{(t)}}^{-1}\mathbf{b}^{(t)})$ are equal by the symmetry between patterns, so in contrast to Equation 4.1.4, we only compute the distinct entries.

The equivalence class approach in Equation 4.1.8 does not lead to a reduction in the computational cost of evaluating the expected pattern frequencies compared to the ODE solver approach on the pattern-based system (2.3.5). In fact, it is the opposite. The $\mathcal{O}({L^{(t)}}^2)$ number of operations to solve the initial value problems across equivalence classes (4.1.7) is trivial. However, we must sum over $2^{L^{(t)}}$ equivalence class contributions for each of the $2^{L^{(t)}} - 1$ patterns in $\mathcal{P}^{(t)}$ to form $\mathbf{x}(t+\Delta)$. Although this $\mathcal{O}(2^{2L^{(t)}})$ *unpacking* operation (4.1.8) is lower than the $\mathcal{O}(2^{3L^{(t)}})$ cost of directly forming the solution in Equation 4.1.1, it is still significantly higher than the $\mathcal{O}[L^{(t)}2^{L^{(t)}}C(L^{(t)})]$ cost of the ODE solver approach in Figure 4.1.

4.2 Approximate expected pattern frequencies

We now describe how to accurately approximate the exact expected pattern frequencies (4.1.1) using the theory developed in the previous section. Our equivalence class-based approximation is efficient in the sense that we approach the level of accuracy of the exact pattern-based approach (2.3.5) in significantly fewer operations.

4.2.1 Construction

For pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, we define $S_{\mathbf{p}}^d = \{\mathbf{q} \in \mathcal{P}^{(t)} : d(\mathbf{p}, \mathbf{q}) = d\}$, the set of patterns at Hamming distance d from $\mathbf{p}$. We can rewrite Equation 4.1.8 as

$$x_{\mathbf{p}}(t+\Delta) = \sum_{d=0}^{L^{(t)}} \sum_{\mathbf{q} \in S_{\mathbf{p}}^d} y_{s(\mathbf{p}),d}^{s(\mathbf{q})}(\Delta) x_{\mathbf{q}}(t) + z_{s(\mathbf{p})}(\Delta). \tag{4.2.1}$$

In Equation 2.3.1, we derive the transition probabilities for the frequency $N_{\mathbf{p}}(t)$ of traits displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ across a short interval of length $\mathrm{d}t$. Following a similar argument, for a short effective interval length $\theta = \Delta \max(\mu, \beta)$, a trait displaying pattern $\mathbf{q} \in S_{\mathbf{p}}^d$ at time t must transition through at least d states across the interval $[t, t+\Delta]$ to display pattern $\mathbf{p}$ at time $t+\Delta$, so

$$y_{s(\mathbf{p}),d}^{s(\mathbf{q})}(\Delta) = \mathcal{O}(\theta^d). \tag{4.2.2}$$

For example, if $\mathbf{q} = (0,0,1)$ and $\mathbf{p} = (0,1,0)$, the Hamming distance between them is $d(\mathbf{p}, \mathbf{q}) = 2$; and for a trait h currently displaying pattern $\mathbf{q}$,

$$\mathbb{P}\left[\mathbf{p}^h(t+\Delta) = \mathbf{p} \middle| \mathbf{p}^h(t) = \mathbf{q}\right] \approx \left(1 - \frac{\beta\Delta}{L^{(t)}}\right) \frac{\beta\Delta}{L^{(t)}} \mu\Delta \approx \frac{\beta\Delta}{L^{(t)}} \mu\Delta = \mathcal{O}(\theta^2).$$

We illustrate this property in Figure 4.5. On this basis, we can truncate the outer summation in Equation 4.2.1 and capture the majority of the flow into pattern $\mathbf{p}$ across the interval.

There are $\binom{L^{(t)}}{d}$ patterns at distance d from pattern $\mathbf{p}$ so $S_{\mathbf{p}}^d$ and the inner summation in Equation 4.2.1 grow rapidly as d increases up to $d = \lfloor L^{(t)}/2 \rfloor$. However, for a

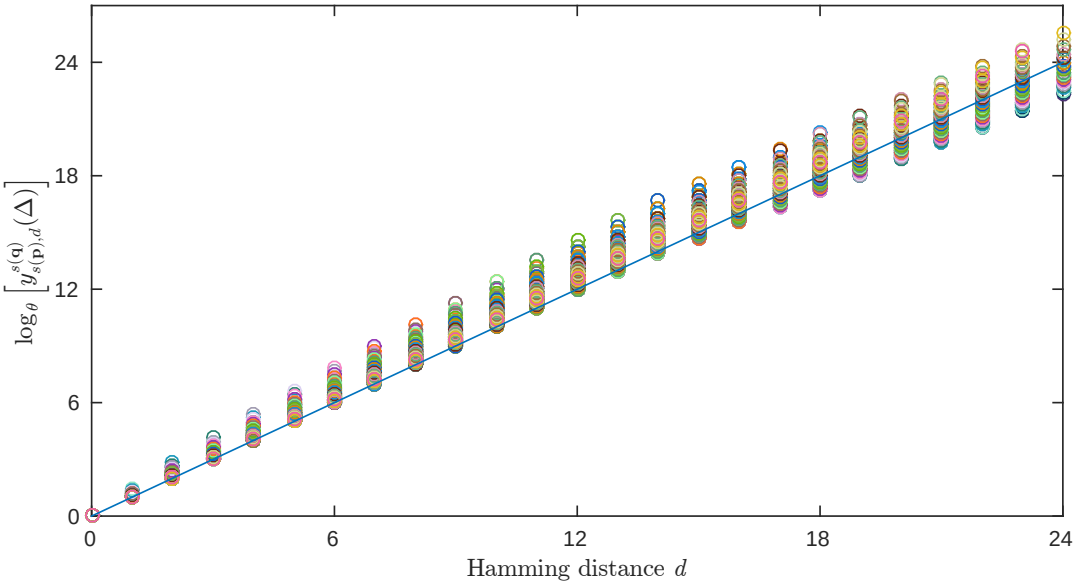


Figure 4.5: The equivalence class solution $y^{s(\mathbf{q})}_{s(\mathbf{p}),d}(\Delta)$ is $\mathcal{O}(\theta^d)$, where $\theta = \Delta \max(\mu, \beta)$. We plot all equivalence class solutions $\mathbf{y}^2, \ldots, \mathbf{y}^{L^{(t)}}$, for $L^{(t)} = 2, \ldots, 24$ lineages. The effective interval length $\theta = 10^{-3}$.

short effective interval length θ, most of the transitions occur between patterns which differ from each other by a single entry. We therefore propose to truncate the outer summation in Equation 4.2.1 at $d = 1$. This approximation has the same $\mathcal{O}(L^{(t)} 2^{L^{(t)}})$ computational cost as a single matrix–vector multiplication in the full pattern–system approach in Equation 2.3.5. Although this is a poor approximation on its own unless the effective interval length θ is close to zero, it forms the basis for a sequence of approximations converging to the true expected pattern frequencies.

We define the sparse matrix $\mathbf{G}_d(\Delta 2^{-k})$ with entry

$$\left[\mathbf{G}_d(\Delta 2^{-k}) \right]_{\mathbf{p},\mathbf{q}} = \begin{cases} y^{s(\mathbf{q})}_{s(\mathbf{p}),d(\mathbf{p},\mathbf{q})}(\Delta 2^{-k}), & d(\mathbf{p},\mathbf{q}) \leq d, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathbf{p}, \mathbf{q} \in \mathcal{P}^{(t)},$$

for distance $d \in \{0, 1, \ldots, L^{(t)}\}$ and index $k \in \mathbb{N}_{\geq 0}$. From Equation 4.1.4, the matrix $\mathbf{G}_d(\Delta)$ is the $\exp(\mathbf{A}^{(t)}\Delta)$ component of Equation 4.1.1 with the $\mathcal{O}(\theta^{d+1})$ and above

terms removed. Furthermore, we define the vector

$$\mathbf{Z}(\Delta) = \left[z_{s(\mathbf{p})}(\Delta) : \mathbf{p} \in \mathcal{P}^{(t)}\right]$$

of contributions to the expected pattern frequencies by the inhomogeneous part of Equation 2.3.5 calculated according to Equation 4.1.7. We define the sequence

$$\mathbf{x}^{(k)}(t+\Delta) = \left[\mathbf{G}_1(\Delta 2^{-k})\right]^{2^k}\mathbf{x}(t) + \mathbf{Z}(\Delta), \quad k = 0, 1, \ldots \tag{4.2.3}$$

of equivalence class-based approximations to the vector of exact expected pattern frequencies $\mathbf{x}(t+\Delta)$.

The computational cost of the kth approximation $\mathbf{x}^{(k)}(t+\Delta)$ is $\mathcal{O}(L^{(t)}2^{k+L^{(t)}})$ operations, so we must make a trade-off between the accuracy of the approximation and the cost of computing it. We first discuss the consistency of the approximation scheme in Equation 4.2.3 then return to this issue below.

4.2.2 Error analysis

The $\mathbf{Z}(\Delta)$ term in Equation 4.2.3 is exact so the error in our approximation $\mathbf{x}^{(k)}(t+\Delta)$ results from the fact that

$$\left[\left[G_d(\Delta 2^{-k})\right]^{2^k}\right]_{\mathbf{p},\mathbf{q}} \leq \left[\left[G_{L^{(t)}}(\Delta 2^{-k})\right]^{2^k}\right]_{\mathbf{p},\mathbf{q}} = \left[\exp(\mathbf{A}^{(t)}\Delta)\right]_{\mathbf{p},\mathbf{q}}, \quad \mathbf{p},\mathbf{q} \in \mathcal{P}^{(t)},$$

for finite k and distance $d < L^{(t)}$. We now use the result in Equation 4.2.2 to demonstrate that the error in our kth equivalence class-based approximation is

$$\begin{aligned}
\mathbf{e}^{(k)} &= \mathbf{x}(t+\Delta) - \mathbf{x}^{(k)}(t+\Delta) \\
&= \left(e^{\mathbf{A}^{(t)}\Delta} - \left[\mathbf{G}_1(\Delta 2^{-k})\right]^{2^k}\right)\mathbf{x}(t) \\
&= \left(\mathbf{G}_{L^{(t)}}(\Delta) - \left[\mathbf{G}_1(\Delta 2^{-k})\right]^{2^k}\right)\mathbf{x}(t) \\
&= \mathcal{O}(\theta^2 2^{-k})\mathbf{x}(t),
\end{aligned}$$

and therefore $\lim_{k\to\infty}[G_1(\Delta 2^{-k}]^{2^k} = \exp(\mathbf{A}^{(t)}\Delta)$.

For the coarsest approximation, when $k = 0$, we have the telescoping sums

$$\mathbf{x}(t+\Delta) - \mathbf{Z}(\Delta) = \Big[\underbrace{\mathbf{G}_0(\Delta)}_{\mathcal{O}(1)} + \underbrace{\mathbf{G}_1(\Delta) - \mathbf{G}_0(\Delta)}_{\mathcal{O}(\theta)} + \underbrace{\mathbf{G}_2(\Delta) - \mathbf{G}_1(\Delta)}_{\mathcal{O}(\theta^2)} + \cdots\Big]\mathbf{x}(t),$$

$$\mathbf{x}^{(0)}(t+\Delta) - \mathbf{Z}(\Delta) = \Big[\underbrace{\mathbf{G}_0(\Delta)}_{\mathcal{O}(1)} + \underbrace{\mathbf{G}_1(\Delta) - \mathbf{G}_0(\Delta)}_{\mathcal{O}(\theta)}\Big]\mathbf{x}(t),$$

and see that the error $\mathbf{e}^{(0)} = \mathcal{O}(\theta^2)\mathbf{x}(t)$. When $k = 1$, we have

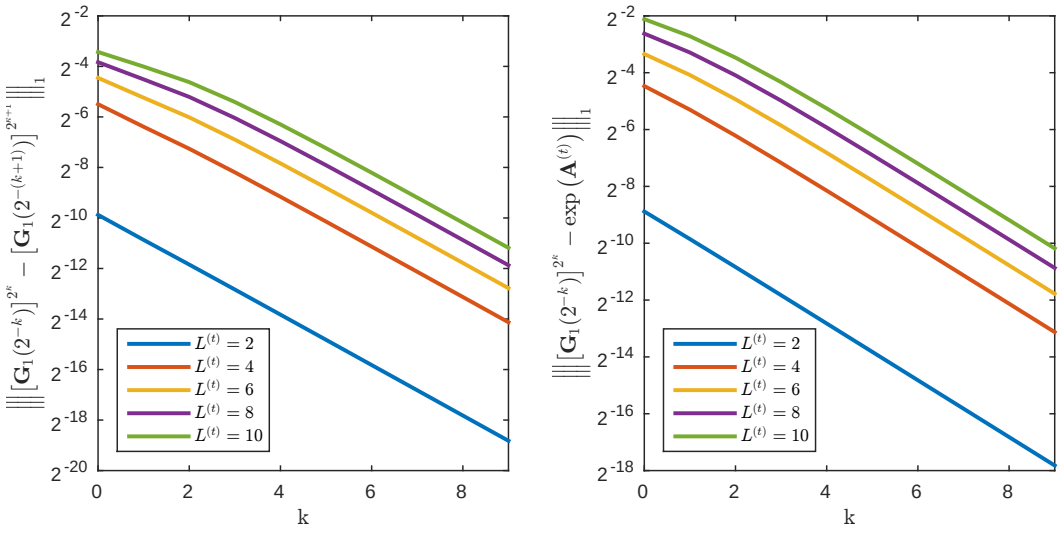
$$\begin{aligned}
&\mathbf{x}(t+\Delta) - \mathbf{Z}(\Delta) \\
&\quad= \Big[\mathbf{G}_0(\tfrac{\Delta}{2}) + \mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2}) + \mathbf{G}_2(\tfrac{\Delta}{2}) - \mathbf{G}_1(\tfrac{\Delta}{2}) + \cdots\Big]^2\mathbf{x}(t), \\
&\quad= \Big[\underbrace{\mathbf{G}_0(\tfrac{\Delta}{2})^2}_{\mathcal{O}(1)} + 2\underbrace{\mathbf{G}_0(\tfrac{\Delta}{2})}_{\mathcal{O}(1)}\underbrace{\big[\mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2})\big]}_{\mathcal{O}(\theta/2)} + \underbrace{\big[\mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2})\big]^2}_{\mathcal{O}(\theta^2/2^2)} \\
&\qquad + 2\underbrace{\overbrace{\mathbf{G}_0(\tfrac{\Delta}{2})}^{\mathcal{O}(1)}\overbrace{\big[\mathbf{G}_2(\tfrac{\Delta}{2}) - \mathbf{G}_1(\tfrac{\Delta}{2})\big]}^{\mathcal{O}(\theta^2/2^2)}}_{\mathcal{O}(\theta^2/2)} + \ldots\Big]\mathbf{x}(t), \\
&\mathbf{x}^{(1)}(t+\Delta) - \mathbf{Z}(\Delta) \\
&\quad= \Big[\mathbf{G}_0(\tfrac{\Delta}{2}) + \mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2})\Big]^2\mathbf{x}(t) \\
&\quad= \Big[\underbrace{\mathbf{G}_0(\tfrac{\Delta}{2})^2}_{\mathcal{O}(1)} + 2\underbrace{\mathbf{G}_0(\tfrac{\Delta}{2})}_{\mathcal{O}(1)}\underbrace{\big[\mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2})\big]}_{\mathcal{O}(\theta/2)} + \underbrace{\big[\mathbf{G}_1(\tfrac{\Delta}{2}) - \mathbf{G}_0(\tfrac{\Delta}{2})\big]^2}_{\mathcal{O}(\theta^2/2^2)}\Big]\mathbf{x}(t)
\end{aligned}$$

and the error $\mathbf{e}^{(1)} = \mathcal{O}(\theta^2/2)\mathbf{x}(t)$. As we consider $\mathbf{e}^{(2)}$, $\mathbf{e}^{(3)}\ldots$ and so on, a pattern emerges whereby our kth approximation scheme misses the term equal to

$$\underbrace{\binom{2^k}{1}}_{2^k}\underbrace{\big[\mathbf{G}_0(\Delta 2^{-k})\big]^{2^k-1}}_{\mathcal{O}(1)}\underbrace{\big[\mathbf{G}_2(\Delta 2^{-k}) - \mathbf{G}_1(\Delta 2^{-k})\big]}_{\mathcal{O}(\theta^2/2^{2k})} = \mathcal{O}(\theta^2 2^{-k})$$

in the expansion of $[G_{L^{(t)}}(\Delta 2^{-k})]^{2^k}$. The error $\mathbf{e}^{(k)} \to \mathbf{0}$ as $k \to \infty$, so our approximation scheme $\mathbf{x}^{(k)}$ (4.2.3) is consistent with the exact solution $\mathbf{x}(t+\Delta)$ (4.1.1, 4.1.8).

We illustrate the convergence of $[\mathbf{G}_1(\Delta 2^{-k})]^{2^k}$ for various $L^{(t)}$ and an effective interval length $\theta = 0.1$ in Figure 4.6. Both the successive differences and errors from $\exp(\mathbf{A}^{(t)}\Delta)$ drop by a factor of 2 as k increases, albeit initially at a slower rate for



(a) $\|\|[\mathbf{G}_1(2^{-k})]^{2^k} - [\mathbf{G}_1(2^{-(k+1)})]^{2^{k+1}}\|\|_1$. (b) $\|\|[\mathbf{G}_1(2^{-k})]^{2^k} - \exp(\mathbf{A}^{(t)})\|\|_1$.

Figure 4.6: Convergence of $[\mathbf{G}_1(2^{-k})]^{2^k}$ in our expected pattern frequency approximation scheme (4.2.3) for various numbers of lineages $L^{(t)}$ and approximation levels k. The effective interval length is $\theta = \Delta\mu = \Delta\beta = 0.1$. The norm $\|\|\cdot\|\|_1$ returns the maximal column sum of its argument.

large $L^{(t)}$. The matrix $G_1(\Delta 2^{-k})$ has the same sparsity pattern as the adjacency matrix for the graph describing the pattern system, the cube in Figure 2.5, for example, so $[G_1(\Delta 2^{-k})]^{2^k}$ has zero entries when $2^k < L^{(t)}$, thereby inflating the discrepancy with the matrix $\exp(\mathbf{A}^{(t)}\Delta)$ of exact transition probabilities.

As the effective interval length θ increases, the error terms are erratic for small k so we must derive conditions where our approximation $\mathbf{x}^{(k)}$ is valid. For the problems we consider, the death rate μ and transfer rate β are typically $\mathcal{O}(10^{-4})$ and the intervals between branching events are typically $\mathcal{O}(10^2)$. The net result here is that the effective interval length $\theta = \mathcal{O}(10^{-2}) \ll 1$ in practice. If θ increases over the course of our MCMC analyses then we can simply increase the order of our approximation to maintain the same level of accuracy in our estimated solution.

From Figure 4.6 and the discussion above, it would appear that we must do double the amount of computation in order to reduce the error in our approximation by a

factor of two. In the following section, we describe how to significantly reduce this error with only negligible additional effort.

4.2.3 Sequence acceleration

Sequence acceleration, as the name suggests, is a branch of numerical analysis which seeks to improve the convergence rate of sequences. Brezinski (2000) provides an overview of sequence acceleration methods, of which Richardson's extrapolation method and Aitken's δ^2 process are two of the best known linear and non-linear methods respectively (Richardson, 1911; Aitken, 1927). No single method works for all sequences (Delahaye and Germain-Bonne, 1980). However, for our equivalence class-based approximation scheme in Equation 4.2.3, an extension of Aitken's δ^2 process for vector sequences works particularly well. We first introduce Aitken's transformation for scalar sequences then describe its extension to vector sequences and how to apply it to our sequence of equivalence class-based approximations.

Suppose we have a scalar sequence $x^{(0)}, x^{(1)}, \ldots$ converging linearly to an unknown limit x; that is,

$$\lim_{k\to\infty} \frac{|x - x^{(k+1)}|}{|x - x^{(k)}|} = \epsilon \in (0, 1).$$

Aitken's δ^2 process is the sequence $v^{(0)}, v^{(1)}, \ldots$ where

$$v^{(k)} = x^{(k+2)} + s^{(k)}\left(x^{(k+2)} - x^{(k+1)}\right), \quad k = 0, 1, \ldots \tag{4.2.4}$$

and the overstep term

$$s^{(k)} = \frac{\lambda^{(k)}}{1 - \lambda^{(k)}} = -\frac{x^{(k+2)} - x^{(k+1)}}{x^{(k)} - 2x^{(k+1)} + x^{(k+2)}}, \quad k = 0, 1, \ldots$$

The ratio of successive differences $\lambda^{(k)} = (x^{(k+2)} - x^{(k+1)})/(x^{(k+1)} - x^{(k)})$ describes how quickly the sequence is converging to x, and $s^{(k)}$ is the corresponding level of

extrapolation. A simple calculation shows that the transformed sequence in Equation 4.2.4 converges to x at a faster rate than the raw sequence.[1]

Returning to the problem at hand, we could apply Aitken's scalar method on an elementwise basis to our vector sequence of equivalence class-based approximations given by Equation 4.2.3. Aitken's method is unstable if $x_{\mathbf{p}}^{(k+2)} - x_{\mathbf{p}}^{(k+1)} \approx x_{\mathbf{p}}^{(k+1)} - x_{\mathbf{p}}^{(k)}$ for any pattern $\mathbf{p} \in \mathcal{P}^{(t)}$, and any error in the estimate of the estimated expected frequency of a single pattern is then propagated down the tree into the expected frequency estimates of other patterns. We cannot guarantee that this behaviour

[1] Aitken's δ^2 process converges to x if $\lim_{k\to\infty} v^{(k)} = x$:

$$\begin{aligned}
\lim_{k\to\infty} x^{(k+2)} - v^{(k)} &= \lim_{k\to\infty} \frac{(x^{(k+2)} - x)^2\left(1 - \frac{x^{(k+1)}-x}{x^{(k+2)}-x}\right)^2}{(x^{(k+2)} - x)\left(1 - \frac{x-x^{(k+1)}}{x-x^{(k+2)}}\right) - (x^{(k+1)} - x)\left(1 - \frac{x-x^{(k)}}{x-x^{(k+1)}}\right)} \\
&= \lim_{k\to\infty} \frac{(x^{(k+2)} - x)^2\left(1 - \frac{1}{\epsilon}\right)^2}{(x^{(k+2)} - x)\left(1 - \frac{1}{\epsilon}\right) - (x^{(k+1)} - x)\left(1 - \frac{1}{\epsilon}\right)} \\
&= \lim_{k\to\infty} \frac{(x^{(k+2)} - x)\left(1 - \frac{1}{\epsilon}\right)}{1 - \frac{x^{(k+1)}-x}{x^{(k+2)}-x}} \\
&= \lim_{k\to\infty} \frac{(x^{(k+2)} - x)\left(1 - \frac{1}{\epsilon}\right)}{1 - \frac{1}{\epsilon}} \\
&= \lim_{k\to\infty} x^{(k+2)} - x \\
&= 0.
\end{aligned}$$

Aitken's transformation accelerates convergence if $\lim_{k\to\infty}(x - v^{(k)})/(x - x^{(k+2)}) = 0$:

$$\begin{aligned}
\lim_{k\to\infty} \frac{x - v^{(k)}}{x - x^{(k+2)}} &= \lim_{k\to\infty} \frac{x - x^{(k+2)} + \frac{(x^{(k+2)}-x)^2\left(1-\frac{x^{(k+1)}-x}{x^{(k+2)}-x}\right)^2}{(x^{(k+2)}-x)\left(1-\frac{x-x^{(k+1)}}{x-x^{(k+2)}}\right)-(x^{(k+1)}-x)\left(1-\frac{x-x^{(k)}}{x-x^{(k+1)}}\right)}}{x - x^{(k+2)}} \\
&= 1 - \lim_{k\to\infty} \frac{\left(1 - \frac{x^{(k+1)}-x}{x^{(k+2)}-x}\right)^2}{\left(1 - \frac{x-x^{(k+1)}}{x-x^{(k+2)}}\right) - \left(\frac{x^{(k+1)}-x}{x^{(k+2)}-x}\right)\left(1 - \frac{x-x^{(k)}}{x-x^{(k+1)}}\right)} \\
&= 1 - \frac{\left(1 - \frac{1}{\epsilon}\right)^2}{\left(1 - \frac{1}{\epsilon}\right) - \frac{1}{\epsilon}\left(1 - \frac{1}{\epsilon}\right)} \\
&= 0.
\end{aligned}$$

will not occur in practice, especially as estimates of the individual expected pattern frequencies converge. In Equation 4.2.3, we form a sequence of approximations which converge linearly to $\exp(\mathbf{A}^{(t)}\Delta)\mathbf{x}(t)$ in Equation 4.1.1. We now make use of this regular behaviour to construct a more efficient and stable acceleration process.

Jennings (1971) considers the case of convergent vector sequences $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ defined by a matrix-iterative process $\mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{M}(\mathbf{x}^{(k)} - \mathbf{x})$. Jennings introduces the sequence $\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \ldots$ where

$$\mathbf{v}^{(k)} = \mathbf{x}^{(k+2)} + s^{(k)}(\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}), \tag{4.2.5}$$

and the overstep term

$$\begin{aligned} s^{(k)} &= \frac{\lambda^{(k)}}{1-\lambda^{(k)}} = -\frac{\mathbf{w}^{(k)\,T}(\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)})}{\mathbf{w}^{(k)\,T}\mathbf{w}^{(k)}}, \\ \mathbf{w}^{(k)} &= \mathbf{x}^{(k)} - 2\mathbf{x}^{(k+1)} + \mathbf{x}^{(k+2)}. \end{aligned} \tag{4.2.6}$$

The quantity $\lambda^{(k)}$, which Jennings refers to as the *predicted eigenvalue*, will form the basis of our error analysis. In contrast to Aitken's scalar acceleration process, the denominator $\mathbf{w}^{(k)\,T}\mathbf{w}^{(k)}$ in the overstep term $s^{(k)}$ in Equation 4.2.6 is zero if and only if $\mathbf{w}^{(k)} = \mathbf{0}$. This only occurs when the sequence has already converged, so issues with numerical instability are much rarer than in Aitken's scalar transformation, and are much easier to catch when they do occur.

We follow Jennings' description here. Suppose the iteration matrix $\mathbf{M}$ has real eigenvalues λ_h and associated right-eigenvectors $\mathbf{q}_h$. If $\mathbf{x}^{(0)} - \mathbf{x} = \sum_h a_h \mathbf{q}_h$ for some constants $a_1, a_2, \ldots$ then $\mathbf{M}\mathbf{q}_h = \lambda_h \mathbf{q}_h$ and

$$\mathbf{x}^{(k)} - \mathbf{x} = \mathbf{M}^k \sum_h a_h \mathbf{q}_h = \sum_h a_h \lambda_h^k \mathbf{q}_h, \quad k = 0, 1, \ldots$$

From Equation 4.2.5, the error in the kth component of the accelerated sequence is

$$\begin{aligned} \mathbf{v}^{(k)} - \mathbf{x} &= \mathbf{x}^{(k+2)} - \mathbf{x} + s^{(k)}(\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}) \\ &= \sum_h a_h \lambda_h^{k+2} \mathbf{q}_h + s^{(k)} \left(\sum_h a_h \lambda_h^{k+1} (\lambda_h - 1) \mathbf{q}_h \right) \end{aligned}$$

$$= \sum_h a_h \lambda_h^{k+2} \mathbf{q}_h \left[1 + s^{(k)} \left(1 - \frac{1}{\lambda_h} \right) \right]$$
$$= \sum_h a_h \lambda_h^{k+2} \mathbf{q}_h \frac{\lambda_h - \lambda^{(k)}}{\lambda_h (1 - \lambda^{(k)})}. \tag{4.2.7}$$

where the predicted eigenvalue $\lambda^{(k)}$ is defined in Equation 4.2.6.

For the vector sequence to converge, the eigenvalues of $\mathbf{M}$ must be inside the unit disc; that is, $\max_h |\lambda_h| < 1$. If $\mathbf{M}$ is symmetric, the predicted eigenvalue $\lambda^{(k)}$ lies between $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, the smallest and largest eigenvalues of $\mathbf{M}$ respectively. To see this, we first rewrite the weight vector $\mathbf{w}^{(k)}$ in Equation 4.2.6 as

$$\mathbf{w}^{(k)} = (\mathbf{x}^{(k)} - \mathbf{x}) - 2(\mathbf{x}^{(k+1)} - \mathbf{x}) + (\mathbf{x}^{(k+2)} - \mathbf{x}) = \sum_h a_h \lambda_h^k \mathbf{q}_h (1 - \lambda_h)^2.$$

The eigenvectors of $\mathbf{M}$ are orthogonal when $\mathbf{M}$ is symmetric so we can substitute the above identity into the definition of $s^{(k)}$ in Equation 4.2.5 to obtain

$$s^{(k)} = -\frac{{\mathbf{w}^{(k)}}^T (\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)})}{{\mathbf{w}^{(k)}}^T \mathbf{w}^{(k)}} = \frac{\sum_h a_h^2 \lambda_h^{2k+1} (1 - \lambda_h)^3}{\sum_h a_h^2 \lambda_h^{2k} (1 - \lambda_h)^4}.$$

The predicted eigenvalue is then

$$\lambda^{(k)} = \frac{s^{(k)}}{1 + s^{(k)}} = \frac{\sum_h a_h^2 \lambda_h^{2k+1} (1 - \lambda_h)^3}{\sum_h a_h^2 \lambda_h^{2k} (1 - \lambda_h)^3} = \frac{\sum_h \lambda_h z_h}{\sum_h z_h},$$

where $z_h = a_h^2 \lambda_h^{2k} (1 - \lambda_h)^3 > 0$ as $|\lambda_h| < 1$ for convergence to occur. As $\lambda^{(k)}$ is a weighted average of the eigenvalues of $\mathbf{M}$, with positive weights which sum to one, it must be the case that $\lambda_{\min}(\mathbf{M}) < \lambda^{(k)} < \lambda_{\max}(\mathbf{M}) < 1$.

If $\mathbf{M}$ is symmetric,

$$\lambda^{(k)} < \frac{\lambda^{(k)}}{\lambda_{\max}(\mathbf{M})} < 1,$$

and therefore, Jennings' transformation (4.2.5) reduces the error (4.2.7) in the direction of the eigenvector corresponding to $\lambda_{\max}(\mathbf{M})$ by a factor

$$\frac{1 - \dfrac{\lambda^{(k)}}{\lambda_{\max}(\mathbf{M})}}{1 - \lambda^{(k)}} < 1$$

compared to $\mathbf{x}^{(k+2)} - \mathbf{x} = \sum_h a_h \lambda_h^{k+2} \mathbf{q}_h$, the error in the unaccelerated sequence.

Our sequence of equivalence class-based approximations (4.2.3) implicitly define a sequence of asymmetric *pseudo-iteration* matrices $\mathbf{M}^{(0)}$, $\mathbf{M}^{(1)}, \ldots$, where

$$\mathbf{M}^{(k)} = \left(\left[G_1(\Delta 2^{-(k+1)}) \right]^{2^{k+1}} - e^{\mathbf{A}^{(t)}\Delta} \right) \left(\left[G_1(\Delta 2^{-k}) \right]^{2^k} - e^{\mathbf{A}^{(t)}\Delta} \right)^{-1}, \quad k = 0, 1, \ldots \tag{4.2.8}$$

satisfies the relation

$$\begin{aligned} \mathbf{x}^{(k+1)}(t+\Delta) - \mathbf{x}(t+\Delta) &= \left(\left[G_1(\Delta 2^{-k+1}) \right]^{2^{k+1}} - e^{\mathbf{A}^{(t)}\Delta} \right) \mathbf{x}(t) \\ &= \mathbf{M}^{(k)} \left[\mathbf{x}^{(k)}(t+\Delta) - \mathbf{x}(t+\Delta) \right] \\ &= \mathbf{M}^{(k)} \left(\left[G_1(\Delta 2^{-k}) \right]^{2^k} - e^{\mathbf{A}^{(t)}\Delta} \right) \mathbf{x}(t). \end{aligned}$$

It is not immediately clear what effect Jennings' transformation will have here as we cannot guarantee that the predicted eigenvalue $\lambda^{(k)} \in [\lambda_{\min}(\mathbf{M}^{(k)}, \lambda_{\max}(\mathbf{M}^{(k)}]$ for all approximations k and effective interval lengths θ. To proceed, we show that the pseudo-iteration matrices are at least asymptotically symmetric and that this convergence is sufficient in practice for Jennings' transformation (4.2.5) to accelerate our sequence of equivalence class-based approximations (4.2.3) to the exact expected pattern frequencies.

Recall from the previous section that, for a sufficiently short effective interval length θ or sufficiently large k,

$$\begin{aligned} \left[[G_1(\Delta 2^{-k})]^{2^k} - \exp(\mathbf{A}^{(t)}\Delta) \right]_{\mathbf{p},\mathbf{q}} &= \mathcal{O}(\theta^2 2^{-k}), \\ \left[[G_1(\Delta 2^{-(k+1)})]^{2^{k+1}} - \exp(\mathbf{A}^{(t)}\Delta) \right]_{\mathbf{p},\mathbf{q}} &= \mathcal{O}(\theta^2 2^{-(k+1)}), \end{aligned} \qquad \mathbf{p}, \mathbf{q} \in \mathcal{P}^{(t)},$$

a property we illustrate in Figure 4.6 above. This implies that $\mathbf{x}^{(k+1)} - \mathbf{x} = \mathcal{O}[(\mathbf{x}^{(k)} - \mathbf{x})/2]$, so the pseudo-iteration matrix $\mathbf{M}^{(k)}$ must converge to $\mathbf{I}/2$, the identity matrix scaled by $1/2$, as k increases. To illustrate this result, we return to the example in Figure 4.6. In Figures 4.7 and 4.8, we plot the elementwise convergence of $[G_1(\Delta 2^{-k})]^{2^k} - \exp(\mathbf{A}^{(t)})$ and $\mathbf{M}^{(k)} - \mathbf{I}/2$ to the zero matrix respectively when there are $L^{(t)} = 6$ lineages and the effective interval length $\theta = 0.1$.

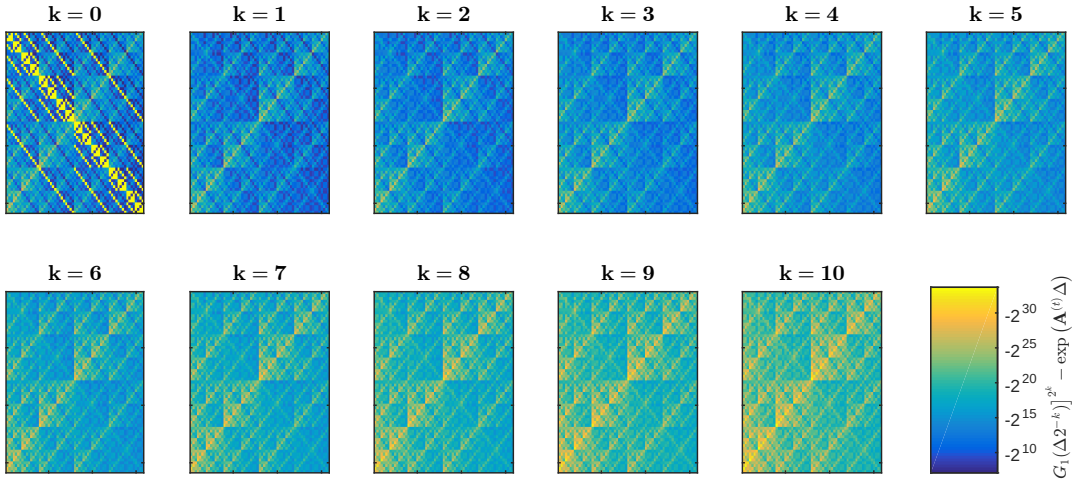


Figure 4.7: Convergence of pseudo-iteration matrix numerator and denominator terms to the zero matrix when $L^{(t)} = 6$ and $\theta = 0.1$.

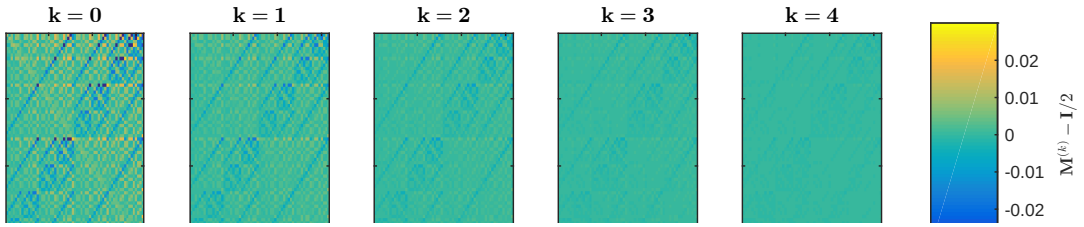
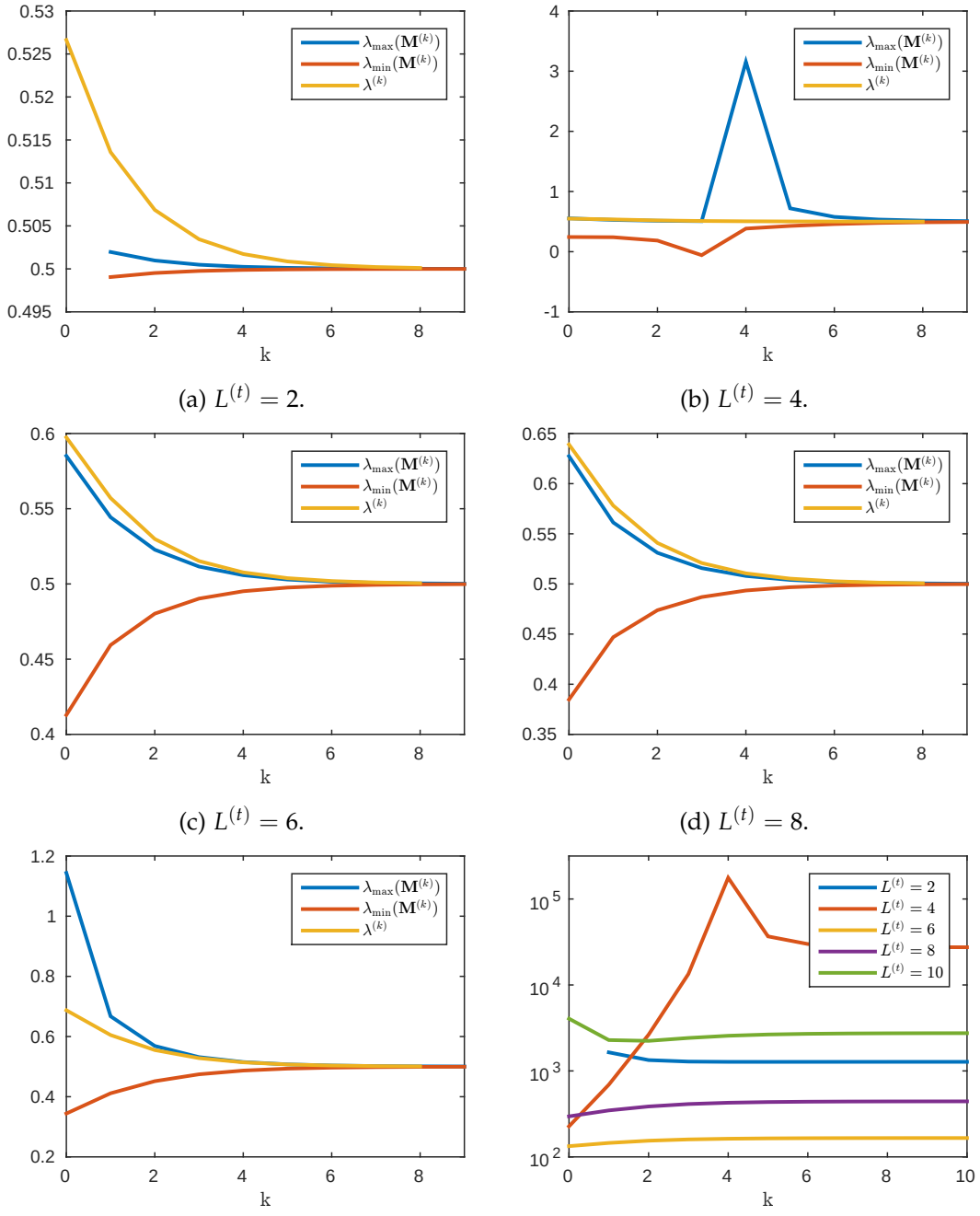


Figure 4.8: Convergence of pseudo-iteration matrix $\mathbf{M}^{(k)}$ to $\mathbf{I}/2$ when $L^{(t)} = 6$ and $\theta = 0.1$.

The characteristic polynomial for the matrix $\mathbf{I}/2$ is $(\lambda - 1/2)^{2^{L^{(t)}}-1}$; therefore, the eigenvalues of the pseudo-iteration matrix $\mathbf{M}^{(k)}$ converge to $1/2$ as $k \to \infty$. Similarly, the predicted eigenvalue $\lambda^{(k)}$ converges to lie between $\lambda_{\min}(\mathbf{M}^{(k)})$ and $\lambda_{\max}(\mathbf{M}^{(k)})$. In Figures 4.9a–e, we plot $\lambda_{\min}(\mathbf{M}^{(k)})$, $\lambda_{\max}(\mathbf{M}^{(k)})$ and $\lambda^{(k)}$ for various $L^{(t)}$ and k. We obtain complex eigenvalues for small k when $L^{(t)} = 2$ and 4, in which case we take the minimum and maximum of the real components. In each case, the predicted eigenvalue $\lambda^{(k)}$ and the eigenvalues of $\mathbf{M}^{(k)}$ converge to $1/2$. We note that even if $\mathbf{M}^{(k)}$ is close to $\mathbf{I}/2$, Jennings' transformation greatly reduces the error in our estimates of the expected pattern frequencies.

Figure 4.9b, we observe erratic behaviour in the computed eigenvalues $\lambda_{\min}(\mathbf{M}^{(3)})$ and $\lambda_{\max}(\mathbf{M}^{(4)})$. The condition number of a matrix, the ratio of its largest and smallest



(a) $L^{(t)} = 2$. (b) $L^{(t)} = 4$.

(c) $L^{(t)} = 6$. (d) $L^{(t)} = 8$.

(e) $L^{(t)} = 10$. (f) Condition number.

Figure 4.9: (a–e) The predicted eigenvalue $\lambda^{(k)}$ in our acceleration scheme (4.2.5) and the largest and smallest eigenvalues of the corresponding pseudo-iteration matrices $\mathbf{M}^{(k)}$ (4.2.8). The initial expected pattern frequencies in the approximation scheme (4.2.3) are independent draws from an Exponential distribution with mean 10. The effective interval length $\theta = 0.1$. (f) The condition number of the matrix $[G_1(\Delta 2^{-k})]^{2^k} - \exp(\mathbf{A}^{(t)}\Delta)$, the numerator in $\mathbf{M}^{(k-1)}$ and denominator in $\mathbf{M}^{(k)}$.

singular values, indicates whether a matrix is close to singular and likely to pose numerical problems if we try to invert it. In Figure 4.9f, we see that $[G_1(\Delta 2^{-3})]^{2^3} - \exp(\mathbf{A}^{(t)}\Delta)$, the numerator in $\mathbf{M}^{(3)}$ and denominator in $\mathbf{M}^{(4)}$, is almost singular compared to the other matrices, suggesting that the sequence is almost converged. This behaviour is not an issue in practice as we can easily control for situations when ${\mathbf{w}^{(k)}}^T\mathbf{w}^{(k)}$, the denominator in the overstep term $s^{(k)}$ in Equation 4.2.6, is close to 0.

We cannot evaluate the matrix $\mathbf{M}^{(0)}$ when $L^{(t)} = 2$ as the denominator, $G_1(\Delta) - \exp(\mathbf{A}^{(t)}\Delta)$, is singular. The Hamming distance from pattern $(1,1)$ to the other patterns in $\mathcal{P}^{(t)}$, $(0,1)$ and $(1,0)$, is 1 so $x_{11}^{(0)} = x_{11}$ is exact. The entries in the corresponding row of $G_1(\Delta) - \exp(\mathbf{A}^{(t)}\Delta)$ are all 0 so the matrix is singular. This highlights a somewhat perverse property of our approximation scheme, also visible in Figure 4.7, that expected frequency estimates for individual patterns may get worse before improving. However, our overall estimates improve as k increases.

For our implementation, at each stage of the expected pattern frequency calculation in Figure 2.6, we compute the sequence of equivalence class-based approximations $\mathbf{x}^{(k)}$, $\mathbf{x}^{(k+1)}, \ldots, \mathbf{x}^{(k+4)}$ to the exact expected pattern frequencies $\mathbf{x}$, where k typically 0 or 1, but may depend on the effective interval length θ or number of lineages $L^{(t)}$. We apply Jennings' transformation (4.2.5) to produce the sequence $\mathbf{v}^{(k)}$, $\mathbf{v}^{(k+1)}$, $\mathbf{v}^{(k+2)}$. We then apply Jennings' transformation a second time to yield $\tilde{\mathbf{x}}$, the estimate of the exact pattern frequencies $\mathbf{x}$ that we use in practice. We illustrate this construction in Figure 4.10.

In Figure 4.11, we illustrate how our approximation to the exact pattern frequencies improves as k increases and through one and two acceleration steps. Jennings' sequence transformation reduces the largest absolute errors by considerable amounts in each case and these reductions in error are consistent as the size of the system increases. The downward trend in errors as the number of lineages $L^{(t)}$ increases is explained by the fact that the space of patterns doubles in size with each additional lineage but the total expected number of patterns increases only linearly. Consequently,

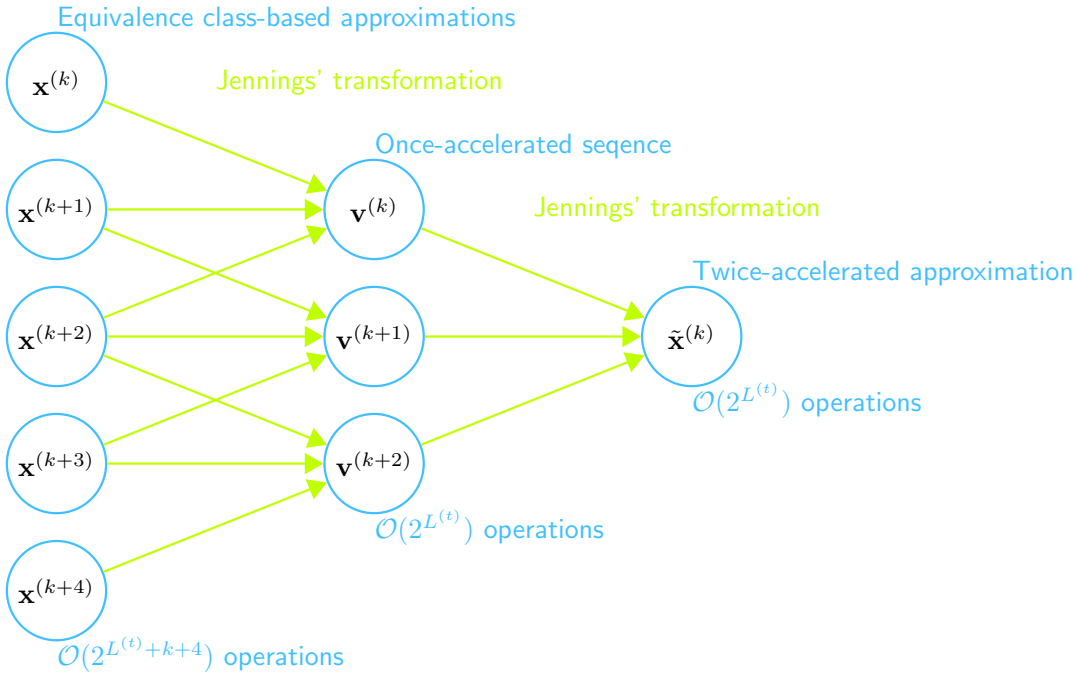


Figure 4.10: Graph depicting the components of our accelerated equivalence class-based approximation to the exact expected pattern frequencies and the computational cost of evaluating each term.

expected pattern frequencies decrease by a factor of two on average, and so do the errors in our approximation.

We can compute the equivalence class-based estimates $\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k+4)}$ in parallel so the computational cost is dominated by $\mathcal{O}[L^{(t)}2^{L^{(t)}}D(k)]$ number of operations required to form $\mathbf{x}^{(k+4)}$, where $D(k) = 2^{k+4}$. The acceleration steps are trivial in comparison as each application of Jennings' transformation requires $\mathcal{O}(2^{L^{(t)}})$ number of operations: one pass through $\mathbf{x}^{(k)}$, $\mathbf{x}^{(k+1)}$ and $\mathbf{x}^{(k+2)}$ to compute the overstep parameter $s^{(k)}$ in Equation 4.2.6, and a second pass applying the transformation in Equation 4.2.5 to produce $\mathbf{v}^{(k)}$. In practice, we perform the transformation from $\mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(k+4)}$ to $\mathbf{v}^{(k)}$, $\mathbf{v}^{(k+1)}$ and $\mathbf{v}^{(k+2)}$ in parallel and simultaneously evaluate the terms in the second acceleration to $\tilde{\mathbf{x}}$. Ignoring any decrease in sampling efficiency of the resulting Markov chain Monte Carlo algorithm — we address this issue later — our $\mathcal{O}(L^{(t)}2^{L^{(t)}}D(k))$ operation to form the doubly accelerated approximation $\tilde{\mathbf{x}}$ compares favourably with

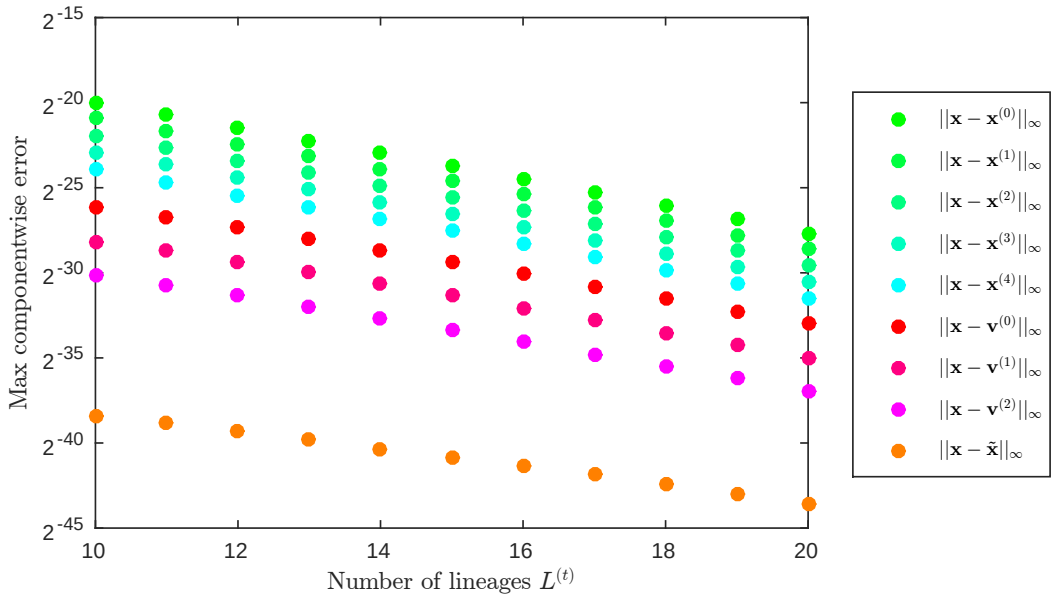


Figure 4.11: Largest componentwise errors between vectors of exact and approximate expected pattern frequencies. We normalise the vectors to sum to 1 and the effective interval length is $\theta = \Delta\mu = \Delta\beta = 10^{-2}$. The norm $\|\cdot\|_\infty$ returns the largest magnitude entry of its argument.

the $\mathcal{O}[L^{(t)}2^{L^{(t)}}C(L^{(t)})]$ cost of the exact, pattern-based approach when $D(k) < C(L^{(t)})$. In Figure 4.1b, $C(L^{(t)}) = \mathcal{O}(10^2)$ so $D(k) < C(L^{(t)})$ when $k \leq 2$. The important message here though is that, in contrast to $C(L^{(t)})$, the cost function $D(k)$ does not grow with the number of lineages $L^{(t)}$ under consideration.

In the following section, we describe how to construct an unbiased estimator of the likelihood using our approximation to the expected pattern frequencies and implement an exact-approximate MCMC inference scheme targeting the exact posterior distribution in Equation 2.5.2.

4.3 The Pseudo-Marginal method

Inspired by recent work on constructing unbiased estimators (Rhee and Glynn, 2012) from a sequence of biased components, we now construct an unbiased estimator of the SDLT model likelihood. The estimator only requires the exact expected pattern

Algorithm 3 The pseudo-marginal Metropolis–Hastings algorithm

1: **Initialise** the chain at state X_0
2: **Draw** an unbiased estimate $\tilde{\pi}_0$ of the target $\pi(X_0)$
3: **For** t = 1 **to** T **do**
4: **Sample** X_t from the randomised Metropolis–Hastings kernel $\tilde{K}(X_{t-1}, \cdot)$
(a) Propose a state $X^* \sim Q(X_{t-1}, \cdot)$ and estimate the target $\pi(X^*)$ by $\tilde{\pi}^*$
(b) Set $(X_t, \tilde{\pi}_t) \leftarrow (X^*, \tilde{\pi}^*)$ with probability

$$\tilde{\alpha}(X_{t-1}, X^*) = \min\left(1, \frac{\tilde{\pi}^*}{\tilde{\pi}_{t-1}} \frac{Q(X^*, X_{t-1})}{Q(X_{t-1}, X^*)}\right),$$

and set $(X_t, \tilde{\pi}_t) \leftarrow (X_{t-1}, \tilde{\pi}_{t-1})$ otherwise

5: **End For**
6: **Return** samples $X_1, \ldots, X_T$

frequencies (2.3.5) to be computed a small fraction of the time. We first discuss the general framework then describe our exact-approximate inference scheme.

4.3.1 Sampling algorithm

Beaumont (2003) describes how to perform exact MCMC inference with an unbiased estimate of the likelihood. That is to say, if we have an unbiased estimator $\tilde{\pi}$ of a target distribution π, we can substitute this estimator into the Metropolis–Hastings transition kernel (3.1.2) and perform exact inference. We summarise this approach in Algorithm 3. Andrieu and Roberts (2009) prove the correctness of this *pseudo-marginal* Metropolis–Hastings algorithm and Andrieu and Vihola (2015) discuss its convergence properties. Andrieu and Vihola (in press) establish a general *convex ordering* framework to compare various pseudo-marginal algorithms. This convex ordering scheme implies ordering by asymptotic variance and mean acceptance probability, amongst others.

At each iteration, we only estimate the target of the candidate state; that is, we do not refresh the estimate of the posterior in the current state. This can lead to poor mixing and *sticky* behaviour when the estimator has high variance, for example, as an overestimate of the posterior will, if the move is accepted, likely lead to a long sequence of rejected moves afterwards. In these situations, one should instead use

a method which refreshes the estimates of both $\pi(X_t)$ and $\pi(X^*)$ in the Metropolis–Hastings acceptance probability in Equation 3.1.1 (Ceperley and Dewing, 1999; Ball et al., 2003; Møller et al., 2006; Murray et al., 2006; Nicholls et al., 2012; Andrieu et al., 2015). We do not pursue these approaches here.

4.3.2 Unbiased estimators

Giles (2008) proposes the *multi-level Monte Carlo* method to determine the optimal allocation of computing resources when estimating the expected value of a function f of a random variable $X = (X(t) : t \geq 0)$ given by the solution of a stochastic differential equation. We cannot simulate an exact continuous sample path of the diffusion but we can approximate it at a sequence of discrete time steps. The most accurate estimates of the function typically require the most computational effort, and Giles describes how to combine lower- and higher-order estimates to minimise this computational cost for a given level of accuracy, although the resulting estimate is biased.

Rhee and Glynn (2012) consider the problem of computing unbiased estimates of $\mathbb{E}[f(X)]$. Let $(X_{h_n} : n \geq 0)$ denote a sequence of discrete-time approximations to X based on steps of size $h_1, h_2, \ldots$ using a method which is consistent in the limit $\lim_{n\to\infty} h_n = 0$. Each estimate $f(X_{h_n})$ is a biased estimate of $f(X)$. However, provided that

$$\mathbb{E}\left[|f(X_{h_0})| + \sum_{k=1}^{\infty} |f(X_{h_n}) - f(X_{h_{n-1}})|\right] < \infty,$$

then for a randomly chosen index variable N which takes values in $\mathbb{N}_0$,

$$f(X_{h_0}) + \sum_{n=1}^{N} \frac{f(X_{h_n}) - f(X_{h_{n-1}})}{\mathbb{P}(N \geq n)} = f(X_{h_0}) + \sum_{n=1}^{\infty} \frac{f(X_{h_n}) - f(X_{h_{n-1}})}{\mathbb{P}(N \geq n)} \mathbf{1}_{\{N \geq n\}}$$

is an unbiased estimator of $\mathbb{E}[f(X)]$. Vihola (2015) presents a general framework which includes both multi-level Monte Carlo and the Rhee–Glynn debiasing technique as special cases.

Let $L^{(e)}(X) = \pi[\mathbf{D}|\mathbf{x}(X)]$ and $L^{(a)}(X) = \pi[\mathbf{D}|\tilde{\mathbf{x}}(X)]$ respectively denote the likelihood component of the posterior (2.5.2) in state X computed using the exact expected pattern frequencies (2.3.5) and their doubly-accelerated approximation in Section 4.2.3. For parameters $p, q, r \in (0,1)$ chosen by the user, we define our unbiased estimator $\tilde{L}$ of the exact likelihood $L^{(e)}$ in a given state to be the random variable with distribution

$$\tilde{L} = \begin{cases} rL^{(a)}, & pq, \\ \dfrac{L^{(e)} - prL^{(a)}}{1-p}, & (1-p)q, \\ L^{(e)}, & 1-q. \end{cases} \tag{4.3.1}$$

We suppose that $r = 1$ for the time being and return to parameters p, q and r in Sections 4.3.3 and 4.3.4 below. It is straightforward to check that this likelihood estimator is unbiased with expectation $\mathbb{E}[\tilde{L}] = L^{(e)}$, and has variance

$$\mathrm{Var}[\tilde{L}] = \frac{pq}{1-p}(L^{(e)} - rL^{(a)})^2. \tag{4.3.2}$$

In practice, we work on a log-scale so, using a lower-case l to denote the log-likelihood, Equation 4.3.1 becomes

$$\tilde{l} = \begin{cases} l^{(a)}, & pq, \\ l^{(a)} + \log\left(e^{l^{(e)} - l^{(a)}} - p\right) - \log(1-p), & (1-p)q, \\ l^{(e)}, & 1-q, \end{cases}$$

our log-estimator of the exact likelihood. Let $\epsilon_{\mathbf{q}} = (x_{\mathbf{q}} - \tilde{x}_{\mathbf{q}})/\tilde{x}_{\mathbf{q}}$, the relative error in our estimate of the expected frequency of pattern $\mathbf{q} \in \mathcal{Q}$ under our doubly-accelerated approximation scheme (4.2.3). From the preceding section, the error $\epsilon_{\mathbf{q}} \approx 0$ so the error in our approximate log-likelihood is

$$\begin{aligned} l^{(e)} - l^{(a)} &= \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log\left(\frac{x_{\mathbf{q}}}{\tilde{x}_{\mathbf{q}}}\right) - \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log\left(\frac{\sum_{\mathbf{q}\in\mathcal{Q}} x_{\mathbf{q}}}{\sum_{\mathbf{q}\in\mathcal{Q}} \tilde{x}_{\mathbf{q}}}\right) \\ &= \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log\left(1 + \frac{x_{\mathbf{q}} - \tilde{x}_{\mathbf{q}}}{\tilde{x}_{\mathbf{q}}}\right) - \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log\left(1 + \frac{\sum_{\mathbf{q}\in\mathcal{Q}} x_{\mathbf{q}} - \tilde{x}_{\mathbf{q}}}{\sum_{\mathbf{q}\in\mathcal{Q}} \tilde{x}_{\mathbf{q}}}\right) \end{aligned}$$

$$= \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log(1+\epsilon_{\mathbf{q}}) - \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}} \log\left(1+\sum_{\mathbf{q}\in\mathcal{Q}} \epsilon_{\mathbf{q}}\right)$$
$$\approx \sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}}\epsilon_{\mathbf{q}} - \left(\sum_{\mathbf{q}\in\mathcal{Q}} n_{\mathbf{q}}\right)\left(\sum_{\mathbf{q}\in\mathcal{Q}} \epsilon_{\mathbf{q}}\right).$$

Under the assumption of a Log-Normally distributed noise term, Doucet et al. (2015) recommend that the variance of the log-likelihood estimator is 1 when the corresponding exact MCMC algorithm is efficient and is 1.7 otherwise. For the likelihood estimator in Equation 4.3.1, this quantity is

$$\begin{aligned}\mathrm{Var}[\log(\tilde{L})] = {} & pq(1-pq)\log^2\left(\frac{L^{(e)}-pL^{(a)}}{L^{(a)}(1-p)}\right) \\ & - 2pq(1-q)\log\left(\frac{L^{(e)}-pL^{(a)}}{L^{(a)}(1-p)}\right)\log\left(\frac{L^{(e)}-pL^{(a)}}{L^{(e)}(1-p)}\right) \\ & + q(1-q)\log^2\left(\frac{L^{(e)}-rL^{(a)}}{L^{(e)}(1-p)}\right).\end{aligned}$$

The noise in our estimator in Equation 4.3.1 is not Log-Normal so we do not attempt to match the variance terms here.

Bardenet et al. (2015) also consider Rhee–Glynn-type estimates of the likelihood. In their setting, they have a large amount of data and a likelihood calculation which scales poorly with the size of the data set. Bardenet et al. start with a sequence of Rhee-Glynn estimates of the likelihood whereby they take a random subsample of the data and add data points to increase the accuracy of their estimates. Unfortunately, the variance of their likelihood estimator is too high to be of any practical use. This is not the case with our approach but there are other issues we must address.

4.3.3 Non-negativity

The likelihood estimator in Equation 4.3.1 is not almost surely non-negative and this poses a problem when we implement it in a pseudo-marginal sampling scheme. The issue arises when we estimate L by $\tilde{L} = (L^{(e)} - pL^{(a)})/(1-p)$ and $L^{(e)} < pL^{(a)}$. Jacob and Thiery (2015) show that we cannot construct almost surely non-negative unbiased

estimators of a function using unbiased estimators of its input. In our context, we cannot guarantee that the estimator in Equation 4.3.1 is almost surely non-negative without a bound on the error in our expected pattern frequency approximation scheme.

To address this issue, Lyne et al. (2015)store the sign of their estimated target $\tilde{\pi}(X_t)$ at each sampled state X_t but use $|\tilde{\pi}(X_t)|$ in the corresponding Metropolis–Hastings acceptance ratio (3.1.1). For $X_1, X_2, \ldots, X_T$ drawn from their sampling scheme, we may estimate expectations of a function $h(X)$ with respect to the correct target distribution π by importance sampling as

$$\frac{\sum_{t=1}^{T} h(X_t)\text{sgn}[\tilde{\pi}(X_t)]}{\sum_{t=1}^{T} \text{sgn}[\tilde{\pi}(X_t)]} \xrightarrow{T\to\infty} \mathbb{E}_{\pi}[h(X)]. \tag{4.3.3}$$

It is difficult to quantify how such a chain will behave in practice as we cannot guarantee that negative estimates of π will not draw the chain away from its target equilibrium. However, almost sure non-negativity of the estimator is not a sufficient condition for the pseudo-marginal algorithm to work: we only require that estimator is unbiased (Sherlock et al., 2015). We shall exploit this fact in the approach below.

In practice, we can ensure that our likelihood estimator in Equation 4.3.1 rarely returns negative estimates over the course of a MCMC run, and that these negative estimates typically coincide with proposed moves to sub-optimal states. First of all, we make a conservative choice of parameter p. This reduces the variance of our estimator and increases $L^{(e)} - pL^{(a)}$. Secondly, we penalise the approximate likelihood $L^{(a)}$ by a factor $r \in (0,1)$ in order to further reduce the chances of obtaining a negative likelihood estimate over the course of a MCMC run. If our estimator (4.3.1) returns a negative likelihood estimate at any stage of an MCMC analysis, we simply restart the chain with a more conservative choice of parameters p, q and r. Alternatively, we could increase the accuracy of our equivalence class-based approximation. If we do not observe a negative likelihood estimate over the course of the MCMC chain then any inference based on the resulting samples is valid. For a simple explanation of

this, the corresponding sign terms in Equation 4.3.3 are all positive so we recover the importance sampling estimator.

These measures come at a cost. Decreasing p and q increases the overall computation time as we must compute the exact pattern frequencies more often during the course of an MCMC run. Furthermore, adding a downward bias r to the approximate likelihood $L^{(a)}$ increases the variance of our estimator (4.3.2). The variance of our estimator (4.3.2) penalises symmetric differences in the exact and approximate likelihoods but, in practice, we place a much heavier penalty on negative likelihood estimates from overestimating the exact likelihood — we must restart the corresponding chain as a result — so we do not worry too much about the increase in variance here. In choosing these parameters, we make a trade-off between the efficiency of the algorithm in terms of computing time and parameter effective sample sizes, and the risk of having to repeat our analyses because of a negative likelihood estimate. For all of our analyses, we set $p = q = 0.95$ and $r = 0.9$ on the basis of trial runs. We illustrate this procedure below.

4.3.4 Parameter choice

In order to choose the parameters of our pseudo-marginal scheme, we fit the SDLT model to a synthetic data set, `SIM-B`, which we describe in Chapter 5. We construct two Metropolis–Hastings chains: the first chain targets the posterior with likelihood component $L^{(e)}$, and the second chain targets posterior with likelihood component $L^{(a)}$ from in the previous section. These chains represent our best and worse attempts at sampling from SDLT model posterior, and so long as the error terms are relatively small, the unbiasedness correction in Equation 4.3.1 will be negligible.

In Figure 4.12, we compare the exact and approximate Metropolis–Hastings log-transition probabilities α and $\hat{\alpha}$ respectively (3.1.1). For a proposed move $X_t \to X^*$

with $\pi(X_t) < \pi(X^*)$ and $\hat{\pi}(X_t) < \hat{\pi}(X^*)$,

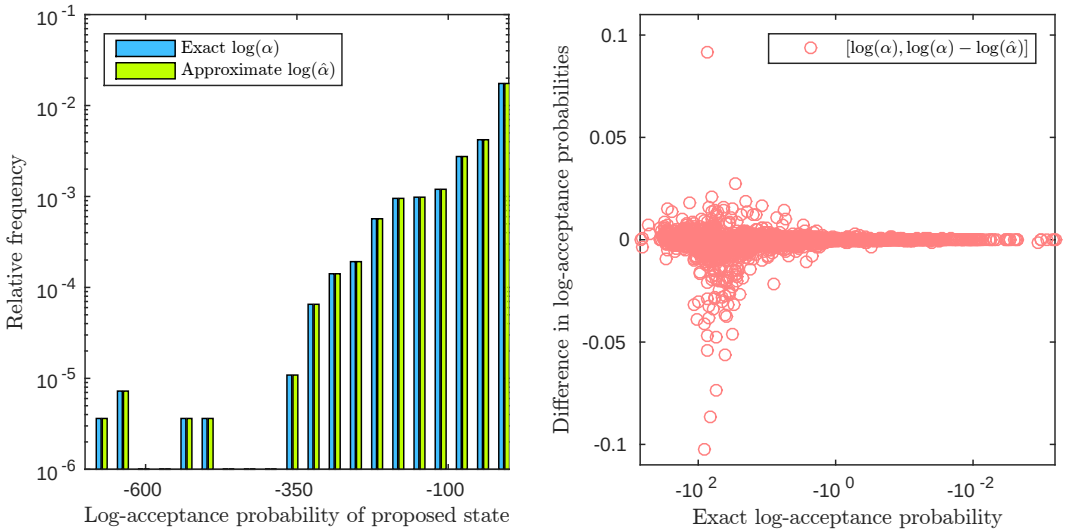
$$\log[\alpha(X_t, X^*)] - \log[\hat{\alpha}(X_t, X^*)] = l^{(e)}(X^*) - l^{(e)}(X_t) - l^{(a)}(X^*) + l^{(a)}(X_t),$$

the differences in exact and approximate log-likelihoods. In general, $L^{(a)}$ accurately estimates $L^{(e)}$, but in both chains we observe absolute errors up to 0.1 in the log-transition probabilities. On this basis, we may set $p = 0.95$ and $r = 0.9$ and be reasonably confident that we will not obtain a negative likelihood estimate over the course of our pseudo-marginal analysis. Verifying our remark earlier, we see that the estimator is least accurate for moves which are far from the current state in terms of the exact acceptance probability.

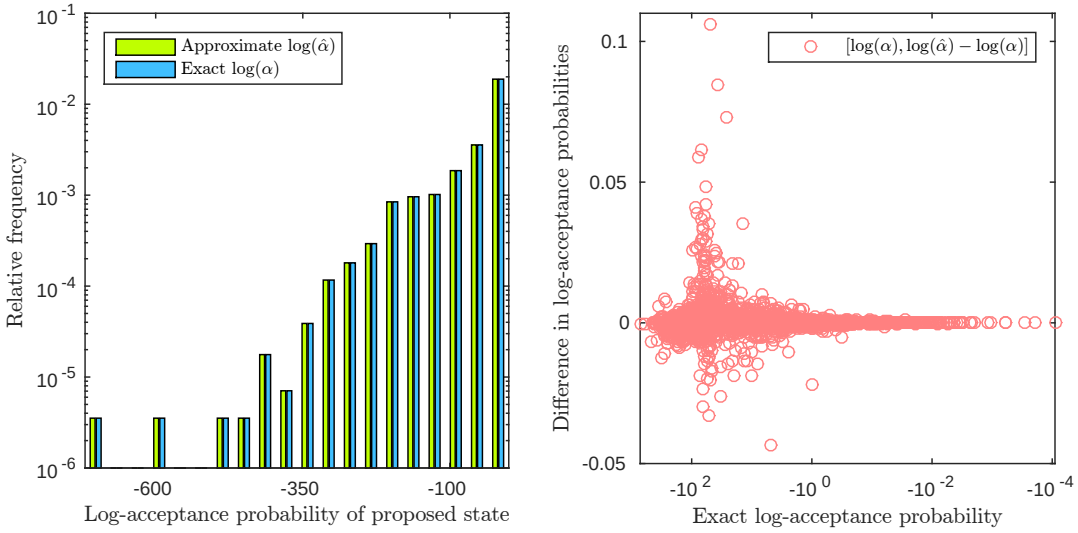
We denote by $\tilde{\alpha}$ the transition probability for a pseudo-marginal Metropolis–Hastings chain targeting the SDLT posterior with likelihood component estimated according to Equation 4.3.1. In Figure 4.13, we compare the exact-approximate and exact transition probabilities at each state of the resulting pseudo-marginal chain. The multimodality in the plot of differences in acceptance probabilities is a result of the likelihood estimator returning a different estimate at each iteration.

Following the results of Peskun (1973) and Andrieu and Vihola (in press), we expect the exact chain to dominate the exact-approximate chain, at least asymptotically, in terms of mean acceptance probability. We record the mean acceptance probabilities for each of the three chains in Table 4.1 and see that this is the case in practice, but not by much. We summarise samples of the root age $-t_1$, death rate μ and relative transfer rate β/μ in Figure 4.14. Here, the exact and exact-approximate chains are indistinguishable in terms of their posterior distributions on parameters and effective sample sizes. The most striking result here is the fact that the exact-approximate method produces a similar number of effectively independent samples as the exact method, but does so in approximately one quarter of the time.

We detect a slight bias in the samples of the root age $-t_1$ using the approximate inference scheme, but otherwise there is little to distinguish the approximate chain



(a) Exact chain.



(b) Approximate chain.

Figure 4.12: Exact and approximate acceptance probabilities at various states of Markov chains targeting the SDLT posterior (2.5.2).

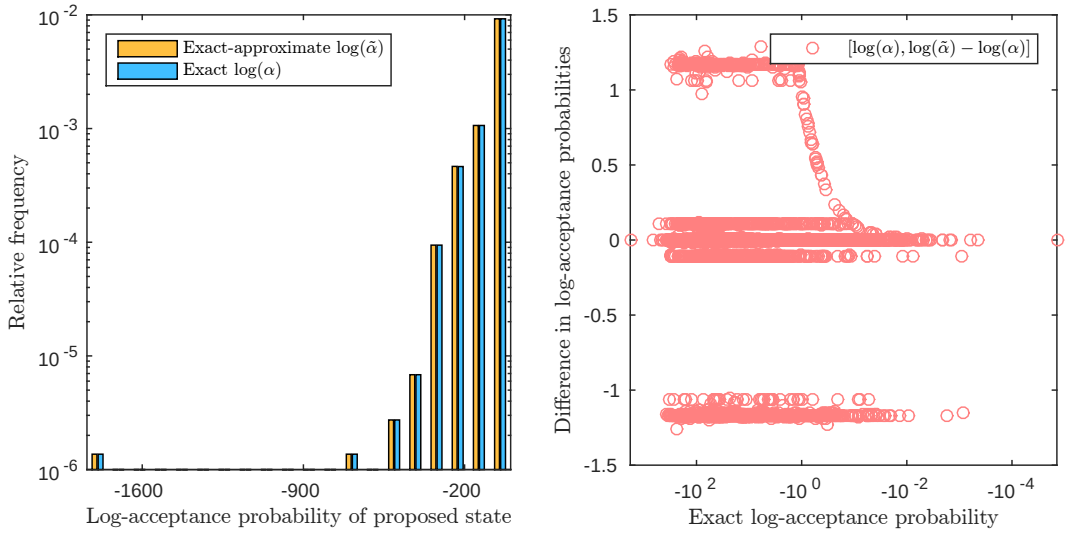


Figure 4.13: Exact-approximate and exact acceptance probabilities at various states of a Markov chain targeting the SDLT posterior (2.5.2) with unbiased estimates of the likelihood components (4.3.1).

Chain	Acceptance probability	
	Mean	Variance
Exact	0.172	0.114
Approximate	0.167	0.110
Exact-approximate	0.157	0.104

Table 4.1: Acceptance probabilities in the exact, approximate and exact-approximate chains.

from the others. Nicholls et al. (2012) use a simple *coupling* argument to assess the quality of their approximate inference. For the exact and approximate chains we describe above, suppose that we actually pass them the same sequence of random numbers. The transition kernels are not identical so eventually the two chains will make different decisions and *decouple*. However, the samples from the approximate chain up to this point are exact as they are identical to the samples from the exact chain. If we perform this experiment in practice, the approximate chain decouples from the exact chain after $\mathcal{O}(10^5)$ iterations. Although this decoupling time is relatively high, MCMC algorithms for phylogenetic models often mix slowly so we require

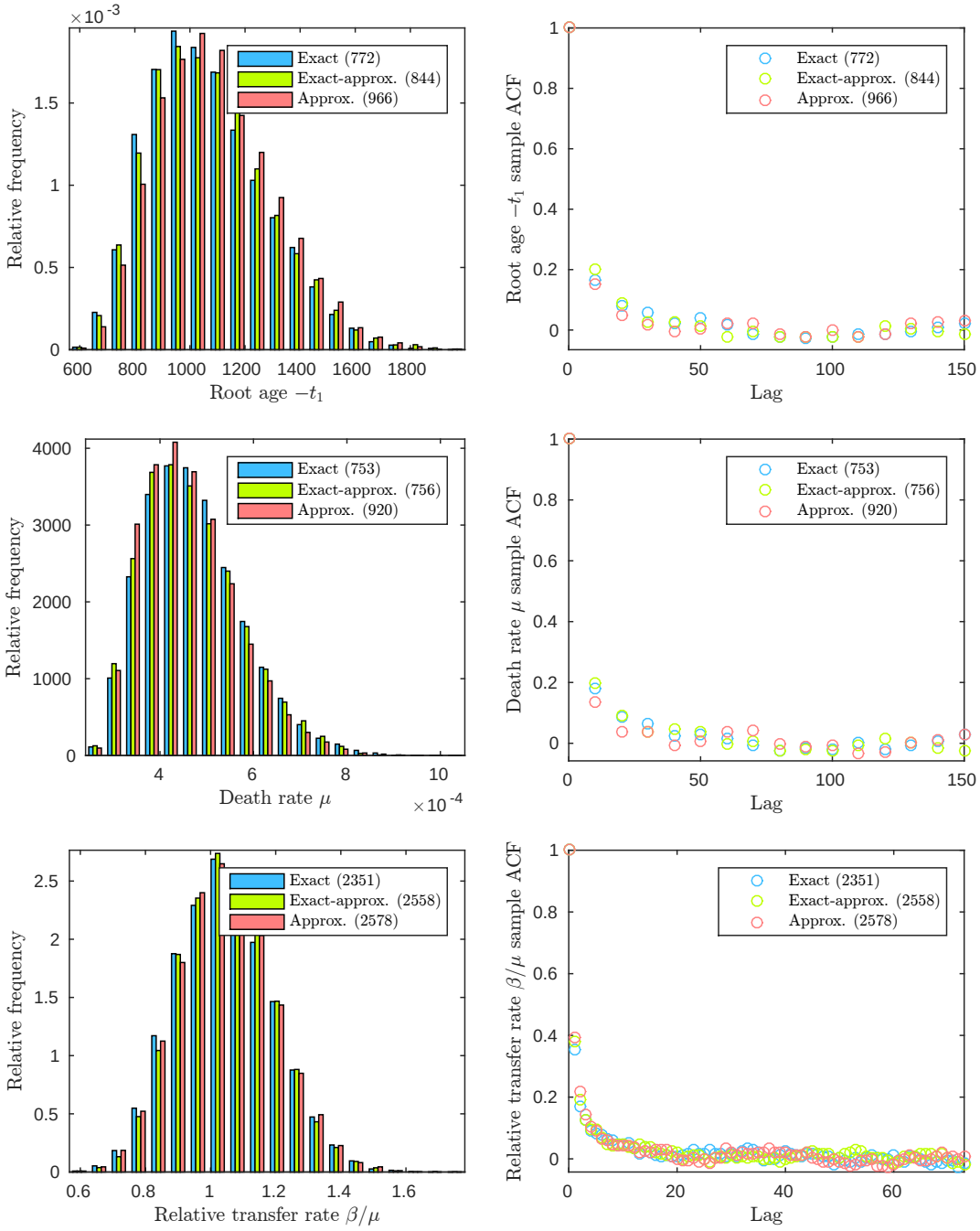


Figure 4.14: Histograms of samples in our analyses of `SIM-B`. Effective sample sizes are in parentheses.

much longer chains in order to obtain adequate effective sample sizes. However, if we are willing to accept a small bias in our results, the approximate chain is even faster than the exact-approximate scheme. On this basis, we implement this approximate algorithm in Chapter 6 in order to analyse a data set of traits in Indo-European languages which is sufficiently large that the exact and exact-approximate approaches we describe are infeasible.

4.4 Discussion

In this chapter, we describe how to exploit symmetry in the initial value problems in Chapter 2 to construct a linearly converging sequence of estimates to the exact expected pattern frequencies. We then transform this sequence to create a new sequence of estimates which converges at a faster rate in practice. From this, we construct an unbiased estimator of the exact likelihood (2.3.6) — that is, the likelihood computed with the exact expected pattern frequencies — but only compute the exact expected pattern frequencies a small proportion of the time. We now remark on a possible extension to our unbiased likelihood estimator and an alternative approach to our pseudo-marginal sampling scheme.

To simplify the analysis in Section 4.3, we restrict our approach to a two-step Rhee–Glynn estimator of the likelihood. However, we have access to a sequence of estimates to the exact expected pattern frequencies with increasing accuracy and computational cost so it would require a straightforward extension of the estimator in Equation 4.3.1 to include these terms. To ensure that the resulting likelihood estimator is non-negative in practice, we could apply a heavier downward bias to the less accurate terms, for example.

As an alternative to the pseudo-marginal approach we describe Section 4.3, we could use the approximate likelihood $L^{(a)}$ in the first step of a two-stage accept/reject procedure (Christen and Fox, 2005, 2010). If the first step is accepted, we then make

a correction using the exact likelihood in the second step so that we target the exact posterior. The exact likelihood must be computed for each accepted move and this can lead to a high computational cost when the tentative acceptance rate is relatively high. For the example above, the algorithm would pass the first stage approximately 16–17% of the time. The proportion of moves which then make it through the second step would depend on the quality of the estimator. In contrast, our pseudo-marginal algorithm requires the exact likelihood to be computed $1 - pq = 9.75\%$ of the time, a significant saving on a two-stage approach.

Chapter 5

Model validation and testing

Chapter overview

Nicholls and Gray (2008) and Ryder and Nicholls (2011) describe tests using synthetic data sets to validate their respective models and inference procedures. In this chapter, we first describe how to assess the goodness-of-fit of the SDLT and SD models and then apply them to three synthetic data sets. We conclude with a brief description of our software to implement the SDLT model.

5.1 Goodness-of-fit

For two competing hypotheses, H_i and $H_{i'}$, the Bayes factor (Jeffreys, 1935; Kass and Raftery, 1995) comparing their marginal posterior support is

$$B_{i',i} = \frac{\pi(R(\mathbf{D})|H_{i'})}{\pi(R(\mathbf{D})|H_i)} = \frac{\int \pi(R(\mathbf{D})|x_{i'}, H_{i'})\pi(x_{i'}|H_{i'})\mathrm{d}x_{i'}}{\int \pi(R(\mathbf{D})|x_i, H_i)\pi(x_i|H_i)\mathrm{d}x_i}, \tag{5.1.1}$$

where x_i and $x_{i'}$ denote the parameters under the respective hypotheses H_i and $H_{i'}$. As evidence against H_i, Kass and Raftery recommend the following interpretation: $\log(B_{i',i}) < 1$ is not worth mentioning; $1 \leq \log(B_{i',i}) < 3$ is positive evidence; $3 \leq \log(B_{i',i}) < 5$ is strong evidence; and $\log(B_{i',i}) > 5$ is very strong evidence; where log denotes the natural logarithm. We now outline two simple Bayes factor-based

tests for goodness-of-fit and model selection, and mention a third which is difficult to implement successfully when comparing the SDLT and SD models on data with lateral transfer.

5.1.1 Relaxing constraints

One measure of goodness-of-fit is how influential the model constraints are on our inference. We can relax a clade constraint in the model and compute a Bayes factor comparing the relaxed and constrained models. Suppose the constraint $\Gamma^{(i)}$ fixes node $i \in V$ and $\Gamma^{(i')}$ denotes its relaxation. We denote by $\Gamma^{C'}$ the calibrated space of phylogenies with $\Gamma^{(i)}$ replaced by $\Gamma^{(i')}$. Now, $\Gamma^{C} \subset \Gamma^{C'}$ so the Bayes factor (5.1.1) is

$$\begin{aligned} B_{i',i} &= \frac{\pi(R(\mathbf{D})|g \in \Gamma^{C'})}{\pi(R(\mathbf{D})|g \in \Gamma^{C})} \\ &= \frac{\pi(R(\mathbf{D})|g \in \Gamma^{C'})}{\pi(R(\mathbf{D})|g \in \Gamma^{C} \cap \Gamma^{C'})} \\ &= \frac{\pi(g \in \Gamma^{C}|g \in \Gamma^{C'})}{\pi(g \in \Gamma^{C}|R(\mathbf{D}), g \in \Gamma^{C'})}, \end{aligned} \tag{5.1.2}$$

a Savage–Dickey ratio of the marginal prior and posterior densities/probabilities that the constraint is satisfied in the relaxed model (Ryder and Nicholls, 2011). A large Bayes factor here indicates a lack of support for the constraint and is a sign of model misspecification.

For all of the tests in this thesis, we relax the time constraint on an internal or leaf node. In each case, the marginal prior on the node time is uniform across its range. We cannot compute the Savage–Dickey ratio (5.1.2) in closed form so, in practice, we estimate the densities by the proportions of sampled node times satisfying the respective constraints.

5.1.2 Predictive scores

We assess the predictive performance of each model on a random splitting of the registered data $R(\mathbf{D})$ into evenly sized training and test sets labelled $\mathbf{D}^{\text{tr}}$ and $\mathbf{D}^{\text{te}}$

respectively. Following Madigan and Raftery (1994), we score each model by its log-posterior predictive probability,

$$\log \pi(\mathbf{D}^{\text{te}}|\mathbf{D}^{\text{tr}}) = \log \int \pi(\mathbf{D}^{\text{te}}|x)\pi(x|\mathbf{D}^{\text{tr}})\mathrm{d}x, \tag{5.1.3}$$

where $x = (g, \mu, \beta, \kappa, \Xi)$, the SDLT model parameters, and $x = (g, \mu, \kappa, \Xi)$ for the SD model. The posterior predictive $\pi(\mathbf{D}^{\text{te}}|\mathbf{D}^{\text{tr}})$ is a marginal likelihood so the difference in predictive scores is a log-Bayes factor measuring the relative success of the models in predicting the test data (Kass and Raftery, 1995).

5.1.3 Reversible jump MCMC

An alternative approach to model comparison is to use reversible jump MCMC (Green, 1995) to sample from a posterior on models and parameters. The SD model is nested within the SDLT model, so it is not difficult to implement such an algorithm in practice. However, for the problems we consider, there is little overlap in posterior support between the SDLT and SD models, so the algorithm struggles to jump between them. We could force the chain to jump between models using the Wang–Landau algorithm from Chapter 4, but we do not attempt that approach here as the methods we describe are sufficient.

5.2 Exact and empirical distributions of pattern frequencies

In addition to manual tests, we now validate our implementation of the exact expected pattern frequency calculation in Equation 2.3.5 and the likelihood in Theorem 1 though simulation. To simulate the SDLT model on a given tree, in a method similar to Gillespie (1977), we directly simulate the dynamics of the jump process in Equation 2.3.1 on intervals between branching events with the transfer operation of Section 2.3.1 propagating trait frequency vectors across branching events.

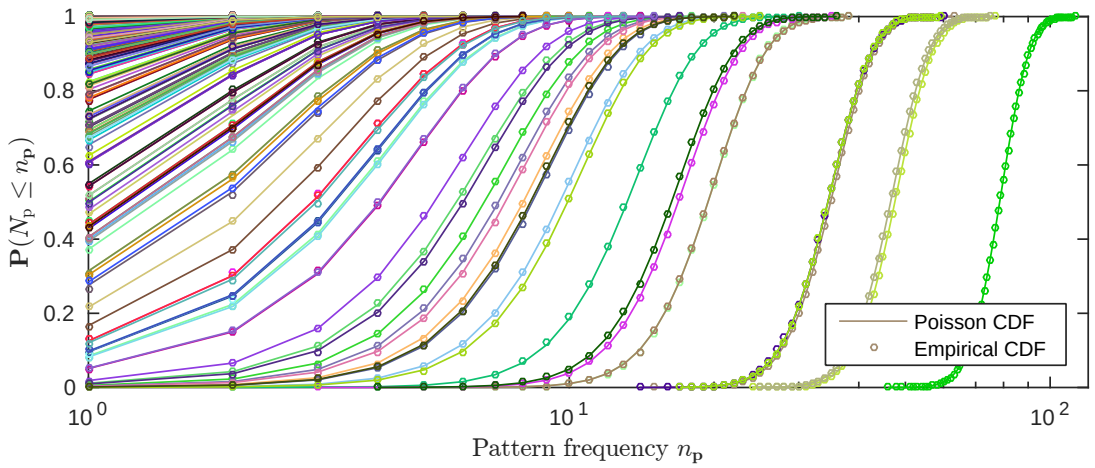


Figure 5.1: Exact and empirical CDFs of patterns in $\mathcal{P}$ in 10^4 draws from the SDLT process on the tree in Figure 5.2 with parameters in Table 5.1.

In Figure 5.1, we compare the exact Poisson cumulative distribution function of each pattern in $\mathcal{P}$ under the SDLT model in Theorem 1 with empirical estimates based on 10^4 draws from the SDLT model on the tree in Figure 5.2 with parameters in Table 5.1. The exact and empirical cumulative distribution functions (CDFs) are in clear agreement so we conclude that our simulation scheme, likelihood and expected pattern frequency calculation are all correct.

5.3 Coupled synthetic data sets

We validate our model and computer implementation using three coupled synthetic data sets. Our reason for coupling the data is to illustrate the effect that discarding known-transferred traits, a common practice, has on our inference.

5.3.1 Construction

The data set `SIM-B` is a draw from the SDLT process on the tree in Figure 5.2 with parameters in Table 5.1. We shall use `SIM-B` to test the identifiability of the SDLT model. The relative transfer rate $\beta/\mu = 1$. Although the shortcomings of the SD model in this setting have already been established (Nicholls and Gray, 2008; Greenhill

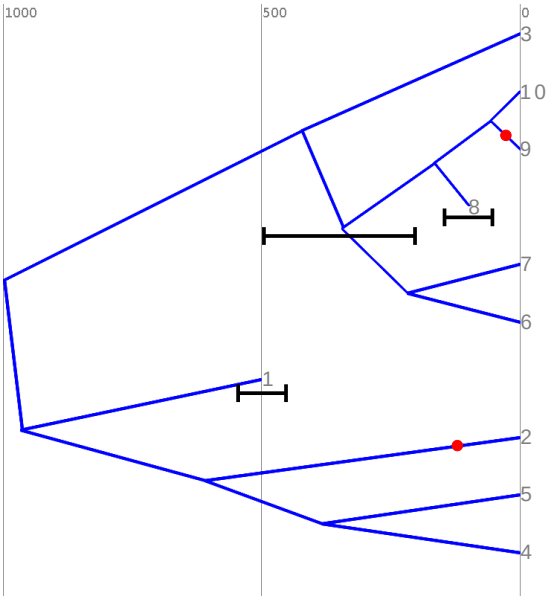


Figure 5.2: The tree underlying the synthetic data sets. Time is in years before the present. Catastrophe locations are marked in red and clade constraints in black. In addition to its time restriction, the constrained internal node is fixed as the most recent common ancestor of the nodes labelled '6' to '10'.

Parameter	Value	Parameter	Value
Trait birth rate	$\lambda = 10^{-1}$	Root time	$t_1 = -10^3$
Trait death rate	$\mu = 5 \times 10^{-4}$	Catastrophe severity	$\kappa = 0.2212$
Trait transfer rate	$\beta = 5 \times 10^{-4}$	Observation probabilities	$\Xi \sim \beta(1, 1/3)^L$

Table 5.1: Model parameters for synthetic data set `SIM-B`. The catastrophe duration $\delta = -\mu^{-1} \log(1 - \kappa) = 500$ years.

et al., 2009), we fit the SD model here to highlight the effect of properly controlling for the effect of lateral transfer. For one of the data sets we analyse in Chapter 6, we infer a relative transfer rate well above 1.

From `SIM-B`, we create two additional data sets: `SIM-N` and `SIM-T`. To form `SIM-N`, we discard any trait copy derived from a lateral transfer event; that is, we do not discard all instances of a given trait, only the copies which transferred. This is equivalent to ignoring all the lateral transfer events in the generation of `SIM-B` so `SIM-N` is an exact draw from the SD process with the useful property that is coupled to `SIM-B` so we can compare our inferences across data sets. The SD model is nested

Data set	Traits	True model	Purpose
SIM-B	678	SDLT	Identifiability
SIM-N	672	SD	Consistency
SIM-T	675	SDLT before time -250, and SD thereafter	Robustness

Table 5.2: Summary of the synthetic data sets.

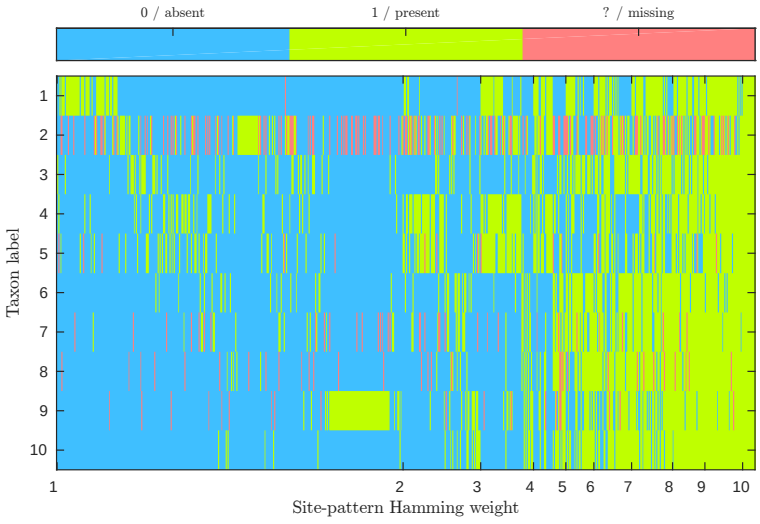
within the SDLT model so we use SIM-N to establish the consistency of the two models when the lateral transfer rate $\beta = 0$.

Recently transferred traits are more readily identified and discarded in practice. This potential bias is a common source of model misspecification. To this end, we create a further data set SIM-T from SIM-B except this time we only discard instances of traits in SIM-B which transferred in the final 250 years. We fit the SDLT and SD models here to test their robustness to this source of model misspecification. We summarise the synthetic data sets in Table 5.2 and plot them in Figure 5.3. From each of the synthetic data sets, we discard traits not marked present in at least one taxon according to the registration process in Section 2.4.4.[1]

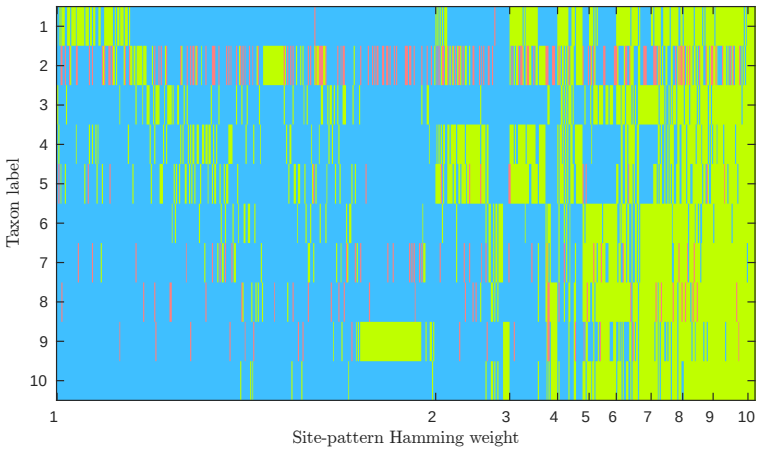
5.3.2 Analyses

In addition to the clade constraints in Figure 5.2, we enforce a minimum root time $\underline{t}_1 = -2000$ years (maximum root age $-t_1$ is 2000 years). We analyse each data set under the exact and exact-approximate SDLT methods, referred to as SDLT-EE and SDLT-EA respectively, and the SD model. For the exact-approximate method, we estimate the expected pattern frequencies by twice accelerating the sequence $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(5)}$. We use the Metropolis–Hastings algorithm of Chapter 3 to fit the models and summarise the results of these analyses in Figures 5.4 and 5.5.

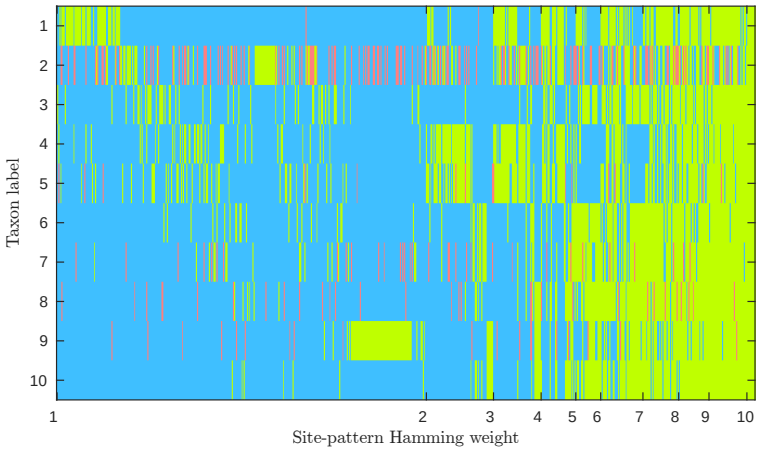
[1]The differing figures in Table 5.2 for the total number of registered traits reflect the fact that some traits are removed completely in creating SIM-N and SIM-T from SIM-B. For the most part, a trait displays a pattern with reduced Hamming weight instead.



(a) `SIM-B`.



(b) `SIM-N`.



(c) `SIM-T`.

Figure 5.3: Synthetic data sets.

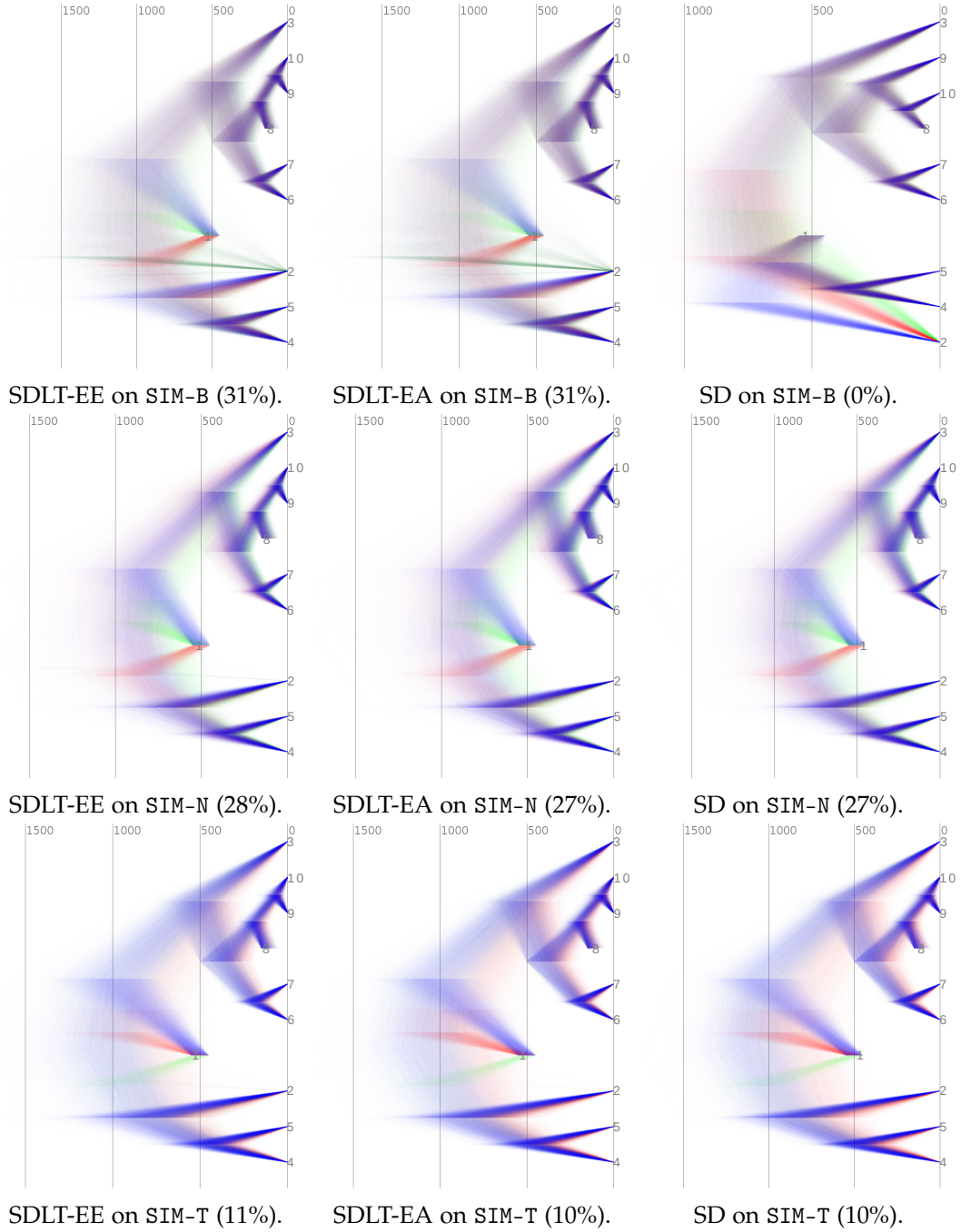


Figure 5.4: `DensiTree` (Bouckaert and Heled, 2014) plots of the marginal tree posteriors for each synthetic dataset and model. Figures in parentheses denote posterior support for the true tree topology. The most frequently sampled topologies in each case are coloured blue, followed by red then green, with the remainder in dark green.

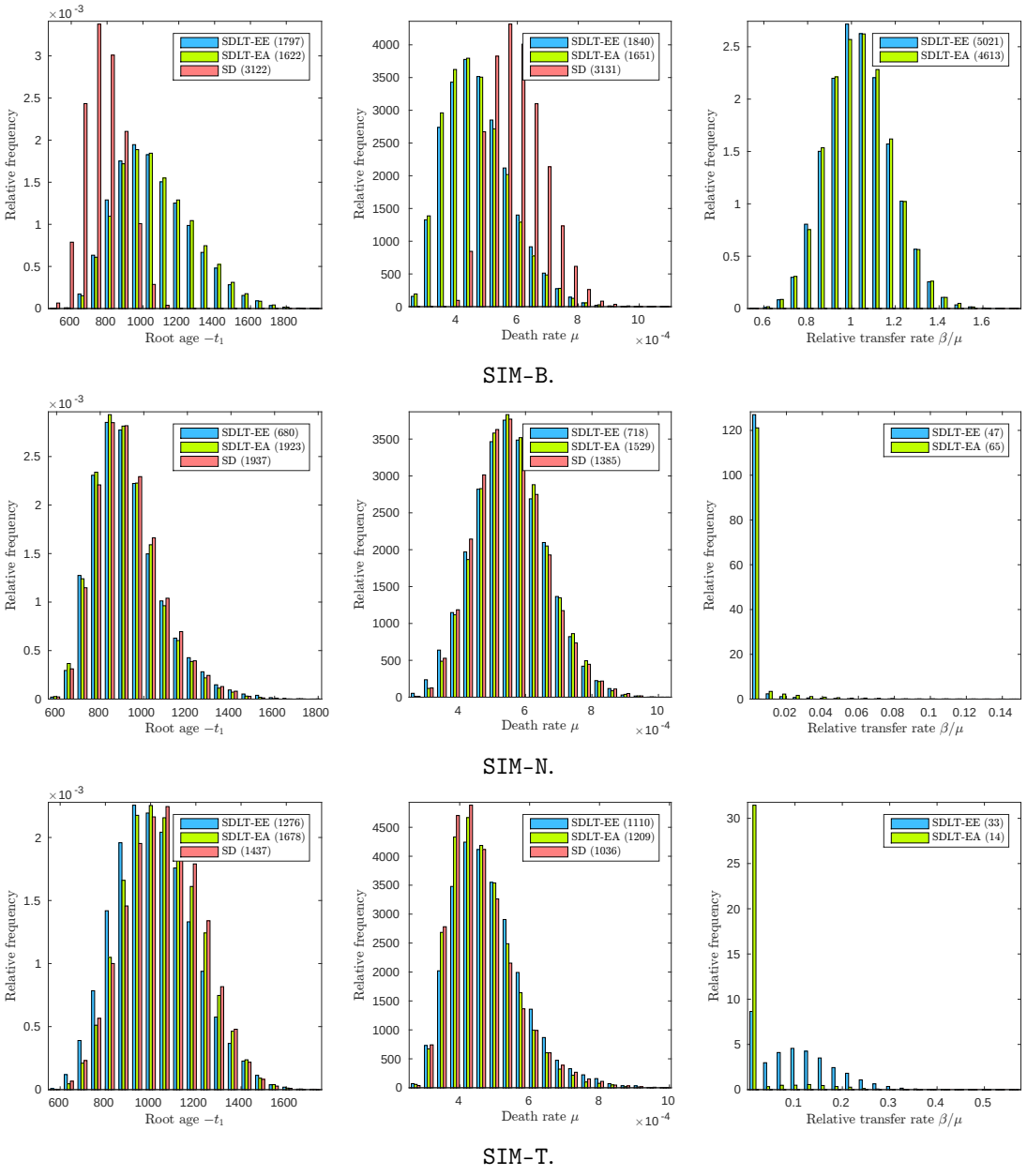


Figure 5.5: Histograms of parameter samples in our analyses of the synthetic data sets. Effective sample sizes are in parentheses.

We first note that our exact and exact-approximate SDLT inferences are, for the large part, indistinguishable from each other. Both methods return similar tree and parameter posterior distributions and the exact method does not dominate the exact-approximate scheme in terms of effective sample size. The two sampling schemes in our analysis of `SIM-T` return different distributions on the relative transfer rate β/μ. However, the effective sample size is so low here that we can put this difference down to sampling error. On this basis, we focus on the differences between the exact SDLT and SD models from here on.

Of particular interest among the marginal tree posteriors in Figure 5.4 is the contrasting supports for the true topology in each case, particularly `SIM-B` where the SD model focuses on the wrong topology entirely. Unsurprisingly, there is little to distinguish the SDLT and SD marginal tree posteriors on `SIM-N` and `SIM-T`, suggesting that the SDLT model is consistent with the SD model on `SIM-N` and that neither model is seriously misspecified on `SIM-T`. The difference in posterior support for the true topology between the SDLT model fit to `SIM-B` and `SIM-T` is remarkable, and likewise for the SD model fit to `SIM-N` and `SIM-T`. The traits which did not transfer in the final 250 years are having a clear, if subtle, effect on our inference.

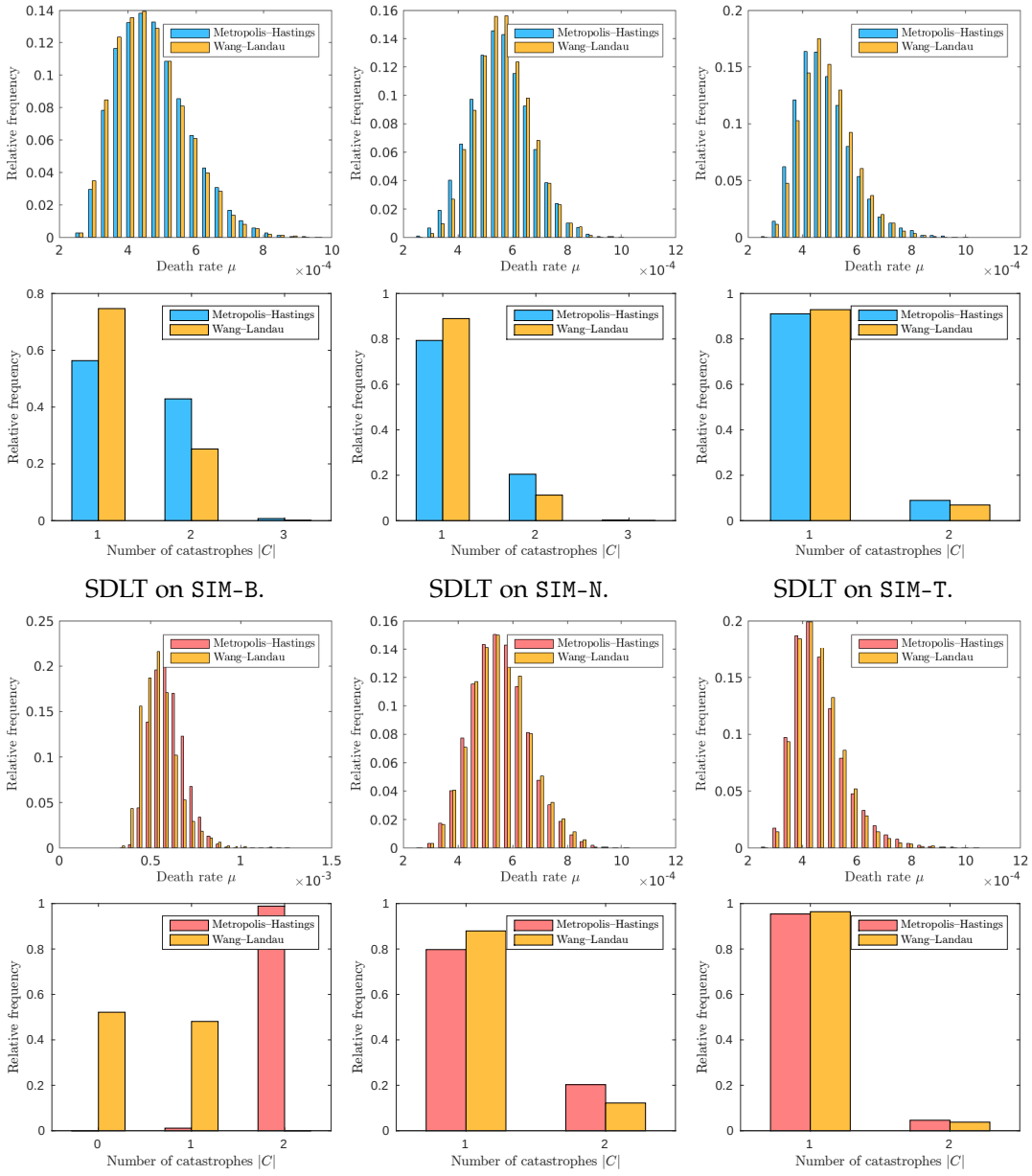
In Figures 5.4 and 5.5, we see that the SD model fit to `SIM-B` reports a much lower root age $-t_1$ than the SDLT model. Similarly, the death rate μ is inflated under the SD model as it attempts to account for the laterally transferred traits. Both models fit well to `SIM-N`. Posterior estimates of the relative transfer rate β/μ in Figure 5.5 are consistent with their true values in the generation of `SIM-B` and `SIM-N`. On `SIM-T`, the posteriors resemble mixture distributions, which is not surprising given the nature of the data. There is no cause for concern among the histogram plots for the remaining parameters, which we defer to Appendix B.1, nor the trace and sample autocorrelation plots either. We now assess the quality of our analyses using the Wang–Landau algorithm from Chapter 3 and the goodness-of-fit tests in Section 5.1.

i	Node constraint	Time constraint	
		$\Gamma^{(i)}$	$\Gamma^{(i')}$
1	'1'	$[-550, -450]$	$[-800, -200]$
2	'8'	$[-150, -50]$	$[-400, +200]$
3	pa('6', ..., '10')	$[-500, -200]$	—

Table 5.3: Clade constraints in our goodness-of-fit analyses of the synthetic data sets. For constraint 3, we remove the specific time constraint and leave the ancestry constraint.

The catastrophe severity κ is below the threshold we set in Chapter 2 so we do not expect to infer more than two catastrophes, the number on the true tree. We now use the Wang–Landau algorithm to assess whether our Metropolis–Hastings inferences are based on chains which mimic the overall trait process through catastrophes. We summarise samples of the death rate μ and total number of catastrophes $|C|$ for each combination of model, data set and sampling algorithm in Figure 5.6. For the samples generated by the Wang–Landau algorithm, we estimate the histogram bin frequencies according to the importance sampling estimator in Equation 3.3.7 with the bin penalties in Figure 3.3. As we remark previously, the SD model is seriously misspecified on `SIM-B` and the Metropolis–Hastings algorithm appears unable to escape a mode with two catastrophes, with a small resulting effect on the death rate μ. For the remainder of the figures, there is little to distinguish between the Metropolis–Hastings and Wang–Landau samples.

Each clade constraint in Figure 5.2 restricts the time t_i of a node $i \in V$ to an interval $[\underline{t}_i, \bar{t}_i] \subset \mathbb{R}$. We label the clades 1, 2 and 3, and summarise these constraints and their relaxations for assessing goodness-of-fit in Table 5.3. We plot histograms of the node ages under each model in Figure 5.7 and report the corresponding Bayes factors (5.1.2) in Figure 5.8. Unsurprisingly, the SD model fit to `SIM-B` poorly predicts the leaf constraints. There is little to distinguish between the models fit to the other combinations of data sets and constraints as in each case they accurately predict the constraints in the analyses above.



SDLT on SIM-B. SDLT on SIM-N. SDLT on SIM-T.

SD on SIM-B. SD on SIM-N. SD on SIM-T.

Figure 5.6: Histograms of parameter samples in our Metropolis–Hastings and Wang–Landau analyses of the synthetic data sets.

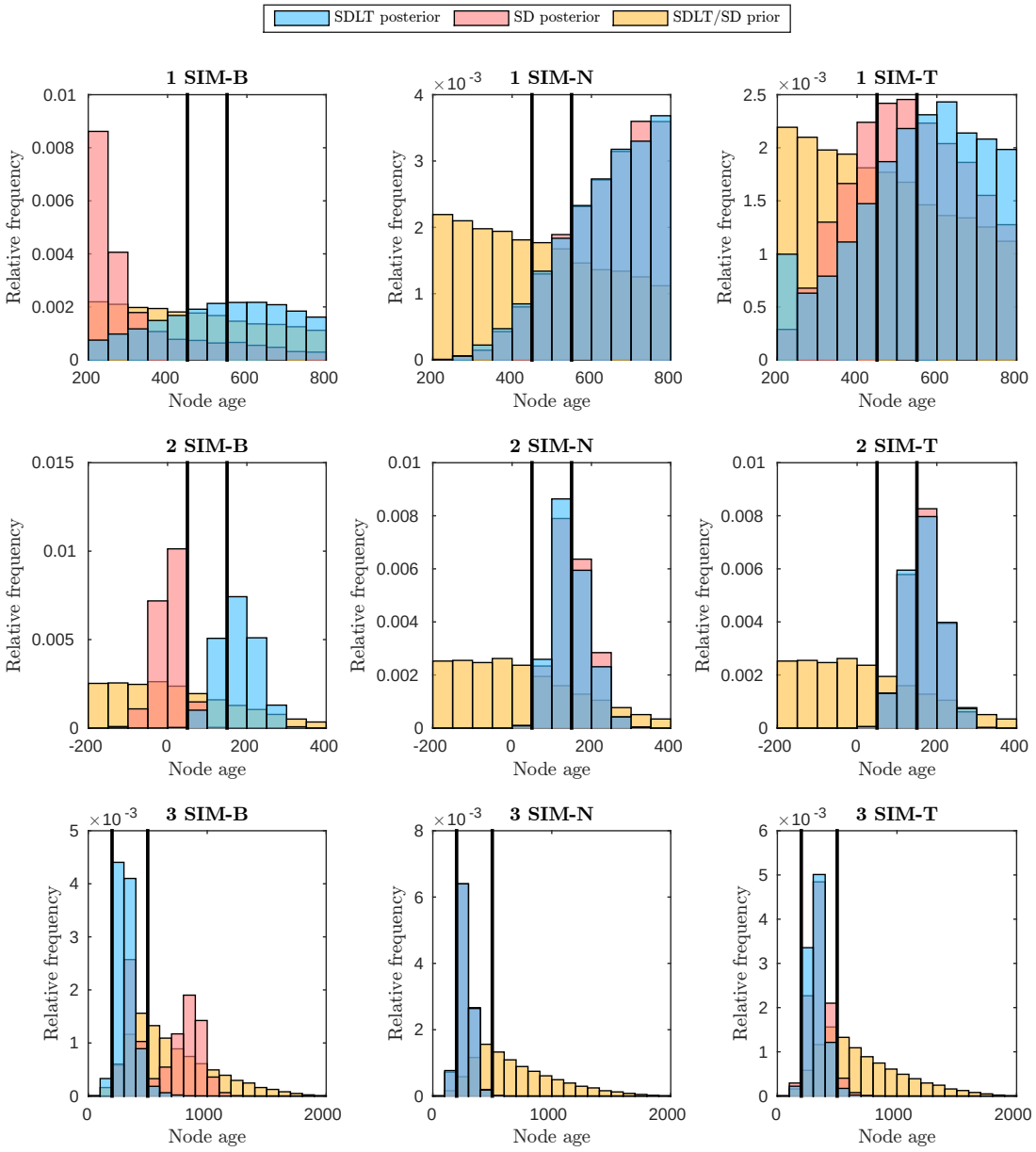


Figure 5.7: Histograms of marginal leaf times in our goodness-of-fit analyses of the synthetic data sets. The plot title describes the constraint and data set under analysis in each case. Vertical lines represent the time constraints in Figure 5.2.

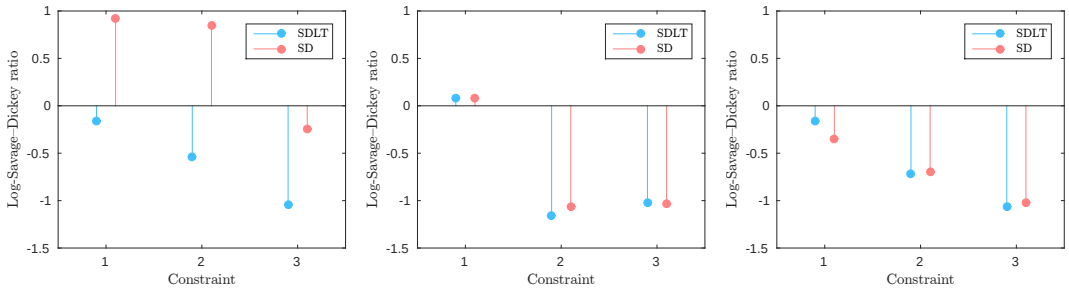


Figure 5.8: Bayes factors describing the lack of support for the clade constraints in the SDLT and SD models fit to the synthetic data sets.

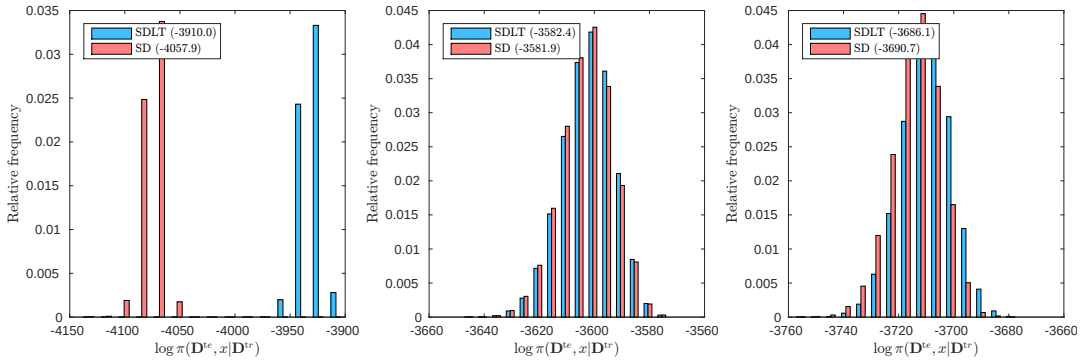


Figure 5.9: Posterior predictive model evaluation on the models fit to the synthetic data sets. We plot the joint posterior distribution of the test data and parameters given the training data. Predictive scores (5.1.3) are in parentheses.

We repeat the predictive performance checks in Section 5.1 when the registered data $R(\mathbf{D})$ is randomly split into evenly sized training and test portions. For the plots in Figure 5.9, the training sets are `SIM-B`, `SIM-N` and `SIM-T` in the analyses above with the corresponding coupled test sets drawn from the same distributions. The more flexible SDLT model outperforms the SD model on each data set, albeit marginally in the case of `SIM-N`.

On the basis of these analyses, we are satisfied that our inference scheme is correct, and that the SDLT model is identifiable, consistent with the SD model in the absence of lateral transfer and robust to a common form of model misspecification. In each case, parameter samples from the SDLT model are consistent with their true values in

Figure 5.2 and Table 5.1. We cannot say the same for the SD model fit to SIM-B, and this is confirmed by our goodness-of-fit assessment. Our analyses of SIM-B and SIM-T are particularly relevant for the following chapter where we analyse a data set with known-transferred traits present and another data set where the known-transferred traits have been removed.

5.4 TraitLab software

We implement the Stochastic Dollo with Lateral Transfer model in TraitLab (Nicholls et al., 2013), a freely available software package for fitting the Stochastic Dollo model and its extension to missing data and rate heterogeneity (Nicholls and Gray, 2008; Ryder and Nicholls, 2011). We shall publicly release this code upon publication of the corresponding paper (Kelly and Nicholls, 2016). In short, the software reads in the data set and clade constraints in a Nexus-formatted file, and the user specifies the remaining constraints on the MCMC algorithm either in an instruction file or through the graphical user interface. The output files are in a standard format and may be analysed in TraitLab and, with some slight modifications, in other phylogenetic software packages such as BEAST (Drummond et al., 2012) and DensiTree (Bouckaert and Heled, 2014).

The code for the exact expected pattern frequencies calculation is primarily in Matlab while we call the C code for the accelerated equivalence class-based approximation through the MEX interface in Matlab. On top of the standard C libraries, we use OpenMP to compute the equivalence class-based approximations (4.2.3) in parallel. On the basis of extensive profiling, we are satisfied that the C code is thread- and memory-safe and does not leak memory.

For the exact and approximate expected pattern frequency calculations, we compare their computed values with separate implementations and manual calculations on relatively small trees with every combination of isochronous and offset leaves,

catastrophes and missing data. We also implement the approximate expected pattern frequencies calculation in `Matlab`. The approximate expected pattern frequencies $\tilde{\mathbf{x}}$ in `C` differ from their `Matlab` counterparts by negligible amounts due to the different ordinary differential equation solvers in use. These differences are well within the specified relative (10^{-3}) and absolute (10^{-6}) error tolerances. The approximate `Matlab` code produces identical output to the exact expected pattern frequencies when we include equivalence class contributions from all patterns on an interval. The `Matlab` and `C` likelihood calculations provide identical output given the same input expected pattern frequencies. Ryder (2009) describes exhaustive tests of the SD likelihood calculation in `TraitLab`. When the lateral transfer rate $\beta = 0$, the output of the SDLT likelihood computation using the exact expected pattern frequencies is identical to the SD likelihood.

Chapter 6

Applications

Chapter overview

Computational phylogenetic methods have outgrown their original biological setting and are now used to estimate the phylogenies of everything from Iranian textiles (Tehrani et al., 2010) to folktales (Ross et al., 2013) and Paleoindian points (O'Brien et al., 2001). In this thesis, we focus on their application to problems in historical linguistics, namely inferring language family phylogenies from lexical trait data. We illustrate our model on two data sets. The first is relatively small and allows us to perform a thorough analysis of the data and inference schemes in Chapter 3. The second data set is much larger and we perform approximate inference here according to the method we describe in Chapter 4.

6.1 Language family phylogenies

Swadesh (1950, 1952, 1955) presents lists of words which, in his opinion, we may reasonably expect to observe in every language, such as those in the example in Section 2.1 and the data sets we analyse in this chapter. Swadesh (1950, 1952, 1953, 1955), and many others since, compares words in these lists across languages as to

infer their phylogeny using two now widely discredited methods: *lexicostatistics* and *glottochronology*.

Lexicostatistics is a distance-based method. To perform a lexicostatistical analysis, we first form an array C containing the percentages of shared traits between each pair of languages. We then cluster the languages to form a tree (Dyen et al., 1992). Glottochronology is an attempt to extend the lexicostatistic approach to infer dated trees. Under the assumption that languages retain a constant percentage r of characteristics from one time interval to the next, Swadesh (1955) estimates the times t_i and t_j since two languages i and j diverged from a common ancestor as

$$t_i + t_j = \frac{\log(C_{i,j})}{\log(r)}, \tag{6.1.1}$$

where $C_{i,j}$ is the corresponding entry in the lexicostatistic data array.

Hoijer (1956) and Bergsland and Vogt (1962) criticise lexicostatistics and glottochronology from a linguistic perspective. Chrétien (1962) attempts to debunk glottochronology mathematically but Dobson et al. (1972) criticise his attempts at mathematical rigour. In spite of this, there are many criticisms which can level at lexicostatistics and glottochronology and the assumptions they methods make. In addition to the criticisms of non-model-based methods in Chapter 1, much information is lost in going from the raw data to an array of percentages, and we cannot account for missing data in a principled manner. Most relevant to this thesis is the inability of these methods to properly control for lateral transfer. For example, Swadesh (1955) proposes to include a term $\bar{s}$ for the average degree of separation between languages so that Equation 6.1.1 becomes $t_i + t_j = \log(C_{i,j})/[\bar{s}\log(r)]$. The argument then becomes somewhat circular as we cannot calculate $\bar{s}$ without $t_i + t_j$.

The application of Bayesian phylogenetic methods in historical linguistics is a more recent innovation. The *restriction site* model is a two-state vertical model of trait evolution which allows for *homoplasy*, or *back mutation*, and is a feature of many phylogenetic analyses of linguistic data sets (Gray and Atkinson, 2003; Chang et al.,

2015; and many others). Under this model, the total number of traits is fixed, and a trait may transition between states 0 and 1 without restriction. For trait presence/absence data, this means that an instance of a trait may transition between present and absent states multiple times on a tree. This model does not explicitly account for lateral transfer but may generate trait histories synonymous with it. In spite of Farris (1983), who recommends the use of parsimonious phylogenetic models which minimise the requirement of "ad hoc hypotheses of homoplasy", the restriction site model remains a popular tool when inferring language phylogenies.

Gray and Atkinson (2003) apply the restriction site model of character evolution to a set of presence/absence data on Indo-European languages. The authors claim their inference is robust to laterally transferred traits in the data and that their inferred root time supports the *Anatolian Hypothesis* that civilisation entered Europe through Asia Minor approximately 9000 years BP. The restriction site model does not allow for the birth of new trait classes or the extinction of existing traits, something which occurs in practice, so it is potentially a poor model of lexical trait data. The Stochastic Dollo model is much more realistic in this setting. The analyses of Nicholls and Gray (2008) and Ryder and Nicholls (2011) both return root age ranges which do not contradict Gray and Atkinson.

The *covarion model* is an extension of the restriction site model to fast- and slow-evolving traits. Bouckaert et al. (2012, 2013) implement both the covarion and Stochastic Dollo models in their *phylogeographic* study of Indo-European languages. We do not comment on the geographic aspect of their study other than to say that their results also support the Anatolian Hypothesis.

Chang et al. (2015) propose an *ancestry-constrained* model whereby internal nodes may appear in the data — Latin as the root of the Romance languages, for example. The authors instead fit the restriction site and covarion models, and claim that their results support the *Steppe Hypothesis* that civilisation entered Europe from Russia at a much later date than that proposed by the Anatolian Hypothesis, approximately

6000 years BP. As we establish above, the restriction site model and, by extension, the covarion model are poor models for lexical trait data. Chang et al. are unable to fit the Stochastic Dollo model without lateral transfer here because traits marked absent in an ancestral state but present in both descendent and non-descendent leaves violate the Dollo parsimony assumption.

Other studies of note include Skelton's analysis of Linear B dialects, although the methods she uses are not model based (Skelton, 2008); the restriction site and Stochastic Dollo analyses of Semitic languages by Kitchen et al. (2009) and Nicholls and Ryder (2011) respectively; and the phylogeographic analysis by Gray et al. (2009) which claims that the Austronesian language family originated in Taiwan.

Lateral transfer is rife in language diversification (Greenhill et al., 2009) yet a common feature of the above studies is that the authors discard known-transferred traits and fit a vertical model to the remainder. This is a primary criticism among linguists of statistical approaches to inferring language phylogenies. Although some authors claim to use traits which are resistant to lateral transfer (Pagel et al., 2013), they do not address the sampling bias inherent in this approach. To address these issues, we now apply our Stochastic Dollo with Lateral Transfer model to two data sets, one with laterally transferred traits present in the data and the other with known-transferred traits discarded.

6.2 Eastern Polynesian

Eastern Polynesia is a collection of archipelagos in the South Pacific ranging over a vast area from New Zealand in its south-western corner, Rapanui (Easter Island) in the south east and Hawaii at its northern end. Unlike Western Polynesia, the order and timing of human settlement in Eastern Polynesia is a matter of debate.

In the standard subgrouping of the Eastern Polynesian languages, Rapanui diverges first, followed by the split leading to the Marquesic (Hawaiian, Mangarevan,

Marquesan) and Tahitic (Manihiki, Maori, Penrhyn, Rarotongan, Rurutuan, Tahitian, Tuamotuan) languages (Marck, 2000). This theory has recently been challenged in light of new linguistic and archaeological evidence. In an implicit phylogenetic network study of lexical traits, Gray et al. (2010) detect non-tree-like signals in the data and the Tahitic and Marquesic languages do not form clean clusters. From a meta-analysis of radiocarbon dates, Wilmshurst et al. (2011) claim that the islands of Eastern Polynesia were settled in two distinct phases: the Society Islands (of which the largest is Tahiti) between 900 and 1000 years before the present (BP) and the remainder between 700 and 900 years BP. These dates are much later than previously thought (Spriggs and Anderson, 1993) and do not allow much time for the development of the Eastern Polynesian language subgroups. Conte and Molle (2014) present evidence of human settlement in the Marquesas Islands approximately 1100 years BP. On the basis of the above and further evidence of lateral transfer in primary source material, Walworth (2014) disputes Marquesic and Tahitic as distinct language subgroups.

The early Polynesians were able sailors and navigators (Lewis, 1964; Irwin, 2008) so the vast distances separating the archipelagos of Eastern Polynesia did not pose a barrier to contact between the various populations in the region. This statement is supported by archaeological evidence (Walter and Sheppard, 1996; Weisler and Kirch, 1996; Weisler, 1998). It is therefore unlikely that the languages of Eastern Polynesia evolved in isolation and possess a tree-like phylogeny. To add to this debate, we illustrate the SDLT model on lexical traits in eleven Eastern Polynesian languages drawn from the approximately 1200 languages in the Austronesian Basic Vocabulary Database (Greenhill et al., 2008). We compare our results with the SD model to highlight the effect of the laterally transferred traits in the data. The data is a subset of the Polynesian language data in the study of Gray et al. (2010).

In Figure 6.1, we plot the data encoded according to Section 2.1. We analyse the 968 traits marked present in at least one of the eleven languages, hereafter referred to as `POLY-0`. The languages are modern so the data are isochronous. Consistent with

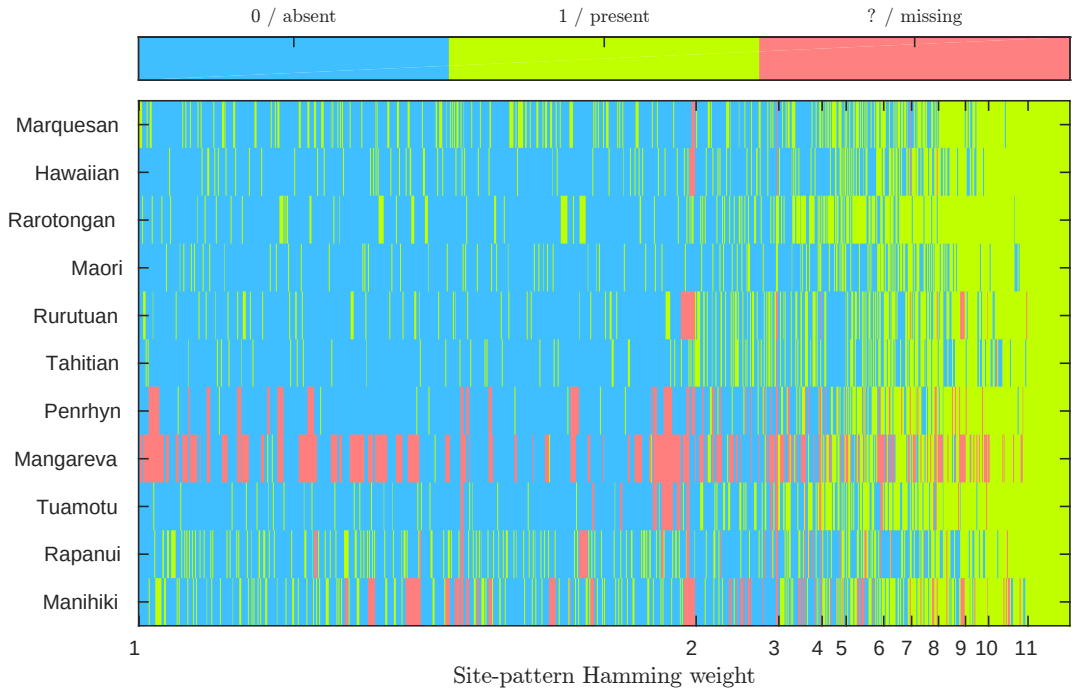


Figure 6.1: Eastern Polynesian language data set `POLY-0`.

Gray et al. (2009), the sole clade constraint limits the root of the tree to be between 1150 and 1800 years BP. We fit the exact SDLT model according to Chapter 3, the exact-approximate SDLT of Chapter 4, and the SD model of Ryder and Nicholls (2011) with the addition our proper prior on the death rate μ and the threshold on the catastrophe severity κ.

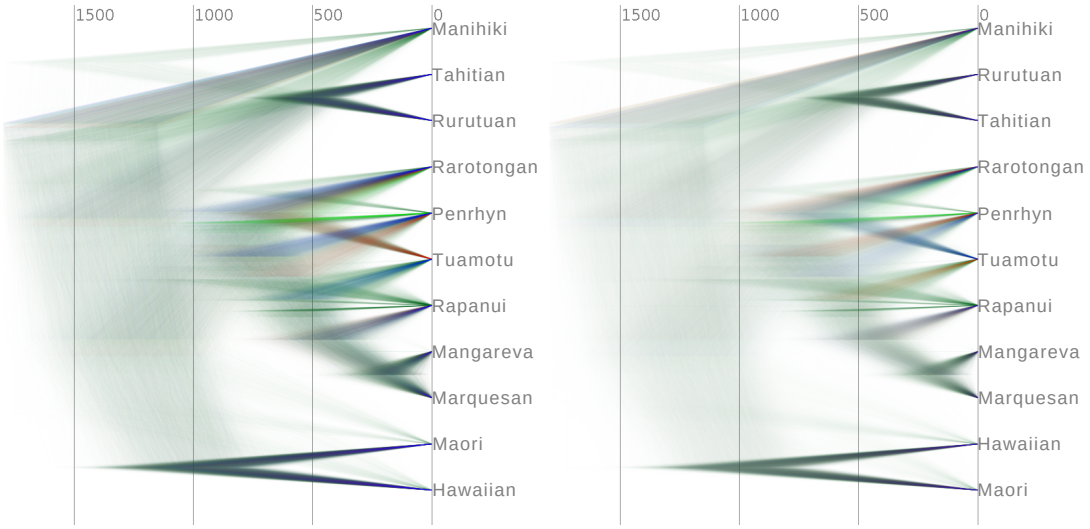
We plot samples from the marginal tree posterior under the respective models in Figure 6.2. We summarise these distributions with *majority rule consensus trees* in Appendix B. Once again, our samples from the exact-approximate chain are indistinguishable from the exact chain, with the exception of marginally higher effective sample sizes in the exact chain, which we expect anyway. We focus on the exact SDLT and SD models from here on.

In agreement with Gray et al. (2010) and Walworth (2014), we do not infer the traditional Marquesic and Tahitic subgroupings of Marck (2000) as distinct subtrees in either model. In contradiction with the above studies, Rapanui does not appear as an

outgroup in either model. The majority of the uncertainty under the SDLT model is in the topology of the subtree containing Rarotongan, Penrhyn, Tuamotu, Rapanui, Mangareva and Marquesan. This subtree also has 100% posterior support under the SD model, but here most of the uncertainty is in the relationships further up the tree. We obtain the 95% highest posterior probability sets for the tree topologies using `BEAST` (Drummond et al., 2012). This set comprises 135 topologies for the SDLT model and 19 for the SD model. This level of confidence in so few topologies is a likely result of the SD model's misspecification on the laterally transferred traits.

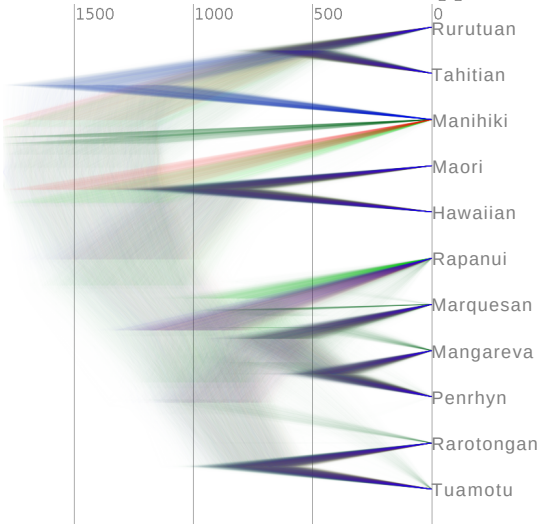
The effect of the laterally transferred traits is evident again in the histograms of samples from the marginal posterior distributions of the death rate μ and relative transfer rate β/μ in Figure 6.3. The death rate is approximately 50% higher under the SD model as traits must be born further up the tree and killed off at a higher rate to explain the variation in the data. Recall from Section 2.5.1 the relative transfer rate β/μ, the expected number of times that an instance of a trait attempts to transfer before dying. The posterior distribution of β/μ in Figure 6.3 is well informed and centred on 1.35, so it is unsurprising then that the SD model struggles here. We report histograms for the remaining parameters in Appendix B, as well as the trace and autocorrelation plots to diagnose convergence.

The SDLT posterior distribution on the root time t_1 in Figures 6.2 and 6.3 is approximately uniform across its predefined range, as there is little information in the data to push this term away from its approximately uniform prior distribution in Section 2.5.1. This is our only time restriction on an ancestral node in the model so, coupled with the lack of rate-time identifiability in the prior, we cannot support or refute any claims about the absolute timing of events. We can infer relative timings, however. Were we to restrict our model to the root time range that Wilmshurst et al. (2011) propose, namely 900–1000 years BP, then it is plausible that we would infer that the splits leading to the main subgroups in the SDLT tree in Figure 6.2 — Manihiki; Tahitian and Rurutuan; Maori and Hawaiian; and Rarotongan, Penrhyn, Tuamotu,



Exact SDLT model.

Exact-approximate SDLT model.



SD model.

Figure 6.2: `DensiTree` (Bouckaert and Heled, 2014) plots of the marginal tree posteriors in our analyses of `POLY-0`. Time is in years BP and the most frequently sampled topology is coloured blue, followed by red then green and the remainder in dark green.

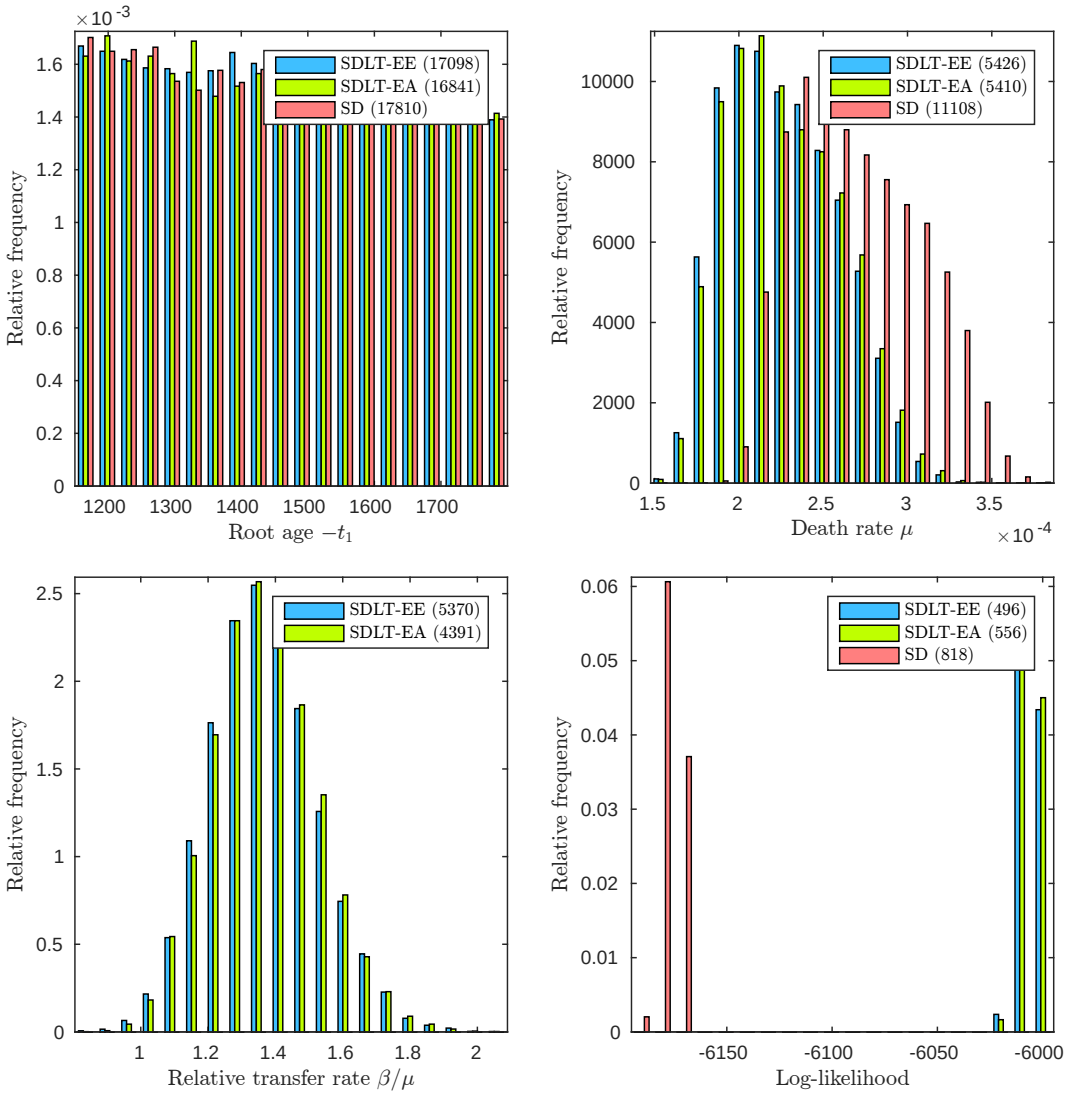


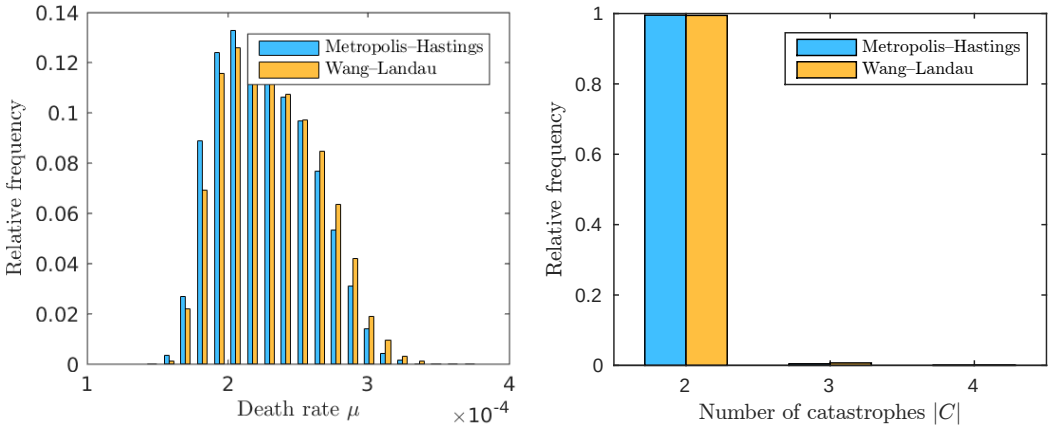
Figure 6.3: Marginal posterior distributions of the root age $-t_1$, death rate μ, relative transfer rate β/μ and model likelihood in our analyses of `POLY-0`. Effective sample sizes are in parentheses.

Rapanui, Mangareva and Marquesan — occurred between 700 and 900 years BP. One could argue that a pulse-pause model of expansion such as that which Gray et al. (2009) describe would resolve this problem without restricting the root time. If the data is unable to inform the root time, then it is likely that any inference about pulses and pauses would be dominated by our choice of prior. As this inference is entirely subjective and others have reported dates which conflict with Wilmshurst et al., we do not make any further remarks on this subject.

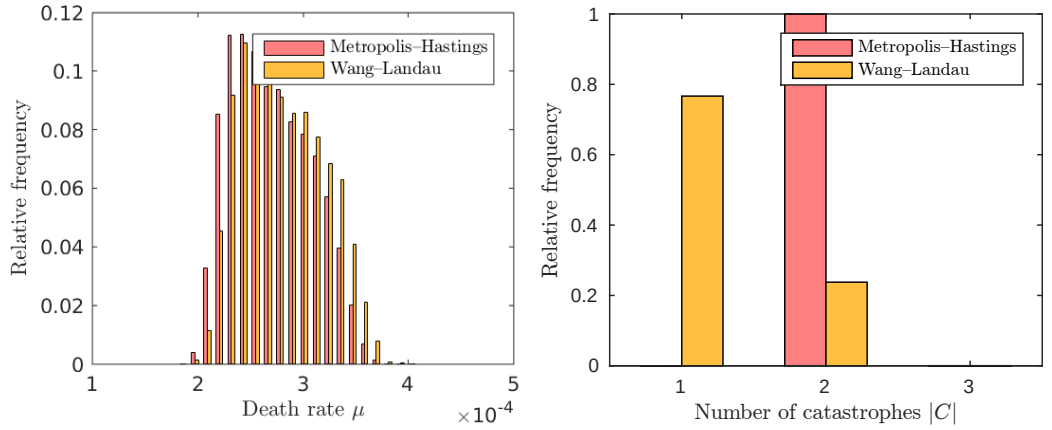
For our first assessment of goodness-of-fit, we repeat our analyses using the Wang–Landau algorithm from Chapter 4. We compare histograms of samples from each combination of model and sampling algorithm in Figure 6.4. As we would expect, the SD model appears unable to escape a local mode with two catastrophes when sampling under the Metropolis–Hastings algorithm and it infers a slightly lower death rate μ as a result.

We cannot relax the calibration constraint on the root time without compromising our overall inference. Instead, we relax each of the leaf constraints in turn and compute a Bayes factor according to Equation (5.1.2) to compare the corresponding relaxed and constrained models. The constraint $\Gamma^{(i)} = \{g \in \Gamma : t_i = 0\}$ fixes leaf $i \in V_L$ at time 0 and $\Gamma^{(i')} = \{g \in \Gamma : -10^3 \leq t_i \leq 10^4\}$ denotes its relaxation to a wide interval around the present. Once again, a large Bayes factor here indicates a lack of support for the leaf constraint and is a sign of model misspecification.

In Figure 6.5, we report log-Savage–Dickey ratios corresponding to the marginal prior and posterior leaf time distributions in Figure 6.6. We cannot compute the Savage–Dickey ratio in closed form so we estimate the densities in Equation 5.1.2 by the proportions of sampled leaf times in the range $[-50, 50]$ around the present, time 0. We note that varying this time range does not significantly affect our results. The SD model does not support the leaf time constraints on Manihiki and Marquesan so we replace the corresponding Bayes factors by lower bounds. In addition to this, the



SDLT.



SD.

Figure 6.4: Histograms of parameter samples in our Metropolis–Hastings and Wang–Landau analyses of POLY-0. For the samples generated by the Wang–Landau algorithm, we estimate the histogram bin frequencies according to the importance sampling estimator in Equation 3.3.7.

SD model poorly predicts the constraint on Rapanui, returning a log-Bayes factor of 2.5. In contrast, the SDLT model performs well in predicting all of the constraints.

Traits marked present in a single language are often deemed unreliable and removed in the registration process. To address this concern, we repeat our analyses on the data set POLY-1 which we form by discarding the *singleton* patterns from POLY-0. In Figure 6.7, we report the results of the posterior predictive tests we describe in Chapter 5 on random splittings of POLY-0 and POLY-1 into training and test sets. The SDLT model outperforms the SD model in each case. Although the outcome of this

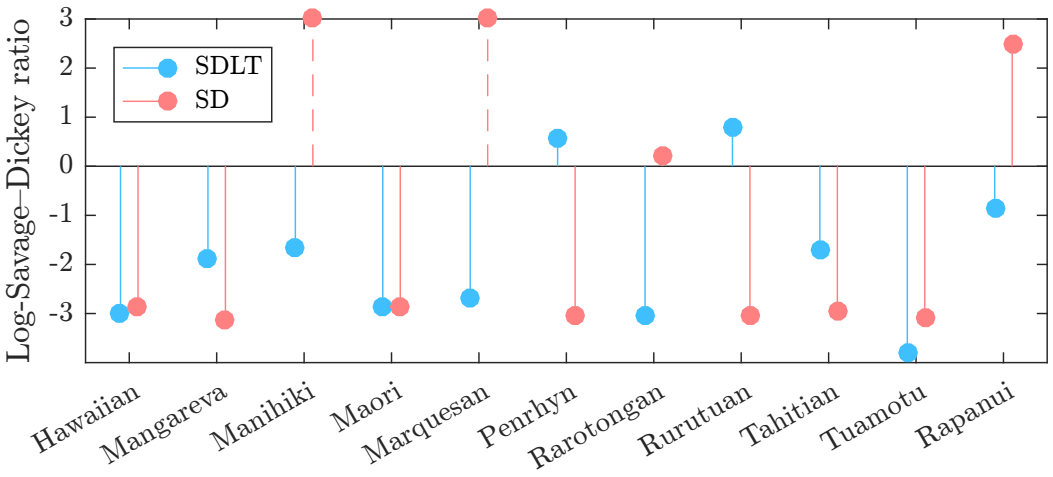


Figure 6.5: Bayes factors comparing the support for the leaf constraints used to fit the SDLT and SD models to `POLY-0`. The SD model does not support the constraints on Manihiki and Marquesan so we plot lower bounds on the Bayes factors instead.

model selection step is unchanged, these singleton patterns play an important role in SDLT model inference because the their removal affects the inferred parameter credible intervals.

To conclude this section, on the basis of our goodness-of-fit tests, the SDLT model fits the Eastern Polynesian data of Greenhill et al. (2008). The SD model performs poorly in comparison and strongly rejects three of the eleven constraints in our goodness-of-fit test in Figure 6.5. We leave as future work to investigate further the inferred bimodal distribution on the age of Rapanui in Figure 6.6 and to repeat our analyses with different choices of constraints, such as Rapanui as an outgroup, for example.

6.3 Indo-European

Gray and Atkinson (2003) analyse the 84-language Indo-European data set of Dyen et al. (1992), who originally study it in a lexicostatistical setting, with the addition of three ancient languages: Hittite, Tocharian A and Tocharian B. Ringe et al. (2002) describe a data set on 24 Indo-European languages, 20 of whom are ancient. A

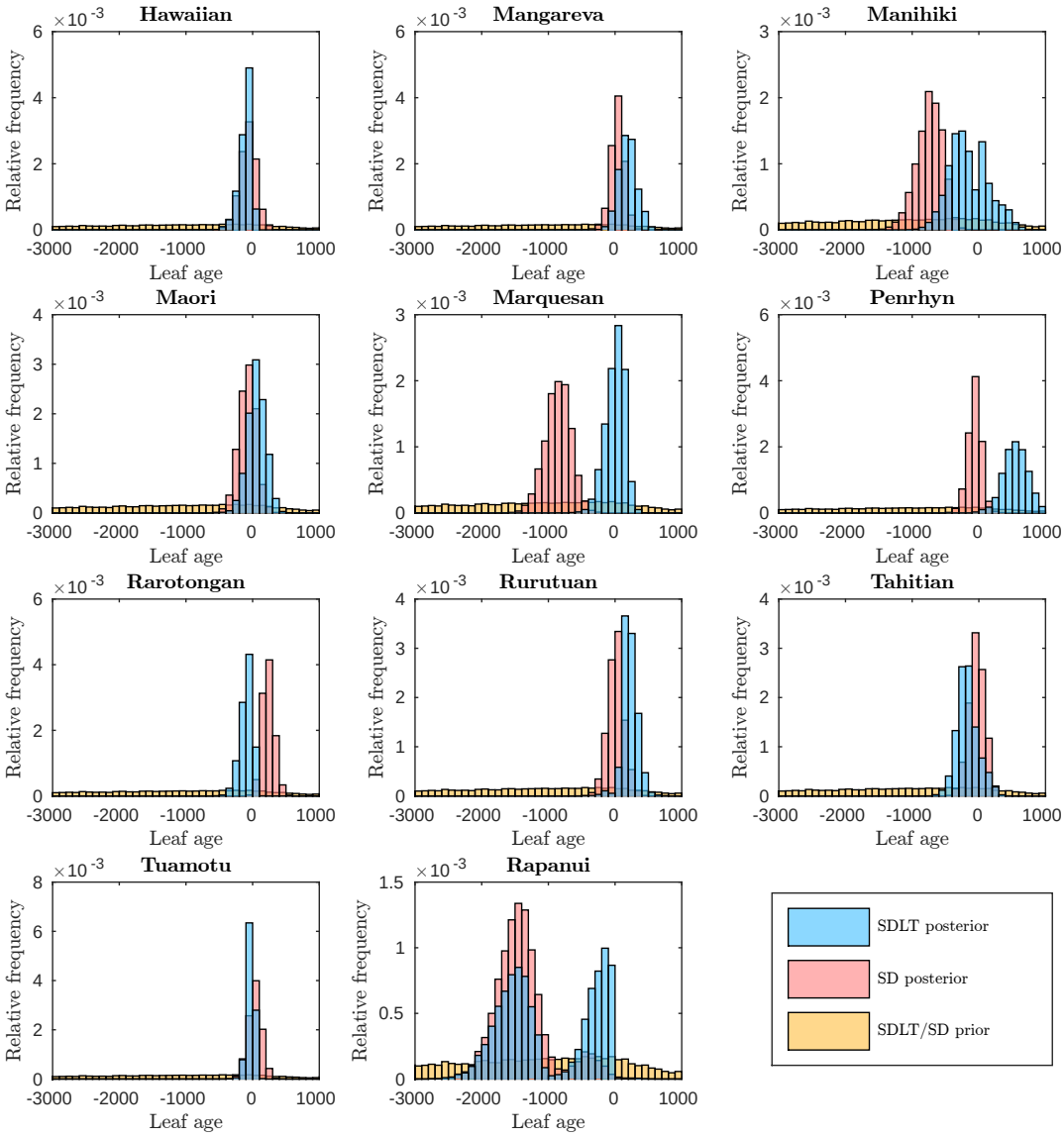
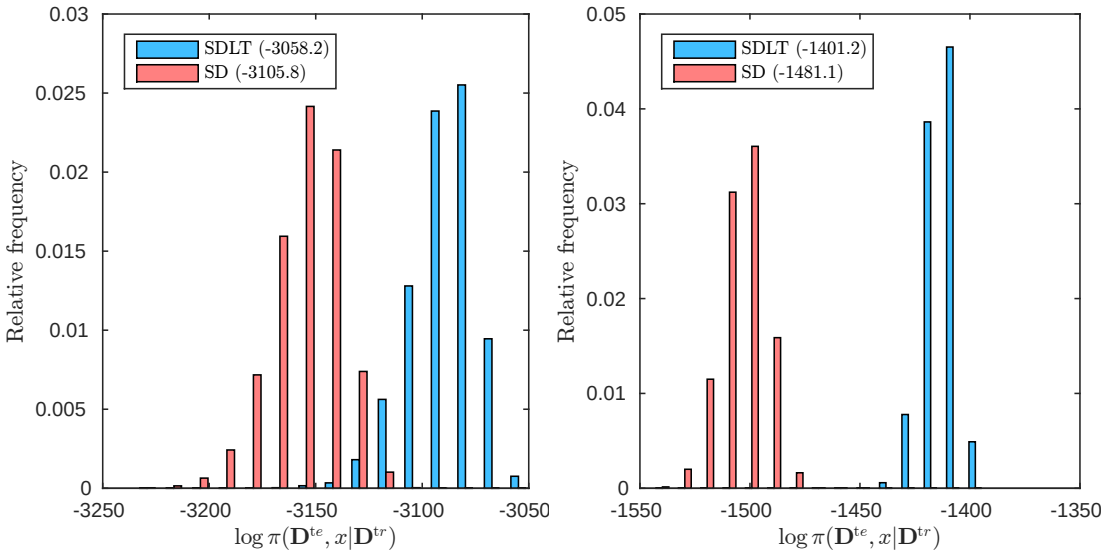


Figure 6.6: We relax the constraint on each leaf in turn and compute the histogram of its time under the prior and posterior for each model fit to `POLY-0`. Time in years is on the horizontal axis and relative frequency on the vertical axis.



(a) (b)

Figure 6.7: Assessing posterior predictive performance of the SDLT and SD models fit to (a) `POLY-0` and (b) `POLY-1`. In each case, we plot the joint posterior distribution of the test data and parameters given the training data. The predictive scores, obtained upon integrating out the parameters, are in parentheses.

common feature of both of these data sets is the absence of known-transferred traits. Nicholls and Gray (2008) and Ryder and Nicholls (2011) analyse both of these data sets under their respective Stochastic Dollo models. The data base `IELex`[1] combines and expands upon these data sets. Bouckaert et al. (2012) analyse data from `IELex` without known-transferred traits in their phylogeographic study of Indo-European. Chang et al. (2015) also analyse data from `IELex`. In addition to repeating the analyses of Bouckaert et al., they analyse the data with and without known-transferred traits. Chang et al. infer a slightly lower root age when known-transferred traits are included.

In this section, we illustrate our approximate inference procedure from Chapter 4 on the same Indo-European data set (Ringe et al., 2002) as Nicholls and Gray (2008) and Ryder and Nicholls (2011). We refer to this data set as `IE` below. The data records the state of 872 traits across $L = 24$ taxa. This number of taxa is essentially the limit of

[1]The current web address for `IELex`, http://ielex.mpi.nl, is marked out-of-date as of December 2015.

what we can analyse under the approximate SDLT model on the current hardware at our disposal. This is also the reason why we do not repeat the analyses of Chang et al. (2015) as their data sets contain many more taxa, and subsampling the taxa may distort our analysis, particularly with lateral transfer.

With 20 offset leaves in the data set, we must compute the expected pattern frequencies across at least $20 + L - 1$ intervals between branching, branch death and catastrophe events at every iteration of the MCMC analysis. In each step of the likelihood parameter calculation, we compute estimates $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ to the exact pattern frequencies $\mathbf{x}$ and then perform one Aitken–Jennings acceleration step to produce the estimate $\tilde{\mathbf{x}} = \mathbf{v}^{(0)}$ we use in practice. The time to approximate the likelihood parameters with method is approximately eight seconds. In contrast, the time to compute the likelihood parameters exactly using the pattern-based ODE approach is over five minutes on the same hardware. From the representation in Figure 6.8, we see that the majority of the traits are observed in a single language and that there is a large amount of missing data, particularly in Luvian, Old Persian, Umbrian, Lycian and Oscan; and many traits are missing in all of these languages. This further increases the cost of the likelihood calculation as we must sum up contributions from 2^c binary patterns to compute the expected frequency of traits displaying a pattern with c missing entries (2.4.2).

For our MCMC analyses, we place time and ancestry constraints on ten of the 22 ancestral nodes below the root, and discard traits not marked present in at least one taxon. Due to the increase in computation time, we draw fewer samples from the SDLT posterior here than in the analyses of the synthetic data sets in Chapter 5 and Eastern Polynesian data sets in Section 6.2. In addition to this, the mixing times for the MCMC chains increase as the size of the tree increases, so our parameter effective sample sizes are lower again.

We report marginal posterior distributions on trees in Figure 6.9, marginal parameter posterior distributions in Figure 6.10. Although known-transferred traits are

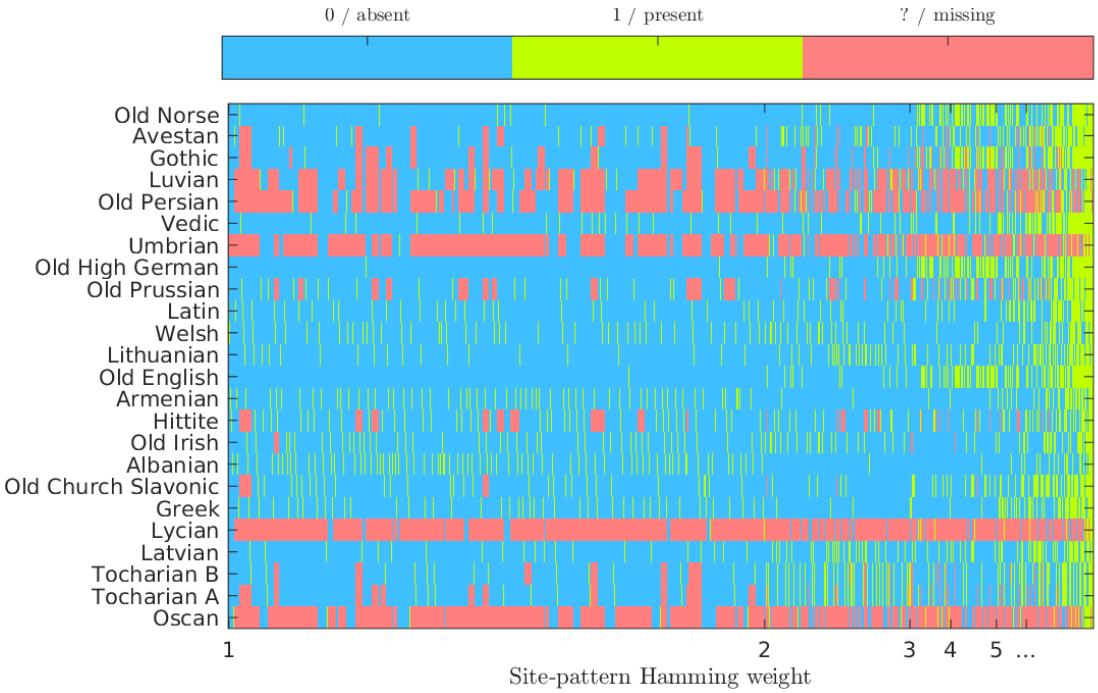
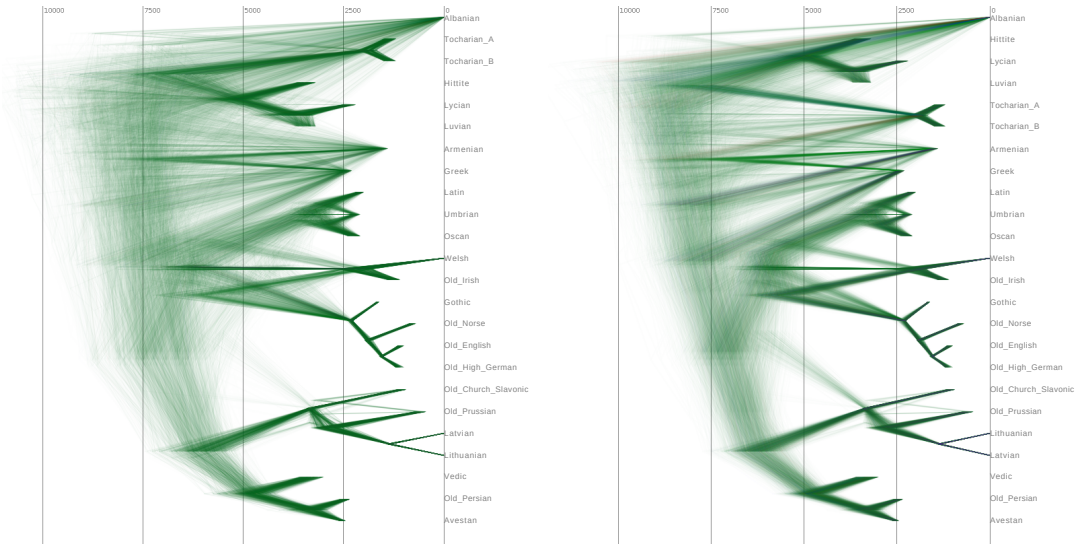


Figure 6.8: Indo-European data set `IE` of Ringe et al..

absent from the data, we infer a relative transfer rate centred on 0.15. This suggests the presence of unidentified transfers in the data. The primary statistic of interest here is the root age $-t_1$, however. The SDLT model posits a slightly lower root age than the SD model but not significant to support the Anatolian hypothesis. Therefore, much further analysis is required before we can make a definitive statement on the validity of the Anatolian and Kurgan hypotheses with respect to the SDLT model.

With so few samples from the SDLT model and a large number of clade constraints, we cannot effectively distinguish between the respective tree distributions in Figure 6.9 or the consensus trees in Appendix B. The SD model here returns fewer catastrophes and a slightly greater root age than in the analysis by Ryder and Nicholls (2011), possibly due to the threshold we enforce on the catastrophe severity κ.

The computational cost of the goodness-of-fit test in Section 5.1 whereby we relax each clade constraint in turn is prohibitive for this data set. Instead, we restrict ourselves to comparing the predictive scores (5.1.3) under the approximate SDLT and



Approximate SDLT model. Exact SD model.

Figure 6.9: `DensiTree` (Bouckaert and Heled, 2014) plots of the marginal tree posteriors in our analyses of `IE`. Time is in years BP and the most frequently sampled topology is coloured blue, followed by red then green and the remainder in dark green.

exact SD models. Although the SD model simulations on the training data set are exact, we estimate the likelihood parameters in this test according to the approximation scheme we outline above. We report the results of this test in Figure 6.11, and we observe a significantly better fit under the approximate SDLT model.

To conclude, our primary aim in this section is to illustrate our approximate inference scheme rather than make inferences about the origins of the Indo-European language family. As such, these results represent a tentative, if promising, step towards a full analysis of the Ringe et al. data set, something that would take upwards of a year to complete on current hardware.

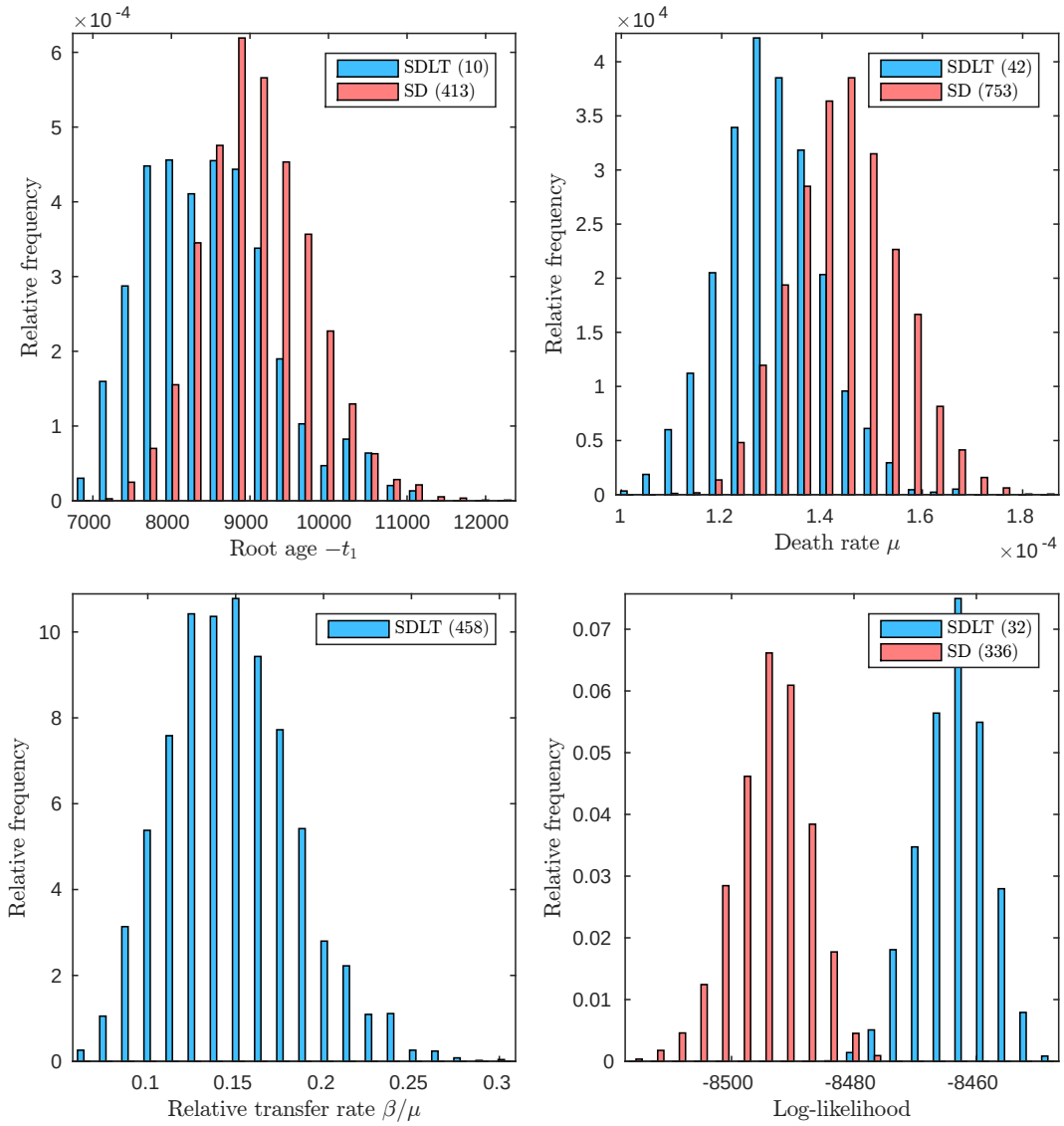


Figure 6.10: Marginal posterior distributions of the root age $-t_1$, death rate μ, relative transfer rate β/μ and model likelihood in our analyses of IE. Effective sample sizes are in parentheses.

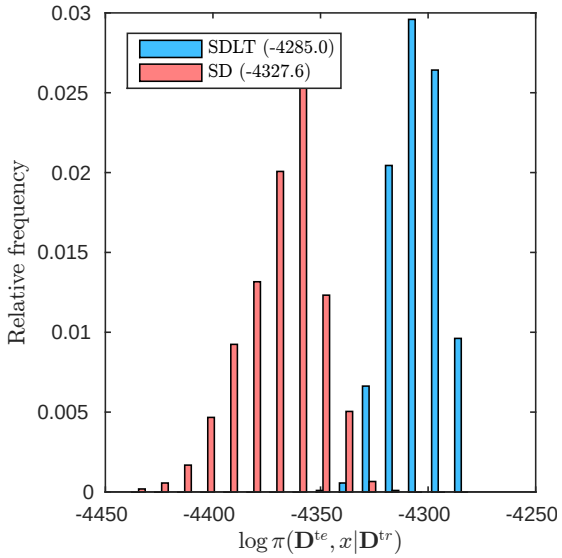


Figure 6.11: Posterior predictive model evaluation of the models fit to the IE. We plot the joint posterior distribution of the test data and parameters given the training data. The predictive scores (5.1.3) are in parentheses.

Concluding remarks

Lateral transfer is an important problem in phylogenetics but practitioners lack the statistical tools to perform fully likelihood-based inference for phylogenies in this setting. We address this issue with a novel model of species diversification which extends the Stochastic Dollo model for trait presence/absence data. To our knowledge, the method we describe is the first fully likelihood-based approach to control for lateral transfer in reconstructing a rooted phylogenetic tree. The second major contribution of this thesis is the inference procedure whereby we integrate out the locations of the trait birth, death and transfer events through a sequence of initial value problems on the tree. Our third major contribution is the scheme to efficiently and accurately approximate the likelihood parameters and perform unbiased MCMC inference. The primary drawbacks of the model we describe are:

- This computational cost of fitting the model is exponential in the number of taxa under consideration
- We assume that there are no unobserved lineages contributing to the evolutionary process
- We assume that traits evolve independently of each other

We first comment on these issues then consider how we might further improve upon our model.

The approximation we derive to the exact expected pattern frequencies reduces the computational cost of evaluating the likelihood by a constant factor and leads to an

almost commensurate gain in the number of effectively independent samples we can draw from our model per unit time. However, we do not overcome the exponential increase in computational cost as the number of taxa increases. This cost arises because the transition rates between pattern states depend on the Hamming weight of the pattern and the number of extant lineages.

One possible approach to this problem is to redefine the model such that pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ transitions to state $\mathbf{q} \in S^+_{\mathbf{p}}$ at rate β rather than $\beta s(\mathbf{p})/L^{(t)}$. It would appear that we can now effectively ignore the intermediate pattern states that a trait realises and decouple the expected pattern frequency calculation across branches. However, we still need to ensure that the trait does not display the empty pattern at any point in time. A second approach, which also changes the definition of the model, is to limit the number of events which can affect a trait on a branch — for example, we may state that a trait may only undergo one transition on each branch. This is similar to compartmental models in epidemiology, for example. Much further work is required to assess whether these models represent viable alternatives to our SDLT model.

A third possible approach to the exponential computational cost of fitting our model is to restrict the global regime of lateral transfer. Rather than allow traits to transfer between all contemporary lineages, we could instead attempt to infer disjoint clusters of lineages which communicate through trait transfers. At first glance this would appear to lead to a decrease in the computational burden as we need only model the expected evolution of the trait process across decoupled subgroups of branches. This is true for traits born after a split into communicating subgroups. However, as we must also account for traits born above any split into subgroups, this approach does not lead to a decrease in computational cost unless the largest subgroup of communicating lineages is small relative to the overall number of lineages.

In Chapter 2, we remark that when we calculate the likelihood, we must account for both the contributions of traits which survive to be recorded in the data as well as those which do not. Returning to the Poisson likelihood in Equation 2.3.6, this

means that we must calculate the total expected number of patterns and the expected frequency of each pattern we observe in the data. Møller et al. (2006) propose a clever trick to perform MCMC inference for distributions with intractable normalising constants. As we can directly simulate patterns from the SDLT model, we can apply their method here to cancel the terms involving the total expected pattern frequencies in the corresponding Metropolis–Hastings acceptance ratio (3.1.1). We now need only to compute the expected frequencies of patterns we actually observe. We leave this problem for future work.

The model we describe is not *projective* in the sense that we cannot marginalise the effect of unobserved lineages out of our inference analytically. Consequently, as the number of unobserved lineages increases, the probability of a trait transferring to a sampled lineage decreases; and a trait which previously died out on the sampled lineages may transfer back into the system from an unobserved lineage in conflict with our *Dollo* parsimony assumption. Dirichlet diffusion trees Neal (2001b) and Gibbs fragmentation trees McCullagh et al. (2008) define projective priors for trees evolving forwards in time. Investigating whether these methods may be adapted to include projective lateral transfer processes is a topic for future research. Szöllősi et al. (2013) study the effect of unsampled lineages in the context of the lateral transfer model of Szöllősi et al. (2012) which we describe in Chapter 1. The authors introduce additional unobserved *ghost* lineages to their tree and conclude that lateral transfers in their data primarily originate in unobserved lineages. Szöllősi et al. do not assess the goodness-of-fit of their inference — they do not consider whether their conclusions are a result of model misspecification, for example — so it is difficult to draw any concrete conclusions from this study. The primary difficulty in introducing ghost lineages to the SDLT model is the increased computational cost of inference.

In its current form, the SDLT model assumes that traits evolve independently of each other. This assumption may not always hold in practice as the presence of one trait in a taxon may decrease the likelihood of it possessing another trait. One

possible approach to this problem is provided by the extension of the SDLT model for multiple character states. For example, if a species may possess only one trait among a group of traits then, similar to Alekseyenko et al. (2008), we could model transitions between traits on top of the birth-death-transfer process we describe in this thesis. Naïvely implementing a model with K possible character states (equivalently, traits in categories) would lead to a sample space with $\mathcal{O}[(1+K)^{L^{(t)}}]$ possible patterns. Again, we leave this problem for future work.

One avenue for future work is to partition the data across a mixture of models and trees. Returning to the historical linguistics problems in Chapter 6, we could, for example, allow core vocabulary to transfer at a much slower rate than non-core vocabulary. Alternatively, the assignment of traits to models could be inferred as part of the overall inference. Pagel and Meade (2004) propose a finite mixture of models and data partitions to address this problem. Lartillot and Philippe (2004) use a Dirichlet process prior to model an infinite number of models and partitions.

Another direction for future work, related to the above idea of partitioning the data across models and of interest to practitioners, is to infer the most likely histories of the traits in a data set under the posterior. Compared to the cost of fitting the SDLT model, this would be a relatively simple optimisation problem whereby we would fix the tree and trait parameters and find the most likely trait history displaying a given pattern. A simpler idea would be to attempt to group patterns and, by extension, the traits which display them, *a posteriori* into those patterns which are likely to be a result of lateral transfer on the tree, and those which are not.

We allow for bursts of activity in the form of catastrophes as a surrogate for rate heterogeneity. To allow trait event rate parameters to vary across time and lineages would require a combination of time- and state-dependent differential equations to describe the evolution of the expected pattern frequencies in the likelihood. A simpler approach would be to allow the catastrophe severity to vary across catastrophes. We could also limit lateral transfer to occur at catastrophes only and restrict the set of

branches which can transfer traits at a given catastrophe, and thereby attempt to infer the time and direction of lateral transfer.

There are many open problems which have been ignored due to the expense of fitting models that account for lateral transfer. In Chapter 6, we describe one such example in the model of Chang et al. (2015). Our method provides a fully model-based solution to this problem and many others.

References

S.S. Abby, E. Tannier, M. Gouy, and V. Daubin. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinform.*, 11(1):324, 2010.

A.C. Aitken. On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinb.*, 46:289–305, 1927.

A.V. Alekseyenko, C.J. Lee, and M.A. Suchard. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.*, 57(5):772–784, 2008.

C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.*, pages 697–725, 2009.

C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal markov chain monte carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 2015.

C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, in press.

C. Andrieu, A. Doucet, and S. Yildirim. On an alternative class of pseudo-marginal algorithms. *Manuscript in preparation*, 2015.

Y.F. Atchadé and J.S. Liu. The Wang-Landau algorithm in general state spaces: applications and convergence analysis. *Statist. Sinica*, pages 209–233, 2010.

R.C. Ball, T.M. Fink, and N.E. Bowler. Stochastic annealing. *Phys. Rev. Lett.*, 91(3): 030201–1–4, 2003.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *ArXiv:1505.02827*, 2015.

M. Barlow. What antimicrobial resistance has taught us about horizontal gene transfer. In M.B. Gogarten, J.P. Gogarten, and L.C. Olendzenski, editors, *Horizontal Gene Transfer: Genomes in Flux*, pages 397–411. Humana Press, Totowa, 2009.

M.A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

R.G. Beiko and N. Hamilton. Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.*, 6(1), 2006.

K. Bergsland and H. Vogt. On the validity of glottochronology. *Curr. Anthropol.*, 3(2): 115–153, 1962.

L.J. Billera, S.P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.*, 27(4):733–767, 2001.

E.W. Bloomquist and M.A. Suchard. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. *Syst. Biol.*, 59(1):27–41, 2010.

A. Bouchard-Côté and M.I. Jordan. Evolutionary inference via the Poisson Indel Process. *Proc. Natl. Acad. Sci. USA*, 110(4):1160–1166, 2013.

A. Bouchard-Côté, S. Sankararaman, and M.I. Jordan. Phylogenetic inference via sequential Monte Carlo. *Syst. Biol.*, 61(4):579–593, 2012.

R. Bouckaert and J. Heled. DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*, 2014.

R. Bouckaert, P. Lemey, M. Dunn, S.J. Greenhill, A.V. Alekseyenko, A.J. Drummond, R.D. Gray, M.A. Suchard, and Q.D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012.

R. Bouckaert, P. Lemey, M. Dunn, S.J. Greenhill, A.V. Alekseyenko, A.J. Drummond, R.D. Gray, M.A. Suchard, and Q.D. Atkinson. Corrections and Clarifications. *Science*, 342(6165):1446, 2013.

C. Brezinski. Convergence acceleration during the 20th century. *J. Comput. Appl. Math.*, 122(1):1–21, 2000.

D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, 21(2):255–265, 2004.

G. Casella and C.P. Robert. Post-processing accept-reject samples: recycling and rescaling. *J. Comput. Graph. Statist.*, 7(2):139–157, 1998.

D.M. Ceperley and M. Dewing. The penalty method for random walks with uncertain energies. *J. Chem. Phys.*, 110(20):9812–9820, 1999.

W. Chang, C. Cathcart, D. Hall, and A. Garrett. Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis. *Language*, 91(1):194–244, 2015.

C.D. Chrétien. The mathematical models of glottochronology. *Language*, 38(1):11–37, 1962.

J.A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Stat.*, 14(4):795–810, 2005.

J.A. Christen and C. Fox. A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Anal.*, 5(2):263–281, 2010.

E. Conte and G. Molle. Reinvestigating a key site for Polynesian prehistory: new results from the Hane dune site, Ua Huka (Marquesas). *Archaeol. Oceania*, 49(3): 121–136, 2014.

F.W. Crawford, R.E. Weiss, and M.A. Suchard. Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. *Ann. Appl. Stat.*, 9(2):572, 2015.

G.B. Cybis, J.S. Sinsheimer, T. Bedford, A.E. Mather, P. Lemey, and M.A. Suchard. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.*, 9(2):969–991, 2015.

C. Darwin. *On the origins of species by means of natural selection*. John Murray, London, 1859.

V. Daubin, M. Gouy, and G. Perrière. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, 12(7):1080–1090, 2002.

J.P. Delahaye and B. Germain-Bonne. Résultats négatifs en accélération de la convergence. *Numer. Math*, 35(4):443–457, 1980.

V. Dinh and F.A. Matsen IV. The shape of the one-dimensional phylogenetic likelihood function. *ArXiv:1507.03647*, 2015.

A.J. Dobson, J.B. Kruskal, D. Sankoff, and L.J. Savage. The mathematics of glottochronology revisited. *Anthropol. Linguist.*, pages 205–212, 1972.

A. Doucet, M.K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, pages 295–313, 2015.

A.J. Drummond, G.K. Nicholls, A.G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

A.J. Drummond, M.A. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, 29(8):1969–1973, 2012.

I. Dyen, J.B. Kruskal, and P. Black. An indoeuropean classification: A lexicostatistical experiment. *T. Am. Philos. Soc.*, 82(5):iii–132, 1992.

A.W.F. Edwards. Estimation of the branch points of a branching diffusion process. *J. Roy. Stat. Soc. B*, 32(2):155–174, 1970.

P.L. Erdõs, M.A. Steel, L.A. Székely, and T.J. Warnow. A few logs suffice to build (almost) all trees: Part I. *Random Struct. Algor.*, 14(2):153–184, 1999.

J.S. Farris. Phylogenetic analysis under Dollo's law. *Syst. Biol.*, 26(1):77–88, 1977.

J.S. Farris. The logical basis of phylogenetic analysis. In N.I. Platnick and V.A. Funk, editors, *Advances in Cladistics*, volume 2, pages 7–36. Columbia University Press, New York, 1983.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

J. Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, pages 1–15, 1985.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, pages 457–472, 1992.

C.J. Geyer. Practical Markov chain Monte Carlo. *Statist. Sci.*, 7(4):473–483, 1992.

C.J. Geyer and E.A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, 90(431):909–920, 1995.

M.B. Giles. Multilevel Monte Carlo Path Simulation. *Oper. Res.*, 56(3):607–617, 2008.

D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

R.D. Gray and Q.D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

R.D. Gray, A.J. Drummond, and S.J. Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913):479–483, 2009.

R.D. Gray, D. Bryant, and S.J. Greenhill. On the shape and fabric of human history. *Philos. T. R. Soc. B*, 365(1559):3923–3933, 2010.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

S.J. Greenhill, R. Blust, and R.D. Gray. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evol. Bioinform. Online*, 4:271–283, 2008.

S.J. Greenhill, T.E. Currie, and R.D. Gray. Does horizontal transmission invalidate cultural phylogenies? *Proc. Roy. Soc. B*, 276(1665):2299–2306, 2009.

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, third edition, 2001.

M.W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.*, 8(7):1–9, 2007.

D. Hall and D. Klein. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039. Association for Computational Linguistics, 2010.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

J. Heled and A.J. Drummond. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.*, 61(1):138–149, 2012.

L.S.T. Ho, J. Xu, F.W. Crawford, V.N. Minin, and M.A. Suchard. Birth (death)/birth-death processes and their computable transition probabilities with statistical applications. *ArXiv:1603.03819*, 2016.

H. Hoijer. Lexicostatistics: A Critique. *Language*, 32(1):49–60, 1956.

R. R Hudson. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, 23(2):183–201, 1983.

D.H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23(2):254–267, 2006.

D.H. Huson and C. Scornavacca. A survey of combinatorial methods for phylogenetic networks. *Genome Biol. Evol.*, 3:23–35, 2011.

G. Irwin. Pacific seascapes, canoe performance, and a review of lapita voyaging with regard to theories of migration. *Asian Perspect.*, 47(1):12–27, 2008.

P.E. Jacob and R.J. Ryder. The Wang–Landau algorithm reaches the flat histogram criterion in finite time. *Ann. Appl. Probab.*, 24(1):34–53, 2014.

P.E. Jacob and A.H. Thiery. On nonnegative unbiased estimators. *Ann. Stat.*, 43(2): 769–784, 2015.

R. Jain, M.C. Rivera, and J.A. Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA*, 96(7):3801–3806, 1999.

H. Jeffreys. Some tests of significance, treated by the theory of probability. *Math. Proc. Cambridge Philos. Soc.*, 31(2):203–222, 1935.

A. Jennings. Accelerating the convergence of matrix iterative processes. *J. Inst. Maths. Applics.*, 8(1):99–110, 1971.

R.E. Kass and A.E. Raftery. Bayes factors. *J. Am. Stat. Assoc.*, 90(430):773–795, 1995.

F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

L.J. Kelly and G.K. Nicholls. Lateral transfer in Stochastic Dollo models. Manuscript submitted for publication, ArXiv:1601.07931, 2016.

J.F.C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13(3):235–248, 1982.

J.F.C. Kingman. *Poisson Processes*. Clarendon Press, Oxford, 1992.

A. Kitchen, C. Ehret, S. Assefa, and C.J. Mulligan. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. Roy. Soc. B*, 276(1668):2703–2710, 2009.

E.V. Koonin, K.S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annu. Rev. Microbiol.*, 55(1):709–742, 2001.

L.S. Kubatko. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.*, 58(5):478–488, 2009.

N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109, 2004.

G.M. Lathrop. Evolutionary trees and admixture: phylogenetic inference when some populations are hybridized. *Ann. Hum. Genet.*, 46(3):245–255, 1982.

D. Lewis. Polynesian navigational methods. *J. Polynesian. Soc.*, 73(4):364–374, 1964.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On Russian Roulette Estimates for Bayesian inference with Doubly-Intractable Likelihoods. *Statist. Sci.*, 30(4):443–467, 2015.

D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.*, 89(428):1535–1546, 1994.

J.C. Marck. *Topics in Polynesian Language and Culture History*, volume 504. Pacific Linguistics, Canberra, 2000.

P. McCullagh, J. Pitman, and M. Winkel. Gibbs fragmentation trees. *Bernoulli*, pages 988–1002, 2008.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087, 1953.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.*, 20(4):801–836, 1978.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003.

J. Møller, A.N. Pettitt, R. Reeves, and K.K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

P.A.P. Moran. Random processes in genetics. *Math. Proc. Cambridge Philos. Soc.*, 54(1): 60–71, 1958.

I. Murray, Z. Ghahramani, and D.J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

L. Nakhleh, D. Ringe, and T. Warnow. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language*, 81(2):382–420, 2005.

R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001a.

R.M. Neal. Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, University of Toronto, 2001b.

G.K. Nicholls and R.D. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *J. Roy. Stat. Soc. B*, 70(3):545–566, 2008.

G.K. Nicholls and R.J. Ryder. Phylogenetic models for Semitic vocabulary. In D. Conesa, A. Forte, A. López-Quílez, and F. Muñoz, editors, *Proceedings of the International Workshop on Statistical Modelling*, pages 431–436, 2011.

G.K. Nicholls, C. Fox, and A. Muir-Watt. Coupled MCMC with a randomized acceptance probability. *ArXiv:1205.6857*, 2012.

G.K. Nicholls, R.J. Ryder, and D. Welch. *TraitLab: a MatLab Package for Fitting and Simulating Binary Trait-Like Data*, 2013.

J.R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

M.J. O'Brien, J. Darwent, and R.L. Lyman. Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. *J. Archaeol. Sci.*, 28(10):1115–1136, 2001.

J. Oldman, T. Wu, L. van Iersel, and V. Moulton. TriLoNet: Piecing Together Small Networks to Reconstruct Reticulate Evolutionary Histories. *Mol. Biol. Evol.*, 33(8): 2151–2162, 2016.

M. Pagel and A. Meade. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53(4):571–581, 2004.

M. Pagel, Q.D. Atkinson, A.S. Calude, and A. Meade. Ultraconserved words point to deep language ancestry across Eurasia. *Proc. Natl. Acad. Sci. USA*, 110(21):8471–8476, 2013.

P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol. Biol. Evol.*, 5(5):568–583, 1988.

N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3): 1065–1093, 2012.

P.H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3): 607–612, 1973.

J.K. Pickrell and J.K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.*, 8(11):e1002967, 2012.

B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, 43(3):304–311, 1996.

B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.

M.D. Rasmussen and M. Kellis. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.*, 17(12):1932–1942, 2007.

B.D. Redelings and M.A Suchard. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3):401–418, 2005.

C.-H. Rhee and P.W. Glynn. A New Approach to Unbiased Estimation for SDE's. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, editors, *Proceedings of the Winter Simulation Conference*, 2012.

L.F. Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philos. T. Roy. Soc. A*, 210:307–357, 1911.

D. Ringe, T. Warnow, and A. Taylor. Indo-European and computational cladistics. *T. Philol. Soc.*, 100(1):59–129, 2002.

G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.

S. Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE ACM T. Comput. Bi.*, 3(1):92–94, 2006.

S. Roch and S. Snir. Recovering the treelike trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis. *J. Comput. Biol.*, 20(2):93–112, 2013.

R.M. Ross, S.J. Greenhill, and Q.D. Atkinson. Population structure and cultural geography of a folktale in europe. *Proc. Roy. Soc. B*, 280(1756):20123065, 2013.

R.J. Ryder. *Phylogenetic Models of Language Diversification*. PhD thesis, University of Oxford, 2009.

R.J. Ryder and G.K. Nicholls. Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *J. Roy. Stat. Soc. C*, 60(1):71–92, 2011.

C. Sherlock, A.H. Thiery, G.O. Roberts, and J.S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 02 2015.

J. Sjöstrand, A. Tofigh, V. Daubin, L. Arvestad, B. Sennblad, and J. Lagergren. A Bayesian method for analyzing lateral gene transfer. *Syst. Biol.*, 63(3):409–420, 2014.

C. Skelton. Methods of using phylogenetic systematics to reconstruct the history of the Linear B script. *Archaeometry*, 50(1):158–176, 2008.

M. Spriggs and A. Anderson. Late colonization of East Polynesia. *Antiquity*, 67(255): 200–217, 1993.

M.A. Steel, L. Székely, and E. Mossel. Phylogenetic information complexity: Is testing a tree easier than finding it? *J. Theor. Biol.*, 258(1):95–102, 2009.

M.A. Suchard. Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics*, 170(1):419–431, 2005.

M.A. Suchard, R.E. Weiss, K.S. Dorman, and J.S. Sinsheimer. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple changepoint model. *J. Am. Stat. Assoc.*, 98(462):427–437, 2003.

M. Swadesh. Salish internal relationships. *Int. J. Am. Linguist.*, 16(4):157–167, 1950.

M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *P. Am. Philos. Soc.*, 96(4):452–463, 1952.

M. Swadesh. Archeological and linguistic chronology of Indo-European groups. *Am. Anthropol.*, 55(3):349–352, 1953.

M. Swadesh. Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.*, 21 (2):121–137, 1955.

G.J. Szöllősi, B. Boussau, S.S. Abby, E. Tannier, and V. Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. USA*, 109(43):17513–17518, 2012.

G.J. Szöllősi, E. Tannier, N. Lartillot, and V. Daubin. Lateral Gene Transfer from the Dead. *Syst. Biol.*, 62(3):386–397, 2013.

J.J. Tehrani, M. Collard, and S.J. Shennan. The cophylogeny of populations and cultures: reconstructing the evolution of Iranian tribal craft traditions using trees and jungles. *Philos. T. R. Soc. B*, 365(1559):3865–3874, 2010.

M. Vihola. Unbiased estimators and multilevel Monte Carlo. *ArXiv:1512.01022*, 2015.

R. Walter and P.J. Sheppard. The Ngati Tiare Adze Cache: Further evidence of prehistoric contact between West Polynesia and the Southern Cook Islands. *Archaeol. Oceania*, 31(1):33–39, 1996.

M. Walworth. Eastern Polynesian: The linguistic evidence revisited. *Ocean. Ling.*, 53 (2):256–272, 2014.

F. Wang and D.P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64: 056101, 2001a.

F. Wang and D.P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86(10):2050, 2001b.

L. Wang, A. Bouchard-Côté, and A. Doucet. Bayesian Phylogenetic Inference using a Combinatorial Sequential Monte Carlo Method. *J. Am. Stat. Assoc.*, 110(512): 1362–1374, 2015.

T.J. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. In P. Forster and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, pages 75–90. McDonald Institute for Archaeological Research, 2006.

M.I. Weisler. Hard Evidence for Prehistoric Interaction in Polynesia. *Curr. Anthropol.*, 39(4):521–532, 1998.

M.I. Weisler and P.V. Kirch. Interisland and interarchipelago transfer of stone tools in prehistoric Polynesia. *Proc. Natl. Acad. Sci. USA*, 93(4):1381–1385, 1996.

D. Wen, Y. Yu, and L. Nakhleh. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLOS Genet.*, 12(5):e1006006, 2016.

C. Whidden and F.A. Matsen IV. Quantifying MCMC Exploration of Phylogenetic Tree Space. *Syst. Biol.*, 64(3):472–491, 2015.

C. Whidden and F.A. Matsen IV. Ricci-Ollivier Curvature of the Rooted Phylogenetic Subtree-Prune-Regraft Graph. In *Proceedings of the Thirteenth Workshop on Analytic Algorithmics and Combinatorics*, pages 106–120, 2016.

A. Willis. Confidence sets for phylogenetic trees. *ArXiv:1607.08288*, 2016.

J.M. Wilmshurst, T.L. Hunt, C.P. Lipo, and A.J. Anderson. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc. Natl. Acad. Sci. USA*, 108(5):1815–1820, 2011.

I.J. Wilson and D.J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150(1):499–510, 1998.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14(7):717–724, 1997.

Appendix A

Proof of Theorem 1

Theorem 1 states that under the SDLT model, the pattern frequencies $\mathbf{N}(t) = (N_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)})$ at time t on a tree g is a vector of independent Poisson-distributed random variables with rate parameters $\mathbf{x}(t) = (x_{\mathbf{p}}(t) : \mathbf{p} \in \mathcal{P}^{(t)}) = \mathbb{E}[\mathbf{N}(t)|g, \lambda, \mu, \beta]$ given by the sequence of initial value problems in Equation 2.3.5.

Proof of Theorem 1. Starting at the Adam node at time $t_0 = -\infty$, the trait process evolves forwards in time along the branches of the tree. The pure birth-death trait process $N_1(t)$ is in equilibrium when it reaches the root at time t_1 by construction so $N_1(t_1^-) \sim \text{Poisson}(x_1(t_1^-))$ where $x_1(t_1^-) = \lambda/\mu$. Following Property T1 in Section 2.2,

- The pattern $(1,1)$ is consistent with the branching event at time t_1 so $N_{11}(t_1) \equiv N_1(t_1^-)$ and $x_{11}(t_1) \equiv x_1(t_1^-)$
- The patterns $(0,1)$ and $(1,0)$ are inconsistent with the branching event at the root so $N_{01}(t_1) \equiv N_{10}(t_1) \equiv 0$ and $x_{01}(t_1) \equiv x_{10}(t_1) \equiv 0$.

This provides us with the initial condition for the start of the next interval $[t_1, t_2^-)$ between branching events. We can calculate the expected pattern frequencies at any time t by alternatively solving the initial value problems in Equation 2.3.5 and computing initial conditions using the initialisation operators $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(L-1)}$.

To complete the proof, we derive the Kolmogorov forward equation describing the temporal evolution of $p_{\mathbf{n}}(t) = \mathbb{P}(\mathbf{N}(t) = \mathbf{n}|g, \lambda, \mu, \beta)$ for an integer vector $\mathbf{n} = (n_{\mathbf{p}} : \mathbf{p} \in \mathcal{P}^{(t)})$ and show that it is equivalent to the time-derivative of the hypothesised Poisson probability mass function

$$\pi_{\mathbf{n}}(t) = \prod_{\mathbf{p} \in \mathcal{P}^{(t)}} \frac{x_{\mathbf{p}}(t)^{n_{\mathbf{p}}} e^{-x_{\mathbf{p}}(t)}}{n_{\mathbf{p}}!}. \tag{A.1}$$

For patterns $\mathbf{p}$ and $\mathbf{q} \in \mathcal{P}^{(t)}$, we require the operators $\mathbf{U_{p0}}$, $\mathbf{U_{pq}}$ and $\mathbf{U_{0q}}$ which applied to $\mathbf{n}$ yield

$$\begin{aligned}
\mathbf{U_{p0}n} &= (\ldots, n_{\mathbf{p}-1}, n_{\mathbf{p}} - 1, n_{\mathbf{p}+1}, \ldots \qquad\qquad\qquad\quad), \\
\mathbf{U_{pq}n} &= (\ldots, n_{\mathbf{p}-1}, n_{\mathbf{p}} - 1, n_{\mathbf{p}+1}, \ldots, n_{\mathbf{q}-1}, n_{\mathbf{q}} + 1, n_{\mathbf{q}+1}, \ldots), \\
\mathbf{U_{0q}n} &= (\qquad\qquad\qquad\qquad \ldots, n_{\mathbf{q}-1}, n_{\mathbf{q}} + 1, n_{\mathbf{q}+1}, \ldots),
\end{aligned}$$

where we have abused notation and used $\mathbf{p} - 1$ and $\mathbf{p} + 1$ to index the entries either side of $n_{\mathbf{p}}$ in $\mathbf{n}$. These operators respectively correspond to the change in $\mathbf{n}$ observed if: a trait displaying pattern $\mathbf{p}$ becomes extinct (T3); a trait which displayed pattern $\mathbf{p}$ transitions to display pattern $\mathbf{q}$ through either a death (T3) or transfer (T4) event; and a trait is born displaying pattern $\mathbf{q}$ (T2). Of course, these transitions may only occur if the patterns communicate. If $\rho(\mathbf{n}, \mathbf{n}')$ denotes the transition rate from state $\mathbf{N}(t) = \mathbf{n}$ to $\mathbf{n}'$, then from Section 2.3.1,

$$\begin{aligned}
\rho(\mathbf{n}, \mathbf{U_{p0}n}) = n_{\mathbf{p}}\lambda_{\mathbf{p0}} \quad \text{where} \quad \lambda_{\mathbf{p0}} &= \begin{cases} \mu, & s(\mathbf{p}) = 1, \\ 0, & \text{otherwise}, \end{cases} \\
\rho(\mathbf{n}, \mathbf{U_{pq}n}) = n_{\mathbf{p}}\lambda_{\mathbf{pq}} \quad \text{where} \quad \lambda_{\mathbf{pq}} &= \begin{cases} \mu, & \mathbf{q} \in S_{\mathbf{p}}^{-}, \\ \beta\frac{s(\mathbf{p})}{L}, & \mathbf{q} \in S_{\mathbf{p}}^{+}, \\ 0, & \text{otherwise}, \end{cases} \\
\rho(\mathbf{n}, \mathbf{U_{0q}n}) = \lambda_{\mathbf{0q}} \quad \text{where} \quad \lambda_{\mathbf{0q}} &= \begin{cases} \lambda, & s(\mathbf{q}) = 1, \\ 0, & \text{otherwise}. \end{cases}
\end{aligned} \tag{A.2}$$

The forward equation We derive the forward equation from first principles. For a short interval of length $\mathrm{d}t$, from Equations 2.3.1 and A.2, we obtain

$$\begin{aligned}
p_{\mathbf{n}}(t+\mathrm{d}t) &= p_{\mathbf{n}}(t)\Bigg[1-\Bigg(\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}\rho(\mathbf{n},\mathbf{U}_{\mathbf{pq}}\mathbf{n})+\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\rho(\mathbf{n},\mathbf{U}_{\mathbf{p0}}\mathbf{n}) \\
&\qquad +\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}\rho(\mathbf{n},\mathbf{U}_{\mathbf{0q}}\mathbf{n})\Bigg)\mathrm{d}t+o(\mathrm{d}t)\Bigg] \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{pq}}\mathbf{n}}(t)\left[\rho(\mathbf{U}_{\mathbf{pq}}\mathbf{n},\mathbf{n})\mathrm{d}t+o(\mathrm{d}t)\right] \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{p0}}\mathbf{n}}(t)\left[\rho(\mathbf{U}_{\mathbf{p0}}\mathbf{n},\mathbf{n})\mathrm{d}t+o(\mathrm{d}t)\right] \\
&\quad +\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{0q}}\mathbf{n}}(t)\left[\rho(\mathbf{U}_{\mathbf{0q}}\mathbf{n},\mathbf{n})\mathrm{d}t+o(\mathrm{d}t)\right] \\
&= p_{\mathbf{n}}(t)\Bigg[1-\Bigg(\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}n_{\mathbf{p}}\lambda_{\mathbf{pq}}+\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}n_{\mathbf{p}}\lambda_{\mathbf{p0}} \\
&\qquad +\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}\lambda_{\mathbf{0q}}\Bigg)\mathrm{d}t+o(\mathrm{d}t)\Bigg] \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{pq}}\mathbf{n}}(t)\left[(n_{\mathbf{q}}+1)\lambda_{\mathbf{qp}}\mathrm{d}t+o(\mathrm{d}t)\right] \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{p0}}\mathbf{n}}(t)\left[\lambda_{\mathbf{0p}}\mathrm{d}t+o(\mathrm{d}t)\right] \\
&\quad +\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{0q}}\mathbf{n}}(t)\left[(n_{\mathbf{q}}+1)\lambda_{\mathbf{q0}}\mathrm{d}t+o(\mathrm{d}t)\right].
\end{aligned} \tag{A.3}$$

We subtract $p_{\mathbf{n}}(t)$ from both sides of Equation A.3, then divide by $\mathrm{d}t$ and take the limit as $\mathrm{d}t \downarrow 0$ to obtain

$$\begin{aligned}
\dot{p}_{\mathbf{n}}(t) &= -p_{\mathbf{n}}(t)\Bigg[\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}n_{\mathbf{p}}\lambda_{\mathbf{pq}}+\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}n_{\mathbf{p}}\lambda_{\mathbf{p0}}+\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}\lambda_{\mathbf{0q}}\Bigg] \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{pq}}\mathbf{n}}(t)(n_{\mathbf{q}}+1)\lambda_{\mathbf{qp}} \\
&\quad +\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{p0}}\mathbf{n}}(t)\lambda_{\mathbf{0p}}+\sum_{\mathbf{q}\in\mathcal{P}^{(t)}}p_{\mathbf{U}_{\mathbf{0q}}\mathbf{n}}(t)(n_{\mathbf{q}}+1)\lambda_{\mathbf{q0}}.
\end{aligned} \tag{A.4}$$

We shall drop the dependence on t in our notation for the remainder of this section. From the hypothesised probability mass function (A.1), we see that

$$\begin{aligned}
\pi_{\mathbf{U}_{\mathbf{p0}}\mathbf{n}} &= \pi_{\mathbf{n}} \frac{n_{\mathbf{p}}}{x_{\mathbf{p}}}, \\
\pi_{\mathbf{U}_{\mathbf{pq}}\mathbf{n}} &= \pi_{\mathbf{n}} \frac{n_{\mathbf{p}}}{x_{\mathbf{p}}} \frac{x_{\mathbf{q}}}{n_{\mathbf{q}}+1}, \\
\pi_{\mathbf{U}_{\mathbf{0q}}\mathbf{n}} &= \pi_{\mathbf{n}} \frac{x_{\mathbf{q}}}{n_{\mathbf{q}}+1},
\end{aligned}$$

and to simplify our argument later, we suppose that $p_{\mathbf{n}}(t) = \pi_{\mathbf{n}}(t)$ and substitute these identities into Equation A.4 to obtain

$$\begin{aligned}
\dot{p}_{\mathbf{n}} &= -p_{\mathbf{n}} \left[\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{q}\in\mathcal{P}} n_{\mathbf{p}} \lambda_{\mathbf{pq}} + \sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}} \lambda_{\mathbf{p0}} + \sum_{\mathbf{q}\in\mathcal{P}} \lambda_{\mathbf{0q}} \right] \\
&\quad + p_{\mathbf{n}} \sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{q}\in\mathcal{P}} \frac{n_{\mathbf{p}}}{x_{\mathbf{p}}} \frac{x_{\mathbf{q}}}{n_{\mathbf{q}}+1} (n_{\mathbf{q}}+1) \lambda_{\mathbf{qp}} + p_{\mathbf{n}} \sum_{\mathbf{p}\in\mathcal{P}} \frac{n_{\mathbf{p}}}{x_{\mathbf{p}}} \lambda_{\mathbf{0p}} \\
&\quad + p_{\mathbf{n}} \sum_{\mathbf{q}\in\mathcal{P}} \frac{x_{\mathbf{q}}}{n_{\mathbf{q}}+1} (n_{\mathbf{q}}+1) \lambda_{\mathbf{q0}} \\
&= -p_{\mathbf{n}} \left[\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{q}\in\mathcal{P}} n_{\mathbf{p}} \lambda_{\mathbf{pq}} + \sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}} \lambda_{\mathbf{p0}} + \sum_{\mathbf{q}\in\mathcal{P}} \lambda_{\mathbf{0q}} \right] \\
&\quad + p_{\mathbf{n}} \left[\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{q}\in\mathcal{P}} n_{\mathbf{p}} \frac{x_{\mathbf{q}}}{x_{\mathbf{p}}} \lambda_{\mathbf{qp}} + \sum_{\mathbf{p}\in\mathcal{P}} \frac{n_{\mathbf{p}}}{x_{\mathbf{p}}} \lambda_{\mathbf{0p}} + \sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{q}} \lambda_{\mathbf{q0}} \right] \\
&= p_{\mathbf{n}} \left[\sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{q}\in\mathcal{P}} n_{\mathbf{p}} \left(-\lambda_{\mathbf{pq}} + \frac{x_{\mathbf{q}}}{x_{\mathbf{p}}} \lambda_{\mathbf{qp}} \right) + \sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}} \left(-\lambda_{\mathbf{p0}} + \frac{1}{x_{\mathbf{p}}} \lambda_{\mathbf{0p}} \right) \right. \\
&\quad \left. + \sum_{\mathbf{q}\in\mathcal{P}} \left(-\lambda_{\mathbf{0q}} + x_{\mathbf{q}} \lambda_{\mathbf{q0}} \right) \right].
\end{aligned} \tag{A.5}$$

Equation A.5 is the forward equation for $\mathbf{N}(t)$.

Time derivative of the probability mass function Equation 2.3.2 describes the temporal evolution of $x_{\mathbf{p}}(t)$, the expected number of traits displaying pattern $\mathbf{p} \in \mathcal{P}^{(t)}$ at time t, which we can write

$$\dot{x}_{\mathbf{p}}(t) = -x_{\mathbf{p}}(t) \left(\lambda_{\mathbf{p0}} + \sum_{\mathbf{q}\in\mathcal{P}^{(t)}} \lambda_{\mathbf{pq}} \right) + \lambda_{\mathbf{0p}} + \sum_{\mathbf{q}\in\mathcal{P}^{(t)}} x_{\mathbf{q}}(t) \lambda_{\mathbf{qp}}, \tag{A.6}$$

using the identities in Equation A.2 and the fact that $s(\mathbf{p}) = |S_{\mathbf{p}}^-| + \mathbf{1}_{\{s(\mathbf{p})=1\}}$ and $L^{(t)} - s(\mathbf{p}) = |S_{\mathbf{p}}^+|$. Differentiating the hypothesised probability mass function (A.1) with respect to time t, we obtain

$$\dot{\pi}_{\mathbf{n}}(t) = \pi_{\mathbf{n}}(t)\frac{\mathrm{d}}{\mathrm{d}t}\log(\pi_{\mathbf{n}}(t)) = \pi_{\mathbf{n}}(t)\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\frac{n_{\mathbf{p}}}{x_{\mathbf{p}}(t)}\dot{x}_{\mathbf{p}}(t) - \pi_{\mathbf{n}}(t)\sum_{\mathbf{p}\in\mathcal{P}^{(t)}}\dot{x}_{\mathbf{p}}(t). \tag{A.7}$$

Dropping the dependence on time t from our notation and substituting Equation A.6 into Equation A.7 yields

$$\begin{aligned}
\dot{\pi}_{\mathbf{n}} &= \pi_{\mathbf{n}}\sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}}\left[-\left(\lambda_{\mathbf{p}0} + \sum_{\mathbf{q}\in\mathcal{P}}\lambda_{\mathbf{pq}}\right) + \frac{1}{x_{\mathbf{p}}}\left(\lambda_{0\mathbf{p}} + \sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{q}}\lambda_{\mathbf{qp}}\right)\right] \\
&\quad + \pi_{\mathbf{n}}\sum_{\mathbf{p}\in\mathcal{P}}\left[x_{\mathbf{p}}\left(\lambda_{\mathbf{p}0} + \sum_{\mathbf{q}\in\mathcal{P}}\lambda_{\mathbf{pq}}\right) - \lambda_{0\mathbf{p}} - \sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{q}}\lambda_{\mathbf{qp}}\right] \\
&= \pi_{\mathbf{n}}\sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}}\left[\sum_{\mathbf{q}\in\mathcal{P}}\left(-\lambda_{\mathbf{pq}} + \frac{x_{\mathbf{q}}}{x_{\mathbf{p}}}\lambda_{\mathbf{qp}}\right) - \lambda_{\mathbf{p}0} + \frac{1}{x_{\mathbf{p}}}\lambda_{0\mathbf{p}}\right] \\
&\quad + \pi_{\mathbf{n}}\left[\sum_{\mathbf{p}\in\mathcal{P}}\left(-\lambda_{0\mathbf{p}} + x_{\mathbf{p}}\lambda_{\mathbf{p}0}\right)\right] + \pi_{\mathbf{n}}\sum_{\mathbf{p}\in\mathcal{P}}\left[x_{\mathbf{p}}\sum_{\mathbf{q}\in\mathcal{P}}\lambda_{\mathbf{pq}} - \sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{q}}\lambda_{\mathbf{qp}}\right] \\
&= \pi_{\mathbf{n}}\left[\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{q}\in\mathcal{P}} n_{\mathbf{p}}\left(-\lambda_{\mathbf{pq}} + \frac{x_{\mathbf{q}}}{x_{\mathbf{p}}}\lambda_{\mathbf{qp}}\right) + \sum_{\mathbf{p}\in\mathcal{P}} n_{\mathbf{p}}\left(-\lambda_{\mathbf{p}0} + \frac{1}{x_{\mathbf{p}}}\lambda_{0\mathbf{p}}\right)\right. \\
&\qquad \left. + \sum_{\mathbf{p}\in\mathcal{P}}\left(-\lambda_{0\mathbf{p}} + x_{\mathbf{p}}\lambda_{\mathbf{p}0}\right)\right] + \pi_{\mathbf{n}}\left[\sum_{\mathbf{p}\in\mathcal{P}} x_{\mathbf{p}}\sum_{\mathbf{q}\in\mathcal{P}}\lambda_{\mathbf{pq}} - \sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{q}\in\mathcal{P}} x_{\mathbf{q}}\lambda_{\mathbf{qp}}\right].
\end{aligned} \tag{A.8}$$

The final term in Equation A.8 is 0, so it matches the forward equation in Equation A.5. We therefore conclude that Equation A.1 with parameters given by Equation 2.3.5 correctly describes the distribution of the pattern frequencies.

□

Appendix B

Supporting figures

This chapter contains figures of secondary importance to support the analyses in Chapters 5 and 6.

B.1 Chapter 5 — Model validation and testing

In Chapter 5, we fit the exact SDLT, exact-approximate SDLT and exact SD models to three synthetic data sets: `SIM-B`, `SIM-N`, `SIM-T`. We plot histograms of parameter samples in Figures B.1 to B.3. Figures in parentheses represent parameter effective sample sizes.

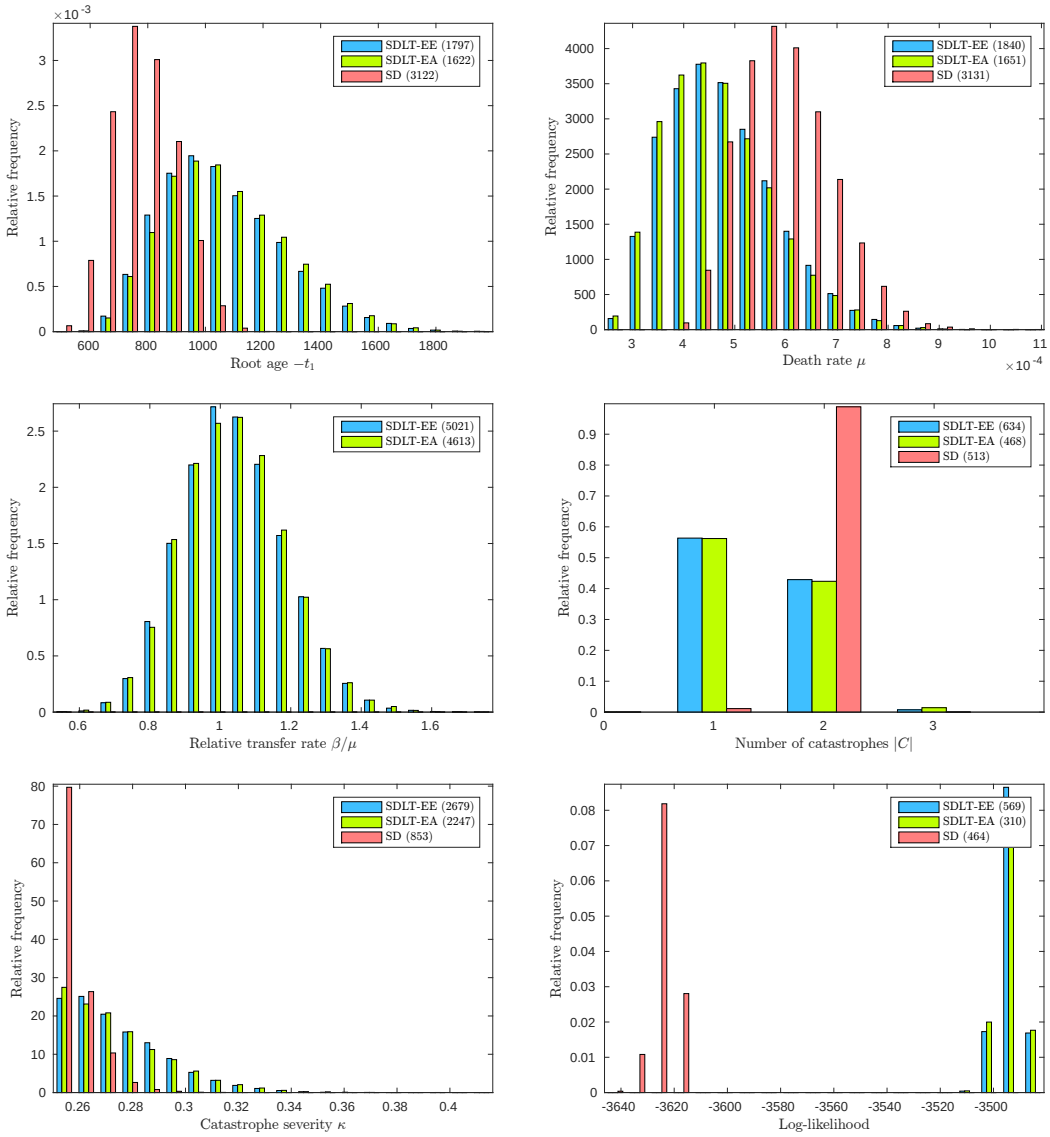


Figure B.1: Histograms of samples in our analyses of `SIM-B`.

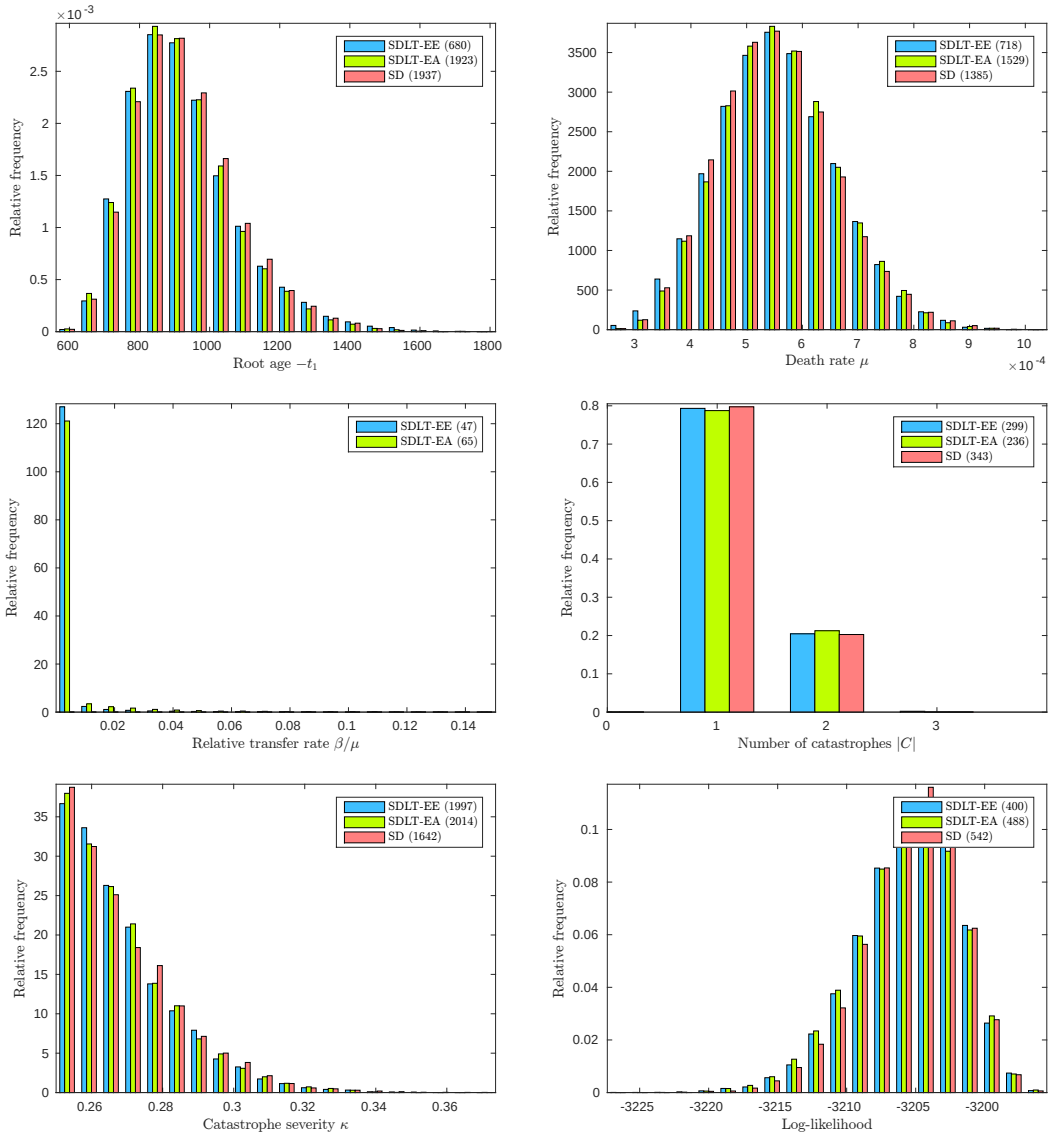


Figure B.2: Histograms of samples in our analyses of `SIM-N`.

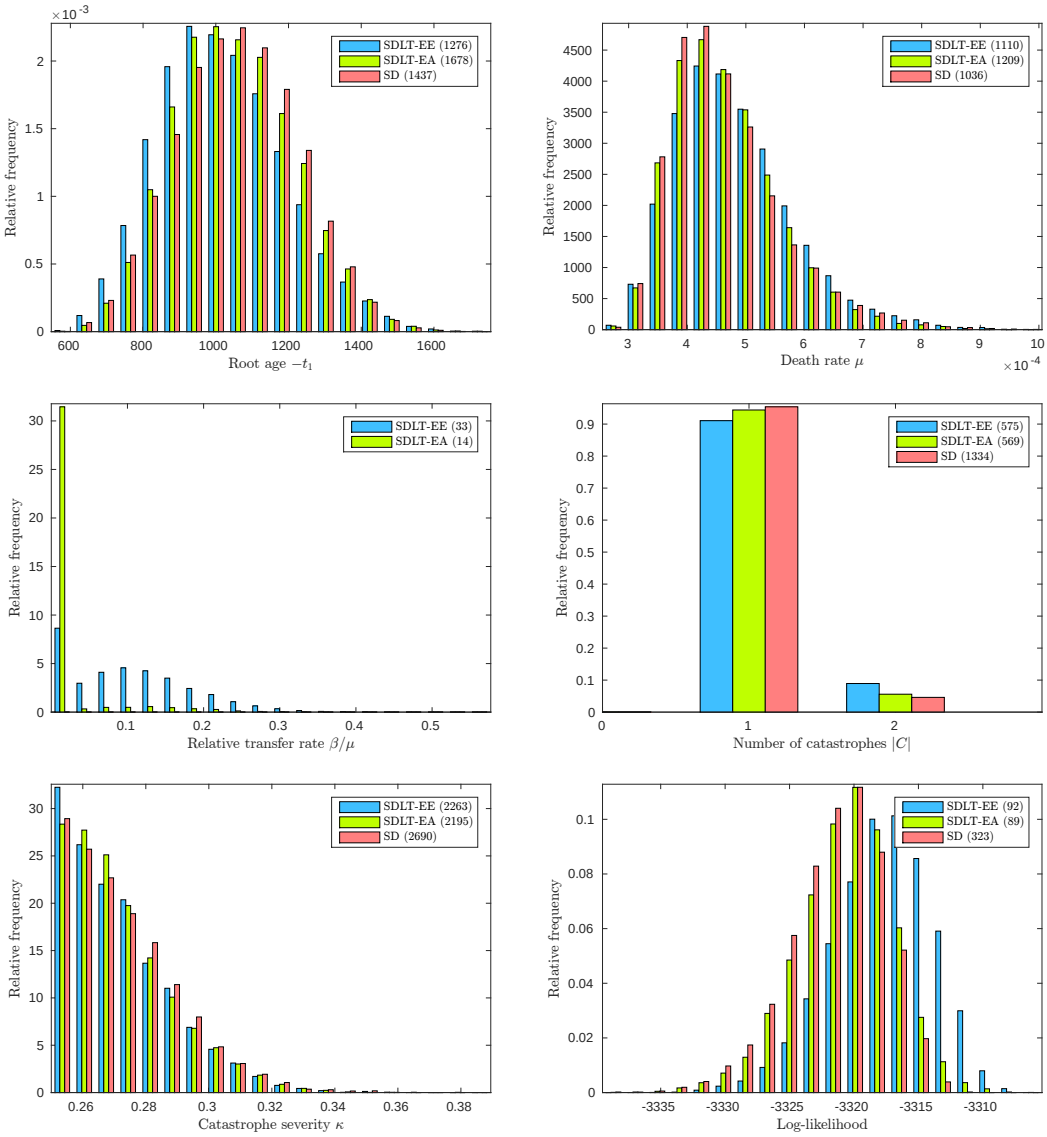
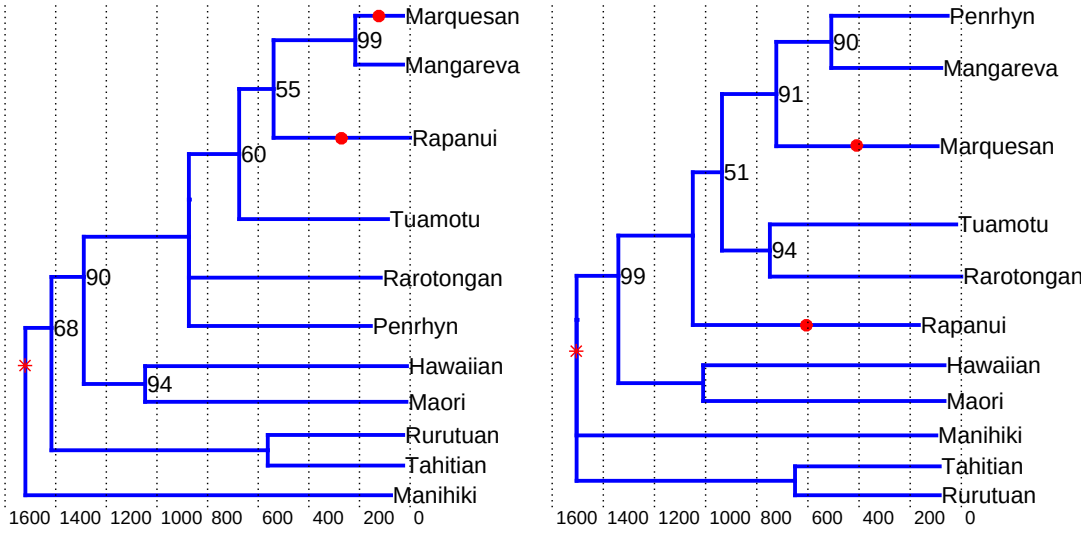


Figure B.3: Histograms of samples in our analyses of `SIM-T`.

B.2 Chapter 6 — Applications

B.2.1 Eastern Polynesian

This section contains figures to support our analyses of the Eastern Polynesian data set POLY-0 in Chapter 6. We report majority-rule consensus trees in Figure B.4, histograms in Figure B.5, trace plots in Figure B.6, and sample autocorrelation plots in Figure B.7. Figures in parentheses denote parameter effective sample sizes.



SDLT. SD.

Figure B.4: Majority-rule consensus trees in our analyses of POLY-0. A majority-rule consensus tree comprises the ancestry relationships which appear in a majority of the samples. Time is in years BP and we depict in red the catastrophes with support greater than 50% on the branches in the consensus tree. Figures on internal nodes represent their posterior support.

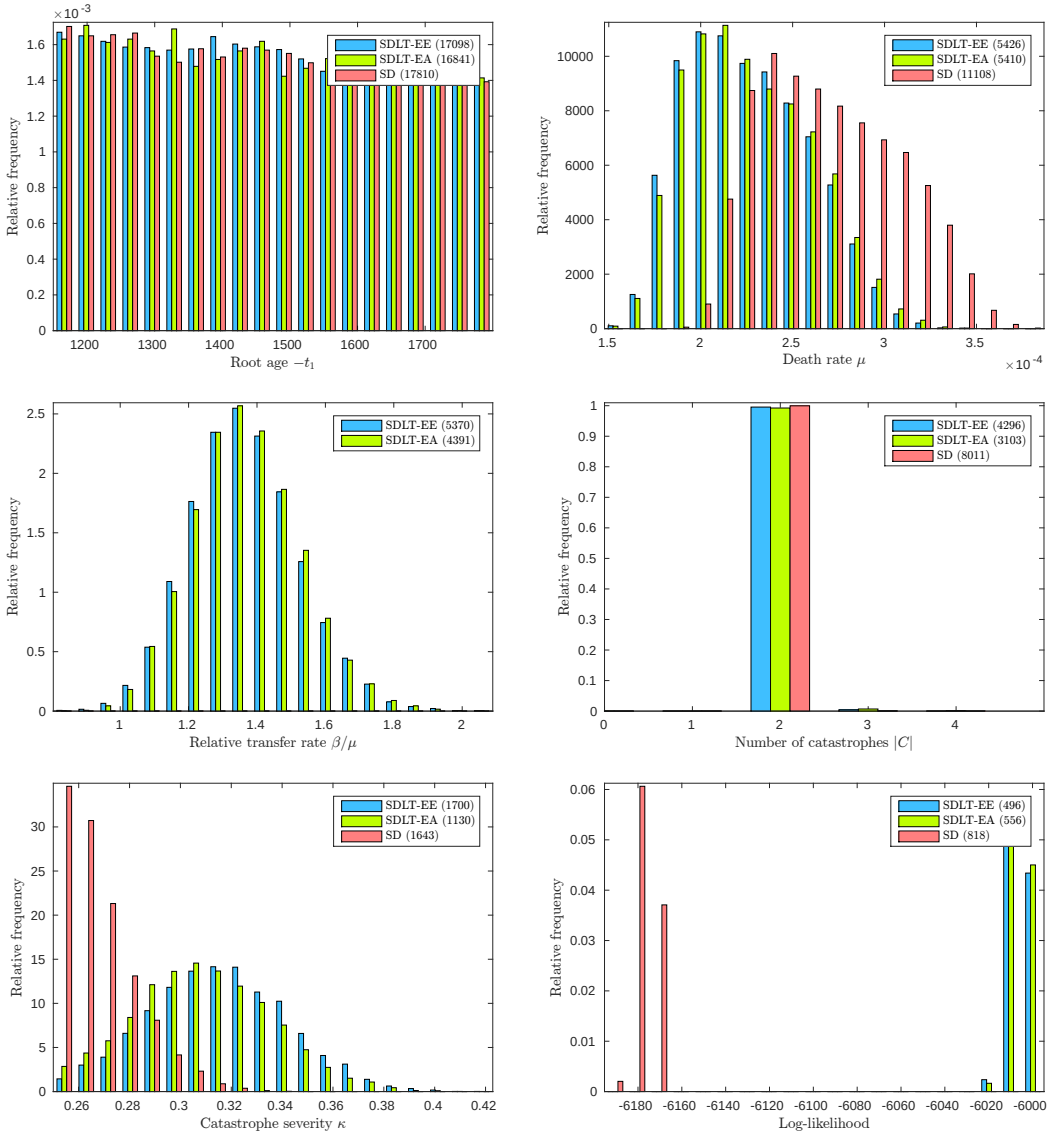


Figure B.5: Histograms of samples in our analyses of `POLY-0`.

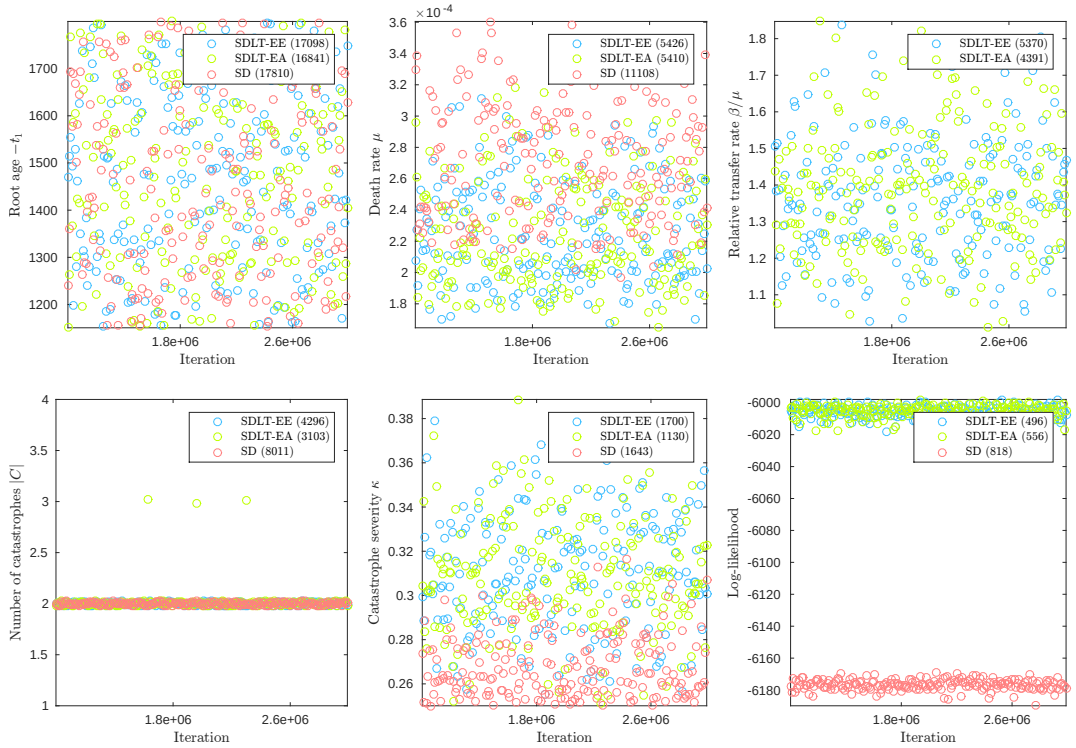


Figure B.6: Trace plots of samples in our analyses of `POLY-0`.

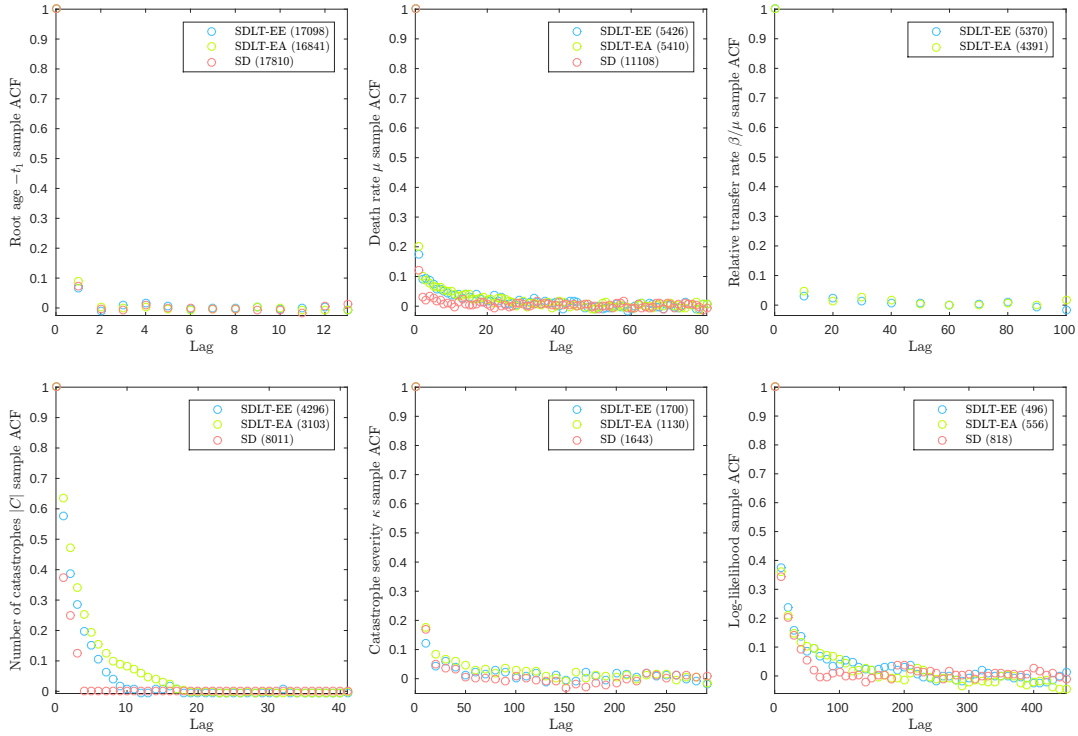
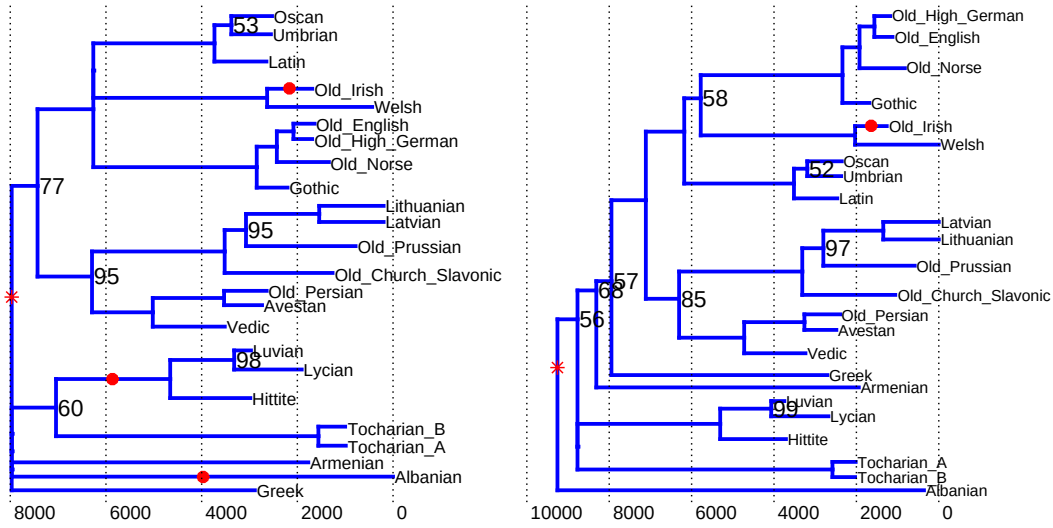


Figure B.7: Sample autocorrelation plots in our analyses of `POLY-0`.

B.2.2 Indo-European

In this section, we report figures to support our analyses of the Indo-European data set `IE` in Chapter 6. We display consensus trees in Figure B.8, histograms of parameter samples in Figure B.9, trace plots in Figure B.10, and sample autocorrelation plots in Figure B.11.



Approximate SDLT. Exact SD.

Figure B.8: Majority-rule consensus trees in our analyses of `IE`. A majority-rule consensus tree comprises the ancestry relationships which appear in a majority of the samples. Time is in years BP and we depict in red the catastrophes with support greater than 50% on the branches in the consensus tree. Figures on internal nodes represent their support in the posterior.

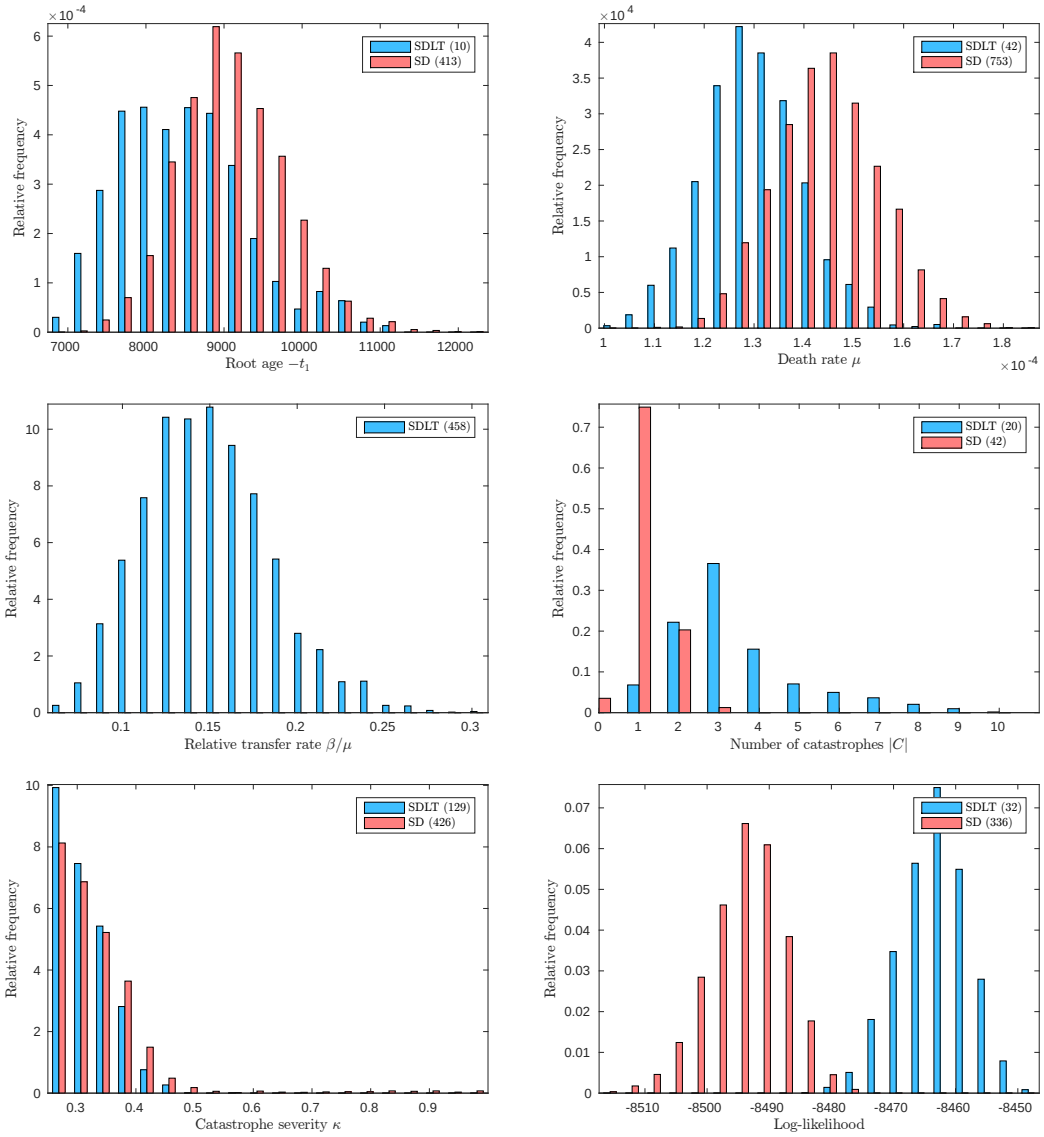


Figure B.9: Histograms of samples in our analyses of IE.

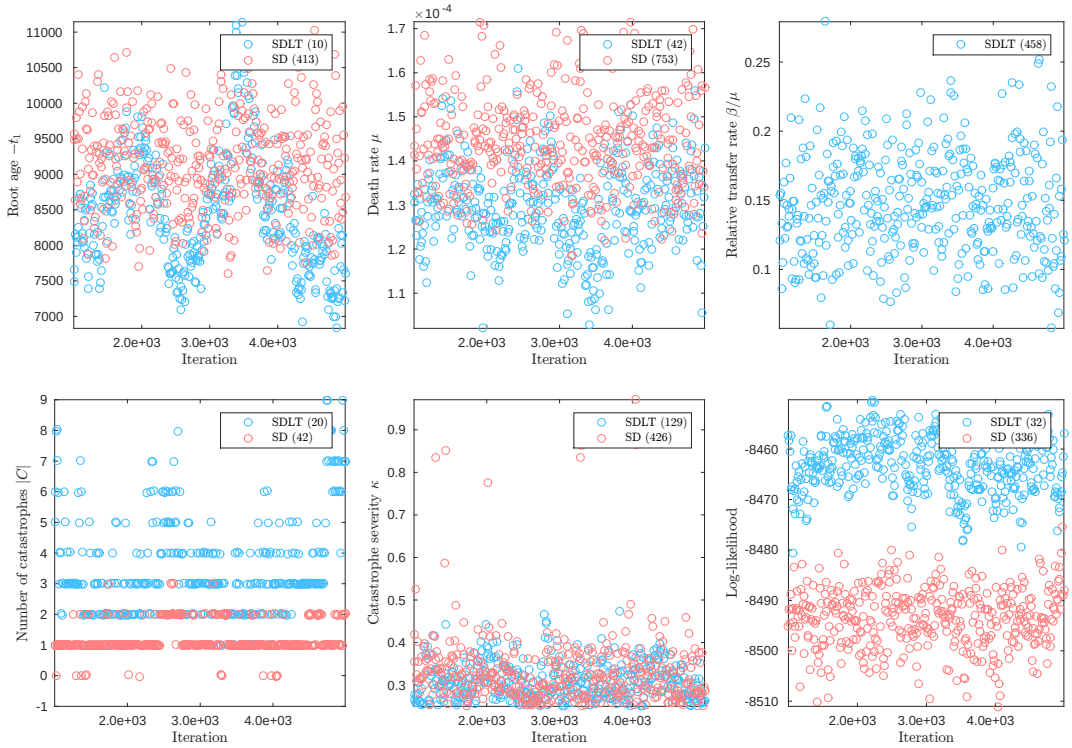


Figure B.10: Trace plots of samples in our analyses of IE.

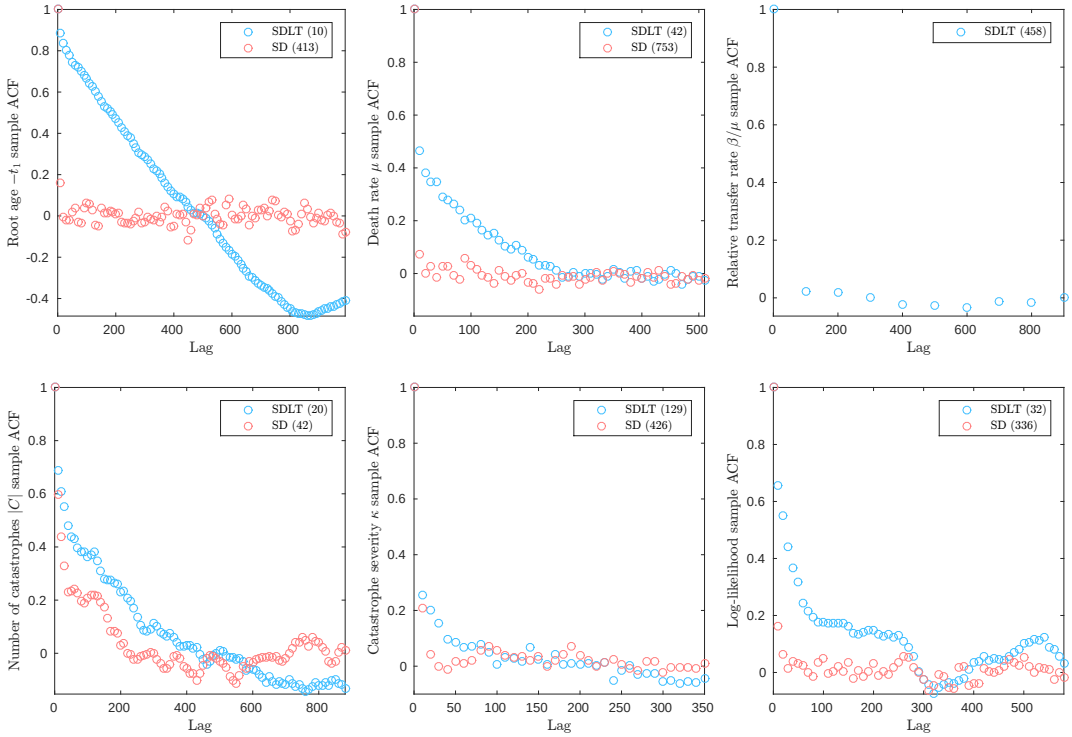


Figure B.11: Sample autocorrelation plots in our analyses of IE.