UNIVERSITY OF CALIFORNIA

Los Angeles

Scalable Inference in Bayesian Phylogenetics

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

by

Alexander A. Fisher

2021

ABSTRACT OF THE DISSERTATION

Scalable Inference in Bayesian Phylogenetics

by

Alexander A. Fisher
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2021
Professor Marc A. Suchard, Chair

Phylogenetic models with lineage-specific parameter characterizations provide a flexible framework to model ancestral changes in diffusion and evolution processes. However, increased taxonomic sampling challenges inference under these models as the number of unknown parameters grows with the number of taxa. To solve this problem, I develop scalable inference machinery as well as scalable models to permit the study of increasingly massive trees within a Bayesian phylogenetic framework. First, I introduce a method to compute the gradient of the trait data log-likelihood of the popular relaxed random walk model of trait diffusion with computational complexity that is linear with the number of tips in the tree. I use this gradient to build an efficient Hamiltonian Monte Carlo (HMC) sampler that simultaneously samples all branch-specific model parameters with high acceptance probability. Next, I propose a new, auto-correlated molecular clock rate model together with scalable inference methods. My approach permits estimating both the presence and location of local clocks without a priori knowledge of their placement and avoids inordinately shrinking clock-rates. Finally, I develop a shrinkage-based adaptive shift model that automatically detect the number and placement of shifts in adaptive trait optima along a tree. Leveraging recent fast closed-form gradient calculations, I build an efficient HMC sampler that scales inference under this new model. I demonstrate the speed and utility of each method via a range of applications, including the study of viral evolution and phenotypic trait data.

The dissertation of Alexander A. Fisher is approved.

Van Savage

Donatello Telesca

Janet Sinsheimer

Marc A. Suchard, Committee Chair

University of California, Los Angeles

2021

For Mom, my first teacher.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I thank Marc Suchard for his reliably sage advice, exceeding kindness, and patient faith in me. I am fortunate to have such an invested advisor and cannot imagine a better one. Moreover, I wish to thank Marc for instilling an esprit de corps among Suchard group members in the Gonda lab space. In this last, unprecedented year of remote work, I have severely missed the routine lunches, camaraderie, and inspiring research discussions with the group. I am exceedingly grateful to previous postdocs in the group, Xiang Ji, Andrew Holbrook and Aki Nishimura who often served as second mentors and helped me grow as a statistician. Additionally, I thank members of my dissertation committee, Janet Sinsheimer, Van Savage and Donatello Telesca for their thoughtful questions and comments. They have inspired me to more carefully consider implicit assumptions in phylogenetic models and helped guide ideas about future work. I especially want to thank my committee for their excellent classroom instruction in topics foundational and germane to this dissertation.

I am extremely grateful to my fellow bruins. I thank Gabe Hassler and Zhenyu Zhang for deep diving into phylogenetic topics and troubleshooting BEAST with me. Furthermore, I cannot imagine my time at UCLA without friendships forged in the fires of study, in particular, my Biomath cohort, Jason Lin, Benjamin Chu and Alfonso Landeros. Homemade milk tea, hikes through the Los Angeles mountains and a cold night in Joshua Tree have made my time at UCLA significantly better than it otherwise could have been. Outside of my cohort, pull-ups with Sam Christensen and mathematical philosophizing sessions with Bhaven Mistry truly enriched the days. Additionally, I thank Biomath staff Martha Riemer, Emily Fitch, Mark Lucas and David Tomita for helping navigate the logistics of being a graduate student in the 21st century.

During my time at UCLA I had the privilege to work with and learn from many excellent educators. I thank Alan Garfinkel and Jane Shevstov for introducing me to student-focused methods in pedagogy. I thank Rita Cantor for the opportunity to teach statistics to genetics students. I am especially grateful to Tony Friscia for the opportunity to teach for the UCLA

Cluster program.

I must also thank my many mentors at Florida State University that guided my path towards graduate school. In particular, I am grateful to David Ekrut for first inspiring me to pursue Biomathematics and to teach math at the ACE tutoring center. I thank Daniel Weingard for actively pushing me to pursue doctoral studies. I am especially grateful to Harsh Jain for believing in me, so very early on. Further, I wish to thank Frank Johnson for enthusiastically accepting me into his lab and for our many exciting discussions.

Three chapters of this dissertation resulted in or are anticipated to result in co-authored publications. Chapter 3 is a version of Fisher AA, Ji X, Zhang Z, Lemey P, Suchard MA. Relaxed random walks at scale. *Systematic Biology*. 70(2):258–267 (2021). Chapter 4 is a version of Fisher AA, Ji X, Nishimura A, Lemey P, Suchard MA. Shrinkage-based random local clocks with scalable inference. *Under review*. Chapter 5 presents a current project with Paul Bastide, Philippe Lemey and Marc Suchard.

I am grateful for support I received during the writing of this dissertation. I was supported by Systems and Integrative Biology Training Grant (NIH T32 GM008185) as well as graduate student research funds from Marc Suchard.

Finally, I wish to thank my family to whom I am indebted and whose unwavering support for me these last five years has meant the world. I am extremely grateful to my Mom for her tireless efforts to provide a work-haven this past year. Additionally, her active interest in my education from an early age has undoubtedly been one of the greatest gifts. I thank my brother, Michael, for always offering a listening ear and providing perspective that keeps me sane. I appreciate my sister, Kira, and her numerous home-cooked meals and desserts that fueled many days of dissertation work. I thank my nephew, Lucan, for graciously sitting through math lessons with me. I appreciate my sister, Natasha, and brother, Arturo, for their reliable positivity that always uplifts me. I thank my daughter, Gabriella, for always being happy to see me, even during my most stressful days. Furthermore, I am exceedingly indebted to my wife, Molly, whose loving patience supports me in immeasurable ways each and every day. Throughout these five years I have relied on Molly for so much. From keeping

me healthy and well-fed, to watching our daughter Gabriella while I write for hours on end, I can confidently say that this dissertation would be impossible without my rock, my wife, Molly. Last, I thank God for directing my steps so that I might arrive at this particular place, at this particular time, on this particular day. Thank you for my Rose garden.

VITA

2012-2016	B.S. Biomathematics and Psychology Florida State University Tallahassee, FL
2016-2018	Systems and Integrative Biology Training Grant (NIH)
2016-2018	Graduate Dean's Scholar Award
2017	Carol Newton Travel Award
2016-2018	M.S. Biomathematics University of California, Los Angeles Los Angeles, CA
2018	Teaching Assistant Mathematics for Life Sciences, University of California, Los Angeles
2019	Instructor Statistics for Geneticists Bootcamp Universiy of California, Los Angeles
2019-2020	Teaching Assistant Evolution of the Cosmos and Life University of California, Los Angeles
2020	Instructor (Teaching Associate) Stochastic Beasts: An Introduction to Probability and Statistics University of California, Los Angeles

PUBLICATIONS

Fisher AA, Ji X, Nishimura A, Lemey P, and Suchard MA. Shrinkage-based random local clocks with scalable inference. *Under review*

Fisher AA, Ji X, Zhang Z, Lemey P, and Suchard MA (2021). Relaxed random walks at scale. *Systematic Biology*, 70(2):258–267.

Dellicour, S., Lequime, S., Vrancken, B., Gill, M. S., Bastide, P., Gangavarapu, K., Matteson, N., Tan, Y., Du Plessis, L., Fisher, A. A., Nelson, M. I., Gilbert, M., Suchard, M. A., Andersen, K. G., Grubaugh, N. D., Pybus, O. G., and Lemey, P. (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*.

CHAPTER 1

Introduction

Phylogenetic methods offer a powerful framework to reconstruct evolutionary histories, order sequences of past events, or learn about trait dispersion whilst controlling for evolutionary history. Fundamentally, phylogenetic methods are a tool to understand the past. Phylogenetic models exhibit such utility in this domain that they are employed to study everything from the emergence and spread of infectious diseases (Pybus et al., 2012) to the history of our cosmos (Jofré et al., 2017). Not only is the scope of phylogenetic models quickly growing but also the complexity. Phylogenetic models with lineage-specific parameter characterizations are increasingly used to describe dynamic evolutionary processes that unfold along a tree. Two popular phylogenetic models of this flavor are the relaxed random walk (RRW) of trait data diffusion (Lemey et al., 2010) and the relaxed local clock model of molecular sequence substitution (Drummond et al., 2006). Each model describes a data generative process for separate types of data, however they both employ branch-specific parameters to capture the effects of ancestral biological or environmental changes that influence the generative processes. I revisit both of these models later in this work.

While increasingly complex phylogenetic models often generate increasingly closer approximations to the observed world, they almost always come at the cost of slowed inference. Inference speed may be explicitly hindered by model complexity or, in many cases, by the size of the data. This is particularly true of models with parameter space that grows with the number of taxa under study. The burden of large data sets is compounded by recent advances in portable genome sequencing (Quick et al., 2016) that make data collection more affordable and practical than ever before.

In this dissertation, I present methods to scale complex, biologically informed statistical

models to the study massive phylogenetic data sets. I begin Chapter 2 with motivation for studying phylogenetics. I continue in Chapter 2 to develop the basic phylogenetic framework used throughout this dissertation and discuss previously developed models and methods that comprise the bedrock upon which I build my contributions. Followed is a brief discussion of the Bayesian approach to statistical phylogenetics. I complete the chapter with proofs of two simple statistical results that are crucial to the mathematical developments of later chapters.

Chapters 3, 4, 5 stand as independent projects and can be read as such. In Chapter 3, I introduce the relaxed random walk model (RRW) of trait evolution developed by Lemey et al. (2010). The RRW is a phylogenetic Brownian diffusion model where traits at each node are multivariate normally distributed about their parent's value with branch-specific variance. Since branch-specific parameters grow with the number of tips in the tree, increased sampling challenges inference under the RRW. I address this challenge by developing a scalable method to efficiently fit RRWs and infer branch-specific variations in a Bayesian framework. I demonstrate the speed and scalability of my approach in studies of West Nile virus phylogeography and the evolution of mammalian life history traits.

In Chapter (4) I present a new autocorrelated molecular clock rate model together with scalable inference machinery to estimate the number, magnitude, and location of heritable local clock rate changes on an unknown phylogenetic tree. Previous state-of-the-art approaches assume the location of local clocks a priori (Worobey et al., 2014) or search through exhaustively large parameter space to determine whether each discrete branch starts a local clock (Drummond and Suchard, 2010). I leverage recent efficient gradient computation on the clock rate space (Ji et al., 2019) to implement an efficient Hamiltonian Monte Carlo sampler that scales my model to large trees. I test the speed of my model on a simulated dataset and further apply it to study local clocks in interspecies influenza strains as well as the adaptive radiation of rodents and other placental mammals.

In Chapter (5) I introduce the phylogenetic Ornstein-Uhlenbeck (OU) model of adaptive evolution towards an optimal trait value (Hansen, 1997). When an organism's environment dynamically shifts, the fitness landscape can shift as well. Optimal trait values are not fixed

in time. Previous approaches to detecting shifts are challenged by the combinatorial nature of their model induced parameter space (Bastide, 2017), ignore trait covariation (Khabbazian et al., 2016) or assume no convergent evolution (Uyeda and Harmon, 2014). To better model this phenomena, I propose a branch-specific optima OU, where each branch of the tree has its own optimal trait vector. Building off the developments of Chapters 2 and 3, I additionally develop an efficient Hamiltonian Monte Carlo sampler to learn about branch-specific optima at scale. I demonstrate the efficacy of my method by estimating the number and magnitude of shifts in 4-dimensional trait optima present in the ancestry of 100 Anolis lizard species.

Finally, Chapter (6) explores future directions and extensions to the models of the previous three chapters. I outline a general approach to incorporate traditional epidemiological data such as case counts within a phylogenetic framework. Furthermore, I propose a path towards implementing efficient inference under grouped shrinkage priors within the context of detecting shifts under a phylogenetic OU.

CHAPTER 2

Background

2.1 Why study trees?

In some cases, the utility of examining evolutionary trees is obvious. If you wish to learn about the relatedness of a set of species or date a viral spill-over event from an animal reservoir, constructing an evolutionary tree or finding the height of a rooted time tree undoubtedly undergirds the question of interest. In other cases, the utility of a phylogeny is more subtle. For example, one may wish to learn how fast a viral pathogen is diffusing across a geographic landscape or learn about correlation structure between a set of phenotypic traits. For the first case, suppose SARS-CoV-2 is sampled in Los Angeles and subsequently in New York City, it may only be concluded that the virus traveled the distance between the two cities (as opposed to there being a novel introduction) if the two samples are genetically correlated. Ignoring evolutionary history can lead to erroneous claims about speed of geographic diffusion. For the second case, consider two phenotypic traits, Felsenstein (1985) offers the example of brain size and body size. The topology of phylogenetic history informs us the degree to which the taxa under study have evolved as the same organism. This is important because we expect traits of recently diverged species to bear correlation that is an artifact of their evolutionary history. Thus phylogenetic trees offer insight into the effective sample size of our taxonomic observations and it may be fatal to ignore this possible source of confounding.

In summary, each of the seemingly disparate examples above benefits from what is broadly referred to as the phylogenetic comparative method. The phylogenetic comparative method, introduced by Cavalli-Sforza and Edwards (1967) and formalized and popularized by Felsen-

stein (1985) enables the study of trait data covariance whilst controlling for shared evolutionary history; see Paradis (2014); Cornwell and Nakagawa (2017) for an expository introduction. Here "trait data" describes any quantitative measurement associated with the taxa under study. Under this definition, "trait" is a broad term and may include, as previously noted, heritable morphological features and geographic coordinates of viral isolates, among others.

2.2 Phylogenetic framework

2.2.1 Notation

Consistent with the notation of the sequel, let $\mathscr{F}$ represent the topological structure of a rooted, bifurcating tree that describes the evolutionary history of N biological entities i.e. "taxa". Thus, $\mathscr{F}$ contains $2N-1$ nodes in total: a root node, together with N tips and $N-2$ internal nodes (see Figure 2.1). Throughout this work, tips number $\{1,\ldots,N\}$, I index internal nodes $\{N+1,\ldots 2N-2\}$ and $2N-1$ is reserved for the root. Let branch length t_i connects node $i \in \{1,\ldots,2N-2\}$ with its parent node $\text{pa}(i)$. Note, multifurcations may be cast into this framework via 0-length branches. Here, branch length t_i denotes actual time node i evolves independently.

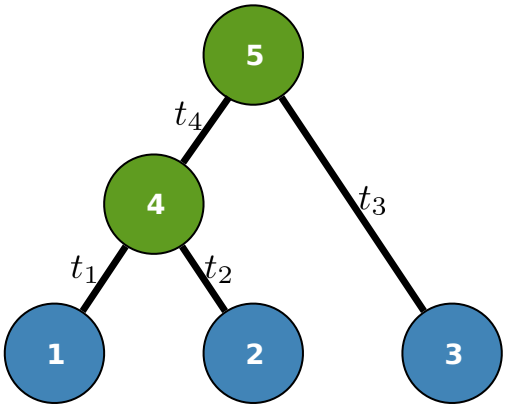


Figure 2.1: Example rooted bifurcating tree with $N = 3$ tips (blue). Internal (ancestral) nodes are green. Branch lengths t_i (units time) connect nodes.

2.2.2 Two data generative models

Within the setting of phylogenetic comparative methods, there are, broadly, two data generative processes at play. The first is concerned with the evolution of molecular sequence data $\mathbf{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$ while the second describes the diffusion of trait data $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$. Continuous time Markov chains (CTMCs) with finite state space (e.g. nucleotide characters, amino acids or codons) underlie most modern models of molecular evolution (O'Meara, 2012). CTMCs offer a flexible framework to model realistic biological processes such as multiple substitutions at a site, adjusted rates of silent mutations and site specific mutation speeds (Yang, 2014). Lange (2003) provides a concise review of CTMCs in the context of phylogenetic reconstruction. Drummond et al. (2006) introduce a popular model of molecular evolution, that I revisit in Chapter (4), the relaxed local clock (RLC). Under the RLC, sites evolve independently according to a CTMC defined by infinitesimal rate matrix $\mathbf{Q}$ and branch-specific clock rates $\mathbf{r} = \{r_1, \ldots, r_{2N-2}\}$ that represent the expected number of substitutions per unit time on each branch i in $\mathscr{T}$. The RLC is "relaxed" in the sense that variable mutation rates are allowed on each branch of the phylogeny. $\mathbf{Q}$ may be a 4×4, 20×20 or 61×61 matrix that describes the substitution process of nucleotides, amino acids, or codons respectively (Yang, 2014). The RLC accommodates possible site specific rate heterogeneity, such as the gamma distributed "hot spot" identifier of Yang (1996). All together, at site k, on branch i, the transition probability $\mathbf{P}_{ik}$ unwinds via the matrix exponential,

$$\mathbf{P}_{ik} = \exp\left\{r_i t_i s_k \mathbf{Q}\right\}. \tag{2.1}$$

where s_k denotes rate variation specific to site k. See table (2.1) to quickly reference relevant notation. At the heart of both likelihood-based and Bayesian phylogenetics, computing the likelihood of the observed sequence data $p(\mathbf{S})$ is an important task. To compute the full likelihood of observed data $\mathbf{S}$, we must marginalize over the latent sequence data at each internal node $i \in \{N+1, \ldots, 2N-1\}$ and root node $2N-1$,

$$p(\mathbf{S}) = \sum_{\mathbf{S}_{N+1}} \cdots \sum_{\mathbf{S}_{2N-1}} p(\mathbf{S}, \mathbf{S}_{N+1}, \ldots, \mathbf{S}_{2N-1}), \tag{2.2}$$

where

$$p(\mathbf{S}, \mathbf{S}_{N+1}, \ldots, \mathbf{S}_{2N-1}) = p(\mathbf{S}_{2N-1}) \prod_{i=1}^{2N-2} p(\mathbf{S}_i \,|\, \mathbf{S}_{\text{pa}(i)}). \tag{2.3}$$

A major theme of this dissertation is that direct computation of phylogenetic likelihoods, as well as other mathematical quantities of interest, are inefficient and greatly benefit from careful study. In equation (2.2) above, at each site, there are ℓ^{N-1} possible states (one for each of the $N-1$ internal nodes) where $\ell \in \{4, 20, 61\}$ for nucleotide, amino acid and codon models respectively. Thus, the possible combinations of ancestral states grows exponentially with the number of tips in the tree. Felsenstein (1973, 1981) introduces the tree-pruning algorithm, also known as the "peeling" algorithm that achieves $\mathcal{O}(N)$ computational complexity by grouping the sums. For example, in the three-tipped tree given by figure (2.1), the grouping unwinds,

$$p(\mathbf{S}) = \sum_{\mathbf{S}_5} p(\mathbf{S}_5) \left[\sum_{\mathbf{S}_4} p(\mathbf{S}_4 \,|\, \mathbf{S}_5) p(\mathbf{S}_1 \,|\, \mathbf{S}_4) p(\mathbf{S}_2 \,|\, \mathbf{S}_4) \right] p(\mathbf{S}_3 \,|\, \mathbf{S}_5), \tag{2.4}$$

so that all computations concerning node i are collected together and each node need be visited only once. For trees of arbitrary size, we begin at the tips and work our way up the tree by grouping sums according to parent-child node triples, $\sum_{\mathbf{S}_k} p(\mathbf{S}_k \,|\, \mathbf{S}_{\text{pa}(k)}) p(\mathbf{S}_i \,|\, \mathbf{S}_k) p(\mathbf{S}_j \,|\, \mathbf{S}_k)$ where i and j are sister daughters of k.

Equation (2.4) is but one example of simplifying cumbersome computation by exploiting the structure of the phylogenetic topology that the data generative process unfolds upon. Similar peeling-style algorithms greatly improve the computation of trait data likelihoods on trees as well. One such example is the popular Brownian diffusion model of trait evolution. Under the simple Brownian diffusion model, P-dimensional trait vector $\mathbf{Y}_i$ evolves according to a multivariate normal diffusion process,

$$\begin{aligned} p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\text{pa}(i)}) &= \text{MVN}\left(\mathbf{Y}_i;\ \mathbf{Y}_{\text{pa}(i)}, t_i \boldsymbol{\Sigma}\right) \\ p(\mathbf{Y}_{2N-1}) &= \text{MVN}\left(\mathbf{Y}_{2N-1};\ \boldsymbol{\mu}, \kappa_0^{-1} \boldsymbol{\Sigma}\right) \end{aligned} \tag{2.5}$$

where $\boldsymbol{\Sigma}$ is a $P \times P$ matrix that describes covariance between traits, $\boldsymbol{\mu}$ is the prior mean at the root and κ_0 reflects prior sample size. See figure (2.2) for a 2-dimensional diffusion example. I note here that covariance between traits is distinct from covariance induced by the tree, Ψ. Ψ is an $N \times N$ matrix with tip-to-root distance on its diagonals and the amount

Table 2.1: Glossary of phylogenetic notation.

Symbol	**Meaning**	**Sequence model**	**Trait model**
$\mathbf{S}$, $(\mathbf{S}_i)$	molecular sequence data (at node i)	✓	
$\mathbf{Y}$, $(\mathbf{Y}_i)$	trait data (at node i)		✓
t_i	length of branch i in units time	✓	✓
r_i	clock rate on branch i	✓	
s_k	among site rate variation at site k	✓	
$\mathbf{Q}$	infinitesimal rate matrix	✓	
$\mathbf{\Sigma}$	trait covariance		✓
$\mathscr{F}$	phylogenetic topology, often includes branch lengths	✓	✓
Ψ	tree variance, determined by topology	✓	✓

of time i and j evolved as one unitary entity (i.e. their tree-induced covariance) on its off-diagonals. For example, the tree covariance of figure (2.1),

$$\Psi = \begin{bmatrix} t_1 + t_4 & t_4 & 0 \\ t_4 & t_2 + t_4 & 0 \\ 0 & 0 & t_3 \end{bmatrix}. \tag{2.6}$$

Since the trait data evolves in a multivariate normal fashion, one can write the full joint distribution of likelihood $p(\mathbf{Y})$ by stacking the observed trait vectors $\mathbf{Y}_i$ into one long NP vector with stacked $\boldsymbol{\mu}$ mean,

$$\text{vec}(\mathbf{Y}) = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_N \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix}, \Psi \otimes \mathbf{\Sigma} \right), \tag{2.7}$$

where $\otimes$ denotes the Kronecker product. Naive evaluation of this full likelihood requires finding the determinant of $(\Psi \otimes \mathbf{\Sigma})^{-1}$, an $\mathcal{O}(N^3 P^3)$ task. Pybus et al. (2012) develop an

approach to compute $p(\mathbf{Y})$ with $\mathcal{O}(NP^3)$ complexity, an immense speed-up considering $P << N$.

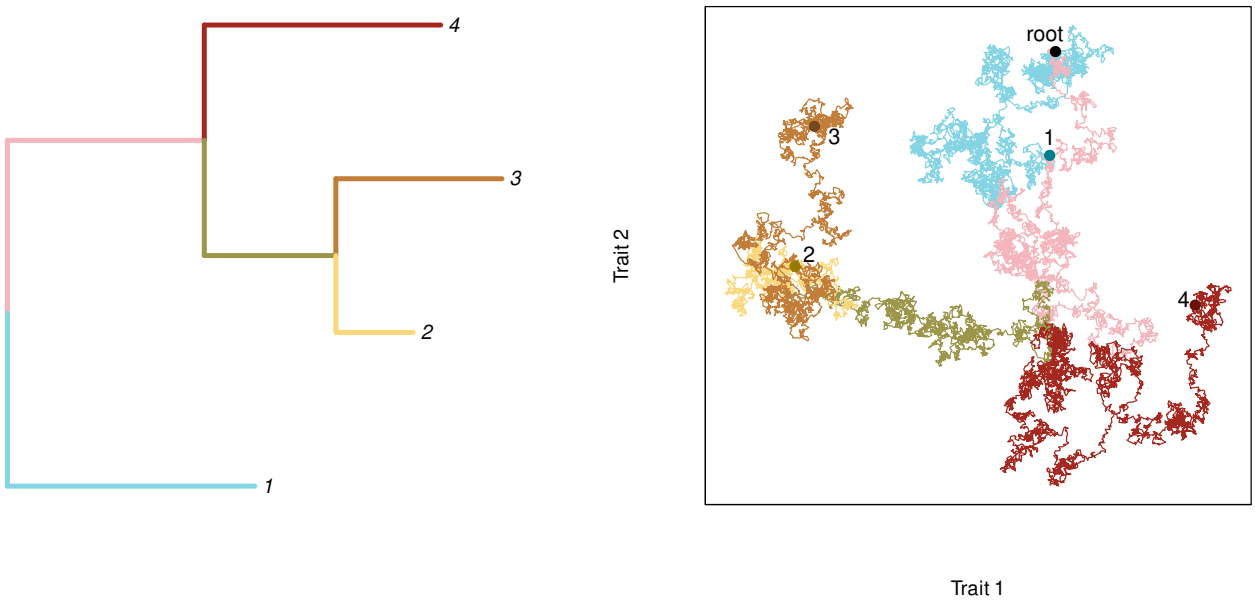


Figure 2.2: Four tipped tree (right). Example of simple Brownian diffusion (random walk) of two independent traits unfolding on a tree with four tips (left). Diffusion colors match branches. Colored circles indicate trait values at tips $i \in \{1, 2, 3, 4\}$ and receive corresponding labels.

Here I detail the fast trait data likelihood computation of Pybus et al. (2012) since I revisit and extend their technique in subsequent Chapters. To begin, let $\mathbf{Y}_{\lfloor i \rfloor}$ be the set of observed trait data descendent from node i. For example in figure (2.1), $\mathbf{Y}_{\lfloor 4 \rfloor} = \{\mathbf{Y}_1, \mathbf{Y}_2\}$. At the tips, $\mathbf{Y}_{\lfloor i \rfloor} = \mathbf{Y}_i$ while at the root, $p(\mathbf{Y}_{\lfloor 2N-1 \rfloor}) = p(\mathbf{Y})$. Just like in the sequence data likelihood, the trait data likelihood requires integration over latent internal node states,

$$p(\mathbf{Y}) = \int_{\mathbf{Y}_{N+1}} \cdots \int_{\mathbf{Y}_{2N-1}} p(\mathbf{Y}, \mathbf{Y}_{N+1}, \ldots, \mathbf{Y}_{2N-1}), \tag{2.8}$$

where

$$p(\mathbf{Y}, \mathbf{Y}_{N+1}, \ldots, \mathbf{Y}_{2N-1}) = p(\mathbf{Y}_{2N-1}) \prod_{i=1}^{2N-2} p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\text{pa}(i)}). \tag{2.9}$$

We group terms under the integrals in the spirit of pruning and notice each integrand sim-

plifies to contain three terms under our data subset notation,

$$\int_{\mathbf{Y}_k} p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_k) p(\mathbf{Y}_{\lfloor j \rfloor} \,|\, \mathbf{Y}_k) p(\mathbf{Y}_k \,|\, \mathbf{Y}_{\mathrm{pa}(k)}) \tag{2.10}$$

where again, i and j are sister daughters of k.

Therefore, to compute likelihood, we must unwind the recursively defined conditional partial likelihood

$$p(\mathbf{Y}_{\lfloor k \rfloor} \,|\, \mathbf{Y}_{\mathrm{pa}(k)}) = \int_{\mathbf{Y}_k} \underbrace{p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_k)}_{1} \underbrace{p(\mathbf{Y}_{\lfloor j \rfloor} \,|\, \mathbf{Y}_k)}_{2} \underbrace{p(\mathbf{Y}_k \,|\, \mathbf{Y}_{\mathrm{pa}(k)})}_{3}, \tag{2.11}$$

by noting that when i and j are tip nodes, each term of the integrand in (2.11) is multivariate normal due to our Brownian diffusion modeling assumption (see (2.5)). Since terms 1 and 2 share the same mean, namely $\mathbf{Y}_k$, their product is proportional to a Gaussian of equal dimension (see section (2.4.1)). Furthermore, the resulting Gaussian convolves with term 3 so that $p(\mathbf{Y}_{\lfloor k \rfloor} \,|\, \mathbf{Y}_{\mathrm{pa}(k)})$ is also proportional to a Gaussian, see section (2.4.2) for a general proof. To compute the likelihood $p(\mathbf{Y})$, we integrate over the latent state at the root,

$$\begin{aligned} p(\mathbf{Y}) &= p(\mathbf{Y}_{\lfloor 2N-1 \rfloor}) \\ &= \int_{\mathbf{Y}_{2N-1}} p(\mathbf{Y}_{\lfloor 2N-1 \rfloor} \,|\, \mathbf{Y}_{2N-1}) p(\mathbf{Y}_{2N-1}) \end{aligned} \tag{2.12}$$

where, for any node k in $\mathscr{T}$, it follows from careful bookkeeping that

$$p(\mathbf{Y}_{\lfloor k \rfloor} \,|\, \mathbf{Y}_k) \propto \mathrm{MVN}\left(\mathbf{Y}_k;\ \mathbf{m}_k, \mathbf{P}_k^{-1}\right), \tag{2.13}$$

where inverse-precision

$$\mathbf{P}_k^{-1} = \begin{cases} 0 \times \mathbf{I} & \text{if k is a tip} \\ (\mathbf{P}_i^* + \mathbf{P}_j^*)^{-1} & \text{otherwise,} \end{cases} \tag{2.14}$$

and

$$\mathbf{P}_i^* = \left(\mathbf{P}_i^{-1} + t_i \boldsymbol{\Sigma}\right)^{-1} \text{ for all i,} \tag{2.15}$$

and mean

$$\mathbf{m}_k = \begin{cases} \mathbf{Y}_k & \text{if k is a tip} \\ \mathbf{P}_k^{-1}(\mathbf{P}_i^* \mathbf{m}_i + \mathbf{P}_j^* \mathbf{m}_j) & \text{otherwise.} \end{cases} \tag{2.16}$$

To complete the likelihood calculation, we plug conditional partial likelihood (2.13) and root prior (2.5) into likelihood (2.12) and perform the convolution to find

$$p(\mathbf{Y}) \propto \mathrm{MVN}\left(\mathbf{m}_{2N-1};\ \boldsymbol{\mu}, \kappa_0^{-1}\boldsymbol{\Sigma} + \mathbf{P}_{2N-1}^{-1}\right). \tag{2.17}$$

2.3 Bayesian approach

Throughout this dissertation, I take a Bayesian approach to inference where the primary focus is to learn about the posterior distribution of model parameters $\boldsymbol{\theta}$, given all relevant sequence data $\mathbf{S}$ and trait data $\mathbf{Y}$. Bayesian statistical modeling provides a rich unified theory to approach inference problems where uncertainty in parameter estimates can be readily quantified and interpreted. Bayes theorem states the posterior

$$p(\boldsymbol{\theta} \,|\, \mathbf{S}, \mathbf{Y}) \propto \underbrace{p(\mathbf{S}, \mathbf{Y} \,|\, \boldsymbol{\theta})}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \tag{2.18}$$

with proportionality constant

$$\frac{1}{\int p(\mathbf{S}, \mathbf{Y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\ d\boldsymbol{\theta}}. \tag{2.19}$$

In Bayesian phylogenetic studies, model parameter space is often very large. For example, one might wish to learn about the tree topology and its corresponding branch lengths, parameters that characterize the molecular substitution model, including possibly branch-specific clock-rates, covariance between traits, branch-specific diffusion scalars and more. Therefore it is not surprising that computing the marginal likelihood (2.19) is unfeasible. For this reason, standard practice in Bayesian inference is to approximate the posterior, typically via Markov chain Monte Carlo (MCMC) sampling (Metropolis et al., 1953; Hastings, 1970); see Brooks et al. (2011) for an excellent review of several modern approaches. While MCMC is undoubtedly the most ubiquitous method to approximate Bayesian phylogenetic posteriors, other techniques, such as variational inference (Blei et al., 2017), are gaining popularity (Zhang and Matsen IV, 2018).

In each of the following Chapters, I study and develop phylogenetic models with branch-specific parameters. While branch-specific parameters often allow a flexible framework to

model increasingly realistic biological processes on trees, the number of parameters grows with the number of taxa under study. Therefore, increased taxonomic sampling slows inference, especially when using traditional inference methods. For this reason, I develop efficient Hamiltonian Monte Carlo (HMC) samplers that generate proposals in all parameter dimensions simultaneously and with high acceptance probability. To achieve this feat, HMC uses gradient information about the target density to generate proposals. Proposals are subsequently accepted or rejected with a traditional Metropolis-Hastings step (Neal, 2011; Betancourt, 2017).

2.4 Statistical background

Herein, I present statistical results that recur several times throughout this dissertation. For this reason, I report and sketch derivations of these basic results to add clarity and completeness.

2.4.1 Product of Gaussian densities

Claim: Let X be a random vector of arbitrary dimension. The product of two Gaussian probability density functions with shared mean (or identically, shared random variable) is proportional to a Gaussian density. More precisely,

$$\mathrm{MVN}(X;\ \boldsymbol{\mu}_1, \Sigma_1) \cdot \mathrm{MVN}(X;\ \boldsymbol{\mu}_2, \Sigma_2) \propto \mathrm{MVN}(X; \boldsymbol{\mu}, \Sigma)\,, \tag{2.20}$$

where

$$\begin{aligned} \Sigma &= \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1} \\ \boldsymbol{\mu} &= \Sigma\left(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2\right) \end{aligned} \tag{2.21}$$

Proof: The proof of (2.20) follows by writing out the product of two Gaussians and com-

pleting the square. Key algebraic steps are detailed below.

$$\begin{aligned}
&\mathrm{MVN}(X;\ \boldsymbol{\mu}_1, \Sigma_1) \cdot \mathrm{MVN}(X;\ \boldsymbol{\mu}_2, \Sigma_2) \\
&\qquad \propto \exp\left\{-\frac{1}{2}\left[(X-\boldsymbol{\mu}_1)^t \Sigma_1^{-1}(X-\boldsymbol{\mu}_1) + (X-\boldsymbol{\mu}_2)^t \Sigma_2^{-1}(X-\boldsymbol{\mu}_2)\right]\right\} \\
&\qquad \propto \exp\left\{-\frac{1}{2}\left[X^t\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)X - 2\left(\boldsymbol{\mu}_1^t\Sigma_1^{-1} + \boldsymbol{\mu}_2^t\Sigma_2^{-1}\right)X\right]\right\} \\
&\qquad \propto \exp\left\{-\frac{1}{2}\left[(X-\boldsymbol{\mu})^t\,\Sigma^{-1}(X-\boldsymbol{\mu})\right]\right\}.
\end{aligned} \tag{2.22}$$

2.4.2 Convolution of Gaussians

Claim: A convolution of independent, equal dimension Gaussians is Gaussian, i.e. let X and Y be of equal dimension,

$$\int \mathrm{MVN}(X;\ \boldsymbol{\mu}_1, \Sigma_1) \cdot \mathrm{MVN}(Y;\ \boldsymbol{\mu}_2, \Sigma_2)\, dX = \mathrm{MVN}(Z;\ \boldsymbol{\mu_3}, \Sigma_3) \tag{2.23}$$

where $Z = X+Y$ and $\boldsymbol{\mu_3} = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$ and $\Sigma_3 = \Sigma_1 + \Sigma_2$. Equivalently, the sum of independent Gaussian random variables is a convolution of Gaussians.

Proof: There are many ways to prove (2.23). The most direct approach is to write out the product of densities and again complete the square. Next, utilize the fact that the integral of a proper density is unity and the result follows. To avoid lengthy algebraic computation, I outline a shorter proof below. First, I equate a convolution of independent Gaussians with the distribution of a sum of Gaussian random variables and then prove the stability of the sum.

To begin, define $Z = X + Y$ and expand the density $p(Z = z)$,

$$\begin{aligned}
p(Z=z) &= p(X+Y=z) \\
&= \int p(X=x, Y=z-x)\ dx \\
&= \int p(Y=z-x \,|\, X=x,)p(X=x)\ dx \\
&= \int p(Y)p(X)\ dX
\end{aligned} \tag{2.24}$$

by independence. Recall $X \sim \text{MVN}(X;\ \boldsymbol{\mu}_1, \Sigma_1)$ and $Y \sim \text{MVN}(Y;\ \boldsymbol{\mu}_2, \Sigma_2)$ with $\Sigma_1 \perp \Sigma_2$ as above. It follows that the characteristic function of Z, $\phi_Z(\mathbf{t}) = \mathbb{E}[\exp\{i\mathbf{t}Z\}]$, is equal to the product of characteristic functions for X and Y,

$$\begin{aligned}
\phi_Z(\mathbf{t}) &= \phi_X(\mathbf{t}) \cdot \phi_Y(\mathbf{t}) \\
&= \mathbb{E}[\exp\{i\mathbf{t}X\}] \cdot \mathbb{E}[\exp\{i\mathbf{t}Y\}] \\
&= \exp\left\{ i\mathbf{t}^t(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{1}{2}\mathbf{t}^t(\Sigma_1 + \Sigma_2)\mathbf{t} \right\}.
\end{aligned} \tag{2.25}$$

Since the characteristic function uniquely defines the density, $p(Z) = \text{MVN}(Z;\ \boldsymbol{\mu_3}, \Sigma_3)$. Thus $\int p(Y)p(X)\ dX = p(Z) = \text{MVN}(Z;\ \boldsymbol{\mu_3}, \Sigma_3)$.

CHAPTER 3

Relaxed random walks at scale

3.1 Introduction

Phylogenetic comparative methods are an indispensable tool to study the evolution of biological traits across taxa while controlling for their shared evolutionary history that confounds the inference of trait correlation (Felsenstein, 1985). Modern comparative methods usually entertain continuous, multivariate traits, although extensions to mixed discrete and continuous outcomes are readily available (Ives and Garland Jr, 2009; Cybis et al., 2015). Approaches typically model trait evolution as a Brownian diffusion or "random walk" process that acts conditionally independently along the branches of a known or random phylogeny. Specifically, the observed or unobserved (latent) trait value of a node in a phylogeny arises from a multivariate normal distribution centered on the latent trait value of its ancestral node with variance proportional to the units of time between nodes. A strict Brownian diffusion model, however, is unable to accommodate the overdispersion in trait data that often emerges from real biological processes (Schluter et al., 1997). One such example arises when examining the dispersal rate of measurably evolving viral pathogens (Biek et al., 2007). For example, if birds serve as the viral host, migratory patterns may induce inhomogeneous dispersal rates over time (Pybus et al., 2012). In such cases, a strict Brownian diffusion model fails to capture, and therefore can also fail to predict, the spatial dynamics of an emerging epidemic.

Lemey et al. (2010) relax the strict Brownian diffusion assumption by introducing branch-rate multipliers that scale the variance of the Brownian diffusion process along each branch of the phylogeny. This "relaxed random walk" (RRW) model requires estimating $2N - 2$

correlated branch-rate multipliers, where N is the number of taxa in the phylogeny. Lemey et al. (2010) take a Bayesian approach to parameter estimation where they infer the posterior distribution of the branch-rate multipliers via Markov chain Monte Carlo (MCMC) employing a simple univariable Metropolis-Hastings (UMH) proposal distribution (Hastings, 1970). Since the rates remain correlated in the posterior, a random-scan (Liu, 2008) of UMH proposals inefficiently explores branch-rate space. Specifically, univariable samplers force accepted proposals to be very close together to avoid a large number of rejection steps in the Markov chain simulation. This results in high correlation between MCMC samples from the posterior, making point estimates of the branch-rate multipliers unreliable and slow to converge. In our study of the West Nile virus herein, the branch-rate multipliers are the slowest parameters to achieve sufficient effective sample sizes and therefore extend total run-time when jointly inferring the phylogeny structure. Furthermore, in our mammalian life history example, a UMH sampler fails to provide reasonable posterior estimates of branch-rate multipliers on a fixed phylogeny after 10 days of run-time. Despite this present drawback, RRWs find many impactful applications, e.g., in phylodynamics and phylogeography (Bedford et al., 2014; Faria et al., 2014).

To ameliorate the difficulties that high dimensional MCMC sampling presents, we propose adopting a geometry-informed sampling approach using Hamiltonian Monte Carlo (HMC). HMC equates sampling from a probability distribution with simulating the trajectory of a puck sliding across a frictionless surface warped by the shape of the distribution (Neal, 2011). To map from this statistical problem to the physical one, we view the MCMC samples of our branch-rate multipliers as the "position" of the puck and, then, for each positional dimension we introduce an associated momentum variable. In this way, we extend a D-dimensional parameter space to 2D-dimensional phase space (Betancourt, 2017) and traverse the 2D phase space via differentiating the Hamiltonian and using a numerical integration method to offer proposal states for our MCMC chain. This numerical integration may introduce small error, so we then accept or reject proposals according to the traditional Metropolis-Hastings algorithm (Hastings, 1970) with high acceptance rates. The major limitation to HMC is calculating the gradient of the log-posterior with respect to all position parameters

simultaneously. Previous approaches for calculating gradients on phylogenies have employed "pruning"-type algorithms (Felsenstein, 1981) that scale quadratically with the number of taxa in the tree (Bryant et al., 2005). Likewise, numerical approaches also scale quadratically.

In this paper, we derive a method to calculate the gradient with computational complexity that scales only linearly with the number of taxa. We implement our method in the BEAST software package (Suchard et al., 2018), a popular tool for the study and reconstruction of rooted, time-measured phylogenies. We demonstrate the speed and accuracy of our linear-order gradient HMC versus previous best practices by examining the spread of the West Nile virus across the Americas in the early 2000s. Finally, we use our technique to apply the RRW model to study the sensitivity of correlation estimates to model misspecification between mammalian adult body mass, litter size, gestation length, weaning age and litter frequency across 3650 mammals, thereby demonstrating the scalability of our HMC implementation to tackling a previously intractable problem.

3.2 Materials and Methods

3.2.1 Model and inference

Consider a known or random phylogeny $\mathscr{F}$ with N sampled tip nodes and $N-1$ internal and root nodes, each with an observed or latent continuous trait value $\mathbf{Y}_i \in \mathbb{R}^P$. To traverse the phylogeny $\mathscr{F}$, let node $\text{pa}(i)$ index the parent of node i with branch length t_i connecting the two nodes. Then under the RRW model,

$$\mathbf{Y}_i \sim \text{MVN}\left(\mathbf{Y}_{\text{pa}(i)}, t_i \mathbf{V}(\phi_i)\right), \tag{3.1}$$

where the $P \times P$ matrix-valued function $\mathbf{V}(\phi_i)$ characterizes the branch-specific multivariate normal (MVN) increment that defines the diffusion process. We parameterize this function in terms of an unknown $P \times P$ positive-definite matrix $\mathbf{\Sigma}$ that describes the covariation between trait dimensions after controlling for shared evolutionary history and an unknown

branch-rate multiplier ϕ_i. Typical choices include

$$\mathbf{V}(\phi_i) = \begin{cases} \phi_i \boldsymbol{\Sigma} & \text{rate-scalar parameterization}, \phi_i > 0, \\ \frac{1}{\phi_i}\boldsymbol{\Sigma} & \text{scale-mixture-of-normals parameterization}, \phi_i > 0, \\ e^{\phi_i}\boldsymbol{\Sigma} & \text{unconstrained parameterization}, \phi_i \in \mathbb{R}, \text{ and} \\ \boldsymbol{\Sigma} & \text{standard Brownian diffusion.} \end{cases} \tag{3.2}$$

To complete the RRW model specification, we adopt a prior density on the unobserved trait at the parentless root node,

$$p(\mathbf{Y}_{2N-1}) = \text{MVN}\left(\boldsymbol{\nu}_0, \kappa_0^{-1}\boldsymbol{\Sigma}\right) \tag{3.3}$$

with prior mean $\boldsymbol{\nu}_0$ and sample-size κ_0.

Letting $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{2N-2})$ and the observed data $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ at the tips, we are interested in learning about the posterior

$$p(\boldsymbol{\phi}, \boldsymbol{\Sigma}, s, \mathscr{T}, \boldsymbol{\theta} \,|\, \mathbf{Y}, \mathbf{S}) \propto \underbrace{p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{T}) p(\mathbf{S} \,|\, \mathscr{T}, \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\phi} \,|\, s) p(s) p(\boldsymbol{\Sigma}) p(\mathscr{T}, \boldsymbol{\theta})}_{\text{priors}}, \tag{3.4}$$

where s is an unknown parameter characterizing our prior on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ represents parameters of a molecular sequence substitution model for the evolution of aligned molecular sequence data $\mathbf{S}$. Note that we follow usual convention (Cybis et al., 2015) and assume that $\mathbf{Y}$ and $\mathbf{S}$ are conditionally independent given $\mathscr{T}$. We follow the example of Lemey et al. (2010) and place a log-normal prior distribution on $\boldsymbol{\phi}$ with mean 1 and standard deviation s. We further assume an exponential prior on s with mean $\frac{1}{3}$. In the examples that follow, we place one of two priors on the covariance structure $\boldsymbol{\Sigma}$. In our first example, we study the West Nile virus and follow the original modeling assumptions of Pybus et al. (2012). We assign a Wishart conjugate prior with scale matrix $\mathbf{I}_P$ and P degrees of freedom to $\boldsymbol{\Sigma}^{-1}$. In our second example, we study correlation between mammalian life history traits and employ a more general "separation strategy" whereby $\boldsymbol{\Sigma}$ is separated into a correlation matrix and diagonal variance matrix (Barnard et al., 2000; Zhang et al., 2006; Caetano and Harmon, 2019). We specify the eponymous "LKJ" prior (Lewandowski et al., 2009) on the correlation matrix and assign the diagonal of marginal variances a log-normal distribution with mean 0 and standard deviation of 4. The LKJ prior is uniform over the space of positive definite

correlation matrices and this is favorable for our purpose of comparing correlation estimates under contrasting models. Efficient application of the LKJ prior in phylogenetics is well described by Zhang et al. (2019).

We use MCMC integration to approximate this posterior using a random-scan Metropolis-within-Gibbs approach (Liu, 2008; Levine and Casella, 2006). One cycle of this scheme consists of sampling $\boldsymbol{\phi}, \boldsymbol{\Sigma}, s$ and then $(\mathscr{F}, \boldsymbol{\theta})$ via

$$p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F}) p(\boldsymbol{\phi} \,|\, s), \tag{3.5a}$$

$$p(\boldsymbol{\Sigma} \,|\, \boldsymbol{\phi}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F}) p(\boldsymbol{\Sigma}), \tag{3.5b}$$

$$p(s \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\boldsymbol{\phi} \,|\, s) p(s), \text{ and} \tag{3.5c}$$

$$p(\mathscr{F}, \boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, s, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{S} \,|\, \mathscr{F}, \boldsymbol{\theta}) p(\mathscr{F}, \boldsymbol{\theta}), \tag{3.5d}$$

where update (3.5d) is unnecessary when $\mathscr{F}$ is fixed, otherwise efficient sampling from the density $p(\mathscr{F}, \boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, s, \mathbf{Y}, \mathbf{S})$ is well described elsewhere, see for example Suchard et al. (2018). Updates (3.5b) and (3.5c) are straightforward due to the conjugate priors chosen in our model. We turn our focus to the remaining component of our scheme, namely sampling from $p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S})$.

3.2.2 Hamiltonian Monte Carlo

We wish to sample $\boldsymbol{\phi}$ jointly to avoid potentially high autocorrelation in the resulting MCMC chain. To this end, we propose using HMC and begin with a brief description of how HMC maps sampling from a probability distribution to simulating a physical system. In classical mechanics, the Hamiltonian is the sum of the kinetic and potential energy in a closed system. To build the connection, we introduce auxiliary momentum variable $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{2N-2})$ and write our Hamiltonian,

$$H(\boldsymbol{\phi}, \boldsymbol{\rho}) = \underbrace{-\log p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S})}_{\text{potential energy}} + \underbrace{\frac{1}{2} \boldsymbol{\rho}^t \mathbf{M} \boldsymbol{\rho}}_{\text{kinetic energy}}, \tag{3.6}$$

where the mass matrix $\mathbf{M}$ weights our momentum variables. The canonical distribution from statistical mechanics relates the joint density of state variables $\boldsymbol{\phi}$ and $\boldsymbol{\rho}$ to the energy in a

system via the relationship,

$$p(\boldsymbol{\phi}, \boldsymbol{\rho} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto e^{-H(\boldsymbol{\phi}, \boldsymbol{\rho})}. \tag{3.7}$$

Substituting our Hamiltonian into (3.7), we observe that $\boldsymbol{\phi}$ and $\boldsymbol{\rho}$ are independent and recognize the marginal density of $\boldsymbol{\rho}$ to be MVN. To start the HMC algorithm, we first sample $\boldsymbol{\rho}$ from this marginal density. Then by differentiating $H(\boldsymbol{\phi}, \boldsymbol{\rho})$, we generate Hamilton's equations of motion,

$$\begin{aligned} \frac{d\phi_i}{dt} &= +\frac{\partial H}{\partial \rho_i}, \text{ and} \\ \frac{d\rho_i}{dt} &= -\frac{\partial H}{\partial \phi_i} = \frac{\partial}{\partial \phi_i} \log p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \quad \text{for all } i = 1, \ldots, 2N-2. \end{aligned} \tag{3.8}$$

We can use the resulting vector field in conjunction with a variety of numerical integration techniques to propose new states of $\boldsymbol{\phi}$ for our MCMC chain. Consistent with typical construction (Neal, 2011), we use the leapfrog method for numerical integration, where we follow the trajectory of $\boldsymbol{\rho}$ for a half-step before updating $\boldsymbol{\phi}$. For a full discussion of HMC, see Neal (2011). Importantly, Hamilton's equations elicit a need to calculate the gradient $\left(\frac{\partial}{\partial \phi_1}, \ldots, \frac{\partial}{\partial \phi_{2N-2}}\right)^t \log p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S})$ at each chain step to traverse phase space and gradient computation can be costly.

3.2.3 Gradient of trait data log-likelihood

A practical HMC sampler demands efficient calculation of $\nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S})$. Differentiating the logarithm of (3.5a), we obtain

$$\frac{\partial}{\partial \phi_i} \log p(\boldsymbol{\phi} \,|\, \boldsymbol{\Sigma}, s, \mathscr{F}, \mathbf{Y}, \mathbf{S}) = \frac{\partial}{\partial \phi_i} \log p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F}) + \frac{\partial}{\partial \phi_i} \log p(\boldsymbol{\phi} \,|\, s). \tag{3.9}$$

Our log-normal prior choice for $\boldsymbol{\phi}$ renders evaluating the second term in Equation (3.9) trivial. Here we develop a general recursive algorithm for calculating $\nabla_{\boldsymbol{\phi}} \log p(\mathbf{Y} \,|\, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F})$. To facilitate this development, consider splitting $\mathbf{Y}$ into two disjoint sets relative to any node i in $\mathscr{F}$. We define $\mathbf{Y}_{\lfloor i \rfloor}$ as the observed data descendant of node i and $\mathbf{Y}_{\lceil i \rceil}$ as the observed data "above" (or not descendent of) node i. For clarity, see Figure (3.1).

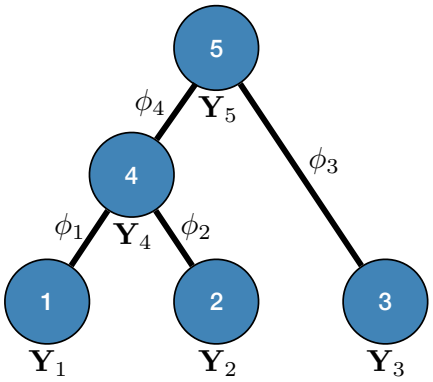


Figure 3.1: Example tree with $N = 3$ tips. Assume trait data $\mathbf{Y}_i$ are fully observed for $i = \{1, 2, 3\}$. We write $\mathbf{Y}_{\lfloor 4 \rfloor}$ and $\mathbf{Y}_{\lceil 4 \rceil}$ to denote the observed data below and above node 4 respectively. Specifically, $\mathbf{Y}_{\lfloor 4 \rfloor} = \{\mathbf{Y}_1, \mathbf{Y}_2\}$ while $\mathbf{Y}_{\lceil 4 \rceil} = \{\mathbf{Y}_3\}$. Partial likelihoods $p(\mathbf{Y}_{\lfloor 4 \rfloor} \,|\, \mathbf{Y}_4) = p(\mathbf{Y}_1, \mathbf{Y}_2 \,|\, \mathbf{Y}_4)$ and $p(\mathbf{Y}_4 \,|\, \mathbf{Y}_{\lceil 4 \rceil}) = p(\mathbf{Y}_4 \,|\, \mathbf{Y}_3)$.

In the following, we drop the dependence of the log-likelihood on $\boldsymbol{\phi}$, $\boldsymbol{\Sigma}$ and $\mathscr{T}$ for notational convenience. To begin,

$$\begin{aligned}
\frac{\partial}{\partial \phi_i} \left[\log p(\mathbf{Y})\right] &= \frac{\partial}{\partial \phi_i} \left[p(\mathbf{Y})\right] \Big/ p(\mathbf{Y}) \\
&= \frac{\partial}{\partial \phi_i} \left[\int p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i) p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) p(\mathbf{Y}_{\lceil i \rceil}) \mathrm{d}\mathbf{Y}_i \right] \Big/ p(\mathbf{Y}) \\
&= \int \frac{\partial}{\partial \phi_i} \left[p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i) p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) p(\mathbf{Y}_{\lceil i \rceil})\right] \mathrm{d}\mathbf{Y}_i \Big/ p(\mathbf{Y}) \\
&= \int p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i) \frac{\partial}{\partial \phi_i} \left[p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})\right] p(\mathbf{Y}_{\lceil i \rceil}) \mathrm{d}\mathbf{Y}_i \Big/ p(\mathbf{Y}).
\end{aligned} \tag{3.10}$$

The last equality above follows from the fact that ϕ_i is associated only with the branch above node i. Therefore when we condition on $\mathbf{Y}_i$, $\mathbf{Y}_{\lfloor i \rfloor}$ is independent of ϕ_i. Similarly, $\mathbf{Y}_{\lceil i \rceil}$ evolves independent of ϕ_i. To proceed with the differential above, we use the fact that $p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})$ follows a MVN distribution with as of yet undetermined mean $\mathbf{n}_i$ and precision $\mathbf{Q}_i$ (see section (3.5.1) for a detailed derivation). We extract the middle term from Equation

(3.10) and find

$$\frac{\partial}{\partial \phi_i}\left[p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})\right] = \frac{1}{2}\left\{(\mathbf{Y}_i - \mathbf{n}_i)^t\, \mathbf{Q}_i\, t_i \frac{\partial}{\partial \phi_i}\left[\mathbf{V}(\phi_i)\right] \mathbf{Q}_i\, (\mathbf{Y}_i - \mathbf{n}_i) \right. \\ \left. - \mathrm{tr}\left[\mathbf{Q}_i\, t_i \frac{\partial}{\partial \phi_i}\left[\mathbf{V}(\phi_i)\right]\right]\right\} p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}), \tag{3.11}$$

using the differential properties

$$\begin{aligned} \mathrm{d}\mathbf{Q}_i &= -\mathbf{Q}_i \left(\mathrm{d}\mathbf{Q}_i^{-1}\right) \mathbf{Q}_i \quad \text{and} \\ \mathrm{d}\left|\mathbf{Q}_i^{-1}\right| &= \left|\mathbf{Q}_i^{-1}\right| \mathrm{tr}\left[\mathbf{Q}_i \mathrm{d}\mathbf{Q}_i^{-1}\right], \end{aligned} \tag{3.12}$$

found in, e.g., Petersen et al. (2008).

To simplify notation, we let function

$$\mathbf{F}(\mathbf{Y}_i) = \frac{1}{2}\left\{(\mathbf{Y}_i - \mathbf{n}_i)^t\, \boldsymbol{\Upsilon}_i\, (\mathbf{Y}_i - \mathbf{n}_i) - \mathrm{tr}[\boldsymbol{\chi}_i]\right\}, \tag{3.13}$$

where $\boldsymbol{\Upsilon}_i = \mathbf{Q}_i t_i \frac{\partial}{\partial \phi_i}\left[\mathbf{V}(\phi_i)\right] \mathbf{Q}_i$ and $\boldsymbol{\chi}_i = \mathbf{Q}_i t_i \frac{\partial}{\partial \phi_i}\left[\mathbf{V}(\phi_i)\right]$. Substituting Equation (3.13) back into Equation (3.10), we observe that

$$\begin{aligned} \frac{\partial}{\partial \phi_i}\left[\log p(\mathbf{Y})\right] &= \int \mathbf{F}(\mathbf{Y}_i)\, p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i) p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) p(\mathbf{Y}_{\lceil i \rceil}) \mathrm{d}\mathbf{Y}_i \Big/ p(\mathbf{Y}) \\ &= \int \mathbf{F}(\mathbf{Y}_i)\, p(\mathbf{Y}_i \,|\, \mathbf{Y}) \mathrm{d}\mathbf{Y}_i \\ &= \mathbb{E}\left[\mathbf{F}(\mathbf{Y}_i) \mid \mathbf{Y}\right]. \end{aligned} \tag{3.14}$$

When $\mathbf{Y}_i$ is fully observed (typically $i \leq N$), this expectation collapses to the direct evaluation of $\mathbf{F}(\mathbf{Y}_i)$. When $i = N+1, \ldots, 2N-2$ or if $\mathbf{Y}_i$ is partially observed for $i = 1, \ldots, N$, we require $p(\mathbf{Y}_i \,|\, \mathbf{Y})$. From Bayes' theorem, $p(\mathbf{Y}_i \,|\, \mathbf{Y}) \propto p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i) p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})$. Partial likelihood $p(\mathbf{Y}_{\lfloor i \rfloor} \,|\, \mathbf{Y}_i)$ is proportional to a MVN density characterized by computable mean $\mathbf{m}_i$ and precision $\mathbf{P}_i$ (Pybus et al., 2012). Using this fact, $p(\mathbf{Y}_i \,|\, \mathbf{Y})$ becomes MVN with mean $\boldsymbol{\mu}_i = \mathbf{Z}_i\left(\mathbf{P}_i \mathbf{m}_i + \mathbf{Q}_i \mathbf{n}_i\right)$ and variance $\mathbf{Z}_i = \left[\mathbf{P}_i + \mathbf{Q}_i\right]^{-1}$. When tip i is partially observed, we partition $\mathbf{Y}_i = \left(\mathbf{Y}_i^{\mathrm{u}},\ \mathbf{Y}_i^{\mathrm{o}}\right)^t$ into its unobserved and observed entries. Using properties of the conditional MVN, $p(\mathbf{Y}_i \,|\, \mathbf{Y})$ becomes degenerate with mean

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mathbf{n}_i^{\mathrm{u}} - \left(\mathbf{Q}_i^{\mathrm{u}}\right)^{-1} \mathbf{Q}_i^{\mathrm{uo}} \left(\mathbf{Y}_i^{\mathrm{o}} - \mathbf{n}_i^{\mathrm{o}}\right) \\ \mathbf{Y}_i^{\mathrm{o}} \end{bmatrix} \tag{3.15}$$

and variance

$$\mathbf{Z}_i = \begin{bmatrix} (\mathbf{Q}_i^{\mathrm{u}})^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.16}$$

Finally, for both partially and completely unobserved cases above,

$$\mathbb{E}[\mathbf{F}(\mathbf{Y}_i) \mid \mathbf{Y}] = \frac{1}{2}\left\{\mathrm{tr}[\mathbf{Z}_i \boldsymbol{\Upsilon}_i] + (\boldsymbol{\mu}_i - \mathbf{n}_i)^t \boldsymbol{\Upsilon}_i (\boldsymbol{\mu}_i - \mathbf{n}_i) - \mathrm{tr}[\boldsymbol{\chi}_i]\right\}. \tag{3.17}$$

Equation (3.17) provides a recipe to compute $\nabla_{\boldsymbol{\phi}} \log p(\mathbf{Y} \mid \boldsymbol{\phi}, \boldsymbol{\Sigma}, \mathscr{F})$ using the means and precisions that characterize partial data likelihoods $p(\mathbf{Y}_i \mid \mathbf{Y}_{\lceil i \rceil})$ and $p(\mathbf{Y}_{\lfloor i \rfloor} \mid \mathbf{Y}_i)$.

3.2.4 Tree Traversals

We introduce post- and pre-order tree traversals to recursively calculate all partial data likelihood means and precisions in computational complexity $\mathcal{O}(N)$ that scales linearly with N. To begin, let nodes i and j be daughters of node k. Following Hassler et al. (2020), let $\boldsymbol{\delta}_k = \mathrm{diag}(\delta_{k1}, \ldots \delta_{kP})$ for $k = 1, \ldots, N$ be a diagonal matrix with indicator elements δ_{kp} that take value 1 if Y_{kp} is observed and 0 if not. For the post-order traversal,

$$p(\mathbf{Y}_{\lfloor i \rfloor} \mid \mathbf{Y}_i) \propto \mathrm{MVN}(\mathbf{Y}_i;\ \mathbf{m}_i, \mathbf{P}_i), \tag{3.18}$$

with post-order mean $\mathbf{m}_i$ and precision $\mathbf{P}_i$. For $k = 1, \ldots, 2N-1$ in post-order, we build the precision via

$$\mathbf{P}_k = \begin{cases} \infty \times \boldsymbol{\delta}_k & \text{if k is a tip} \\ (\mathbf{P}_i^* + \mathbf{P}_j^*) & \text{otherwise,} \end{cases} \tag{3.19}$$

with the definition that $\infty \times 0 = 0$ and

$$\mathbf{P}_i^* = \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \mathbf{V}(\phi_i)\, \boldsymbol{\delta}_i\right)^- \quad \text{and} \quad \mathbf{P}_j^* = \left(\mathbf{P}_j^- + t_j \boldsymbol{\delta}_j \mathbf{V}(\phi_j)\, \boldsymbol{\delta}_j\right)^-, \tag{3.20}$$

where the pseudo-inverse, defined and developed by Hassler et al. (2020); Bastide et al. (2018), is described in Appendix (3.5.2). At the tips, we build the mean $\mathbf{m}_k = \boldsymbol{\delta}_k \odot \mathbf{Y}_k$ where $\odot$ is the elementwise dot product, and for the internal nodes, $\mathbf{m}_k$ is a solution to

$$\mathbf{P}_k \mathbf{m}_k = (\mathbf{P}_i^* \mathbf{m}_i + \mathbf{P}_j^* \mathbf{m}_j). \tag{3.21}$$

For a proof of these post-order updates, see Hassler et al. (2020) (Supplemental Material).

To compute $p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})$, we traverse the tree in pre-order fashion according to our generalized version of the recursive algorithm proposed by Cybis et al. (2015). See section (3.5.1) for a derivation of our generalized pre-order update. For the pre-order traversal,

$$p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) = \text{MVN}(\mathbf{Y}_i;\ \mathbf{n}_i, \mathbf{Q}_i)\,. \tag{3.22}$$

For $i = 2N-1, \ldots, 1$ looking down the tree, we update our pre-order precision,

$$\mathbf{Q}_i = \begin{cases} \kappa_0 \boldsymbol{\Sigma}^{-1} & \text{if i is root} \\ \Big((\mathbf{Q}_i^*)^{-1} + t_i \mathbf{V}(\phi_i) \Big)^{-1} & \text{otherwise} \end{cases} \tag{3.23}$$

at each node where

$$\mathbf{Q}_i^* = \mathbf{P}_j^* + \mathbf{Q}_k. \tag{3.24}$$

We also keep track of the pre-order mean at each node via

$$\mathbf{n}_i = \begin{cases} \boldsymbol{\nu}_0 & \text{if i is root} \\ (\mathbf{Q}_i^*)^{-1} \big(\mathbf{P}_j^* \mathbf{m}_j + \mathbf{Q}_k \mathbf{n}_k \big) & \text{otherwise.} \end{cases} \tag{3.25}$$

Both traversals visit each node exactly once and perform a matrix inversion as their most costly operation, providing an $\mathcal{O}(NP^3)$ algorithm. However, as we observe in Equation (3.2), generally $\mathbf{V}(\phi_i) = g(\phi_i)\,\boldsymbol{\Sigma}$. In this case, we can further reduce the computational complexity to $\mathcal{O}(NP^2)$ by factoring out $\boldsymbol{\Sigma}$. Instead of inverting $\mathbf{V}(\phi_i)$ at each step, we only need to invert $\boldsymbol{\Sigma}$ at most once per likelihood or gradient evaluation.

3.3 Results

3.3.1 West Nile Virus

West Nile virus (WNV) is responsible for more than 1,500 deaths and caused over 700,000 illnesses since first reported in North America in 1999. The virus typically spreads via mosquito bites; however, the primary host is birds. First identified in New York City, WNV

spread to the Pacific coast by 2003 and reached south into Argentina by 2005 (Petersen et al., 2013). We examine whole aligned viral genomes (11,029 nt) and geographic data on 104 cases of WNV collected between 1999 and 2007 (Pybus et al., 2012). In cases where only the year of sampling is known, we set the sampling date to the midpoint of that year. Previous authors have recorded latitude and longitude geographic sampling information by converting zip code locations using ZIPList5. For 27 of the specimens, only the U.S. or Mexican state of discovery is known and so we have augmented sampling data with the coordinates of the centroid of the state (Pybus et al., 2012).

Here we study the simultaneous evolution and dispersal of WNV as it spreads across North America, following the modeling choices of Pybus et al. (2012). We define geographic location as our trait of interest $\mathbf{Y}$ within a RRW and infer rates $\boldsymbol{\phi}$ using our new HMC method. In two separate inference scenarios, we compare the computational efficiency of our method to the random-scan UMH approach employed by Pybus et al. (2012). The UMH kernel proposes new branch-rate multipliers individually by randomly scaling up or down the current ϕ_i. Under the UMH, all dimensions of $\boldsymbol{\phi}$ share the same adaptable tuning constant that controls the scaling range. Additionally, we compare our method to a less naive univariable proposal that provides each dimension of $\boldsymbol{\phi}$ its own adaptable tuning constant. We term this transition kernel multiple Metropolis-Hastings (MMH).

To begin, we set up a RRW model with log-normal prior on rates $\boldsymbol{\phi}$ with mean = 1 and standard deviation s and use a general time-reversible (GTR) + Γ substitution model with a log-normal relaxed molecular clock. We use the UMH transition kernel to run a 250 million state MCMC chain simulation to obtain posterior mean estimates of $\boldsymbol{\Sigma}$ and s. In scenario (**a**) we use these fixed model parameters and a topology drawn from the posterior to strictly sample $\boldsymbol{\phi}$ using HMC and univariable transition kernels. Under this fixed analysis we run our HMC-based chain for 1 million states and UMH/MMH-based chains for 150 million states. We use effective sample size (ESS) of the posterior ϕ_i samples for all i divided by computational runtime to evaluate the performance of each MCMC approach and report densities of ESS/second across all branches in Figure (3.2). ESS/second is averaged across five runs each with uniform (0-10) random initial branch-rate multipliers. The median

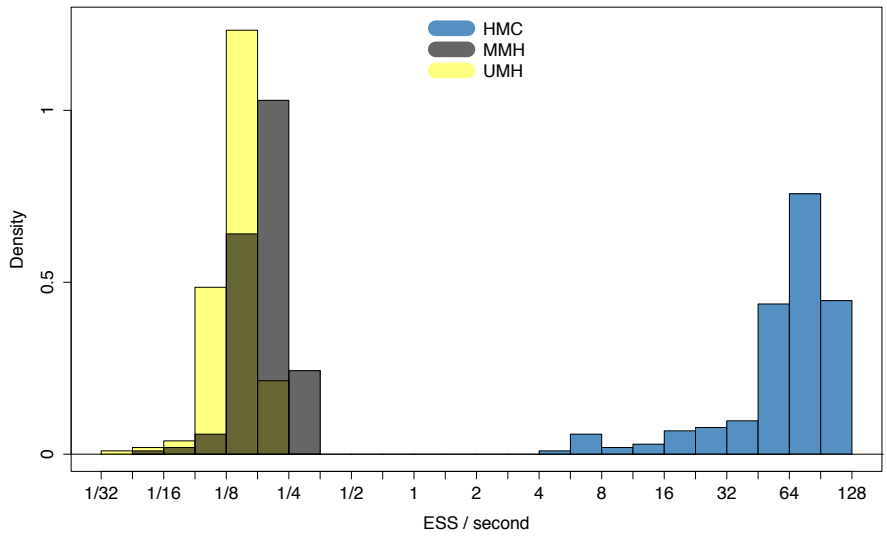


Figure 3.2: Comparing computational efficiency of Hamiltonian Monte Carlo (HMC) to univariable Metropolis-Hastings (UMH) and multiple Metropolis Hastings (MMH) transition kernels through effective sample size (ESS) per unit time in West Nile virus (WNV) phylogeography.

ESS/second across $\boldsymbol{\phi}$ is 71.0, 0.18 and 0.13 for the HMC, MMH and UMH transition kernels respectively. This demonstrates an over 394-fold speed increase. Additionally, the minimum ESS/second is 4.77 with HMC, 0.04 with MMH and 0.05 with UMH, exhibiting a 95-fold speed-up for the "least well" explored ϕ_i.

In scenario (**b**), we use a random starting tree and jointly estimate all parameters ($\boldsymbol{\phi}, \boldsymbol{\Sigma}$, s, $\mathscr{F}$ and $\boldsymbol{\theta}$) of the full posterior (3.4). Since branch-specific ϕ_i are no longer identifiable when $\mathscr{F}$ is random, we compare square jump distance across all $\boldsymbol{\phi}$ between samples from the posterior under both MCMC regimes to compare efficiency. We run HMC chains for 22.5 million states so that we are sampling from the posterior distribution of all parameters and we save the state of BEAST. Subsequently, we run both HMC and UMH chains from the same saved states and compute lag-7 square jump distance to adjust for the relative weight of the transition kernel in the full analysis. Since the UMH sampler updates only one branch-rate multiplier at a time, we compare square jump distance between samples of our HMC chain with samples from the UMH chain that are lagged $(2N-2)\times$ (i.e. $206\times$) farther

apart. We run each MCMC simulation until we obtain 5000 samples from the posterior and report the average median across five separate runs. In this comparison, we find that the average median square jump distances from five separate runs is 1457 and 128.2 for the HMC and UMH chains respectively.

In Figure (3.3) we report the MCC tree, obtained from applying HMC to the RRW model as described in scenario (**b**), where substitution rate variation is accounted for by the molecular clock model. The branch with the highest posterior dispersal rate starts the WN02 lineage identified by Gray et al. (2010). The clade of New York isolates sampled in 1999, however, maintains a much slower dispersal rate. We obtain results under this joint inference test by running an MCMC chain until we observe ESS > 200 for all parameters of interest, namely the height of the tree, substitution model parameters, the diffusion matrix $\boldsymbol{\Sigma}$, prior standard deviation s and 90% of the dimensions of $\boldsymbol{\phi}$. We choose only 90% because many dimensions of $\boldsymbol{\phi}$ exhibit multi-modality and therefore experience poor mixing when the tree is random. Under the UMH transition kernel this analysis takes approximately 45.8 hours. Under HMC, this analysis completes in 7.1 hours, a 6.4-fold speed-up. We report average times across five runs. The ESS-limiting parameters in each case are the multi-modal branch-rate multipliers.

3.3.2 Mammalian Life History

Life history theory aims to explain how traits such as adult body mass, litter size and lifespan evolve to optimize reproductive success (Stearns, 2000). Life history theory finds important applications in determining a species' fecundity and predicting extinction risk in response to changing environmental stimuli (Pacifici et al., 2017; Fritz et al., 2009; de Silva and Leimgruber, 2019), but due to the sparseness of much life history data, it is essential to understand how traits covary to make meaningful predictions (Santini et al., 2016). To determine which traits covary, comparative mammalian life-history studies posit a 'fast–slow' continuum, claiming small mammals are typically 'fast', characterized by early maturation, large litters and shorter lifespans, while larger mammals are typically 'slow' and present

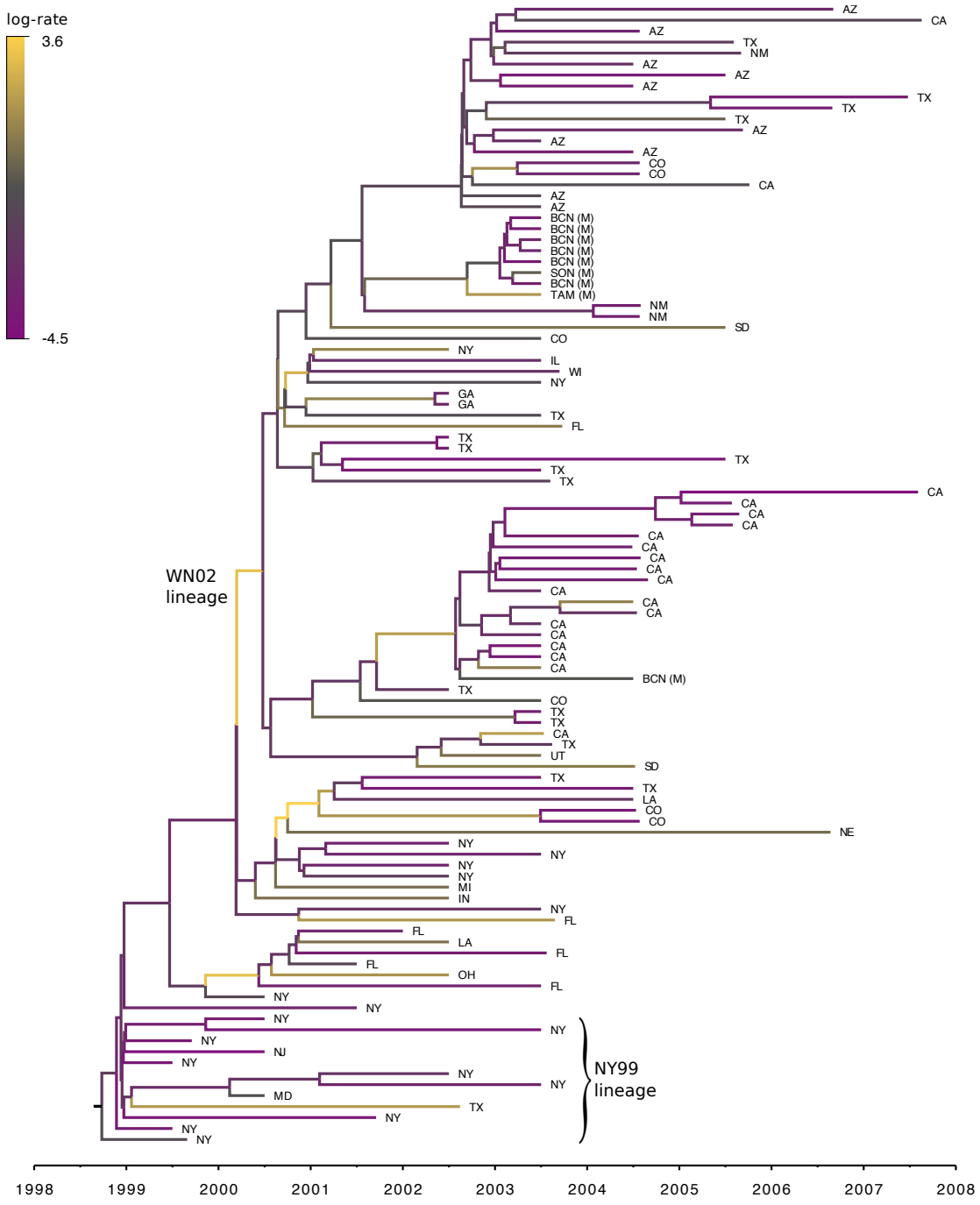


Figure 3.3: Maximum clade credibility (MCC) tree resulting from Hamiltonian Monte Carlo (HMC) inference under phylogeographic relaxed random walk (RRW) of West Nile virus. We color branches by posterior mean branch-rate parameters ϕ. Tips are labeled according to the US or Mexican state of origin.

contrasting characteristics (Oli, 2004; Millar and Zammuto, 1983). Under this framework, certain traits such as gestation length, weaning age and body mass are predicted to be positively correlated, but reported estimates of positive correlation from data may be artifacts of the restrictive assumptions of strict Brownian diffusion modeling. Here we re-evaluate this claim by comparing inferred trait correlation under the strict Brownian diffusion model with estimates under the RRW of trait evolution made tractable through $\mathcal{O}(N)$ HMC sampling.

Under the RRW, we infer correlation between five life history traits from the PanTHERIA data set (Jones et al., 2009), namely body mass, litter size, gestation length, weaning age and litter frequency across 3650 mammalian species related by the fixed supertree of Fritz et al. (2009). To obtain this subset of the supertree, we only consider taxa for which at least one of these five traits is observed. We take the intersection of this set of taxa with those in the fixed supertree of Fritz et al. (2009) and prune all other observations from the tree. We log-transform and standardize the trait measurements and subsequently estimate posterior mean correlations between each pair of traits under the RRW using an HMC-based chain for 300 thousand states. We model diffusion using the rate-scalar parameterization of $\mathbf{V}(\phi_i)$ noted in Equation (3.2). This modeling choice assumes that all taxa share a common correlation structure across the tree. To gauge the effect of a heterogeneous diffusion process on the correlation between traits, we also make inference using the strict Brownian diffusion model where the $\boldsymbol{\phi}$ are all identically 1. Here we perform MCMC inference on the diffusion matrix $\boldsymbol{\Sigma}$ for 50 thousand states. Under the RRW, we find the variance in body mass is 1.17 with 95% high posterior density (HPD) interval $\{1.02, 1.33\}$, gestation length is 0.73 $\{0.62, 0.83\}$, weaning age is 2.94 $\{2.49, 3.39\}$, litter frequency is 5.47 $\{4.62, 6.40\}$, and litter size is 2.82 $\{2.47, 3.17\}$. We report posterior mean estimates of correlation between each pair of traits under both the RRW and strict Brownian diffusion in Figure (3.4). In most cases, the RRW reassuringly confirms analysis under the more limited model. However, in some cases our confidence in the sign of the correlation differs between models and in one instance the sign of the posterior mean correlation disagrees. Under the RRW, we observe positive posterior mean correlation of 0.017 between litter frequency and litter size with posterior odds ratio 1.96 that the correlation is positive. Under the strict Brownian diffusion model

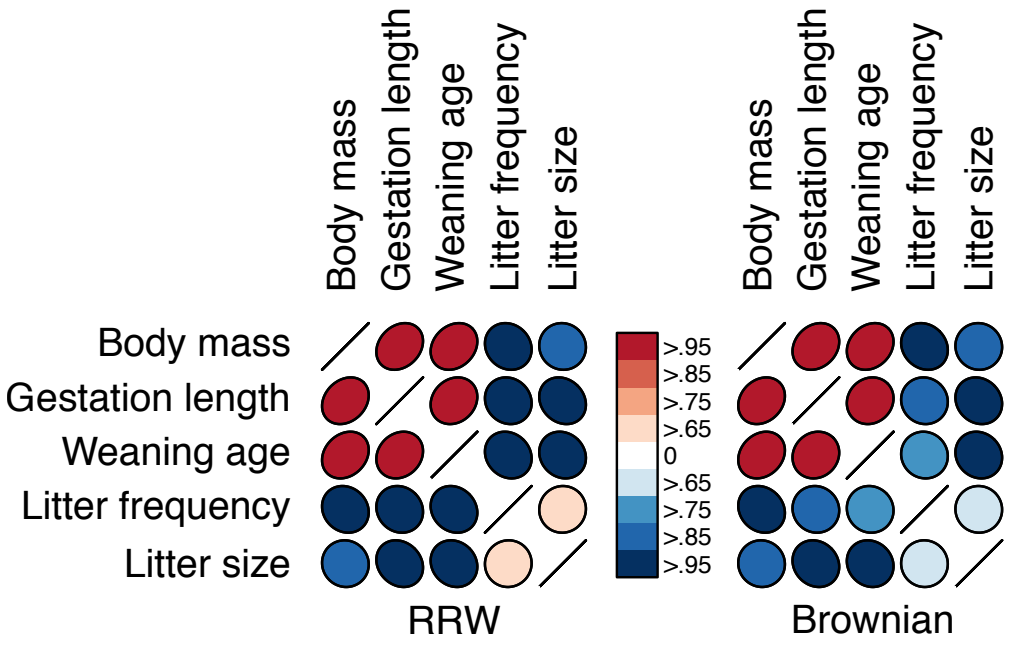


Figure 3.4: Posterior mean correlation between mammalian life history traits under the RRW and strict Brownian diffusion model. Shape of ellipse indicates strength and sign of correlation, while colors indicate the posterior probability that the correlation is positive (red) or negative (blue).

we observe a negative posterior mean correlation of -0.015 with posterior odds ratio 1.90 of being negative, indicating slightly weaker belief in the correlation's sign under the strict model.

3.4 Discussion

Previous MCMC techniques to investigate trait evolution under the RRW model scale poorly with large data sets. Specifically the UMH transition kernel is ineffective for sampling correlated, high dimensional parameter space. We provide a remedy by using an HMC transition kernel to sample all branch-rate multipliers simultaneously. To improve the speed of HMC we derive an algorithm for calculating the gradient of the trait data log-likelihood. This gradient calculation achieves $\mathcal{O}(N)$ computational speed, a vast improvement compared to both numerical and pruning methods for calculating the gradient that typically require $\mathcal{O}(N^2)$.

We observe over 300-fold speed-up when comparing, on a fixed phylogeny, our HMC transition kernel to both MMH and UMH in the spread of the WNV across North America in the early 2000s. Additionally, we note here that HMC on the branch-rate multipliers also improves sampling of hyperparameter s as suggested by Equation (3.5c). HMC offers an over 6-fold speed increase in total run-time when jointly estimating parameters of the RRW, substitution model and phylogeny. The resulting MCC tree reveals that the largest dispersal rate precedes the most recent common ancestor of the WN02 lineage. Subsequently, the dispersal rates slow down through the WN02 clade. This suggests that this clade developed after some rapid geographic displacement. Interestingly, the appearance of smaller branch-rate multipliers within the WN02 lineage is consistent with the slowing speed of sequence evolution as described in Snapinn et al. (2007).

As exhibited in Figure (3.2), ESS from posterior sampling accumulates at variable speed across the branches of the tree. To further improve the sampling of our HMC algorithm, one might use an approximation of the posterior covariance of $\boldsymbol{\phi}$ for the mass matrix $\mathbf{M}$ to appropriately weight momentum updates in the HMC algorithm (Neal, 2011). Possible

approximations include the Hessian of the log-posterior (a local approximation of the curvature of branch-rate multiplier space) or the sample variance across each dimension. An important consideration in choosing an appropriate $\mathbf{M}$ is whether one is studying under a fixed or random phylogeny $\mathscr{F}$. Since varying $\mathscr{F}$ in the posterior often creates multimodal distributions of $\boldsymbol{\phi}$, local approximations such as the Hessian may be of limited assistance in such cases.

We show in our application to mammalian life history that our computationally efficient HMC algorithm imbues the RRW model with the ability to handle large trees with thousands of taxa. By applying the RRW model to this massive example we confirm that large mammals have 'slower' life history characteristics, exhibited by the positive correlation between body mass, gestation length and weaning age, while smaller mammals scale in the opposite manner and tend to have high litter frequency and size, see Figure (3.4). The posterior mean correlation between litter frequency and litter size changes sign under each model, but with low posterior probability reflecting a lack of correlation between these traits. Note that the diffusion variance choices listed in Equation (3.2) all assume that the branch-rate multipliers scale each trait equally. Future modeling work could relax this assumption by letting each element of the diffusion matrix be a function of the branch-rate multipliers. Importantly, our method allows us to obtain posterior estimates for the correlation matrix in 32 hours while the previous UMH method fails to estimate the correlation matrix and branch-rate multipliers with greater than 200 ESS after 10 days.

In a time where biological data are more prolific than ever, scalable approaches to complex models of evolution such as the RRW prove increasingly useful in a variety of applications. From spatial epidemiology where determining the dispersal rate of an infectious disease is crucial, to evolutionary ecology where understanding life-history can provide insight into declining animal populations, the analysis of data is becoming a bottleneck to the scientific process and the need for computationally faster approaches stands evident. We hope that this work will serve to improve the speed of such analyses.

3.5 Appendices

3.5.1 Pre-order partial likelihood

Here we derive a generalized version of the pre-order recursive algorithm proposed by Cybis et al. (2015) to compute $p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil})$ for all i in $\mathcal{O}(N)$. We begin with the law of total probability,

$$p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) \propto \int p(\mathbf{Y}_i \,|\, \mathbf{Y}_k) p(\mathbf{Y}_{\lfloor j \rfloor} \,|\, \mathbf{Y}_k) p(\mathbf{Y}_k \,|\, \mathbf{Y}_{\lceil k \rceil}) \, \mathrm{d}\mathbf{Y}_k \tag{3.26}$$

for node i with parent k and sibling j. Recalling that

$$\begin{aligned} p(\mathbf{Y}_i \,|\, \mathbf{Y}_k) &= \mathrm{MVN}\left(\mathbf{Y}_i;\ \mathbf{Y}_k, t_i \mathbf{V}(\phi_i)\right), \text{ and} \\ p(\mathbf{Y}_{\lfloor j \rfloor} \,|\, \mathbf{Y}_k) &\propto \mathrm{MVN}\left(\mathbf{Y}_k;\ \mathbf{m}_j, (\mathbf{P}_j^*)^{-1}\right), \end{aligned} \tag{3.27}$$

we identify Equation (3.26) as a recursive expression whose solution has the form

$$p(\mathbf{Y}_i \,|\, \mathbf{Y}_{\lceil i \rceil}) = \mathrm{MVN}\left(\mathbf{Y}_i;\ \mathbf{n}_i, \mathbf{Q}_i\right), \tag{3.28}$$

with presently undetermined pre-order mean $\mathbf{n}_i$ and pre-order precision $\mathbf{Q}_i$.

We unravel these quantities by first identifying that $p(\mathbf{Y}_{2N-1} \,|\, \mathbf{Y}_{\lceil 2N-1 \rceil}) = p(\mathbf{Y}_{2N-1})$ and set $\mathbf{n}_{2N-1} = \boldsymbol{\nu}_0$ and $\mathbf{Q}_{2N-1} = \kappa_0 \boldsymbol{\Sigma}^{-1}$. Then proceeding in pre-order fashion for $i = 2N-2, \ldots, 1$

$$\begin{aligned} \mathbf{Q}_i &= \left((\mathbf{Q}_i^*)^{-1} + t_i \mathbf{V}(\phi_i) \right)^{-1} \text{ where} \\ \mathbf{Q}_i^* &= \mathbf{P}_j^* + \mathbf{Q}_k, \text{ and} \\ \mathbf{n}_i &= (\mathbf{Q}_i^*)^{-1} \left(\mathbf{P}_j^* \mathbf{m}_j + \mathbf{Q}_k \mathbf{n}_k \right). \end{aligned} \tag{3.29}$$

3.5.2 Pseudo-inverse

The pseudo-inverse used in the post-order tree traversal and defined by Bastide et al. (2018) and Hassler et al. (2020) is an operation for inverting precision and variance matrices with diagonal entries that take the value ∞. To invert a diagonal precision matrix, $\mathbf{P}_i$ with entries ∞ and 0, we define $\infty^- = 0$ and $0^- = \infty$. To invert the variance matrix, $\left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \mathbf{V}(\phi_i) \, \boldsymbol{\delta}_i \right)$

we invert the block matrix of observed trait covariation and invert the remaining diagonal elements using the convention that $\infty^{-} = 0$.

CHAPTER 4

Shrinkage-based random local clocks with scalable inference

4.1 Introduction

Molecular clock models are ubiquitous phylogenetic instruments for divergence-time estimation with applications ranging from timing placental mammal radiation (Springer et al., 2003) to estimating influenza diversity (Davidson et al., 2014). To capture clock rate variation along the lineages of a phylogeny, Thorne et al. (1998) propose an autocorrelated, or "heritable" rate model while others (Drummond and Suchard, 2010; Yoder and Yang, 2000) assume there exist, at most, a small number of "local" clocks on any given tree. In each case, closely related lineages maintain similar or even identical evolutionary rates. Autocorrelated rate models are computationally appealing due to the induced smooth transition in rate from parent to child node along the tree but may inappropriately shrink large rate changes between adjacent nodes (Smith et al., 2010). On the other hand, local clock models allow large rate changes to exist but can be computationally unpalatable on large problems due to the combinatorial complexity of choosing (or learning) the number and location of local clocks. When these quantities are simultaneously learned with the tree, Drummond and Suchard (2010) call this the random local clock (RLC) model.

Due to these complications, some authors employ uncorrelated relaxed clocks such as the uncorrelated log-normal relaxed molecular clock (Drummond et al., 2006), but this generates excessive rate heterogeneity in cases where clock rate changes are thought to be more punctuated, for example between HIV subtypes (Bletsa et al., 2019). For a more in-depth review of various molecular clock models, see Ho and Duchêne (2014). Here we propose an auto-

correlated clock model where we place a Bayesian bridge shrinkage prior on the increment between parent and child log branch rates. Among various shrinkage priors in the literature, the Bayesian bridge has a unique advantage in having both a collapsed spike-and-slab representation as well as a Gaussian scale-mixture form. The first representation intuitively places large mass near zero reflecting our *a priori* belief that most increments should be zero but has heavy tails that allow for estimating large rate changes in an approximately unbiased manner. Like many other shrinkage priors, the Bayesian bridge includes a "global scale" nuisance parameter about which learning, in the absence of prior information, typically limits the speed of inference. Polson et al. (2014) develop a framework to facilitate efficient Gibbs sampling of this nuisance parameter in a regression context and we utilize their approach here. On the other hand, the second representation of the Bayesian bridge as Gaussian scale-mixture is differentiable almost everywhere. We exploit this feature and develop an efficient Hamiltonian Monte Carlo (HMC) sampler over the space of increments that employs recent work on closed form gradient representations (Ji et al., 2020) to make our shrinkage-clock inference scalable to large trees. Crucial to our inference, we define recursive algorithms to compute the requisite joint gradient of the log posterior in our transformed increment space with computational complexity that scales only linearly with the number of tips in the tree. We implement our method in BEAST (Suchard et al., 2018), a popular software package for reconstructing rooted, time-measured phylogenies. Due to our efficient inference machinery, our shrinkage-clock achieves the tractable benefits of the autocorrelated rate model and simultaneously maintains the flexibility of more punctuated local clock models.

We demonstrate the inference gains of our approach versus the RLC across 20 different simulated data sets of a 40 taxa tree. We additionally compare the accuracy of our shrinkage-clock to the RLC by studying the adaptive radiation of rodents and other mammals, and demonstrate utility of the heavy-tailed Bayesian bridge shrinkage prior by comparing it to the more ubiquitous Laplace prior. Finally, we deploy our shrinkage-clock to estimate the existence, location and magnitude of host-specific clock rates in surface glycoproteins of the influenza A virus.

4.2 Shrinkage-based random local clocks

4.2.1 Setup

Consider a rooted, bifurcating tree $\mathscr{F}$ with N tips and $N-1$ internal (ancestral) nodes. We index tips $i = 1, \ldots, N$ and internal nodes $i = N+1, \ldots, 2N-2$. We designate node $2N-1$ to be the root of the tree. Let $\text{pa}(i)$ denote the parent of the ith node and let branch length t_i connect node i with its parent.

4.2.2 The relaxed clock

Aligned molecular sequence data $\mathbf{S}$ evolve according to a continuous time Markov process defined by infinitesimal rate matrix $\mathbf{Q}$. In our examples, $\mathbf{Q}$ is a 4×4 matrix that describes the relative substitution process between nucleotides along the branches in $\mathscr{F}$, but in general, $\mathbf{Q}$ may be of larger dimension to accommodate alignments at the codon or amino acid resolution, see Yang (2014) for reference. Each site k of $\mathbf{S}$ evolves independently and identically according to $\mathbf{Q}$ but may have its own site-specific rate of evolution s_k. A priori we specify that $\mathbb{E}[s_k] = 1$. Under the relaxed clock model, the transition probability matrix for branch i,

$$\mathbf{P}_i = \exp\left\{\rho_i t_i s_k \mathbf{Q}\right\}, \tag{4.1}$$

where branch-rate multiplier ρ_i is the number of expected substitutions per unit time. To resolve identifiability issues between the height of the tree and branch-rate multipliers $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_{2N-2}\}$, we employ the rescaling proposed by Drummond and Suchard (2010). Under this transform, each branch-rate multiplier is the product of clock rate r_i and location parameter γ scaled by the inverse of total expected substitutions per total tree time,

$$\rho_i = \gamma r_i \frac{\sum_k t_k}{\sum_k r_k t_k}. \tag{4.2}$$

This results in one fewer degree of freedom since

$$\frac{\sum_i \rho_i t_i}{\sum_i t_i} = \gamma. \tag{4.3}$$

For heterochronous data, we estimate γ, but for ultrametric studies where the height of the tree is not identifiable we fix $\gamma = 1$.

4.2.3 Autocorrelated shrinkage-clock

We assume clock rates $\mathbf{r} = \{r_1, \ldots, r_{2N-2}\}$ are autocorrelated and model the incremental difference ϕ_i between branch i's clock rate and its parent lineage clock rate,

$$\log r_i - \log r_{\mathrm{pa}(i)} = \phi_i, \quad \text{for } i \in \{1, \ldots, 2N-2\} \text{ and } r_{2N-1} = 1. \tag{4.4}$$

Under this parameterization, the increments $\boldsymbol{\phi} = \{\phi_1, \ldots \phi_{2N-2}\} \in \mathbb{R}^{2N-2}$ are a linear transformation of $\log \mathbf{r}$. To shrink the total number of rate changes along the tree, we let $\phi_i \overset{\text{iid}}{\sim} P_{\boldsymbol{\phi}}$ such that $\mathbb{E}[\phi_i] = 0$. Typically, $P_{\boldsymbol{\phi}}$ may follow a Gaussian (Thorne et al., 1998) or Laplace distribution. We choose the flexible, heavy-tailed, Bayesian bridge prior (Polson et al., 2014) on the increments,

$$P_{\boldsymbol{\phi}} \propto \exp\left\{ -\left| \frac{\phi_i}{\mu} \right|^{\alpha} \right\}, \tag{4.5}$$

where $\mu > 0$ is termed the "global scale" and $\alpha \in (0, 1]$ changes the shape of $P_{\boldsymbol{\phi}}$ where smaller α places more mass near zero. See Figure (4.1) for a comparison of the bridge to common shrinkage priors. Choosing α to be close to 0 forces the Bayesian bridge prior closer to best subset selection when used in a regression setting, while $\alpha = 1$ matches the Laplace prior. In all examples, we set $\alpha = \frac{1}{4}$ to enforce slightly stronger shrinkage than the default 0.5 employed by Polson et al. (2014). Since increments are independent, the joint prior is simply the product

$$p(\boldsymbol{\phi} \,|\, \mu) \propto \prod_{i=1}^{2N-2} \exp\left\{ -\left| \frac{\phi_i}{\mu} \right|^{\alpha} \right\}. \tag{4.6}$$

4.3 Inference

We follow the computationally efficient sampling approach outlined by Polson et al. (2014) and view the prior on the increments as a scale mixture of normals (West, 1987),

$$p(\phi_i \,|\, \mu) = \int p(\phi_i \,|\, \lambda_i, \mu)\, d\lambda_i, \tag{4.7}$$

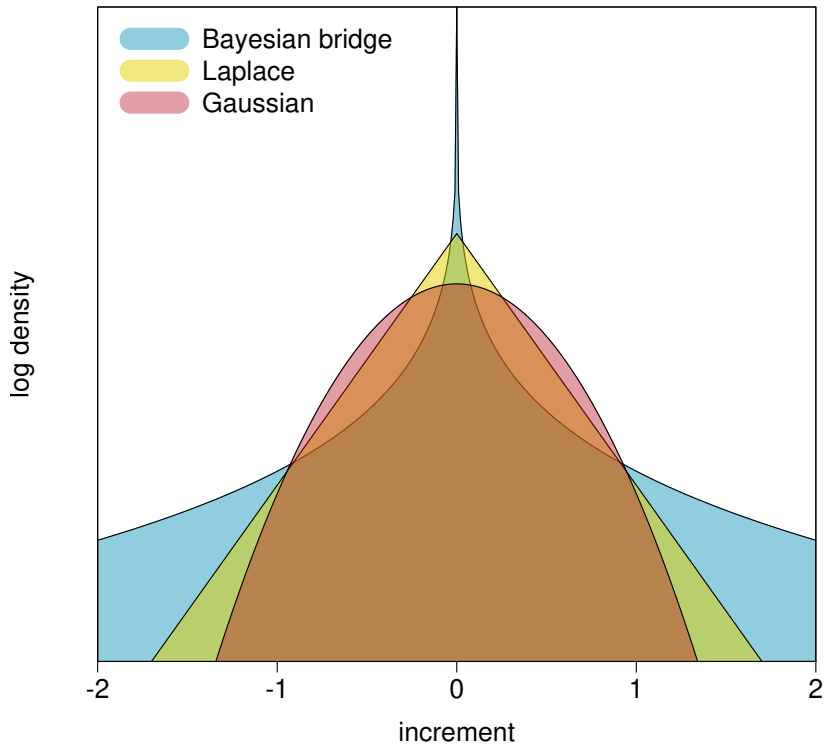


Figure 4.1: The Bayesian bridge prior places more mass near 0 and has heavier tails compared to other common shrinkage priors. The Bayesian bridge reflects our a priori belief that local clocks are rare, but may arbitrarily speed-up or slow-down the rate of molecular evolution.

where the local scale of branch i, $\lambda_i > 0$ and draws from a one-sided stable distribution, see Polson et al. (2014); Nishimura and Suchard (2019) for more details. To improve convergence speed and maintain the benefits of our heavy-tailed prior, we employ the shrunken-shoulder regularization of Nishimura and Suchard (2019) and augment our bridge to have light tails past a reasonably large point. Our scale mixture prior on an increment becomes

$$p(\phi_i \,|\, \lambda_i, \mu) = \mathrm{N}\left(0, \left(\frac{1}{\xi^2} + \frac{1}{\lambda_i^2 \mu^2}\right)^{-1}\right), \tag{4.8}$$

where slab width ξ bounds the variance of increments to ξ^2. In the examples to follow, we set $\xi = 2$, effecting a weakly informative, generous upper bound on clock rate changes. Specifically, a slab width of 2 asserts that there is at most 5% probability for r_i to be greater than $50 \times r_{\mathrm{pa}(i)}$.

We are interested in learning about the posterior,

$$p(\mathbf{r}, \mu, \gamma, \boldsymbol{\theta}, \mathscr{F} \,|\, \mathbf{S}) \propto \int \underbrace{p(\mathbf{S} \,|\, \mathbf{r}, \gamma, \boldsymbol{\theta}, \mathscr{F})}_{\text{likelihood}} \underbrace{p(\mathbf{r} \,|\, \boldsymbol{\lambda}, \mu) p(\boldsymbol{\lambda}) p(\mu) p(\gamma) p(\boldsymbol{\theta}, \mathscr{F})}_{\text{priors}} \, d\boldsymbol{\lambda}, \tag{4.9}$$

where $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_{2N-2}\}$, $\boldsymbol{\theta}$ represents all relevant parameters that describe the molecular substitution model and, again, $\mathscr{F}$ is the phylogenetic tree. We place a relatively uninformative, Gamma prior on $\mu^{-\alpha}$ with shape 1 and scale 2. Additionally, we place the CTMC conditional reference prior of Ferreira and Suchard (2008) on the location γ. We detail the priors on $\boldsymbol{\theta}$ and $\mathscr{F}$ in each example of the sequel.

We use Markov chain Monte Carlo (MCMC) to marginalize over local scale parameters and approximate the posterior (4.9). Specifically, we employ a random-scan Metropolis-within-Gibbs (Liu, 2008; Levine and Casella, 2006) sampling approach to update the full conditional densities implicit in (4.9). Efficient sampling schemes for $\gamma, \boldsymbol{\theta}, \mathscr{F}$ are well described by Suchard et al. (2018), while Polson et al. (2014) outline the efficient scale-mixture approach we use to sample $(\mu, \boldsymbol{\lambda})$. Here, we turn our attention to sampling

$$p(\mathbf{r} \,|\, \mu, \gamma, \boldsymbol{\theta}, \mathscr{F}, \mathbf{S}) \propto \int p(\mathbf{S} \,|\, \mathbf{r}, \gamma, \boldsymbol{\theta}, \mathscr{F}) p(\mathbf{r} \,|\, \boldsymbol{\lambda}, \mu) \, d\boldsymbol{\lambda}. \tag{4.10}$$

Since there are $2N - 2$ correlated branch-rate multipliers, one for each branch of the tree, univariable MCMC sampling schemes for $\mathbf{r}$ scale poorly to large trees. To remedy this difficulty, we employ Hamiltonian Monte Carlo (HMC) to sample all $\mathbf{r}$ simultaneously and with

high acceptance probability. HMC leverages the geometry of the high-dimensional branch-rate multiplier space to propose states that are farther away than traditional proposals but stay within regions of high posterior density. HMC escapes entrapment by local extrema of the posterior by generating random momentum $\boldsymbol{\nu} = \{\rho_1, \ldots, \rho_{2N-2}\}$ in each dimension where typically $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$ (Neal, 2011). Often mass matrix $\mathbf{M} = \mathbf{I}_{2N-2}$ but HMC sampling may be improved by using an alternative $\mathbf{M}$, such as an approximation of the Hessian of the log-posterior (Stan Development Team, 2017; Zhang and Sutton, 2011). For further reading on HMC, see Neal (2011); Betancourt (2017). While HMC samplers offer more efficient posterior exploration, they require computationally expensive gradient calculations that often diminish their usefulness. Here we exploit and extend recent work (Ji et al., 2020) on branch-specific clock rate gradients to facilitate fast inference of $\mathbf{r}$ under our shrinkage model.

4.4 Hamiltonian Monte Carlo increment sampler

We generate proposals in increment space, since $\boldsymbol{\phi}$ are uncorrelated in the prior and we transform back to rate space as described by the linear transformation in Equation (4.4). HMC sampling of the rates requires the gradient of the rate log-posterior,

$$\frac{\partial}{\partial \phi_k} \log p(\mathbf{r} \,|\, \mu, \gamma, \boldsymbol{\theta}, \mathscr{F}, \mathbf{S}) = \int \frac{\partial}{\partial \phi_k} \underbrace{\log p(\mathbf{S} \,|\, \mathbf{r}, \gamma, \boldsymbol{\theta}, \mathscr{F})}_{L\left(\boldsymbol{\rho}(\mathbf{r})\right)} + \frac{\partial}{\partial \phi_k} \log p(\mathbf{r} \,|\, \boldsymbol{\lambda}, \mu)\; d\boldsymbol{\lambda}. \tag{4.11}$$

To compute the gradient of the log-likelihood with respect to the increments, we first find the gradient with respect to clock rates $\mathbf{r}$,

$$\frac{\partial}{\partial r_j} L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) = \gamma \nabla_{\mathbf{r}} \boldsymbol{\rho} L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) \mathbf{T}, \tag{4.12}$$

where we compute all entries in $\nabla_{\mathbf{r}} \boldsymbol{\rho} L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) = (\frac{\partial}{\partial \rho_1}, \ldots, \frac{\partial}{\partial \rho_{2N-2}}) L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right)$ with the computational $\mathcal{O}(N)$ algorithm derived by Ji et al. (2020) and

$$\mathbf{T}_{ij} = \begin{cases} \frac{\sum_k t_k}{\sum_k r_k t_k} - r_i t_i \frac{\sum_k t_k}{(\sum_k r_k t_k)^2} & \text{if } i = j \\ -r_i t_j \frac{\sum_k t_k}{(\sum_k r_k t_k)^2} & \text{if } i \neq j. \end{cases} \tag{4.13}$$

We complete the gradient

$$\begin{aligned}\frac{\partial}{\partial\phi_k}L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) &= \sum_{j=1}^{2N-2}\frac{\partial}{\partial r_j}L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right)\frac{dr_j}{d\phi_k} \quad \text{and} \\ \frac{dr_j}{d\phi_k} &= \begin{cases} r_j & \text{if i ancestral to j} \\ 0 & \text{otherwise,} \end{cases}\end{aligned} \tag{4.14}$$

where transformation $\frac{dr_j}{d\phi_k}$ follows directly from equation (4.4). To preserve the $\mathcal{O}(N)$ gradient computation, we take advantage of the tree structure explicit in Equation (4.14) and accumulate the gradient of the log-likelihood via one post-order traversal of the tree. To begin, let i and j be both daughters of node k in $\mathscr{T}$, then

$$\frac{\partial}{\partial\phi_k}L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) = \begin{cases} r_k \times \frac{\partial}{\partial r_k}\left[L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right)\right] & \text{if } k \text{ is a tip} \\ \left(\frac{\partial}{\partial\phi_i}+\frac{\partial}{\partial\phi_j}+r_k\times\frac{\partial}{\partial r_k}\right)L\left(\boldsymbol{\rho}\left(\mathbf{r}\right)\right) & \text{otherwise.} \end{cases} \tag{4.15}$$

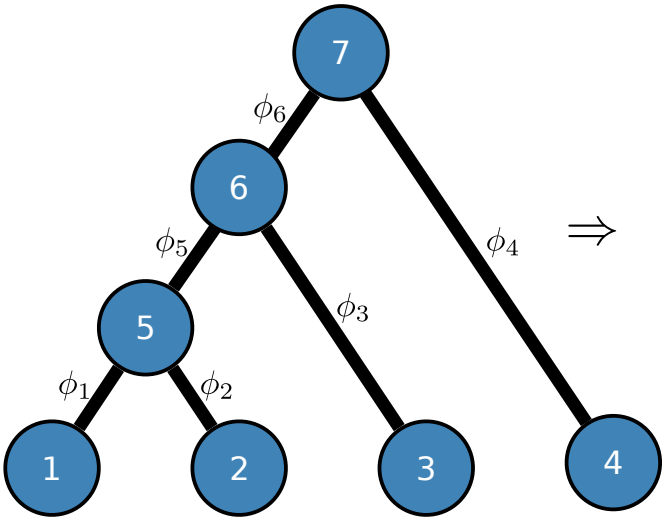
We next turn our attention to the gradient of the log-prior,

$$\frac{\partial}{\partial\phi_k}\log p(\mathbf{r}\,|\,\boldsymbol{\lambda},\mu) = \frac{\partial}{\partial\phi_k}\left[\log p(\boldsymbol{\phi}\,|\,\boldsymbol{\lambda},\mu)+\log\left|\frac{d\boldsymbol{\phi}}{d\mathbf{r}}\right|\right]. \tag{4.16}$$

Since $p(\boldsymbol{\phi}\,|\,\boldsymbol{\lambda},\mu)$ is Gaussian, the first term gradient unwinds,

$$\begin{aligned}\frac{\partial}{\partial\phi_k}\log p(\boldsymbol{\phi}\,|\,\boldsymbol{\lambda},\mu) &= \frac{\partial}{\partial\phi_k}\sum_{j=1}^{2N-2}\log p(\phi_j\,|\,\lambda_j,\mu) \\ &= \frac{\partial}{\partial\phi_k}\sum_{j=1}^{2N-2}\log \mathrm{N}\left(0,\sigma_j^2\right) \\ &= -\phi_k\left(\frac{1}{\xi^2}+\frac{1}{\lambda_i^2\mu^2}\right).\end{aligned} \tag{4.17}$$

Numerical solutions to the second term in (4.11) involve the change-of-variable Jacobian $\frac{\partial}{\partial\phi_k}\log\left|\frac{d\boldsymbol{\phi}}{d\mathbf{r}}\right|$ and appear to necessitate an $\mathcal{O}(N^2)$ sparse determinant computation. To facilitate faster computation of the transform, we index nodes of the tree such that $i<j \implies i$ is not ancestral to j. Under this indexing, $\frac{d\boldsymbol{\phi}}{d\mathbf{r}}$ is an upper triangular matrix with $\frac{1}{r_i}$ along its diagonal, see Figure (4.2) for an example.



$$\Rightarrow \quad \frac{d\boldsymbol{\phi}}{d\mathbf{r}} = \begin{bmatrix} \frac{1}{r_1} & 0 & 0 & 0 & -\frac{1}{r_5} & 0 \\ 0 & \frac{1}{r_2} & 0 & 0 & -\frac{1}{r_5} & 0 \\ 0 & 0 & \frac{1}{r_3} & 0 & 0 & -\frac{1}{r_6} \\ 0 & 0 & 0 & \frac{1}{r_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{r_5} & -\frac{1}{r_6} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{r_6} \end{bmatrix}$$

Figure 4.2: Example tree with corresponding Jacobian matrix. Index $i < j \implies i$ is not ancestral to j thus the Jacobian is upper-triangular and the determinant is the product of diagonal entries.

The gradient of the log determinant,

$$\begin{aligned} \frac{\partial}{\partial \phi_k} \log \left| \frac{d\boldsymbol{\phi}}{d\mathbf{r}} \right| &= \frac{\partial}{\partial \phi_k} \log \prod_{j=1}^{2N-2} \frac{1}{r_j} \\ &= -\frac{\partial}{\partial \phi_k} \sum_{j=1}^{2N-2} \log r_j \\ &= \underbrace{\sum_j \mathbb{1}_{[r_j \text{ depends on } \phi_k]}}_{d_k}, \end{aligned} \tag{4.18}$$

where $\mathbb{1}$ is the indicator function and d_k is the number of descendants of node k. All together,

$$\frac{\partial}{\partial \phi_k} \log p(\mathbf{r} \mid \boldsymbol{\lambda}, \mu) = -\phi_k \left(\frac{1}{\xi^2} + \frac{1}{\lambda_k^2 \mu^2} \right) - d_k, \tag{4.19}$$

and we accumulate d_k in one recursive post-order tree traversal by observing $d_k = d_i + d_j + 1$, where, again, i and j are both daughters of k.

To further improve the proposals of our HMC sampler, we precondition the mass matrix

$\mathbf{M}$ to be the current-state absolute value of the Hessian of the log-prior,

$$\left| \frac{\partial^2}{\partial \phi_i^2} \phi_j \log p(\mathbf{r} \mid \boldsymbol{\lambda}, \mu) \right| = \begin{cases} \left(\frac{1}{\xi^2} + \frac{1}{\lambda_i^2 \mu^2} \right) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{4.20}$$

This diagonal matrix weights momentum draws by prior increment precision. Intuitively, equation (4.20) improves HMC sampling by rescaling increment proposals by the variance of $\boldsymbol{\phi}$, allowing larger steps to be taken in dimensions with larger variance. See Neal (2011) for further discussion on mass matrix transformations.

4.5 Results

4.5.1 Local clocks in three nuclear genes of rodents and other mammals

To verify the accuracy of our model, we turn to a well-studied example of adaptive radiation in mammals and rodents. Huchon et al. (2002) and Douzery et al. (2003) examine the adaptive radiation of 21 rodents compared to 19 other placental mammals and two marsupial outgroups using the first two codon positions for three nuclear genes: ADRA2B, IRBP and vWF (2422 alignment sites). Douzery et al. (2003) establish the presence of clock variability within this set of taxa and report their best fitting model contains five local clocks. Drummond and Suchard (2010) use the RLC model to re-examine this claim and estimate the existence of between 6 and 12 local clocks. Here we employ our shrinkage-clock model to jointly infer the mammalian phylogeny as well as the number and location of local clocks. Because this is an ultrametric example, $\gamma = 1$. We follow the specifications of Drummond and Suchard (2010) and Douzery et al. (2003) and use a general time reversible (GTR) substitution model with a 4 category discrete-Γ site rate model. We run ten separate Markov chains of our shrinkage-clock with ten different starting trees for 30M states and build a maximum clade credibility (MCC) tree from the combined results (Figure 4.3). We further run 100 RLC chains with 100 different starting trees for 30M states and build a MCC tree for comparison. Under the combinatorial parameter space of the RLC, we observe suboptimal mixing and that some chains convergence to different modes, hence our choice

for combining 100 independent chains; the 10 independent chains for the shrinkage-clock simply errs on the side of caution since each independent shrinkage-clock chain converges to the same topology. Incidentally, the shrinkage-clock MCC topology differs from the RLC MCC in two places. First, *Bradypus* attaches to one of two neighbor internal branches deep in the tree. Second, *Anomalurus* is more closely related to the *Dipus* than *Castor* under the shrinkage-clock. This second difference highlights the well-known difficulty in *Anomalurus* placement (Horner et al., 2007). Indeed, Blanga-Kanfi et al. (2009) find *Anomalurus* sits between *Dipus* and *Castor* in an analysis of six nuclear genes. The posterior probabilities of *Bradypus* and *Anomalurus* parent branches under the shrinkage-clock are almost equal (0.64 and 0.49 respectively).

We estimate the existence of four local clocks where we define a local clock on branch i if the posterior odds $\phi_i > 0$ is greater than 10 or less than $\frac{1}{10}$. The posterior odds here is equivalent to a Bayes factor since an increment is equally likely to be positive or negative under the prior. Furthermore, a Bayes factor greater than 10 is suggestive of "strong evidence" against an alternative hypothesis (Kass and Raftery, 1995). In the approach, we do not make assumptions about the magnitude of local clocks on a tree and instead use posterior probability of increment sign to define a clock.

To illustrate the benefits of using a heavy-tailed prior on the increments, we further compare the performance of our Bayesian bridge prior to the more usual Laplace prior for shrinkage (see Figure 4.1). We again fit our shrinkage-clock as described above but remove the slab and fix $\alpha = 1$, thus placing a Laplace prior on each increment. We find the posterior mean of increment variance under both the Laplace and Bridge priors is 0.057 with 95% high posterior density (HPD) intervals $(0.036, 0.089)$ and $(0.035, 0.093)$ respectively. Despite having very similar variance, we find the posterior mean of the absolute maximum increment is 0.84 $(0.57, 1.22)$ and 1.01 $(0.70, 1.54)$ under the Laplace and Bridge priors respectively. This evidences induced smoothing of the clock rates under the exponential tails of a Laplace prior, that on average shrinks the largest increment by approximately 20%.

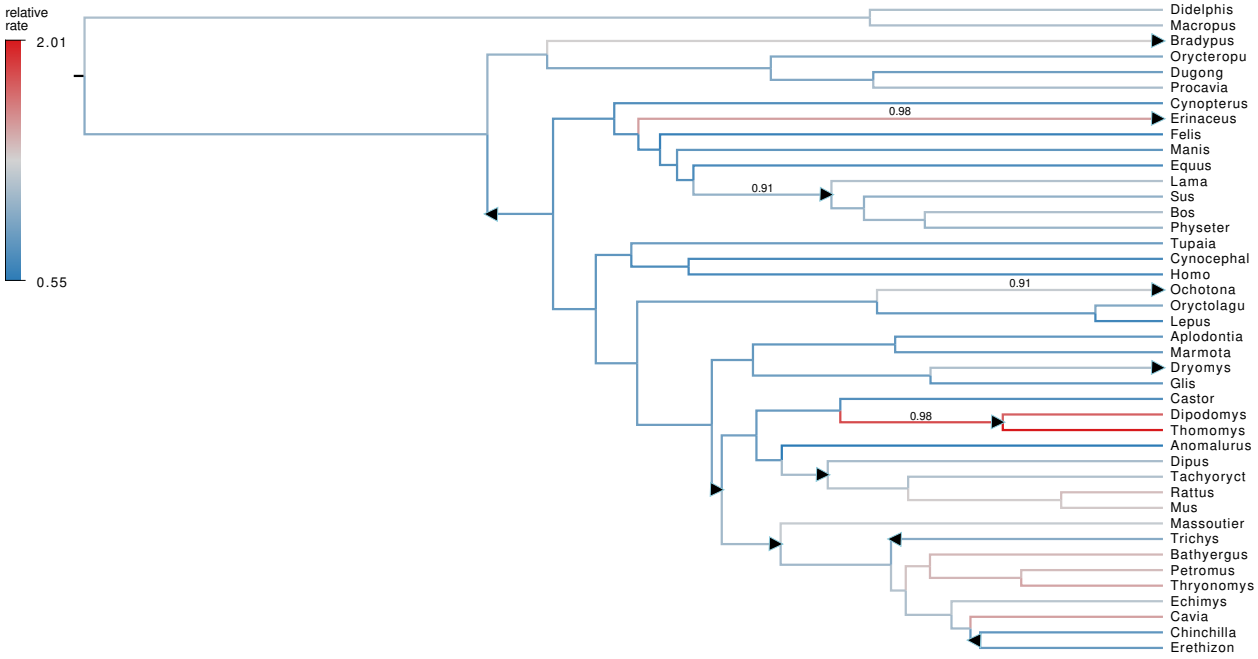


Figure 4.3: Maximum clade credibility (MCC) tree under shrinkage-clock of mammalian and rodent radiation where branches are colored by posterior mean relative clock rates $\mathbf{r}$. If branch i starts a new clock, it is labeled with the posterior probability $\phi_i > 0$. For comparison, local clocks of the random local clock (RLC) model are depicted as black triangles. Triangle direction indicates RLC relative-rate speed-up (right) or slow-down (left). Two local clocks of the RLC are excluded due to topological differences between the RLC and shrinkage-clock MCC trees.

4.5.2 Simulation study

We compare the scalability of our shrinkage-clock to the RLC under a simulated example. We generate 1000 character nucleotide sequences from a fixed 40 tip tree 20 times. In each simulation, there are 4 distinct lineages (A, B, C, D) of 10 taxa each. Time to most recent common ancestor (TMRCA) for each lineage is 40 years and tree height is 80 years. Lineages B, C and D evolve with a relative clock rate of 1.0 while the MRCA of lineage A starts a new clock with relative rate 2.0.

We compare the accumulation of effective sample size (ESS) per unit time of branch-specific clock rates under both our Bayesian bridge shrinkage-clock and the RLC while simultaneously inferring the phylogeny. ESS approximates the number of independent samples from a Markov chain and we use this metric to evaluate how well each inference procedure explores clock rate space. We report the results across all 20 simulated datasets in Figure (4.4). The median ESS/second is 0.49 and 0.13 under the shrinkage-clock and the RLC respectively, exhibiting a 3.8-fold speed increase. Additionally, the "least well" explored clock rate across all simulations accumulated 3.5×10^{-3} and 4.9×10^{-4} ESS/second under the shrinkage-clock and RLC respectively, a 7.1-fold speed-up, for these relatively small taxon-count examples.

We further report here that inference under the shrinkage-clock without preconditioning the mass matrix results in a minimum and median ESS/second of 2.3×10^{-4} and 1.5×10^{-3} respectively. Overall, preconditioning the mass matrix improves ESS/second $15\times$ for the worst-explored clock rate under the shrinkage-clock.

Averaged across all twenty simulations, the Bayes factor for the true clock rate change under our shrinkage clock is 5.25 while the second most likely clock rate has Bayes factor support of 0.56. Comparably, averaged over all runs, the Bayes factor of the one true clock under the RLC is 3.51. Additionally, the true relative clock-rate for the 'A' clade is 2.0 and we estimate 1.51 $(0.90, 2.31)$ and 1.49 $(0.98, 2.36)$ under the shrinkage-clock and RLC respectively.

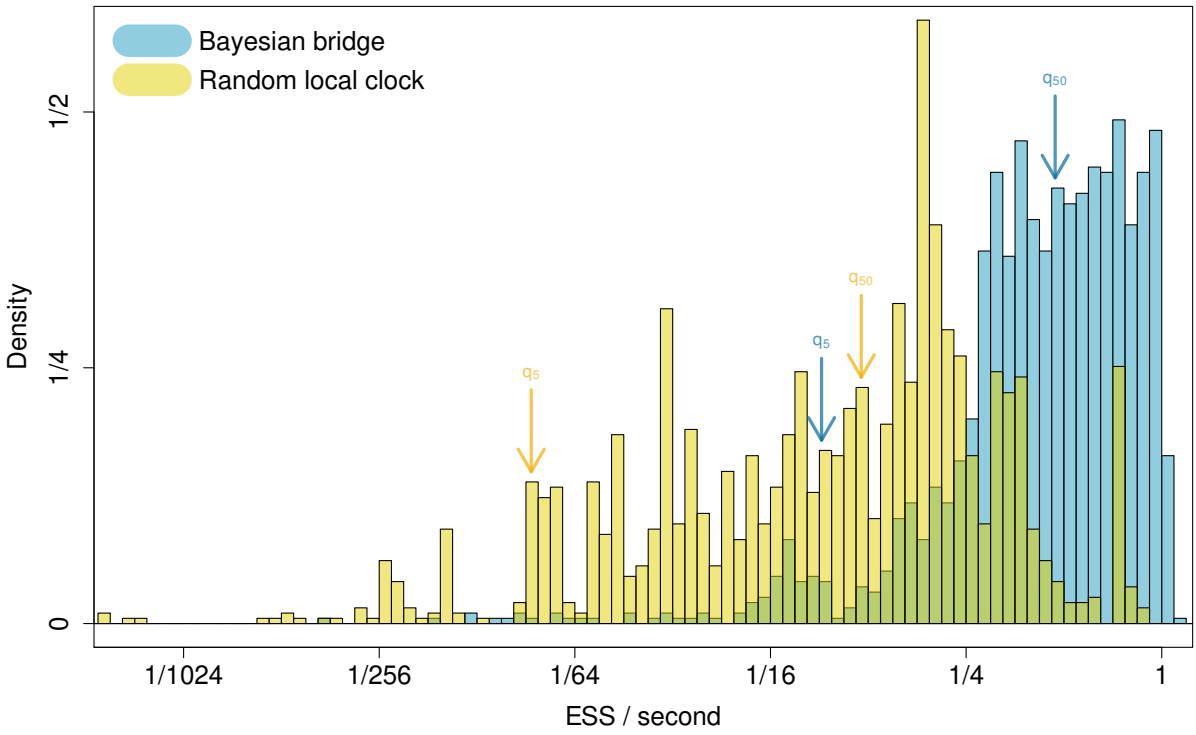


Figure 4.4: Effective sample size of branch-specific clock rates per second of BEAST runtime under the shrinkage-clock and RLC during a full joint phylogenetic analysis. Colored arrows point to 5% (q_5) and 50% (q_{50}) quantiles under each model.

4.5.3 Influenza A virus

We further demonstrate the scalability and utility of our shrinkage-clock model by examining the evolution of two major influenza A virus (IVA) surface glycoprotein subtypes: hemagglutinin (HA) H7 and neuraminidase (NA) N7. Both hemagglutinin and neuraminidase protein mutations impact IVA's epitope and allow IVA to escape adaptive immune responses (McAuley et al., 2019; Wilson and Cox, 1990). Worobey et al. (2014) find divergence time estimation is sensitive to molecular clock model specification. To consistently estimate divergence times, Worobey et al. (2014) allow the clock rates of various glycoprotein subtypes to vary only between viral hosts and find H7 and N7 each evolve slower in equine hosts

than avian hosts. We re-examine this claim with our more general shrinkage-clock model that does not assume the existence of host-dependent clock rates. Specifically, we re-analyze 146 complete gene (1716 nt) sequences of H7 and 92 complete gene (1416 nt) sequences of N7. In each case we follow the model specifications of Worobey et al. (2014) and employ a GTR substitution model with 4 category discrete-Γ site rate model. We depart from their example, however, in our use of tree prior. We employ a Bayesian skygrid prior (Gill et al., 2013) with 50 population size bins and a cutoff of 200 years instead of using the skyride prior (Minin et al., 2008). While the number of taxa here is only approximately double or triple the number in the previous simulation study, the set of all possible local clocks under the RLC grows exponentially with the number of taxa, (31 to 63 orders of magnitude), challenging clock rate inference under the RLC.

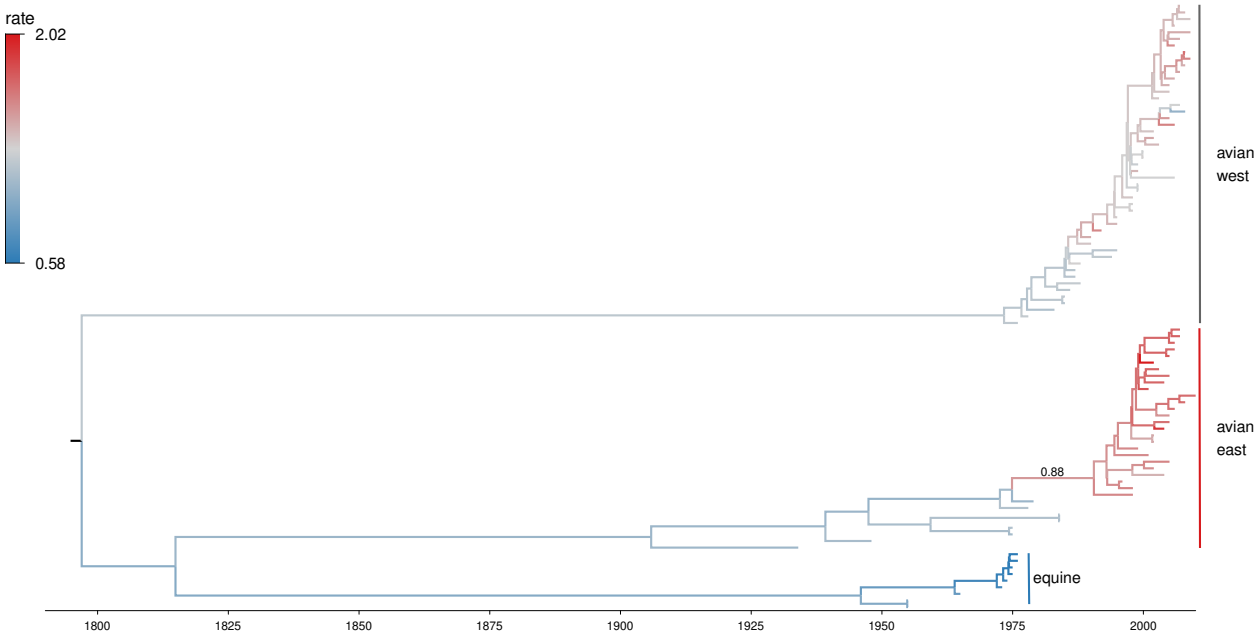


Figure 4.5: Maximum clade credibility tree for influenza A's neuraminidase subtype N7. Branches are colored by posterior clock rates. The most probable local clock is reported and labeled with posterior probability that $\phi_i > 0$. The second most probable clock starts a sub-clade of the equine lineage and has a Bayes factor $\phi_i > 0$ of 0.351.

We find no sharp local clocks exist with Bayes factor > 10 or $< \frac{1}{10}$ on the NA N7 tree

under our shrinkage-clock model but do see evidence for rate heterogeneity. Incidentally, the most likely clock occurs on the branch that begins the Eastern avian clade of the NA N7 tree (Figure 4.5). The second most probable clock is found in a subclade of the equine lineage. On the other hand, we estimate the existence of seven local clocks on the HA H7 tree and report these in Figure (4.6). Overall, the mean posterior clock rate for NA N7 is lower than HA H7. We report the posterior mean and 95% HPD intervals of γ are 2.7 $\{2.1 - 3.3\} \times 10^{-3}$ and 3.3 $\{2.8 - 3.9\} \times 10^{-3}$ for NA N7 and HA H7 respectively. Furthermore, under our shrinkage-clock the posterior mean root heights and 95% HPD intervals of the N7 and H7 trees in absolute time are 1798 (1733-1855) and 1853 (1808 - 1897) respectively.

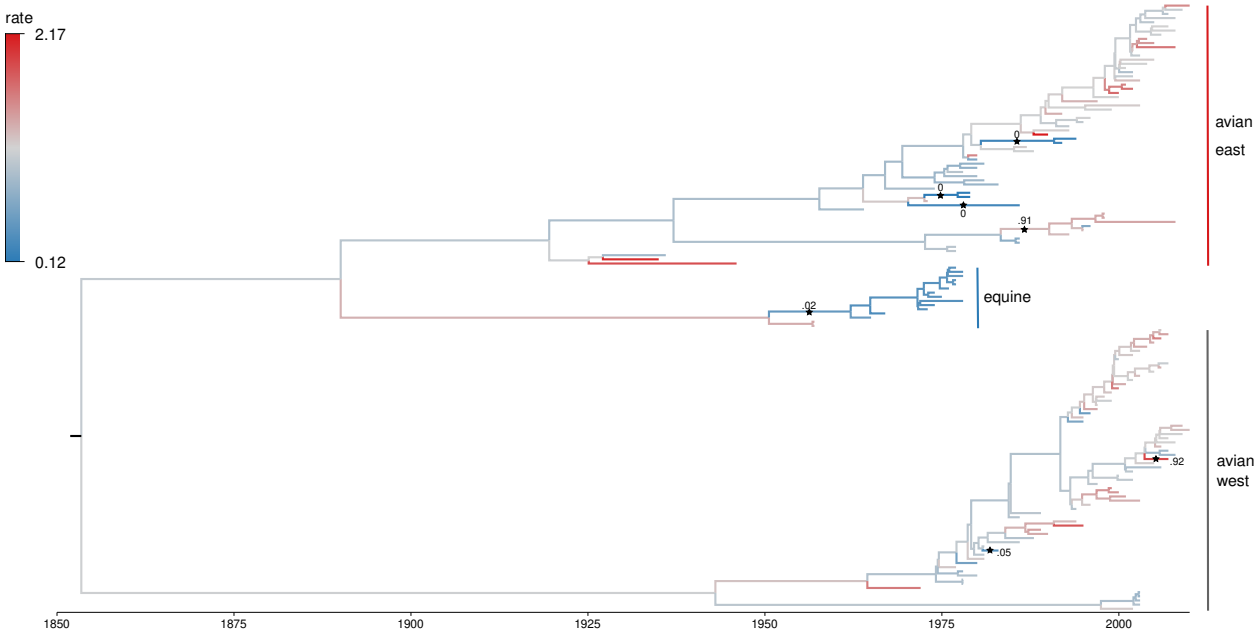


Figure 4.6: Maximum clade credibility tree for influenza A's hemagglutinin subtype H7. Branches are colored by posterior clock rates. Local clocks are labeled with a star and the posterior probability $\phi_i > 0$.

4.6 Discussion

Previous heritable clock models either scale poorly to large trees or excessively shrink clock rates. We develop a robust autocorrelated heritable clock model to overcome these challenges

without specifying *a priori* the number and location of local clocks on a tree. Crucially, we model the incremental difference between log clock rates on the tree as drawing from a Bayesian bridge prior that shrinks most changes to approximately 0 unless the data warrant otherwise. To facilitate scalability, we employ HMC to generate proposals in the independent increment space and derive recursive post-order algorithms to compute the gradient and its requisite transforms. Our recursive algorithms achieve $\mathcal{O}(N)$ computational speed, signifying that they will continue to work well as N grows large. We further improve the speed of our HMC sampler by preconditioning the mass matrix with the Hessian of the log-prior.

In our examination of the adaptive radiation of rodents and other mammals, (see section 4.5.1), our shrinkage-clock recovers the location of four local-clocks estimated under the RLC. This finding is similar to the initial estimate of five clocks reported by Douzery et al. (2003). We choose a Bayes factor of 10 to classify clocks, but shrinkage-clock users may wish to adjust this threshold to increase or decrease clock rate sensitivity. Comparing the statistical properties of different clock-classification schemes remains an important avenue for future work.

We apply our shrinkage-clock model to re-examine the number of local clocks present in IVA surface glycoproteins NA N7 and HA H7 across equine and avian hosts. We confirm the equine slowdown reported by Worobey et al. (2014) but interestingly find that the N7 tree shows marked rate variation (Figure 4.5) between western and eastern hemisphere avian influenza lineages, however, this rate variation is not supported by the Bayes factor cutoff. Root height estimates vary from Worobey et al. (2014) but this may be in part due to the different tree priors. Additionally, we find 7 local clocks under our shrinkage-clock across the H7 tree. Since three of these clocks belong to edges of tip nodes, this may reflect incomplete sampling or sequencing error. Despite inferring six more clocks than the host-specific model, the posterior mean estimate of root height under our shrinkage-clock is within five years of previous estimates (Worobey et al., 2014).

Our shrinkage-clock accumulates ESS/second of the worst explored clock rate 7.1× faster than the RLC across 78 branches of 20 simulated data sets. If ESS is used as stopping criteria for phylogenetic reconstruction, this could save users up to 85% of BEAST runtime. As the

bridge exponent α approaches 0, the bridge prior density is more peaked near zero resulting in sharper increment shrinkage and thus better distinguishable local clocks. Nishimura and Suchard (2019) examine multiple α and report closer to optimal coverage for smaller α but with increasing computational cost due to mixing. For this reason, shrinkage-clock users may find it useful to adjust α depending on desired clock-rate coverage or to tackle even larger tree studies. We make all BEAST XML files used in this work publicly available at https://github.com/suchard-group/shrinking_clocks.

CHAPTER 5

Shrinking shifts on trees

5.1 Introduction

Phylogenetic comparative methods (PCMs) undergird a rich host of trait data studies that describe how "traits", i.e. quantitative measurements belonging to biological entities, change through time. Fundamentally, PCMs are concerned with modeling the evolution of traits on trees. In seminal work on the PCM, Felsenstein (1985) introduces a Brownian motion model to describe phenotypic trait evolution. Consistent with phylogenetic literature, we refer to Brownian motion on a tree as a "random walk" (RW) model (Gill et al., 2017; Pagel, 2002). While popular for its computational simplicity, the RW imposes strong assumptions about the evolution of traits on a tree. The RW assumes traits evolve with variance that strictly grows with time and is unbounded as time tends towards infinity (Butler and King, 2004). Furthermore, while RWs may combine drift with selection (Pagel, 2002), they lack the ability to differentiate evolution towards a primary optimum (Gill et al., 2017) and instead describe the evolution of a secondary optimum (Bastide, 2017).

To generalize the trait data generative process and introduce a stabilizing selective force that bounds the variance, Hansen (1997) introduces the Ornstein-Uhlenbeck (OU) model of trait evolution in a phylogenetic setting that captures adaptive trait evolution towards an optimum. The OU maintains the drift of the RW but adds a deterministic term, often referred to as "call-back" that pushes traits towards a central primary optimum (Hansen, 1997). In its most general, multivariate form, the OU induces variance structure on the tips that requires computing a matrix exponential as well as a Hadamard product, making inference under the full multivariate OU computationally expensive (Clavel et al., 2015;

Bartoszek et al., 2012). For this reason, the foundational work of Hansen (1997) as well as several extensions (Khabbazian et al., 2016; Uyeda and Harmon, 2014) present univariate OU models, avoiding expensive matrix computations altogether.

When an organism's environment dynamically shifts, the fitness landscape can shift as well. Optimal trait values are not fixed in time. To model this phenomena, (Hansen et al., 2008) let the optimum itself evolve according to its own stochastic process and subsequent works try to detect shifts in the optima of multiple traits (Bastide, 2017; Bartoszek et al., 2012). Here, a "shift" refers to a change in optimal traits and is typically posited to be the result of an underlying environment shift. Khabbazian et al. (2016) develop a likelihood-based approach for detecting shifts in univariate (or equivalently multiple independent) traits by setting up a linear model where each branch of a fixed tree has a shift covariate. Khabbazian et al. (2016) subsequently regularize shift covariates via a lasso penalty. Uniquely, their approach allows one to detect convergent evolution. Their method scales well to large trees, but is unable to learn about trait covariance when studying more than one trait and assumes a fixed phylogenetic topology under the likelihood framework. Separately, Uyeda and Harmon (2014) develop a reversible jump Markov chain Monte Carlo (rjMCMC) method to learn about posterior magnitude, number and location of shifts along a tree in a Bayesian setting but are limited to examining univariate traits and are unable to detect convergent evolution. Bastide et al. (2017) develop a maximum-likelihood framework to learn about multivariate shifts and their correlation. To find the number and location of shifts, Bastide et al. (2017) enumerate and search through the space of possible shifts on the tree. Due to the combinatorial complexity of this search, their method is challenged by large trees with many shifts.

Here we propose a new scalable Bayesian framework to detect shifts. We assert each branch of the tree has its own branch-specific optima and encode shifts as significantly non-zero increments between parent and daughter branch optima. We place a Bayesian bridge shrinkage prior on these increments, reflecting our a priori belief that few shifts exist on the tree. Like other shrinkage priors, the Bayesian bridge places large mass near zero but has the unique benefit of heavier than exponential tails that allow shifts to be arbitrarily large. We

follow Fisher et al. (2021) and develop an efficient preconditioned Hamiltonian Monte Carlo sampler to scale inference under our model to large trees. We combine recently developed efficient closed-form gradient expressions for the prior (Fisher et al., 2021) and the OU likelihood (Bastide et al., 2020). We implement our method and make it publicly available within the popular open source Bayesian Evolutionary Analysis Sampling Trees (BEAST) software package (Suchard et al., 2018) where it can be easily matched with state-of-the-art methods to simultaneously learn about trait-specific selection strengths, trait covariance, and the phylogenetic topology. We demonstrate the accuracy of our method by re-examining convergent evolution in four principle component traits that combine 11 phenotypic measurements of Caribbean Anolis lizards (Mahler et al., 2013; Khabbazian et al., 2016).

5.2 Shrinking shifts

Consider a rooted bifurcating phylogeny, $\mathscr{F}$ with N tips and $N-1$ internal nodes. We designate node $2N-1$ to be the root of the tree. Let $\mathbf{Y}_i$ be a P-dimensional vector of latent or observed continuous traits belonging to tip node $i \in \{1,\ldots,N\}$. For each node $j \in \{1,\ldots,2N-1\}$ let $\mathbf{X}_j$ be a $D \leq P$-dimensional latent vector of continuous measurements. See Cybis et al. (2015) and Zhang et al. (2021) for methods to map discrete traits to this continuous framework. Further, let $D \times D$ positive definite matrix $\mathbf{\Sigma}$ describe covariance between traits. Following Khabbazian et al. (2016), we assume there exists an optimal trait vector $\boldsymbol{\upsilon}_j \in \mathbb{R}^D$ for each branch $j \in \{1,\ldots,2N-2\}$ in $\mathscr{F}$. Under the phylogenetic OU of Bastide et al. (2020), latent trait vector $\mathbf{X}_j$ approaches optima vector $\boldsymbol{\upsilon}_j$ independently at a rate prescribed by the diagonal attenuation (selection strength) matrix $\mathbf{A}$. Additionally, branch length t_i connects each node i with its parent node, $\text{pa}(i)$ and describes the duration of attenuation. All together,

$$\begin{aligned} \mathbf{X}_j|\mathbf{X}_{\text{pa}(j)} &\sim \text{MVN}\left(\mathbf{q}_j\mathbf{X}_{\text{pa}(j)} + (\mathbf{I}_D - \mathbf{q}_j)\,\boldsymbol{\upsilon}_j, \mathbf{V}_j\right), \\ \mathbf{Y}_i|\mathbf{X}_i &\sim \text{MVN}\left(\mathbf{L}\mathbf{X}_i, \mathbf{U}_i\right), \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} \mathbf{q}_j &= \exp\left\{-\mathbf{A}t_j\right\}, \\ \mathbf{V}_j &= \boldsymbol{\Sigma} - \mathbf{q}_j \boldsymbol{\Sigma} \mathbf{q}_j^t, \end{aligned} \tag{5.2}$$

and $\mathbf{U}_i$ characterizes the variance associated with tip observations $\mathbf{Y}_i$ such as measurement error. When $D < P$, $\mathbf{U}_i$ may describe the stochastic link function of the phylogenetic factor model where $\mathbf{L}$ maps latent $\mathbf{X}$ to observed traits $\mathbf{Y}$ (Hassler et al., 2020). If $D = P$, then $\mathbf{L} = \mathbf{I}_P$ and $\mathbf{Y}_i|\mathbf{X}_i$ is possibly degenerate. We complete model specification with prior on the parentless root

$$p(\mathbf{X}_{2N-1}) = \text{MVN}(\boldsymbol{\eta}, \Psi), \tag{5.3}$$

where root mean $\boldsymbol{\eta}$ and root variance Ψ are typically unknown.

We re-parameterize branch-specific optimal trait vectors $\boldsymbol{\upsilon} = \{\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_{2N-2}\}$ as increments between child and parent branches,

$$\boldsymbol{\phi}_i = \boldsymbol{\upsilon}_i - \boldsymbol{\upsilon}_{\text{pa}(i)} \text{ for } i \in \{1, \ldots, 2N-2\} \tag{5.4}$$

with root optimum $\boldsymbol{\upsilon}_{2N-1} = \boldsymbol{\gamma}$ and set $\boldsymbol{\gamma} = \mathbf{0}$ when each trait is independently standardized. Under this parameterization, increment ϕ_{ij} is more aptly termed the "shift" in the jth optimal trait on branch i. Let $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{2N-2}\}$. To shrink the number of shifts, we place a Bayesian bridge prior (Polson et al., 2014) on individual shifts,

$$p(\phi_{ij} \,|\, \mu_j) \propto \mu_j^{-1} \exp\left\{ - \left| \frac{\phi_{ij}}{\mu_j} \right|^{\alpha} \right\}, \tag{5.5}$$

where $\mu_j > 0$ is the "global scale" shared amongst shifts in trait j and exponent $\alpha \in (0, 1]$ regulates the prior's mass near zero. Intuitively, smaller α corresponds to higher prior mass near zero whilst $\alpha = 1$ yields the Laplace prior. The full joint prior

$$p(\boldsymbol{\phi} \,|\, \mu_1, \ldots, \mu_D) \propto \prod_{j=1}^{D} \left[{\mu_j}^{-(2N-2)} \exp\left\{ - \sum_{i}^{2N-2} \left| \frac{\phi_{ij}}{\mu_j} \right|^{\alpha} \right\} \right], \tag{5.6}$$

where the product and sum arise from independence in shifts across all traits and branches.

5.3 Inference

Often, we are interested in learning about both trait variation and the phylogenetic topology. Let $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_D\}$, then the full joint posterior

$$p(\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}, \boldsymbol{\theta} \,|\, \mathbf{Y}, \mathbf{S}) \propto \underbrace{p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}) p(\mathbf{S} \,|\, \mathscr{T}, \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\upsilon} \,|\, \boldsymbol{\mu}) p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\mathbf{A}) p(\mathscr{T}, \boldsymbol{\theta})}_{\text{priors}}, \tag{5.7}$$

where $\boldsymbol{\theta}$ represents all parameters of a substitution model for aligned molecular sequence data $\mathbf{S}$. Notice the likelihood of sequence data $\mathbf{S}$ is assumed conditionally independent from trait data $\mathbf{Y}$ given the phylogeny. While this assumption is false for genetically derived traits, it is a practical simplification without prior knowledge of the genetic location of causal genes. In the example to follow, we assume independent shifts in optima as described in Equation (5.6). To facilitate efficient sampling under our shift prior, we work with the Bayesian bridge's data augmented scale mixture of normals representation, popularized by Polson et al. (2014),

$$p(\phi_{ij} \,|\, \lambda_{ij}, \mu_j) = \mathrm{N}\left(0, \left(\frac{1}{\xi^2} + \frac{1}{\lambda_{ij}^2 {\mu_j}^2}\right)^{-1}\right), \tag{5.8}$$

where λ_{ij} is the "local scale" of shifts in optimal trait j on branch i and slab ξ bounds the variance of shifts to ξ^2 to improve sampling speed (Nishimura and Suchard, 2019). The data-augmented joint prior on shifts is simply the double product,

$$p(\boldsymbol{\phi} \,|\, \boldsymbol{\mu}) = \prod_{j=1}^{D} \prod_{i=1}^{2N-2} p(\phi_{ij} \,|\, \lambda_{ij}, \mu_j), \tag{5.9}$$

where we sample λ_{ij} and subsequently marginalize over these nuisance parameters as described by Polson et al. (2014). Crucially, this data augmented bridge prior is differentiable everywhere and thus aids fast gradient-based sampling techniques (see section (5.4)). To complete our bridge prior specification, we follow Fisher et al. (2021) and fix $\xi = 2$ and $\alpha = \frac{1}{4}$. Furthermore, we place a relatively uninformative conjugate Gamma prior on each ${\mu_j}^{-\alpha}$ with shape 1 and scale 2 to Gibbs sample each $\mu_j \in \boldsymbol{\mu}$ (Polson et al., 2014). For the diffusion variance, we place a Wishart prior with scale matrix $\mathbf{I}_D$ and D degrees of freedom on $\boldsymbol{\Sigma}^{-1}$ for computational simplicity but other priors such as the eponymous LKJ Lewandowski

et al. (2009) prior may, in some instances, be favorable (Barnard et al., 2000) and recently Zhang et al. (2021) develop efficient sampling machinery under this alternative prior. For the positive, diagonal selection matrix $\mathbf{A}$, we follow Bastide et al. (2020) and place a relatively uninformative half-normal prior with standard deviation 7 corresponding to the a priori belief that phylogenetic half-life (i.e. time for a trait to travel half way towards its optimum) is larger than 5% of the tree height on a standardized unit height tree, 95% of the time (Hansen, 1997).

We approximate (5.7) via MCMC integration and use a random-scan Metropolis-within-Gibbs sampling approach (Levine and Casella, 2006). Four components comprise one turn of our sampling scheme,

$$p(\boldsymbol{\upsilon} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \mid \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}) p(\boldsymbol{\upsilon} \mid \boldsymbol{\mu}) \tag{5.10a}$$

$$p(\boldsymbol{\mu} \mid \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\boldsymbol{\upsilon} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu}) \tag{5.10b}$$

$$p(\mathbf{A}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \mid \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}) p(\boldsymbol{\Sigma}) p(\mathbf{A}), \text{ and} \tag{5.10c}$$

$$p(\mathscr{F}, \boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \mid \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{F}) p(\mathbf{S} \mid \mathscr{F}, \boldsymbol{\theta}) p(\mathscr{F}, \boldsymbol{\theta}), \tag{5.10d}$$

where update (5.10d) is unnecessary when $\mathscr{F}$ is fixed. When $\mathscr{F}$ is not fixed, Suchard et al. (2018) offer efficient methods to update (5.10d). Bastide et al. (2020) describe efficient sampling of OU parameters $\mathbf{A}$ and $\boldsymbol{\Sigma}$ (update (5.10c)). As we note above, $p({\mu_j}^{-\alpha}) \sim$ Gamma yields efficient Gibbs sampling to update the global scale (equation 5.10b). We focus on efficient sampling of the high dimensional, highly correlated, branch-specific optima (5.10a) using Hamiltonian Monte Carlo (HMC).

HMC is a technique to generate geometrically informed proposal steps that fit within a traditional Metropolis update scheme (Neal, 2011; Betancourt, 2017). By leveraging the geometry of the target distribution, HMC allows us to propose states that update all parameters, are far away from our current chain state and have high acceptance probability. To perform this feat, HMC requires the gradient of the log-posterior. Since increments are uncorrelated in the prior, we differentiate with respect to increments to create proposals in increment space and then map back to trait optima by noticing the 1-1 inverse mapping of

(5.4),

$$\boldsymbol{\upsilon}_i = \boldsymbol{\phi}_i + \sum_{j \in \mathcal{A}(i)} \boldsymbol{\phi}_j + \boldsymbol{\gamma}, \tag{5.11}$$

where $j \in \mathcal{A}(i)$ if j is ancestral to i. All together, the gradient of the log-posterior with respect to the increments,

$$\frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\boldsymbol{\upsilon} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}, \boldsymbol{\theta}, \mathbf{Y}, \mathbf{S}) = \frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}) + \frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\boldsymbol{\upsilon} \,|\, \boldsymbol{\mu}). \tag{5.12}$$

Computing requisite gradients is often the premier computational bottleneck for HMC methods. Numerical gradient calculations that utilize efficient $\mathcal{O}(N)$ likelihood algorithms still scale quadratically ($\mathcal{O}(N^2)$) with the number of tips in the tree since we need to evaluate the likelihood at each of the $2N - 2$ branches. To improve speed of inference, several recent works exploit tree topology structure to develop efficient closed-form gradient solutions under other phylogenetic comparative models (Zhang et al., 2019; Bastide et al., 2020; Fisher et al., 2021; Hassler et al., 2021).

5.4 Gradients

We turn our attention to forging an efficient solution to (5.12). To begin, we unpack the first component of (5.12), the gradient of the log-likelihood,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}) &= \sum_j \frac{d\boldsymbol{\upsilon}_j{}^t}{d\boldsymbol{\phi}_i} \frac{\partial}{\partial \boldsymbol{\upsilon}_j} \log p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}) \\ &= \sum_{j \in \mathcal{A}(j)} \frac{\partial}{\partial \boldsymbol{\upsilon}_j} \log p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T}), \end{aligned} \tag{5.13}$$

since branch-specific Jacobian $\frac{d\boldsymbol{\upsilon}_j{}^t}{d\boldsymbol{\phi}_i}$ unwinds by differentiating equation (5.11),

$$\frac{d\boldsymbol{\upsilon}_j{}^t}{d\boldsymbol{\phi}_i} = \begin{cases} \mathbf{I}_D \text{ if } i \in \mathcal{A}(j) \\ \mathbf{0}_D \text{ otherwise} \end{cases} \tag{5.14}$$

where $\mathbf{0}_D$ denotes a D-dimensional matrix of zeroes. We note the ancestral dependence of terms in (5.13) and accumulate relevant gradient terms in one post-order traversal of the tree. Bastide et al. (2020) provide a general recipe to compute $\frac{\partial}{\partial \boldsymbol{\upsilon}_j} \log p(\mathbf{Y} \,|\, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}, \mathbf{A}, \mathscr{T})$ in $\mathcal{O}(N)$.

Next, we focus on the gradient of the log-prior and adapt the $\mathcal{O}(N)$ solution provided by Fisher et al. (2021) to suit our purposes. We work with the scale mixture of normals representation of the Bayesian bridge prior and let $\boldsymbol{\lambda} = \{\lambda_{ij}\}$ for $i \in \{1, \ldots, 2N-2\}$ and $j \in \{1, \ldots, D\}$. We compute the gradient of the induced log prior on $\boldsymbol{\upsilon}$,

$$\frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\boldsymbol{\upsilon} \,|\, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \frac{\partial}{\partial \boldsymbol{\phi}_i} \left[\log p(\boldsymbol{\phi} \,|\, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \log \left| \frac{d\boldsymbol{\phi}}{d\boldsymbol{\upsilon}} \right| \right]. \tag{5.15}$$

Notice that equation (5.4) defines element-wise shifts as a linear function of optima, $\phi_{ij} = \upsilon_{ij} - \upsilon_{\text{pa}(i)j}$ so that the gradient of the Jacobian, $\frac{\partial}{\partial \boldsymbol{\phi}_i} \log \left| \frac{d\boldsymbol{\phi}}{d\boldsymbol{\upsilon}} \right| = 0$. To complete the gradient of the prior,

$$\frac{\partial}{\partial \boldsymbol{\phi}_i} \log p(\boldsymbol{\phi} \,|\, \boldsymbol{\mu}, \boldsymbol{\lambda}) = -\boldsymbol{\phi}_i^t \cdot \text{diag} \left\{ \frac{1}{\xi^2} + \frac{1}{\lambda_{ij}^2 {\mu_j}^2} \right\} \tag{5.16}$$

follows from (5.8), where diag$\{\cdot\}$ indicates a matrix with D diagonal entries $\{\cdot\}$ that span $j = 1, \ldots D$.

5.5 Results

5.5.1 Lizards

Anolis lizards are a popular example of adaptive and convergent evolution (Losos, 2007). Losos et al. (1997) show lizards placed on islands in the Caribbean evolve directional, non-random distribution of limb length, toe-pad width and body mass over a 10-14 year period in response to the environment they are placed in. The various convergent forms of lizards on these islands are termed ecomorphs (Williams, 1972, 2013). Ecomorphs serve to categorize lizard species with shared phenotypic traits but without shared morphological characteristics in their common ancestor (Williams, 2013). Because of strong evidence for adaptive evolution, many employ an OU PCM to study Anolis lizard trait variation (Bastide, 2017; Khabbazian et al., 2016; Butler and King, 2004; Cressler et al., 2015). Here we re-examine the data set compiled by (Mahler et al., 2013) that contains four phylogenetic principle component (pPC) (Revell, 2009) traits derived from eleven phenotypic measurements important for niche adaptation (Mahler et al., 2010). The first four pPC explain 93% of the variation observed in the eleven traits. The original eleven traits include mean body size, limb size,

tail length and toepad number observed in between 1 to 19 lizards of each species.

Khabbazian et al. (2016) apply their lasso OU (l1OU) model to test for optimum shifts of the four pPC traits of 100 Greater Antillean Anolis species on the fixed, ultrametric tree of Mahler et al. (2013). Khabbazian et al. (2016) assume traits are independent and procedurally estimate trait-specific adaptation, $\mathbf{A}$ and trait specific variance (diagonals of $\mathbf{\Sigma}$) followed by magnitude and location of shifts. Khabbazian et al. (2016) correlate shifts $\boldsymbol{\phi}_i$ in the prior by placing a lasso penalty on the L^2 norm of branch-specific shift vectors. Here, we re-examine shifts on this fixed tree under our shift shrinkage model. We jointly estimate $\mathbf{\Sigma}$ and $\mathbf{A}$ along with shifts in optima $\boldsymbol{\upsilon}$. Although Khabbazian et al. (2016) assume pPC traits are uncorrelated, Bastide et al. (2018) show that this assumption may not hold. We estimate the off-diagonals of $\mathbf{\Sigma}$ to test our model's ability to detect possible covariance between traits. We place four independent Bayesian bridge priors on each trait's vector of increments and jointly learn about the global scale of each. We define a shift in the optimal value of trait j on branch i when the odds of $\boldsymbol{\phi}_{ij} > 0$ is greater than 10 or less than $\frac{1}{10}$. This corresponds to labeling optima shifts when the sign of the shift has Bayes factor support of at least 10 (Kass and Raftery, 1995). Fisher et al. (2021) argue this classification method allows one to categorize non-zero increments when ignorant to their expected magnitude. Under our shift shrinkage model, we run a 3.5 million state MCMC chain in BEAST and report estimated shift locations and magnitudes in figure (5.1). Traits 1 and 2 exhibit Bayes factor support > 10 for eleven and four shifts respectively while traits 3 and 4 do not have strong support for any shifts. Furthermore, we report posterior mean estimates of the diffusion variance for the four pPC traits with 95% high posterior density intervals, in order pPC1,...,pPC4, 0.17 (.13, 0.22), 0.15 (0.1, 0.22), 0.06 (0.05, 0.08) and 0.04 (0.03, 0.05) while covariance estimates are all between two to three orders of magnitude smaller (10^{-4} to 10^{-5}).

5.6 Discussion

We develop a scalable framework to learn the direction, magnitude, number and location of multi-dimensional optimal trait shifts on a tree within a Bayesian phylogenetic setting.

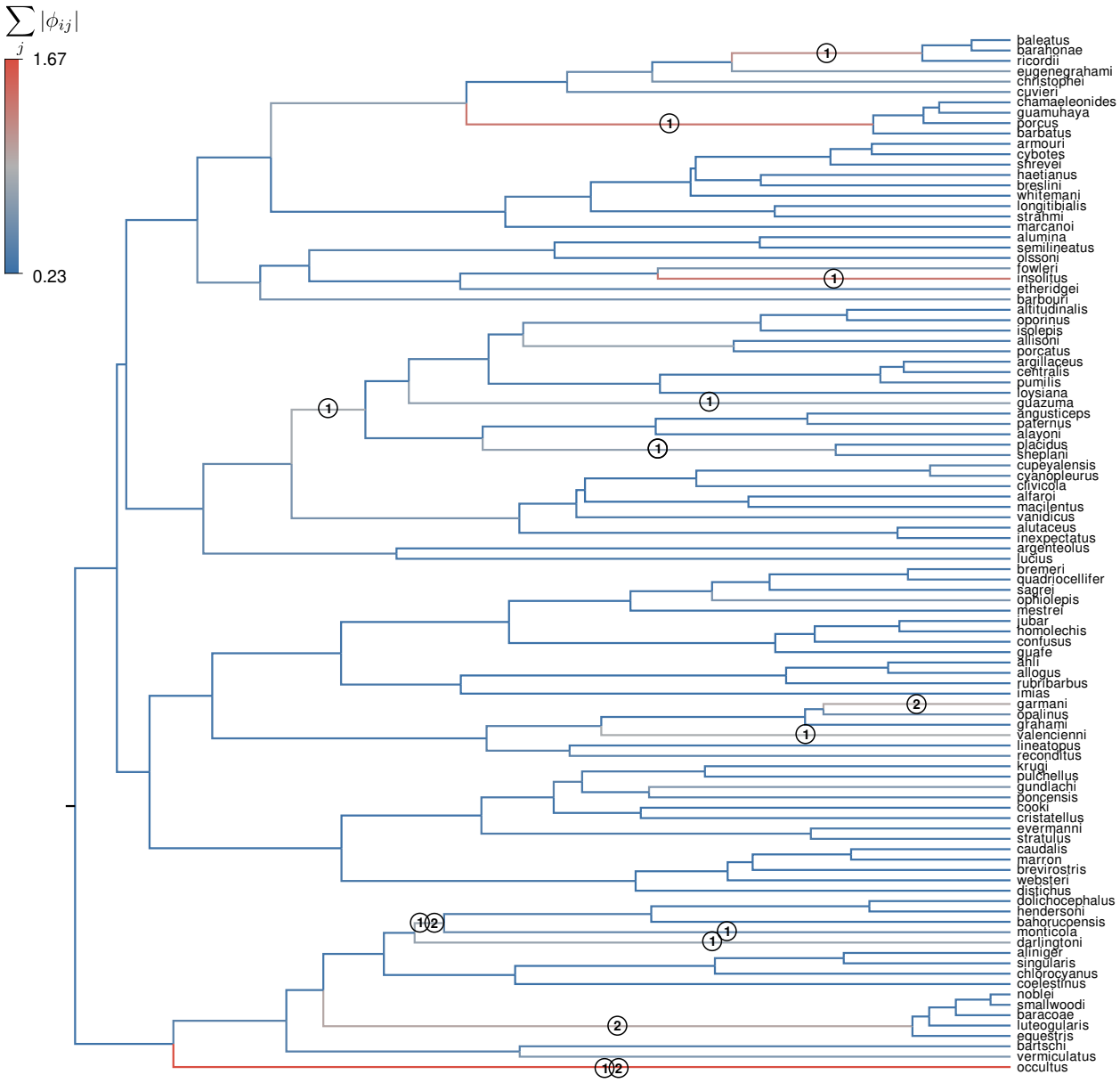


Figure 5.1: Fixed topology from Mahler et al. (2013) annotated by posterior shift estimates under shift shrinkage model. Branches are colored by the total shift magnitude per branch i, i.e. $\sum_j |\boldsymbol{\phi}_{ij}|$. We label branches with shifts that have Bayes factor support > 10 with a label of the principle component trait exhibiting a shift on that branch.

Previous approaches to detecting shifts iteratively mix over the number of shifts (Bastide et al., 2018) limiting scalability, take a maximum-likelihood approach that studies adaptive evolution on a fixed tree topology (Khabbazian et al., 2016; Bastide et al., 2017), assume

constant selection strength across traits (Uyeda and Harmon, 2014; Bastide et al., 2017), or only examine univariate shifts (Uyeda and Harmon, 2014). We assume each branch in the phylogeny maintains branch-specific optima and employ a Bayesian bridge shrinkage prior to reduce optima heterogeneity and identify the most probable shifts. Furthermore, we develop an efficient Hamiltonian Monte Carlo sampler that leverages recent gradient expressions (Fisher et al., 2021; Bastide et al., 2020) to improve the scalability of inference under our model as the number of tips and thus the number of parameters grows. Finally, we estimate the number, location and magnitude of shifts in four principle component traits of Anolis lizards (Mahler et al., 2013) and compare our findings to the l1ou of Khabbazian et al. (2016). We identify 11 shifts in the first pPC, 4 shifts in the second and none in the remaining two. Two of the shifts we identify are overlapping; we detect 13 shifts in total. Comparatively, Khabbazian et al. (2016) detect 16 total shifts throughout the tree. We expect the first two principle components to account for most shifts in the tree since they explain the most variation in their constituent traits. Mahler et al. (2013) report pPC1 accounts for 40% of the variation in the traits that compose it. One notable difference between our approach and that of the l1ou is that we assume optimal trait shifts are independent in the prior while Khabbazian et al. (2016) assumes shifts are coupled between traits on the same branch. Coupling shifts in the prior remains a possible avenue to extend our model in future work. Additionally, we estimate covariance between pPC traits, an important feature since phylogenetic principle components may not always be uncorrelated (Bastide et al., 2018).

In our analysis of the 100 species of Anolis lizards, we weight each of the observed trait vectors at the tips equally to provide a fair comparison to Khabbazian et al. (2016). However, trait measurements are sampled variably across different species of lizards in the fixed lizard tree of Mahler et al. (2013). Some lizard species are sampled only once while others are sampled as many as nineteen times. One could sensibly incorporate observed trait uncertainty in our proposed framework by encoding the observed trait variance matrix $\mathbf{U}_i$ to be inversely proportional to the number of lizard specimens sampled.

We hope our method will find further use enabling shift detection on large unknown

tree topologies. To preserve shift identifiability in this setting, one could track shifts on the branch preceding the most recent common ancestor of some subset of tips. This provides a natural opportunity to quantify shift uncertainty by reporting posterior probabilities of shifts across random trees.

CHAPTER 6

Future directions

Statistical models in phylogenetics, as in most disciplines, grow increasingly complex year by year and thus provide more detailed explanations of the natural world than ever before. However, also like most other disciplines in the 21st century, increasingly massive data sets challenge inference under complex models. For example, recent advances in portable genome sequencing (Quick et al., 2016) increase the accessibility and affordability of sampling taxa, providing unprecedented opportunities to investigate time-sensitive questions in epidemiology, and evolutionary biology with remarkable rapidity. In Chapters 3 through 5, I scale complex phylogenetic models to large trees. Presently, I turn my attention towards modeling extensions that fit within the previously discussed scalable framework and describe increasingly detailed biological hypotheses.

6.1 Mobile random walk

6.1.1 Introduction

Phylogenomics is one of at least two modern approaches to study outbreak dynamics. The more traditional, and also increasingly data-rich approach, is to use epidemiological stochastic compartmental models (Brauer, 2008) and, more recently, granular network-based models (Brauer et al., 2019). The former phylogenetic approach plays a critical role in ruling out viral transmission (Villabona-Arenas et al., 2020), timing zoonotic spillover events (see e.g. Lemey et al. (2003)) and reconstructing the geographic path traveled by an infectious disease (Bielejec et al., 2011). Under the latter compartmental modeling approach, time-series data of case counts, recoveries and population size serve to inform contact rates and incidence

rates. Unfortunately, current state-of-the-art methods are disparately developed and rarely integrate genomic and population-level data simultaneously.

One possible and simple way to marry the two approaches is to perform post-hoc analysis on a spatially explicit relaxed random walk (RRW) of Chapter 3 and regress a series of possible diffusion-related covariates such as population density or human mobility against the branch-rate multipliers of the RRW. In a similar spirit, Dellicour et al. (2020) test the impact of environmental factors on a spatially explicit reconstruction of the West Nile virus diffusing in North America under a RRW. If correlation exists between some covariate (e.g. case count) and branch-rate mulitpliers of the RRW, future work could include altering the diffusion matrix to be a function of relevant additional data. I discuss this modeling approach further in the next section.

6.1.2 Model

Consider a phylogeographic RRW where, as in the West Nile virus example of Chapter 3, we are interested in learning about the spread of an infectious pathogen. To this end, we sample viral isolate i at geographic coordinates $\mathbf{Y}_i \in \mathbb{R}^2$. Under the RRW,

$$\mathbf{Y}_i | \mathbf{Y}_{\text{pa}(i)} \sim \text{MVN}\left(\mathbf{Y}_i; \mathbf{Y}_{\text{pa}(i)}, t_i \mathbf{V}(\phi_i)\right) \tag{6.1}$$

where branch length t_i is the time node i evolves independent of parent $\text{pa}(i)$ and diffusion variance $\mathbf{V}(\phi_i)$ is a function of branch-specific rates ϕ_i.

Various types of epidemiological data may be adapted to this framework. If $\mathbf{V}(\phi_i) = \phi_i \boldsymbol{\Sigma}$ as in Chapter 3, then $\boldsymbol{\phi}$ effectively scale the speed of diffusion. One might posit that diffusion speed

$$\phi_i \propto b_i c_i k \tag{6.2}$$

where b_i is a marker of human mobility in the region where pathogen i is diffusing, c_i is the number of case counts in said region, and k is a measure of infectivity. If one collects local case count data at time of sampling, together with some marker of human mobility e.g. cell-phone movement data (Gonzalez et al., 2008; Zhao et al., 2016), then this is one

possible way to extend the RRW to the study of infectivity. Simultaneously, this framework could help untangle possible correlation between human mobility on pathogen spread.

6.2 Shrinking together and further priors

6.2.1 Coupled shift priors

In Chapter 5, I develop a model that posits independent shifts in optimal traits across all branches. However, when an organism's environment suddenly changes, such as in the placement of Anolis lizards onto various islands in the Caribbean (Mahler et al., 2010), one may consider that several phenotypic trait optima shift together in response to the environmental change. Here, I present an alternative prior formulation to give this scenario more complete consideration. Recall from Chapter 5, ϕ_{ij} is the shift in optimal trait j on branch i. On a tree with N tips, $i \in \{1, \ldots, 2N-2\}$ and $j \in P$, where P is the number of traits under study.

If some subset of the optima $\mathcal{S} \subset \{1, \ldots, P\}$ are thought to shift together, then for each $j \in \mathcal{S}$, the shifts on branch i, ϕ_{ij} should each be identically zero or non-zero. To impose this grouped shrinkage, I propose shrinking a measure of the total shift "magnitude" $f(\boldsymbol{\phi}_i)$ on branch i, where $\boldsymbol{\phi}_i = \{\phi_{i1}, \ldots \phi_{iP}\}$. Two possible examples include branch-specific L^2 and L^1 norms,

$$f(\boldsymbol{\phi}_i) = \begin{cases} \left(\sum_{j \in \mathcal{S}} {\phi_{ij}}^2\right)^{1/2} \\ \sum_{j \in \mathcal{S}} |\phi_{ij}| \end{cases} \tag{6.3}$$

where the first option reflects shrinking the magnitude of branch-specific vector of shifts $\boldsymbol{\phi}_i$, but deflates contributions from shifts < 1. The second option is indifferent towards shift size but is non-differentiable and may require special treatment in gradient-based inference techniques. To shrink the number of branches with shifts but retain the ability to detect large shifts, I place a Bayesian bridge shrinkage prior on one of the shift functions of equation (6.3).

Without loss of generality, assume $\mathcal{S} = \{1, \ldots, P\}$. Under the scale mixture representa-

tion of the Bayesian bridge Polson et al. (2014),

$$p(f(\boldsymbol{\phi}_i) \,|\, \mu, \lambda_i) = \mathcal{N}_{\frac{1}{2}}\big(f(\boldsymbol{\phi}_i);\ 0, \mu^2\lambda_i^2\big) \tag{6.4}$$

where μ is the global scale, λ_i is the local scale on branch i and $\mathcal{N}_{\frac{1}{2}}$ indicates the half-normal distribution since the support of $f(\boldsymbol{\phi}_i)$ is the positive real line. Plugging the L^2 functional into equation (6.4),

$$\begin{aligned} p(f(\boldsymbol{\phi}_i) \,|\, \mu, \lambda_i) &= \frac{\sqrt{2}}{\mu\lambda_i\sqrt{\pi}} \exp\left\{-\frac{1}{2}\frac{1}{\mu^2\lambda_i^2} f(\boldsymbol{\phi}_i)^2\right\} \\ &= \frac{\sqrt{2}}{\mu\lambda_i\sqrt{\pi}} \exp\left\{-\frac{1}{2}\frac{1}{\mu^2\lambda_i^2} \sum_{j\in\mathcal{S}} {\phi_{ij}}^2\right\}. \end{aligned} \tag{6.5}$$

To compute the joint density of shifts on branch i, i.e. $p(\boldsymbol{\phi}_i \,|\, \mu, \lambda_i)$ requires an invertible map between the space of $\boldsymbol{\phi}_i$ and the branch-specific magnitude $f(\boldsymbol{\phi}_i)$. Intuitively, one may view the L^2 $f(\boldsymbol{\phi}_i)$ as the radius of the sphere circumscribed by shift vector $\boldsymbol{\phi}_i$ in P-dimensional parameter space. To complete the spherical coordinate map when $P > 1$, one has a choice of prior over the implicit angle parameters $\theta_{i1}, \ldots, \theta_{iP-1}$ that fully define the transformed space,

$$\begin{aligned} \phi_{i1} &= r_i \cos\theta_{i1} \\ \phi_{i2} &= r_i \sin\theta_{i1} \cos\theta_{i2} \\ &\vdots \\ \phi_{iP-1} &= r_i \sin\theta_{i1} \ldots \sin\theta_{iP-2} \cos\theta_{iP-1} \\ \phi_{iP} &= r_i \sin\theta_{i1} \cdots \sin\theta_{iP-2} \sin\theta_{iP-1}, \end{aligned} \tag{6.6}$$

where $r_i = \left(\sum_{j\in\mathcal{S}} {\phi_{ij}}^2\right)^{1/2}$. For simplicity, I specify uniform priors over the support of each angle,

$$p(\theta_{ij}) = \begin{cases} U(0,\pi) \text{ if } j \in \{1, \ldots P-2\} \\ U(0, 2\pi) \text{ if } j = P-1. \end{cases} \tag{6.7}$$

Under this prior specification, where each angle is independent, shifts $\boldsymbol{\phi}_i$ are equally likely to lie on any edge of the P-dimensional sphere defined by radius r_i. The joint prior on $\boldsymbol{\eta}_i = \{r_i, \theta_1, \ldots, \theta_{iP-1}\}$ is simply the product of independent priors

$$p(\boldsymbol{\eta}_i \,|\, \mu, \lambda_i) = p(r_i \,|\, \mu, \lambda_i) \prod_{j=1}^{P-1} p(\theta_{ij}) \tag{6.8}$$

and since each branch is independent, the full joint prior is the product over all $2N-2$ branches. To compute the prior on shifts $\boldsymbol{\phi}_i$ requires the Jacobian determinant $\left|\frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}}\right|$ where $j, k \in \{1, \ldots, P\}$.

In Chapter 5, I propose a closed-form gradient solution to perform fast Hamiltonian Monte Carlo (HMC) sampling of all branch-specific shifts $\boldsymbol{\phi}$. In order to bring the efficient Hamiltonian Monte Carlo sampler to the grouped shrinkage priors, I will compute the new requisite term $\frac{\partial}{\partial \phi_{ij}} \log \left|\frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}}\right|$ to complete the gradient of the log-prior

$$\begin{aligned}
\frac{\partial}{\partial \phi_{ij}} \log p(\boldsymbol{\phi} \,|\, \mu, \boldsymbol{\lambda}) &= \frac{\partial}{\partial \phi_{ij}} \left[\log \prod_{i=1}^{2N-2} \left(p(r_i \,|\, \mu, \lambda_i) \prod_{j=1}^{P-1} p(\theta ij) \left| \frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}} \right| \right) \right] \\
&= \frac{\partial}{\partial \phi_{ij}} \left[\sum_{i=1}^{2N-2} \log p(r_i \,|\, \mu, \lambda_i) + \sum_{i=1}^{2N-2} \log \left| \frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}} \right| \right] \\
&= \frac{\partial}{\partial \phi_{ij}} \log p(r_i \,|\, \mu, \lambda_i) + \frac{\partial}{\partial \phi_{ij}} \log \left| \frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}} \right| \\
&= -\frac{\phi_{ij}}{\mu^2 \lambda_i^2} + \frac{\partial}{\partial \phi_{ij}} \log \left| \frac{d\boldsymbol{\phi}_{ik}}{d\boldsymbol{\eta}_{ij}} \right|
\end{aligned} \tag{6.9}$$

where the last equality follows from equation (6.5). Additional considerations remain before obtaining efficient sampling under this grouped prior. To facilitate fast sampling of the global scale requires working directly with the exponential power form of the Bayesian bridge (Polson et al., 2014). In future work, I will compute the joint density of the shifts, marginalizing over local scales, to take advantage of global scale Gibbs sampling, an important source of efficiency in Bayesian bridge inference. Xu et al. (2017) offer alternative shift prior grouping approaches to explore. One possible future approach may consider directly utilizing a multivariate Bayesian bridge prior where variance is not independent, however a scale mixture of covariance matrices may prove difficult to work with unless strong assumptions are made about its structure.

6.2.2 Branch time dependent prior

More realistic models of optima shifts in Ornstein-Uhlenbeck phylogenetic comparative methods may assume a priori that shifts are more likely to occur on long branches in the tree (Uyeda and Harmon, 2014). The key principle is that long branches offer more opportunities

for a shift to occur. A very simple way to encode this feature into a prior on the shifts is to make shift variance proportional to branch length. Under the scale-mixture Bayesian bridge representation,

$$p(\phi_{ij} \,|\, \mu, \lambda_i) = \text{MVN}\left(0, t_i \mu^2 \lambda_i^2\right), \tag{6.10}$$

where branch length t_i connects node i with its parent. This extension is compatible with the branch-specific shrinkage described in the previous section. Under other phylogenetic comparative models, such as the relaxed random walk (RRW) of Chapter 3, branch-specific rate multipliers are identifiable even though they only enter the likelihood as a product with branch lengths. Identifiability is maintained in the aforementioned case since strong prior information bounds the magnitude of branch-specific rate multipliers under the RRW and branch length estimates are influenced by the sequence data likelihood as well. In future work, care must be taken to ensure identifiability in the branch time dependent prior described here.

Bibliography

Mahler DL, Ingram T, Revell LJ, and Losos JB (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, 341(6143):292–295.

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

Jofré P, Das P, Bertranpetit J, and Foley R (2017). Cosmic phylogeny: reconstructing the chemical history of the solar neighbourhood with an evolutionary tree. *Monthly Notices of the Royal Astronomical Society*, 467(1):1140–1153.

Lemey P, Rambaut A, Welch JJ, and Suchard MA (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885.

Drummond AJ, Ho SY, Phillips MJ, and Rambaut A (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. (2016). Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232.

Worobey M, Han G.-Z, and Rambaut A (2014). A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257.

Drummond AJ and Suchard MA (2010). Bayesian random local clocks, or one rate to rule them all. *BMC biology*, 8(1):114.

Ji X, Zhang Z, Holbrook A, Nishimura A, Baele G, Rambaut A, et al. (2019). Gradients do grow on trees: a linear-time o(n)-dimensional gradient for statistical phylogenetics. *arXiv preprint arXiv:1905.12146*.

Hansen TF (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351.

Bastide P (2017). *Shifted stochastic processes evolving on trees: application to models of adaptive evolution on phylogenies.* PhD thesis, Université Paris-Saclay.

Khabbazian M, Kriebel R, Rohe K, and Ané C (2016). Fast and accurate detection of evolutionary shifts in ornstein–uhlenbeck models. *Methods in Ecology and Evolution*, 7(7):811–824.

Uyeda JC and Harmon LJ (2014). A novel bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology*, 63(6):902–918.

Felsenstein J (1985). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.

Cavalli-Sforza LL and Edwards AW (1967). Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233.

Paradis E (2014). An introduction to the phylogenetic comparative method. In *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 3–18. Springer.

Cornwell W and Nakagawa S (2017). Phylogenetic comparative methods. *Current Biology*, 27(9):R333–R336.

O'Meara BC (2012). Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics*, 43:267–285.

Yang Z (2014). *Molecular evolution: a statistical approach.* Oxford University Press.

Lange K (2003). *Mathematical and statistical methods for genetic analysis.* Springer Science & Business Media.

Yang Z (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372.

Felsenstein J (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249.

Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

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

Brooks S, Gelman A, Jones G, and Meng X.-L (2011). *Handbook of markov chain monte carlo*. CRC press.

Blei DM, Kucukelbir A, and McAuliffe JD (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Zhang C and Matsen IV FA (2018). Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*.

Neal RM (2011). MCMC using Hamiltonian dynamics. In Brooks S, Gelman A, Jones GL, and Meng X.-L, editors, *Handbook of Markov Chain Monte Carlo*, volume 2. CRC Press New York, NY.

Betancourt M (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Ives AR and Garland Jr T (2009). Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26.

Cybis GB, Sinsheimer JS, Bedford T, Mather AE, Lemey P, and Suchard MA (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The Annals of Applied Statistics*, 9(2):969.

Schluter D, Price T, Mooers AØ, and Ludwig D (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, 51(6):1699–1711.

Biek R, Henderson JC, Waller LA, Rupprecht CE, and Real LA (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences*, 104(19):7993–7998.

Liu JS (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61.

Bryant D, Galtier N, and Poursat M.-A (2005). Likelihood calculation in molecular phylogenetics. In Gascuel O, editor, *Mathematics of Evolution and Phylogeny*, pages 33–62. Oxford Univ. Press.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, and Rambaut A (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Barnard J, McCulloch R, and Meng X.-L (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.

Zhang X, Boscardin WJ, and Belin TR (2006). Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896.

Caetano DS and Harmon LJ (2019). Estimating correlated rates of trait evolution with uncertainty. *Systematic Biology*, 68(3):412–429.

Lewandowski D, Kurowicka D, and Joe H (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Zhang Z, Nishimura A, Bastide P, Ji X, Payne RP, Goulder P, et al. (2019). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *arXiv preprint arXiv:1912.09185*.

Levine RA and Casella G (2006). Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100.

Petersen KB, Pedersen MS, et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Hassler G, Tolkoff MR, Allen WL, Ho L. ST, Lemey P, and Suchard MA (2020). Inferring phenotypic trait evolution on large trees with many incomplete measurements. *Journal of the American Statistical Association*, pages 1–15.

Bastide P, Ané C, Robin S, and Mariadassou M (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology*, 67(4):662–680.

Petersen LR, Brault AC, and Nasci RS (2013). West Nile virus: review of the literature. *Journal of the American Medical Association*, 310(3):308–315.

Gray R, Veras N, Santos L, and Salemi M (2010). Evolutionary characterization of the West Nile virus complete genome. *Molecular Phylogenetics and Evolution*, 56(1):195–200.

Stearns SC (2000). Life history evolution: successes, limitations, and prospects. *Naturwissenschaften*, 87(11):476–486.

Pacifici M, Visconti P, Butchart SH, Watson JE, Cassola FM, and Rondinini C (2017). Species' traits influenced their response to recent climate change. *Nature Climate Change*, 7(3):205.

Fritz SA, Bininda-Emonds OR, and Purvis A (2009). Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology Letters*, 12(6):538–549.

de Silva S and Leimgruber P (2019). Demographic tipping points as early indicators of vulnerability for slow-breeding megafaunal populations. *Frontiers in Ecology and Evolution*, 7:171.

Santini L, Cornulier T, Bullock JM, Palmer SC, White SM, Hodgson JA, et al. (2016). A trait-based approach for predicting species responses to environmental change from sparse data: how well might terrestrial mammals track climate change? *Global Change Biology*, 22(7):2415–2424.

Oli MK (2004). The fast–slow continuum and mammalian life-history patterns: an empirical evaluation. *Basic and Applied Ecology*, 5(5):449–463.

Millar JS and Zammuto RM (1983). Life histories of mammals: an analysis of life tables. *Ecology*, 64(4):631–635.

Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme C. DL, et al. (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648–2648.

Snapinn KW, Holmes EC, Young DS, Bernard KA, Kramer LD, and Ebel GD (2007). Declining growth rate of West Nile virus in North America. *Journal of Virology*, 81(5):2531–2534.

Springer MS, Murphy WJ, Eizirik E, and O'Brien SJ (2003). Placental mammal diversification and the cretaceous–tertiary boundary. *Proceedings of the National Academy of Sciences*, 100(3):1056–1061.

Davidson I, Fusaro A, Heidari A, Monne I, and Cattoli G (2014). Molecular evolution of H9N2 avian influenza viruses in Israel. *Virus Genes*, 48(3):457–463.

Thorne JL, Kishino H, and Painter IS (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657.

Yoder AD and Yang Z (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17(7):1081–1090.

Smith SA, Beaulieu JM, and Donoghue MJ (2010). An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proceedings of the National Academy of Sciences*, 107(13):5897–5902.

Bletsa M, Suchard MA, Ji X, Gryseels S, Vrancken B, Baele G, et al. (2019). Divergence dating using mixed effects clock modelling: An application to HIV-1. *Virus Evolution*, 5(2):vez036.

Ho SY and Duchêne S (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, 23(24):5947–5965.

Polson NG, Scott JG, and Windle J (2014). The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733.

Ji X, Zhang Z, Holbrook A, Nishimura A, Baele G, Rambaut A, et al. (2020). Gradients do grow on trees: a linear-time o(n)-dimensional gradient for statistical phylogenetics. *Molecular Biology and Evolution*, 37(10):3047–3060.

West M (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Nishimura A and Suchard MA (2019). Shrinkage with shrunken shoulders: inference via geometrically/uniformly ergodic gibbs sampler. *arXiv preprint arXiv:1911.02160*.

Ferreira MA and Suchard MA (2008). Bayesian analysis of elapsed times in continuous-time markov chains. *Canadian Journal of Statistics*, 36(3):355–368.

Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0*.

Zhang Y and Sutton C (2011). Quasi-newton methods for markov chain monte carlo. *Advances in Neural Information Processing Systems*, 24:2393–2401.

Huchon D, Madsen O, Sibbald MJ, Ament K, Stanhope MJ, Catzeflis F, et al. (2002). Rodent phylogeny and a timescale for the evolution of glires: evidence from an extensive taxon sampling using three nuclear genes. *Molecular Biology and Evolution*, 19(7):1053–1065.

Douzery EJ, Delsuc F, Stanhope MJ, and Huchon D (2003). Local molecular clocks in three nuclear genes: divergence times for rodents and other mammals and incompatibility among fossil calibrations. *Journal of Molecular Evolution*, 57(1):S201–S213.

Horner DS, Lefkimmiatis K, Reyes A, Gissi C, Saccone C, and Pesole G (2007). Phylogenetic analyses of complete mitochondrial genome sequences suggest a basal divergence of the enigmatic rodent anomalurus. *BMC Evolutionary Biology*, 7(1):1–12.

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, and Huchon D (2009). Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evolutionary Biology*, 9(1):1–12.

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

McAuley JL, Gilbertson BP, Trifkovic S, Brown LE, and McKimm-Breschkin JL (2019). Influenza virus neuraminidase structure and functions. *Frontiers in Microbiology*, 10:39.

Wilson IA and Cox NJ (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annual Review of Immunology*, 8(1):737–787.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA (2013). Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724.

Minin VN, Bloomquist EW, and Suchard MA (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Gill MS, Ho T, Si L, Baele G, Lemey P, and Suchard MA (2017). A relaxed directional random walk model for phylogenetic trait evolution. *Systematic Biology*, 66(3):299–319.

Pagel M (2002). Modelling the evolution of continuously varying characters on phylogenetic trees. *Morphology, shape and phylogeny*, 269:286.

Butler MA and King AA (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695.

Clavel J, Escarguel G, and Merceron G (2015). mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, 6(11):1311–1319.

Bartoszek K, Pienaar J, Mostad P, Andersson S, and Hansen TF (2012). A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology*, 314:204–215.

Hansen TF, Pienaar J, and Orzack SH (2008). A comparative method for studying adaptation to a randomly evolving environment. *Evolution: International Journal of Organic Evolution*, 62(8):1965–1977.

Bastide P, Mariadassou M, and Robin S (2017). Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society: Series B*, page np.

Fisher AA, Ji X, Nishimura A, Lemey P, and Suchard MA (2021). Shrinkage-based random local clocks with scalable inference. *arXiv preprint arXiv:2105.07119*.

Bastide P, Ho L. ST, Baele G, Lemey P, and Suchard MA (2020). Efficient bayesian inference of general gaussian models on large phylogenetic trees. *arXiv preprint arXiv:2003.10336*.

Zhang Z, Nishimura A, Bastide P, Ji X, Payne RP, Goulder P, et al. (2021). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics*, 15(1):230–251.

Fisher AA, Ji X, Zhang Z, Lemey P, and Suchard MA (2021). Relaxed random walks at scale. *Systematic Biology*, 70(2):258–267.

Hassler GW, Gallone B, Aristide L, Allen WL, Tolkoff MR, Holbrook AJ, et al. (2021). Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis. *arXiv preprint arXiv:2107.01246*.

Losos JB (2007). Detective work in the west indies: integrating historical and experimental approaches to study island lizard evolution. *Bioscience*, 57(7):585–597.

Losos JB, Warheitt KI, and Schoener TW (1997). Adaptive differentiation following experimental island colonization in anolis lizards. *Nature*, 387(6628):70–73.

Williams EE (1972). The origin of faunas. evolution of lizard congeners in a complex island fauna: a trial analysis. In *Evolutionary biology*, pages 47–89. Springer.

Williams EE (2013). 15. ecomorphs, faunas, island size, and diverse end points in island radiations of anolis. In *Lizard ecology*, pages 326–370. Harvard University Press.

Cressler CE, Butler MA, and King AA (2015). Detecting adaptive evolution in phylogenetic comparative analysis using the ornstein–uhlenbeck model. *Systematic Biology*, 64(6):953–968.

Revell LJ (2009). Size-correction and principal components for interspecific comparative studies. *Evolution: International Journal of Organic Evolution*, 63(12):3258–3268.

Mahler DL, Revell LJ, Glor RE, and Losos JB (2010). Ecological opportunity and the rate of morphological evolution in the diversification of greater antillean anoles. *Evolution: International Journal of Organic Evolution*, 64(9):2731–2745.

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

Brauer F, Castillo-Chavez C, and Feng Z (2019). *Mathematical models in epidemiology*, volume 32. Springer.

Villabona-Arenas CJ, Hanage WP, and Tully DC (2020). Phylogenetic interpretation during outbreaks requires caution. *Nature Microbiology*, 5(7):876–877.

Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, and Vandamme A.-M (2003). Tracing the origin and history of the hiv-2 epidemic. *Proceedings of the National Academy of Sciences*, 100(11):6588–6592.

Bielejec F, Rambaut A, Suchard MA, and Lemey P (2011). Spread: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics*, 27(20):2910–2912.

Dellicour S, Lequime S, Vrancken B, Gill MS, Bastide P, Gangavarapu K, et al. (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*, 11(1):1–11.

Gonzalez MC, Hidalgo CA, and Barabasi A.-L (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Zhao K, Tarkoma S, Liu S, and Vo H (2016). Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1911–1920. IEEE.

Xu Z, Schmidt DF, Makalic E, Qian G, and Hopper JL (2017). Bayesian sparse global-local shrinkage regression for selection of grouped variables. *arXiv preprint arXiv:1709.04333*.