Title
Big Bayesian Phylogenetic Comparative Methods

Author
Hassler, Gabriel William

Publication Date
2022

UNIVERSITY OF CALIFORNIA

Los Angeles

Big Bayesian Phylogenetic Comparative Methods

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

by

Gabriel Hassler

2022

ABSTRACT OF THE DISSERTATION

Big Bayesian Phylogenetic Comparative Methods

by

Gabriel Hassler
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2022
Professor Marc Adam Suchard, Chair

Phylogenetic comparative methods seek to untangle the complex web of selective pressures driving biological evolution. These methods seek to identify associations between different biological traits over evolutionary history. Statistical models of phenotypic evolution need to account for the shared evolutionary history between different species, and accounting for this non-independence poses computational challenges. These challenges are compounded by missing observations, high-dimensional traits and highly-structured data. Here, I develop computational and modeling approaches that dramatically improve the computational efficiency and scalability of these models to enable Bayesian phylogenetic comparative analysis of unprecedentedly large data sets. First, I develop an algorithm that analytically marginalizes missing observations in a (relatively) simple model of phenotypic evolution. This algorithm is broadly applicable beyond this simple model and allows scalable inference under a variety of model extensions. These extensions include models that accommodate residual variance, allowing measurement of phylogenetic heritability, and linear dimension reduction, allowing phylogenetic comparative analyses for high-dimensional traits. I combine this work into a generalizable modeling framework that allows researchers to build flexible, highly structured models that remain scalable for both large number of taxa and many observations per taxon. This work achieves increases in computation speed by more than two orders of magnitude across several contexts, bringing computation time down from weeks or months to minutes or hours in multiple real-world applications.

The dissertation of Gabriel Hassler is approved.

Damla Şentürk

Janet S. Sinsheimer

Eric M. Sobel

Marc Adam Suchard, Committee Chair

University of California, Los Angeles

2022

To my parents, who inspired my love of knowledge.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to acknowledge my advisor, Marc Suchard, without whom this work would not be possible. Marc has been a source of academic inspiration throughout my time at UCLA, and his statistical intuition has (when followed) led to my most fruitful research and (when ignored) taught me valuable lessons. He created a space where we could disagree and I could fail, with only a very polite “I told you so” when the two coincided. Marc takes is roles as a mentor as seriously as his research, and it has been clear to me that he sees his trainees not as extensions of himself to multiply his research output, but as future researchers and scientists who can act independently.

I thank the members of my committee Damla Şentürk, Eric Sobel and Janet Sinsheimer. Damla provided the clearest classroom instruction I received at UCLA, which was much needed as a Biomathematics graduate student (with almost no math background) in the process of realizing that I really wanted to study statistics. Eric treated me as an equal during my time as student representative to the faculty for the Biomathematics program. He included me in important conversations and genuinely listened to my input. I thank Janet for seeing my potential as an applicant to the Biomathematics program, despite the fact that I hadn’t had any formal training beyond univariate calculus. Janet was the first person to make me feel that I belonged in the program despite my non-traditional background. She has remained supportive throughout the program, writing numerous letters of recommendation (often on short notice) and giving me my first opportunity to teach at the graduate level.

I thank the members of the Suchard group that have spanned my time at UCLA. In particular, I thank Zhenyu Zhang and Alex Fisher who have shared this journey with me. They were always available whether I needed help with BEAST, a sharp mind to bounce ideas off, or just to commiserate. I thank Yuxi Tian for his invaluable help in writing my eventually successful NIH F31 application. I thank Jianxiao Yang for keeping me humble and reminding me that, as hard as I think my work is, hers is harder. I thank Xiang Ji for

his tireless efforts to keep BEAST running despite my efforts to break it. He is a BEAST tamer second only to Marc himself and my first stop when I have exhausted my own abilities. I thank Aki Nishimura for his sage advice on all things HMC, and his compliment on the quality of my homemade kombucha remains one of the greatest achievements of my life. I thank Andrew Holbrook for being a second mentor to me. He has helped me through some of the thornier statistical problems I have enountered and offered some of the best carreer advice I've received. I thank Fan Bu, Karthik Gangavarapu and Andy Magee for gracefully filling the Suchard group's post-doc void and serving as mentors during the latter years of my dissertation research. Finally, I thank Max Tolkoff for inspiring much of this work. This dissertation is largely an extension of the work that Max started, and up to the day I am writing this I still find new parts of his dissertation that I wish I had read earlier.

Chapters 2, 3 and 4 are versions of published or accepted manuscripts, and I would like to thank all of my collaborators on this work. Chapter 2 is a version of "Data integration in Bayesian phylogenetics" in the *Annual Review of Statistics and its Application* with co-authors Andrew Magee, Zhenyu Zhang, Guy Baele, Philippe Lemey, Xiang Ji, Mathieu Fourment and Marc A Suchard. Chapter 3 is a version of "Inferring phenotypic trait evolution on large trees with many incomplete measurements" in the *Journal of the American Statistical Association* with co-authors Max R Tolkoff, William L Allen, Lam Si Tung Ho, Philippe Lemey and Marc A Suchard. Chapter 4 is a version of "Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis" in *Methods in Ecology and Evolution* with co-authors Brigida Gallone, Leandro Aristide, William L Allen, Max R Tolkoff, Andrew J Holbrook, Guy Baele, Philippe Lemey and Marc A Suchard. I would like to specially thank Philippe Lemey for his repeatedly recognizing suitable biological problems on which to test these methods and doing the hard work of collecting and processing much of the real-world data analyzed in this dissertation. I would also like to thank Sarah-Sophie Weil, William Allen and Leandro Aristide for providing the motivating problems that have led to much of this work. I particularly thank Sarah-Sophie Weil for serving as an unofficial beta-tester for PhylogeneticFactorAnalysis.jl, which was much improved by her input. Finally, I thank Paul Bastide, who has authored some of the most impressive and impactful recent work in phylogenetic comparative methods, without which my work would not have been possible.

I am grateful for the camaraderie and support from the students in the Biomathematics Graduate Program. Sam Christensen and I were the only two in our cohort, and my success in coursework my first year was only possible due to his passion for math and willingness to share it. Beyond that, his friendship has helped keep my spirits up throughout my graduate work. I looked forward to the Biomath pub nights every Thursday at Barney's, and I particularly thank the "regulars" for making it something I could rely on each week to decompress. In particular, I thank Bhaven Mistry, Stephanie Lewkiewicz, Tim Stutz, Renaud Desalles, Mauricio Cruz Loya, Alfonso Landeros, Rachel Mester, Mariana Harris, Christine Craib and Gary Zhou. The nights we spent celebrating, commiserating, working through problems together and just having a good time are my best memories of my time at UCLA. I thank Jason Lin for getting on board with my hare-brained plan to run the LA Marathon together, and for his support through training and injuries. I'm not sure I would have finished the first one without him, and the discipline, routine and activity helped keep me sane and productive throughout the pandemic as I trained for more. I thank Paheli Desai-Chowdhry for keeping it real. I thank Vicky Kelley for her friendship and for getting me out of the house more than I otherwise would have.

I am grateful for the funding I have received that has allowed me to pursue this work with a degree of financial independence. This includes funding from UCLA Department of Biomathematics, the NIH Systems and Integrative Biology Training Grant (T32-GM008185), the NIH Genomic Analysis Training Program (T32-HG002536), the NIH F31 Predoctoral Fellowship (F31-AI154824) and several other sources of funding through my advisor, Marc Suchard.

Finally, I thank my family. My wife, Ali, encouraged me pursue this work knowing that she would bear the primary financial burden of supporting our household. Her hard work has meant that I could focus on my research without having to worry about bills or rent. She has patiently listened as I talk about the virtues of Bayesian statistics or the challenges of latent factor models. Most importantly, regardless or whether I have had a good or bad day working on research, I know that I will come home at the end of the day to (or, as was the case for much of the last two years, never leave) a place where I am loved. My cats, Gatsby and Furiosa, have likely contributed more than a few typos to this dissertation, and I thank them for not correcting the record when I blame them for the ones that aren't their fault. My in-laws, Lisa, Chloe, Doug, Rui, Daniel, Danice, Julian and Natalie, have given me a family away from home, and I thank them for the many meals we have shared together and their support throughout this process. My sister, Grace, has consistently been an example to me of someone can simultaneously enjoy their work yet find meaning and joy outside of it. That lesson brought me through some of the more challenging periods of this dissertation and allowed me to maintain my self worth even when I struggled with work. Ultimately, I would like to thank my parents for simultaneously giving me all the resources I could ever need to get to where I am today and allowing me find my own way. They have treated me like an adult for far longer than I likely deserved it and have given me the space to make mistakes and grow. If the goal of parents is to prepare their children to live fulfilling lives, then I believe they have succeeded.

VITA

2014 B.A. (Anthropology and Biology), Washington University in St. Louis

2014–2016 Teacher, St. John's College High School, Washington DC. Designed curricula and taught 9th grade biology and 12th grade Honors Anatomy and Physiology.

2021–2022 Fellow, NIAID Ruth L. Kirschstein NRSA for Individual Predoctoral Fellowship (F31). Conducted research into scalable Bayesian inference for continuous trait models in statistical phylogenetics with applications related to pathogen-host interactions.

PUBLICATIONS

Gabriel W Hassler, Brigida Gallone, Leandro Aristide, William L Allen, Max R Tolkoff, Andrew J Holbrook, Guy Baele, Philippe Lemey, and Marc A Suchard. "Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis". In: *Methods in Ecology and Evolution* (2022). https://doi.org/10.1111/2041-210X.13920.

Gabriel W Hassler, Andrew Magee, Zhenyu Zhang, Guy Baele, Philippe Lemey, Xiang Ji, Mathieu Fourment, and Marc A Suchard. "Data integration in Bayesian phylogenetics". In: *Annual Review of Statistics and its Application* (2022). in press.

Sarah-Sophie Weil, Laure Gallien, Sebastien Lavergne, Luca Borger, **Gabriel W Hassler**, Michael PJ Nicolai, and William L Allen. "The role of chameleons' traits in biogeographic long-distance dispersal". In: *Ecography* (2022). in press.

Tetyana I Vasylyeva, Courtney E Fang, Michelle Su, Jennifer L Havens, Edyth Parker, Jade C Wang, Mark Zeller, Anna Yakovleva, **Gabriel W Hassler**, Moinuddin A Chowdhury, et al. “Introduction and Establishment of SARS-CoV-2 Gamma Variant in New York City in Early 2021”. In: *Journal of Infectious Diseases* (2022). in press.

Alexander A Fisher, **Gabriel W Hassler**, Xiang Ji, Guy Baele, Marc A Suchard, and Philippe Lemey. “Review: Scalable Bayesian phylogenetics”. In: *Philosophical Transactions of the Royal Society B* (2022). in press.

Jiumeng Sun, Xinxin Li, Yan Ziqing, Jiang Zhiwen, Shouzhi Sheng, **Gabriel W Hassler**, Dongyan Li, Zhang Letian, Wan-Ting He He, Xiang Ji, and Shuo Su. “Genome sequencing and assembly at the chromosome-level of Hystrix brachyura provides insight into longevity and immune adaptations”. In: (2021). under review.

Anderson F Brito, Elizaveta Semenova, Gytis Dudas, **Gabriel W Hassler**, et al. “Global disparities in SARS-CoV-2 genomic surveillance”. In: *medRxiv* (2021). https://www.medrxiv.org/content/10.1101/2021.08.21.21262393v1.

Gabriel W Hassler, Max R Tolkoff, William L Allen, Lam Si Tung Ho, Philippe Lemey, and Marc A Suchard. “Inferring phenotypic trait evolution on large trees with many incomplete measurements”. In: *Journal of the American Statistical Association* 117 (538 2020). https://doi.org/10.1080/01621459.2020.1799812.

Parker VanValkenburgh, Sarah Kennedy, Carol Rojas Vega, and **Gabriel W Hassler**. “El Contrato del Mar: Colonial Life and Maritime Subsistence at Carrizales, Zaña Valley, Peru”. In: *Maritime Communities of the Ancient Andes*. Ed. by Gabriel Prieto and Daniel Sandweiss. Gainesville, FL: University Press of Florida, 2019.

CHAPTER 1

Introduction

The biological diversity we see around us is the product of numerous evolutionary forces acting simultaneously over the last ∼4 billion years (Dodd et al., 2017). Natural selection, however, rarely acts on a single biological phenotype or trait but rather causes multiple traits to evolve simultaneously. Physical or energetic constraints can also causes traits that are not the target of selective pressure to co-evolve with other traits that are. Particular traits (or groups of traits) may be the targets of multiple selective forces, often acting in opposing directions. Phylogenetic comparative methods (Felsenstein, 1985b), the subject of this dissertation, seek to untangle this complex web of evolutionary pressures.

Phylogenies are tree structures that capture the evolutionary history of a group of organisms. Nodes on the tree represent individual taxa or species, with those at the tips of the tree representing existing or observable taxa and internal nodes representing ancestral taxa. These trees are often time-resolved, with distances between nodes corresponding to calendar time, and rooted, with there being a single node ancestral to all taxa on the tree.

Phylogenetic comparative methods assume some statistical data-generative process on the tree from which each species' traits arise. For continuous traits, these generative processes typically involve some form of diffusion and result in Gaussian likelihoods on the observed data. The parameter of primary scientific interest in these models is typically the between-trait covariance matrix that captures the relationship between different biological traits as they evolve on the tree. Learning about this between-trait covariance structure is complicated by the complex evolutionary relationships between the different species.

As advances in sequencing technology, computation and statistical methods have enabled inference of increasingly large phylogenetic trees, the computational challenges in performing inference in these large trait-based models have compounded. The work discussed below seeks to ease the computational burden of a sub-set of these phylogenetic comparative methods as well as expand the flexibility of these models in a way that maintains scalability. Chapters 2, 3, 4 and 5 were written as stand-alone manuscripts and can be read as such.

Chapter 2 is a review of recent advances in Bayesian phylogenetics generally, with special emphasis on data integration. As much of the review focuses on topics outside the scope of the rest of this dissertation, I point readers to Sections 2.1.1, 2.3.1 and 2.3.3 as particularly relevant to the remaining chapters.

Chapter 3 addresses the challenge of inference in a common phylogenetic comparative model when some observations are missing. With complete data, likelihood calculations scale linearly in the number of taxa N. However, existing procedures to calculate the likelihood with complete data cannot accommodate missing data, and inference with missing data requires data augmentation or imputation. Bayesian inference in this context then scales as $\mathcal{O}(N^2)$, which can be computationally intractable for large data sets. I develop an algorithm that analytically integrates out the missing observations in the likelihood and can calculate the likelihood of the observed data only in $\mathcal{O}(N)$. This procedure also permits a broad range of previously intractable model extensions. For example, I implement a model that accommodates measurement error or within-species variation. Inference in this model previously scaled as $\mathcal{O}(N^2)$ regardless of whether any data were missing, while my approach allows likelihood calculations and inference in $\mathcal{O}(N)$ time. The flexibility of this likelihood calculation algorithm is the basis for the remainder of the dissertation.

Chapter 4 leverages the likelihood calculation algorithm of the previous chapter to improve the computational efficiency of phylogenetic factor analysis (PFA; Tolkoff et al., 2018). Most comparative methods seeking to understand the evolutionary correlations between groups of traits scale at best quadratically and often cubically in the number of traits P. As such, inference in these models quickly becomes intractable as the number of traits grows. PFA addresses this challenge by mapping high-dimensional traits to low-dimensional latent factors, where the low-dimensional latent factors evolve along the phylogenetic tree. While PFA scales linearly in the number of traits P, is scales scales quadratically in the number of taxa as originally implemented, leaving researchers with no available Bayesian methods that scale to big-N, big-P data sets. My work in Chapter 4 applies the likelihood calculation algorithm from Chapter 3 to decrease inference times from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, enabling Bayesian phylogenetic comparative inference in the big-N, big-P context. I further address challenges with latent factor models generally, including developing procedures for post-processing data that do not rely on an implicit ordering of the traits as well as a data-driven approach for selecting an appropriate number of latent factors.

Chapter 5 generalizes the work of the previous two chapters and introduces a flexible framework for developing structured and scalable phylogenetic comparative models. This approach, dubbed phylogenetic structural equation modeling (SEM) allows researchers to partition their data across multiple sub-models, each of which can map the observed data to the phylogenetic tree in a different way. Assuming these sub-models meet certain criteria, all inference machinery from the previous two chapters holds, and these structured models scale linearly both in the number of taxa N and traits P. The main challenge to Bayesian inference in these models is sampling from an unusually constrained correlation matrix with structural zeros. I develop a new method for sampling from such constrained correlation matrices to enable this work. There are additional identifiability challenges in this model and latent factor models generally that I also address in this work.

Finally, in Chapter 6, I summarize the methodological advancements and scientific insights achieved through this dissertation. I also briefly discuss future directions for this research.

CHAPTER 2

Literature review

2.1 Introduction

All living things on the planet share a common evolutionary history. Phylogenetic trees capture the evolutionary relationships between groups of organisms (Baldauf, 2003). At the extremes, these phylogenies can describe the evolution of all life on earth spanning ~ 4 billion years or that of a viral lineage over weeks. Statistical phylogenetics gives researchers the tools to study these evolutionary processes and can be used to answer both fundamental biological questions, such as “which species of ape is most closely related to humans and when did our evolutionary histories diverge?” (Bradley, 2008) and more practical ones such as “how effective are various interventions at controlling the spread of a viral epidemic?” (Dellicour et al., 2018). Researchers typically rely on molecular sequences (e.g. DNA, RNA, amino acids) to infer the phylogeny itself and commonly incorporate additional sources of data to answer specific questions. For example, toward the end of this review in Section 2.4 we examine a case study where researchers investigate the early spread of SARS-CoV-2, the virus that causes COVID-19, across the world (Lemey et al., 2020). This analysis incorporates viral genetic sequences, sample collection dates and locations, individual-level travel history, global air traffic patterns, local SARS-CoV-2 case counts and within-host infection dynamics into a coherent statistical model that allows researchers to reconstruct the early pathways along which SARS-CoV-2 spread early in the pandemic.

From a statistical perspective, phylogenetics offers a rich array of complex hierarchical models for both inferring the phylogeny itself as well as parameters associated with the underlying evolutionary processes of interest (Nascimento et al., 2017). The complexity of these models, however, can result in theoretical and computational challenges to inference that limit their scalability. These challenges have led to the development of statistical methods with broad utility beyond the field of phylogenetics itself. In this review, we first introduce the fundamental statistical approaches to phylogenetics in Section 2.1.1 and the advantages of the Bayesian approach in Section 2.1.2 below. We then discuss modern methods for inferring phylogenetic trees in Section 2.2 and data integration in Section 2.3. As mentioned previously, we examine in Section 2.4 a case study that relies on many of the methods discussed in earlier sections.

2.1.1 Molecular evolution on a phylogenetic tree

Let the phylogenetic tree $\mathcal{F}$ be a bifurcating directed acyclic graph with N degree-one terminal/tip nodes $\nu_1, \ldots, \nu_N$, $N-2$ degree-three internal nodes $\nu_{N+1}, \ldots, \nu_{2N-2}$ and one degree-two root node ν_{2N-1}. With the exception of the root node, there is an edge connecting each node ν_i to its parent $\nu_{\text{pa}(i)}$ with length t_i. See Figure 2.1 for a simple example. Depending on the statistical model, these edge lengths are typically proportional to either the amount of time or expected number of genetic changes separating nodes ν_i and $\nu_{\text{pa}(i)}$. While some parameterizations permit multifurcations/polytomies (i.e. nodes with more than two children), we focus on the bifurcating case without loss of generality as multifurcations can be represented via bifurcations with edge lengths equal to zero. Note that some parameterizations assume unrooted trees where the degree-two root node is omitted. In the unrooted case, the phylogeny is no longer directed and there are no fixed parent/child relationships between nodes.

Likelihood-based phylogenetic inference typically relies on molecular sequences $\mathbf{S}$ to inform the phylogenetic tree. The tree $\mathcal{F}$ parameter space is divided into a discrete topology space (i.e. the bifurcating tree structure without the edge lengths) and a continuous edge

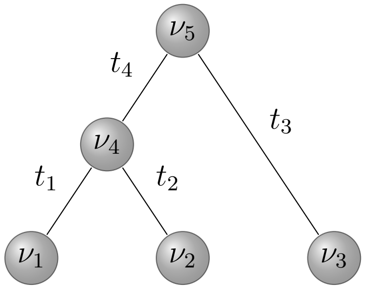


Figure 2.1: Simple phylogeny with $N = 3$ degree-one tip nodes $\nu_1, \ldots, \nu_3$, $N-2 = 1$ degree-three internal node ν_4 and degree-two root node ν_5. The edge connecting each node ν_i to its parent $\nu_{\text{pa}(i)}$ has length t_i. The phylogeny is a directed acyclic graph. It is directed in that there is a parent/child relationship between all nodes connected by an edge, and it is acyclic in that there are no cycles or loops in the graph. Each node has exactly one parent (except for the root which has none).

length space. The edge lengths inhabit a (non-negative) continuous $(2N-2)-$dimensional space, $(t_1, \ldots, t_{2N-2}) \in \mathbb{R}_{\geq 0}^{2N-2} = \{(x_1, \ldots, x_{2N-2}) : x_i \geq 0\}$. The space of tree topologies is unordered, discrete, and grows combinatorially in the number of tips, with $(2N-3)!! = \prod_{i=1}^{N-1} 2i - 1$ possible tree topologies for N tips.

There are many ways to specify the likelihood $p(\mathbf{S} \,|\, \mathcal{F})$ that are beyond the scope of this review (see Felsenstein (2004); Sullivan and Joyce (2005); Lemey et al. (2009) for overviews). However, it is useful to sketch a common form of these likelihoods. Let us assume that we have DNA characters, comprising the nucleotides A, C, G and T (the building blocks of DNA). We make the standard assumption that the molecular sequences $\mathbf{S}$ are aligned into an $N \times M$ matrix, where M is the number of nucleotides in a sequence alignment. Each column, called a site, in this alignment represents a homology assumption, in that all characters in a column share a single common ancestor somewhere back in time. We also commonly assume that each site evolves independently and identically (with the other sites) along the tree according to a four-state continuous-time Markov process with the instantaneous rate matrix $\mathbf{Q}$. Let s_i^m be the nucleotide at site m for node ν_i. The transition probability of observing s_i^m given the parent nucleotide state $s_{\text{pa}(i)}^m$ and edge length t_i is $p_{s_i^m s_{\text{pa}(i)}^m}$, such that $\mathbf{P} = \{p_{\ell m}\} = \exp(t_i \mathbf{Q})$ forms the transition probability matrix.

The clear challenge to computing likelihoods under this model is that we have not observed any sequence data associated with the internal nodes $\nu_{N+1}, \ldots, \nu_{2N-2}$ or the root node ν_{2N-1} and so must marginalize over their values. Assuming independence between sites and a prior $p\big(s^m_{2N-1}\big)$ on the root, the likelihood can then be expressed as

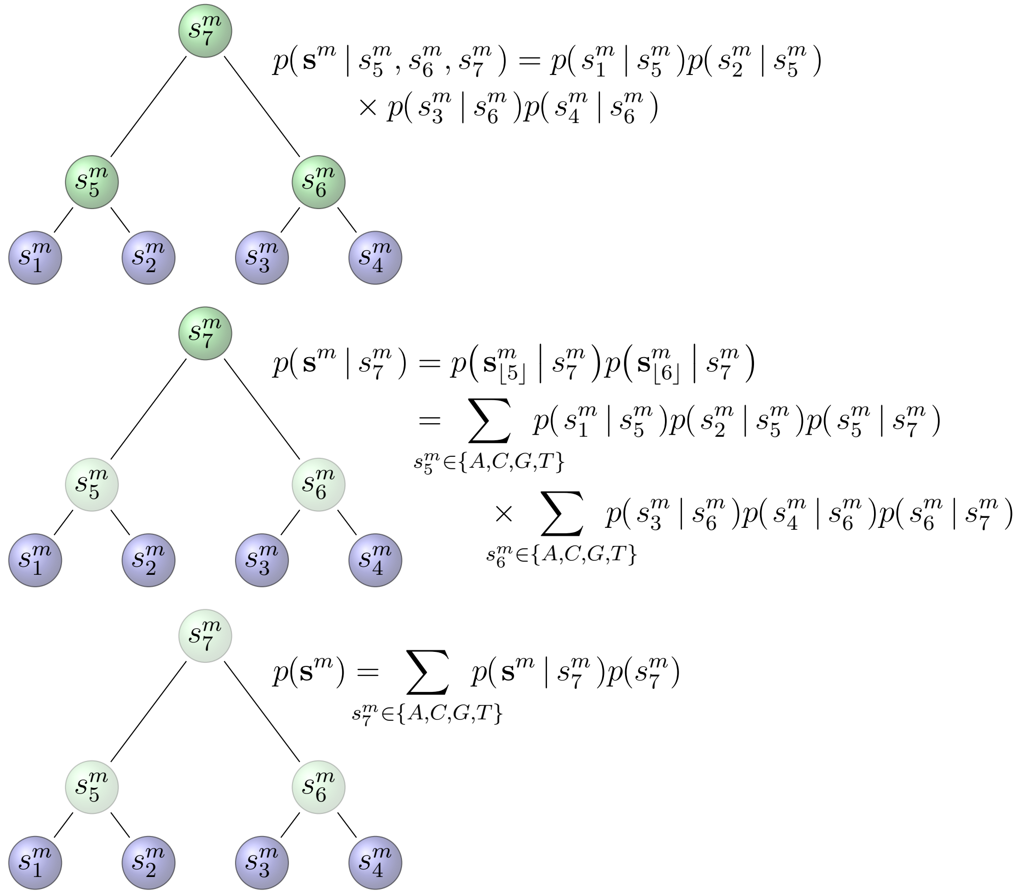
$$p(\mathbf{S} \,|\, \mathcal{F}) = \prod_{m=1}^{M} \sum_{s^m_{N+1} \in \{A,C,G,T\}} \cdots \sum_{s^m_{2N-1} \in \{A,C,G,T\}} p(s^m_{2N-1}) \prod_{i=1}^{2N-2} p\big(s^m_i \,\big|\, s^m_{\mathrm{pa}(i)}, t_i\big). \tag{2.1}$$

Naive computation of the above equation requires summing over 4^{N-1} unobserved states and is computationally intractable. Felsenstein's pruning algorithm (Felsenstein, 1973a, 1981), however, uses a post-order traversal of the tree to compute this likelihood in $\mathcal{O}(N)$ time, and all modern implementations of this likelihood calculation rely on that basic approach. The fundamental approach of this pruning algorithm is based on dynamic programming and has found repeated rediscovery in the message-passing algorithm (Pearl, 1982) and sum-product algorithm (Kschischang et al., 2001).

Let $\mathbf{s}^m$ be the nucleotides at site m associated with all tip nodes. The pruning algorithm relies on recursively computing the probability mass function $p\Big(\mathbf{s}^m_{\lfloor i \rfloor} \,\Big|\, s^m_i, \mathcal{F}_{\lfloor i \rfloor}\Big)$, where $\mathcal{F}_{\lfloor i \rfloor}$ is the sub-tree with root node ν_i, and $\mathbf{s}^m_{\lfloor i \rfloor}$ is the sub-vector of $\mathbf{s}^m$ restricted to the tips in $\mathcal{F}_{\lfloor i \rfloor}$. At the root node ν_{2N-1}, $\mathcal{F}_{\lfloor i \rfloor} = \mathcal{F}$ and $\mathbf{s}^m_{\lfloor 2N-1 \rfloor} = \mathbf{s}^m$, and the pruning algorithm computes $p\big(\mathbf{s}^m \,\big|\, s^m_{2N-1}, \mathcal{F}\big) = p\Big(\mathbf{s}^m_{\lfloor 2N-1 \rfloor} \,\Big|\, s^m_{2N-1,}, \mathcal{F}_{\lfloor 2N-1 \rfloor}\Big)$ via the following recursive relationship:

$$\begin{aligned} p\big(\mathbf{s}^m_{\lfloor i \rfloor} \,\big|\, s^m_i, \mathcal{F}_{\lfloor i \rfloor}\big) &= p\big(\mathbf{s}^m_{\lfloor j \rfloor} \,\big|\, s^m_i, \mathcal{F}_{\lfloor i \rfloor}\big) p\big(\mathbf{s}^m_{\lfloor k \rfloor} \,\big|\, s^m_i, \mathcal{F}_{\lfloor i \rfloor}\big) \\ &= \sum_{s^m_j \in \{A,C,G,T\}} p\big(\mathbf{s}^m_{\lfloor j \rfloor} \,\big|\, s^m_j, \mathcal{F}_{\lfloor j \rfloor}\big) p\big(s^m_j \,\big|\, s^m_i, t_j\big) \\ &\quad \times \sum_{s^m_k \in \{A,C,G,T\}} p\big(\mathbf{s}^m_{\lfloor k \rfloor} \,\big|\, s^m_k, \mathcal{F}_{\lfloor k \rfloor}\big) p\big(s^m_k \,\big|\, s^m_i, t_k\big), \end{aligned} \tag{2.2}$$

where nodes ν_j and ν_k are the children of node ν_i. When the recursion reaches tip nodes $i = 1, \ldots, N$, $p\Big(\mathbf{s}^m_{\lfloor i \rfloor} \,\Big|\, s^m_i, \mathcal{F}_{\lfloor i \rfloor}\Big) = 1_{\left\{\mathbf{s}^m_{\lfloor i \rfloor} = s^m_i\right\}}$, and the actual computations of computing



s_7^m
$p(\mathbf{s}^m \mid s_5^m, s_6^m, s_7^m) = p(s_1^m \mid s_5^m)p(s_2^m \mid s_5^m) \times p(s_3^m \mid s_6^m)p(s_4^m \mid s_6^m)$
s_5^m
s_6^m
s_1^m
s_2^m
s_3^m
s_4^m
s_7^m
$p(\mathbf{s}^m \mid s_7^m) = p\big(\mathbf{s}^m_{\lfloor 5 \rfloor} \mid s_7^m\big)p\big(\mathbf{s}^m_{\lfloor 6 \rfloor} \mid s_7^m\big) = \sum_{s_5^m \in \{A,C,G,T\}} p(s_1^m \mid s_5^m)p(s_2^m \mid s_5^m)p(s_5^m \mid s_7^m) \times \sum_{s_6^m \in \{A,C,G,T\}} p(s_3^m \mid s_6^m)p(s_4^m \mid s_6^m)p(s_6^m \mid s_7^m)$
s_5^m
s_6^m
s_1^m
s_2^m
s_3^m
s_4^m
s_7^m
$p(\mathbf{s}^m) = \sum_{s_7^m \in \{A,C,G,T\}} p(\mathbf{s}^m \mid s_7^m)p(s_7^m)$
s_5^m
s_6^m
s_1^m
s_2^m
s_3^m
s_4^m

Figure 2.2: Example of how Felsenstein's pruning algorithm marginalizes over the ancestral sequences. Tip nodes in blue represent observed sequence data, while green internal nodes represent latent ancestral sequences. Pale nodes have been marginalized. We do not explicitly condition on the tree $\mathcal{F}$ for notational simplicity.

the likelihood are performed via a post-order traversal of the tree (i.e. tips to root). The algorithm marginalizing over the root sequences

$$p(\mathbf{s}^m \mid \mathcal{F}) = \sum_{s^m_{2N-1} \in \{A,C,G,T\}} p\big(\mathbf{s}^m \mid s^m_{2N-1}, \mathcal{F}\big)p\big(s^m_{2N-1}\big) \tag{2.3}$$

and calculating $p(\mathbf{S} \mid \mathcal{F}) = \prod_{m=1}^{M} p(\mathbf{s}^m \mid \mathcal{F})$ is shown in Figure 2.2 on a simple example.

2.1.2 Why Bayesian?

In Bayesian phylogenetic inference, a common goal is to compute the posterior distribution of the phylogenetic tree given our sequence data,

$$p(\mathcal{F} \mid \mathbf{S}) \propto p(\mathbf{S} \mid \mathcal{F})p(\mathcal{F}). \tag{2.4}$$

The tree prior $p(\mathcal{F})$ typically falls into one of two biologically-motivated families. Coalescent models (Kingman, 1982; Strimmer and Pybus, 2001; Minin et al., 2008; Müller et al., 2017; Faulkner et al., 2020) are based on population genetic abstractions of sampling a (relatively) small number of sequences from a large population. Birth-death models (Thompson et al., 1975; Nee et al., 1994; Stadler, 2010; Höhna et al., 2019; Barido-Sottani et al., 2020; MacPherson et al., 2022) provide a forward-in-time model for the origination and termination of entire lineages. Bayesian approaches offer several advantages which we discuss below.

2.1.2.1 Quantifying uncertainty

Bayesian phylogenetics grew largely from the need to quantify and accommodate uncertainty in the phylogenetic tree (Sinsheimer et al., 1996; Rannala and Yang, 1996). Measuring uncertainty in the phylogenetic tree is a fundamentally challenging problem as the primary parameter of interest is often the tree topology: a high-dimensional, unordered, tip-labeled discrete parameter. Typical uncertainty estimates focus on estimating the statistical support for a specific monophyletic clade (i.e. a group of taxa comprising all the descendants of a given ancestor). Prior to the advent of Bayesian phylogenetic inference, phylogenetic uncertainty had been addressed with non-parametric bootstrapping (Felsenstein, 1985a) with much confusion as to interpretation of the bootstrap p-value (see Hillis and Bull, 1993; Felsenstein and Kishino, 1993; Efron et al., 1996; Berry and Gascuel, 1996). Bayesian posterior probabilities provided both an intuitive and statistically coherent method of addressing this uncertainty (Alfaro et al., 2003).

2.1.2.2 Time-resolved trees

Early phylogenetic models focused on the case where branch lengths are measured in genetic distances and thus unconstrained by time. However, Bayesian approaches can naturally accommodate the time-constrained case in a hierarchical model. As the bulk of the review assumes such models, we briefly consider the structure of a time-calibrated phylogenetic model. First, a tree arises from the tree prior $p(\mathcal{F})$. The branch lengths $t_1, \dots, t_{2N-2}$ of $\mathcal{F}$ are in calendar time. For each branch is a branch rate θ_i, such that the probability of changes along the branch is given by $\exp(t_i \theta_i \mathbf{Q})$. The prior on all branch rates $p(\theta_1, \dots, \theta_{2N-2})$ is known as the (molecular) clock model (Zuckerkandl and Pauling, 1962). Clock models typically either assume all branch rates are independent and identically distributed (Drummond et al., 2006) or that rates themselves evolve along the tree according to a correlated process (Thorne et al., 1998; Drummond and Suchard, 2010).

2.1.2.3 Tree as nuisance parameter

Phylogenetic methods offer opportunities to do more than just reconstruct the evolutionary history of a group of organisms. The branching patterns in trees themselves can be informative about patterns and processes governing biodiversity, such as mass extinctions (Stadler, 2011; May et al., 2016), or the rate of spread of infectious diseases (Stadler et al., 2012, 2013). When combined with other information, such as the locality of samples or evolutionary traits, phylogenetic models provide a powerful framework for studying the spatiotemporal spread of both species and diseases, as well as the evolution of important traits (see Section 2.3). In many such cases, the tree itself is a nuisance parameter. Bayesian inference via Markov chain Monte Carlo (MCMC) provides a natural approach to numerically marginalize over the phylogenetic tree and study processes that condition on the tree independent of any single fixed tree's influence (Huelsenbeck et al., 2000, 2001; Suchard et al., 2001).

2.2 Modern phylogenetics: big trees and complex models

Early practitioners of Bayesian phylogenetics naturally used MCMC to sample from the posterior distribution of phylogenetic trees. Since it is relatively straightforward to marginalize over continuous nuisance parameters (e.g. the molecular substitution rate matrix $\mathbf{Q}$), attention quickly turned to improving the efficiency with which the Markov chain explores tree space (Yang and Rannala, 1997; Larget and Simon, 1999; Mau et al., 1999; Li et al., 2000; Huelsenbeck and Ronquist, 2001). This in turn gave rise to the observation that navigating tree space is hard (Lakner et al., 2008; Höhna and Drummond, 2012; Whidden and Matsen IV, 2015; Harrington et al., 2021).

We explore several solutions to this problem below. In Section 2.2.1, we discuss approaches to improving the efficiency of MCMC-based methods. We then discuss in Section 2.2.2 alternatives to MCMC inspired by phylogenetic problems. As these approaches permit researchers to more efficiently explore the space of phylogenetic trees, we revisit in Section 2.2.3 the problem of assessing uncertainty in the phylogeny estimates.

2.2.1 MCMC-based approaches

MCMC is the workhorse of Bayesian phylogenetic inference. The efficiency of MCMC depends on two factors: the auto-correlation between parameter proposals and the speed at which proposals are made and evaluated. Researchers have relied on and contributed to numerous innovative computational and statistical methods in search of MCMC approaches that efficiently explore the high-dimensional tree space.

2.2.1.1 Faster likelihood calculations

In the absence of known conjugate priors, efficient likelihood calculations are critical for efficient MCMC. As common models of sequence evolution assume conditional independence between different sites in the genome, parallelization is a natural approach toward fast computation. The BEAGLE (Suchard and Rambaut, 2009; Ayres et al., 2012, 2019) and PLL (Izquierdo-Carrasco et al., 2013; Flouri et al., 2015) libraries leverage the computational power of multi-core processors, including graphics processing units (GPUs) in the former case, to massively parallelize likelihood calculations and accelerate computation. These libraries also cache calculations on sub-trees such that unnecessary calculations are not repeated when, for example, a branch length on one part of the tree is updated that does not influence the partial likelihood of other parts of the tree.

2.2.1.2 Sampling from high-dimensional posterior distributions

The dimensionality of many continuous parameters (e.g. the branch lengths) scales with the size of the phylogenetic tree. Phylogenetic analyses commonly partition genetic sequences into different genes (or some other genetic unit) that evolve independently conditional on a tree. Modern Bayesian phylogenetic analyses include trees with thousands of tips (e.g. Lemey et al., 2021) and, as such, require inference of the joint posterior of thousands of highly-correlated parameters.

Baele et al. (2017) develop an adaptive Metropolis (AM) algorithm (Haario et al., 2001) that leverages the parallel computing to take advantage of the conditional independence of the genetic partitions. The AM algorithm is a modification of MCMC where proposal distributions are informed by the empirical posterior distribution up to that point in the chain. While AM is non-Markovian, it remains ergodic under weak assumptions (Roberts and Rosenthal, 2009). Baele et al. (2017) update the chain via partition-specific multivariate Gaussian proposals with covariance influenced by the empirical posterior covariance of relevant parameters. The conditionally independent parameter blocks allow parallel likelihood computations, and the multivariate Gaussian proposals informed by the posterior have higher acceptance probability than naive multivariate proposals.

Hamiltonian Monte Carlo (HMC) is now a standard tool across Bayesian statistics for sampling from high-dimensional posterior distributions. At its core, HMC also uses information about the posterior to generate high-dimensional parameter proposals with high acceptance probability. As the aforementioned information originates from the gradient of the log-posterior with respect to the parameters of interest, efficient gradient calculations are essential for efficient HMC. Ji et al. (2020) develop an $\mathcal{O}(N)$ algorithm for computing the gradient of the log-posterior with respect to all branch lengths simultaneously. These gradient calculations are also parallelizable using existing libraries (see Section 2.2.1.1) and result in an order of magnitude increase in computational efficiency.

2.2.1.3 Navigating tree space

The discrete tree topology with $(2N-3)!!$ possible states is often the most difficult model parameter to efficiently sample. As many other parameters, including the branch lengths and latent data associated with internal nodes, are only identifiable in the context of a particular tree, MCMC proposals that make large changes to the tree topology frequently have very low acceptance probability.

HMC is a standard tool for sampling from high-dimensional, highly correlated, continuous parameter spaces, but the discrete, combinatorial nature of the tree topology does not permit traditional HMC approaches. Dinh et al. (2017) develop probabilistic path HMC (PPHMC) to sample from spaces that form an orthant complex. Essentially, they sample the branch lengths via HMC in a way that branch lengths may approach 0. When HMC causes a branch length to cross 0, PPHMC randomly selects from one of the three equivalent topologies resulting from the zero branch length. To reduce error from the leapfrog approximation crossing non-differentiable orthant boundaries, they introduce a smoothing function at these boundaries, which dramatically increases the accuracy of the approximation of the Hamiltonian trajectory and Metropolis-Hastings acceptance probability. Similar work outside of the phylogenetic context includes that of Pakman and Paninski (2013); Mohasel Afshar and Domke (2015) and Nishimura et al. (2020).

More recently, Meyer (2021) has developed a series of AM procedures for efficiently navigating the space of unrooted tree topologies. Like other AM algorithms, these approaches rely on statistics of the posterior sample up to a point in a chain to inform future parameter proposals. In the context of tree topologies, the relevant statistics rely on the fact that each branch splits the taxa into two groups. The Meyer (2021) approach relies on the posterior frequency of these splits for each possible group of taxa, with topology proposals more likely to disrupt low-frequency splits than high-frequency splits. Similarly, Zhang et al. (2020) use parsimony (i.e. the minimum number of genetic changes necessary to account for the observed genetic diversity) to inform tree proposals, with highly parsimonious (i.e. few changes) proposals more likely than less parsimonious ones.

2.2.2 Beyond MCMC

2.2.2.1 Sequential Monte Carlo

Teh et al. (2007) propose sequential Monte Carlo (SMC) for inferring tree-structured models. Due to the hierarchical structure of the model, the intermediate distributions are defined over forests (i.e. groups of sub-trees) over the observed sequences, and hence the dimension of the target distributions increases over each iteration. Based on this idea, Bouchard-Côté et al. (2012) propose an efficient framework, based on partially ordered set structures, which imposes restrictions on proposal distributions so that the final iteration results in valid phylogenetic trees. Since this phylogenetic SMC is restricted to jointly estimate tree topology and branch length distributions, Wang et al. (2015) propose particle MCMC which combines a combinatorial SMC within an MCMC in order to jointly approximate other continuous parameters such as the parameters of the substitution rate matrix $\mathbf{Q}$. Borrowing ideas from annealed importance sampling, Wang et al. (2020) put forward an annealed SMC algorithm to approximate the full phylogenetic model and, as other SMC-based methods, enable the computation of the marginal likelihood.

SMC has also been investigated in an online setting in which a posterior sample of trees is already available from a previous analysis (e.g. MCMC or SMC) and one wishes to directly update the posterior approximation with additional sequences. Dinh et al. (2017) show consistency of online SMCs in terms of weak convergence while Fourment et al. (2018) develop sophisticated proposals that better match the proposal density to the posterior.

2.2.2.2 Variational inference

Until recently, variational inference (VI) has received limited attention in the field of phylogenetics, perhaps due to 1) the absence of conjugate prior distributions in the nearly all phylogenetic models and 2) the difficulty of analytically calculating the gradient of complex joint distributions. Dang and Kishino (2019) develop a computationally efficient VI-based method to approximate a model which allows different equilibrium frequencies across sequence sites. Since the likelihood of this model is in the exponential family, most of the expectations required for optimization are obtained in closed form. This method is restricted to unrooted trees and the authors used closed-form coordinate ascent and stochastic VI algorithms for solving the optimization problem. Fourment et al. (2020) use VI to approximate the marginal likelihood of fixed unrooted topologies using stochastic gradient ascent with analytical derivatives. Using the Stan language (Carpenter et al., 2017) and its automatic differentiation library, Fourment and Darling (2019) propose a framework for approximating complex models, including time-calibrated phylogenies with tree priors (e.g. coalescent models), molecular clock, and discrete phylogeography models.

The methods described so far only approximate continuous parameters of a fixed topology and therefore evade the combinatorial problem of the discrete topology space. The first approach developed to tackle this problem was introduced by Zhang and Matsen IV (2018a) using a general Bayesian network formulation for tree probability estimation. Given a set of topologies, this structure provides an accurate and rich distribution over the topology space. Subsequently, the same authors (Zhang and Matsen IV, 2018b) build on the Bayesian network idea and propose jointly approximating the tree structure and the branch length distributions. This method also necessitates a set of topologies to define the structure of the Bayesian network, however dynamic construction of the network is an active area of research. Moretti et al. (2021) propose a hybrid method using VI and combinatorial SMC to approximate posteriors defined on the space of phylogenetic trees. The main advantage of this method is that it does not require precomputing a set of topologies. With the exception of the Stan-based method which allows approximating a posterior using a multivariate normal distribution, every method described so far uses meanfield approximation thereby ignoring correlation between parameters. Since parameters in phylogenetic models tend to be highly correlated, Zhang (2020) proposes to use normalizing flows to improve the expressiveness of the approximate distribution.

Recently, Ki and Terhorst (2022) synthesized this VI-based work with phylodynamic methods to fit a complex epidemiological model with thousands of sequences. The authors showed that their method was order of magnitude faster than an MCMC-based approaches and was able recover acceptable parameter estimates.

2.2.3 Uncertainty in tree space revisited

As discussed in Section 2.1.2.1, Bayesian phylogenetic methods conveniently quantify uncertainty in the tree. Many evolutionary questions can be phrased as "is there a subtree in the phylogeny which contains all of (some set of) sequences and no other sequences?" With MCMC samples in hand, we can easily obtain this probability by counting MCMC samples with the subtree. The fact that this estimate can carry substantial Monte Carlo error is often ignored. For continuous random variables, Monte Carlo error is typically addressed using the effective sample size (ESS, i.e. the number of independent samples which would yield the same standard error of the mean). Trees, however, are more complex objects.

Gaya et al. (2011) introduce one approach that focuses on taxa splits (i.e. bi-partitions of the tips by cutting the tree at a given edge). The tree is reduced to a series of indicator variables denoting whether a given split is present or absent in each tree. Uncertainty in the probability of specific splits can then be expressed via the ESS of these indicators. Fabreti and Höhna (2021) observe, however, that this approach has difficulty with splits whose probabilities approach 0 or 1. They also note that the Gaya et al. (2011) ESS incorrectly assumes that splits are independent. Regardless, Fabreti and Höhna (2021) find evidence via simulation that the Gaya et al. (2011) approach may remain robust.

Lanfear et al. (2016) propose an ESS for the phylogeny itself. They suggest two approaches based on distances between trees. One such approach is the pseudo-ESS, where for each posterior tree sample the distance is computed to all other tree samples. The overall tree ESS is taken to be the median of the ESSs of these distance metrics. Lanfear et al. (2016), however, do not establish any link between this pseudo-ESS and Monte Carlo error.

Magee et al. (2021) develop several additional approaches for computing the ESS of a phylogeny. One such approach employs Fréchet generalizations of covariance such that the generalized auto-correlation ρ_t between trees can be computed and the following standard identity can be applied: $\text{ESS} = n/(\sum_{t=-\infty}^{\infty} \rho_t)$. Additionally, Magee et al. (2021) propose a simulation-based approach to test whether a putative tree ESS is useful for quantifying Monte Carlo error in the tree. They find that most tested tree ESS measures can capture Monte Carlo error in the probabilities of splits, as well as other important summaries of the posterior distribution. The tree ESS approaches additionally do not appear to suffer from the difficulties Fabreti and Höhna (2021) identified with low and high probability splits.

2.3 Data integration

In many cases the phylogenetic tree is actually a nuisance parameter and not of scientific interest itself (see Section 2.1.2.3). Rather, there is some other process (e.g. rate of viral transmission between two locations, strength of natural selection) that is separate from yet dependent on the evolutionary history that researchers would like to explore. In these cases, researchers frequently seek to integrate varying sources of data into a single, coherent statistical model of evolution. These additional sources of data frequently include time (see Section 2.1.2.2) and geographic location (Lemey et al., 2009, 2010).

Before discussing specific statistical models for integrating varying types of data, we first introduce a general framework in which to orient these models in Section 2.3.1. We then examine models and inference methods associated with integrating both discrete and continuous data into phylogenetic models in Sections 2.3.2 and 2.3.3, respectively. While we briefly discuss applications in the sections below, Baele et al. (2017) offer a more thorough overview of the different kinds of data integrated into these phylogenetic models.

2.3.1 A unified modeling framework

There are myriad statistical models for integrating additional data into these phylogenetic models. While each model is naturally tailored to a specific application, most share a common, general framework (see Section 2.3.4.3 for a notable exception). Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})^t$ be a vector of latent traits associated with node ν_i for $i = 1, \ldots, 2N-1$. Similarly, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iP})$ be the data associated with tip nodes $\nu_1, \ldots, \nu_N$. For tips $i = 1, \ldots, N$, we posit a possibly stochastic link function $\mathbf{y}_i = f(\mathbf{x}_i)$.

These models describe a data generative process where the distribution of each $\mathbf{x}_i$ conditional on the trait values of its parent $\mathbf{x}_{\text{pa}(i)}$ are distributed with density or mass function $p\big(\mathbf{x}_i \,\big|\, \mathbf{x}_{\text{pa}(i)}\big) = g\big(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, \boldsymbol{\theta}_i, \boldsymbol{\Theta}\big)$, where $\boldsymbol{\theta}_i$ represents branch-specific parameters, and $\boldsymbol{\Theta}$ represent universal model parameters. Typically, $\boldsymbol{\theta}_i$ includes at the very minimum the branch length t_i. By placing a prior on the root $p(\mathbf{x}_{2N-1} \,|\, \boldsymbol{\theta}_{2N-1})$, we can define a likelihood over the data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^t$:

$$p(\mathbf{Y} \,|\, f, g, \mathcal{F}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{2N-1}, \boldsymbol{\Theta}). \tag{2.5}$$

See Figure 2.3 for a model schematic.

While this framework seems (and indeed is) incredibly generalizable, all models resulting from it share a critical property: once lineages diverge they evolve independently. To formalize this notion, assume two nodes ν_i and ν_j that share a common parent $\nu_{\text{pa}(i)} = \nu_{\text{pa}(j)} = \nu_k$. Let $\mathbf{Y}_{\lfloor i \rfloor}$ and $\mathbf{Y}_{\lfloor j \rfloor}$ be the data associated with all tip nodes descended from node ν_i and ν_j, respectively. By construction, $\mathbf{Y}_{\lfloor i \rfloor} \,\big|\, \mathbf{x}_k$ and $\mathbf{Y}_{\lfloor j \rfloor} \,\big|\, \mathbf{x}_k$ are independent. This conditional independence is a defining feature of these phylogenetic models that statisticians routinely exploit to increase computational efficiency of statistical inference.

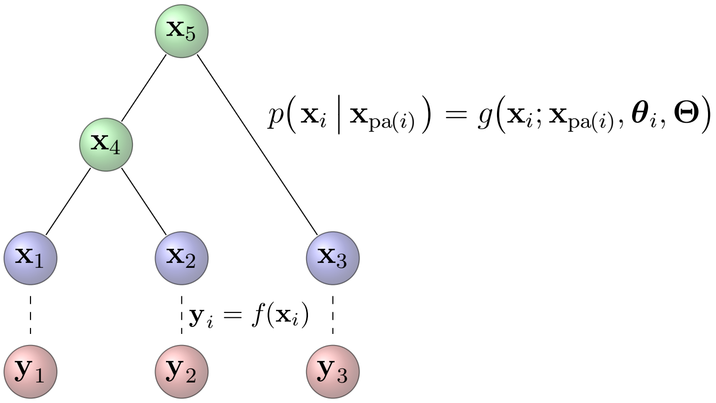

$\mathbf{x}_5$
$p\big(\mathbf{x}_i \,\big|\, \mathbf{x}_{\text{pa}(i)}\big) = g\big(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, \boldsymbol{\theta}_i, \boldsymbol{\Theta}\big)$
$\mathbf{x}_4$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{y}_i = f(\mathbf{x}_i)$
$\mathbf{y}_1$
$\mathbf{y}_2$
$\mathbf{y}_3$

Figure 2.3: Schematic of a generalized phylogenetic model. The data $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (red nodes) are assumed to have arisen from the latent traits $\mathbf{x}_1, \ldots, \mathbf{x}_N$ (blue nodes) at the respective tips via the possibly stochastic link function $f(.)$. The latent tip traits $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and latent internal traits $\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{2N-2}$ arise from some evolutionary process on the phylogenetic tree where the traits of each child node $\mathbf{x}_i$ are drawn from a distribution with density $p\big(\mathbf{x}_i \,\big|\, \mathbf{x}_{\text{pa}(i)}\big) = g\big(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, \boldsymbol{\theta}_i, \boldsymbol{\Theta}\big)$.

Readers may note that the model of molecular sequence evolution described in Section 2.1.1 fits neatly within this more general framework. Specifically, the data $\mathbf{Y}$ are comprised of discrete nucleotides (e.g. $y_{ij} \in \{A, C, G, T\}$), the link function $f(\mathbf{x}_i) = \mathbf{x}_i$, and the probability mass function $g\big(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, t_i, \mathbf{Q}\big) = \prod_{m=1}^{M} \exp(t_i \mathbf{Q})_{x_{\text{pa}(i)m} x_{im}}$.

As noted above, Bayesian methods (specifically MCMC) offer a to-date unmatched ability to study evolutionary processes without conditioning on an particular evolutionary history. This follows simply from the fact that researchers can easily sample from the marginal density of a parameter of interest from a realized MCMC simulation. Let $\boldsymbol{\Phi}$ represent all parameters associated with nucleotide evolution (e.g. the substitution rate matrix $\mathbf{Q}$) and let

$\mathbf{\Psi} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{2N-1}, \mathbf{\Theta}\}$ be the parameters associated with some separate trait-evolutionary process. One can then sample from the posterior

$$p(\mathcal{F}, \mathbf{\Phi}, \mathbf{\Psi} \,|\, \mathbf{S}, \mathbf{Y}\,) \propto p(\,\mathbf{S} \,|\, \mathcal{F}, \mathbf{\Phi}\,)p(\,\mathbf{Y} \,|\, \mathcal{F}, \mathbf{\Psi}\,)p(\mathcal{F})p(\mathbf{\Phi})p(\mathbf{\Psi}) \tag{2.6}$$

via a Metropolis-within-Gibbs approach (Gelfand, 2000) where one iteratively samples from $p(\,\mathbf{\Phi} \,|\, \mathcal{F}, \mathbf{S}\,)$, $p(\,\mathbf{\Psi} \,|\, \mathcal{F}, \mathbf{Y}\,)$, and $p(\,\mathcal{F} \,|\, \mathbf{S}, \mathbf{Y}, \mathbf{\Phi}, \mathbf{\Psi}\,)$. This compartmentalization of the inference procedure means that methods for sampling from the nucleotide substitution parameters $\mathbf{\Phi}$ are not influenced by the trait-evolutionary model and vice versa. The sections below focus on the conditional posterior $p(\,\mathbf{\Psi} \,|\, \mathcal{F}, \mathbf{Y}\,)$.

2.3.2 Discrete character integration

Many processes of interest can be modeled as the evolution of discrete traits on the tree (Ronquist, 2004). Perhaps the most common discrete outcome of interest is location in phylogeographic models (Sanmartín et al., 2008; Comas et al., 2013; Lemey et al., 2020). However, other discrete characters of interest include pathogen host species (Ward et al., 2014; Dearlove et al., 2016; Latinne et al., 2020) and ecological habitat (Bryja et al., 2014; Terra-Araujo et al., 2015; Sánchez-Baracaldo et al., 2017). See Baele et al. (2017), Table 1 for a more thorough list of discrete-trait analyses.

The most common model of discrete-character evolution is essentially the same as the continuous-time Markov model of nucleotide evolution introduced in Section 2.1.1. The states can be arbitrarily defined to be whatever discrete character is evolving along the tree.

2.3.2.1 Developments in Markov jump processes

Problems of both genetic sequence and discrete trait evolution have motivated much work on Bayesian networks, hidden Markov models, endpoint-conditioned Markov jump processes and Markov reward processes to infer the number of times specific trait changes occur or the length of time a trait is realized along an evolution history. Siepel et al. (2006), for example, analytically derive the probability mass function of the total number of Markov jumps in an endpoint-conditioned continuous-time Markov chain along a graph with arbitrary rate matrix. Similarly, Minin and Suchard (2008a,b) analytically calculate the moments of the number of jumps between each pair of states. Sometimes, expectations are insufficient and simulation is required to answer the question of interest. Hobolth and Stone (2009) provide several approaches for simulating endpoint-conditioned continuous-time Markov chains. Minin and Suchard (2008a) and Hobolth and Jensen (2011) develop computationally efficient, simulation-free methods for calculating the moments of Markov reward processes (e.g. the average amount of time spent in a particular state of a continuous-time Markov chain).

Phylogenetics has also motivated the development of statistical theory related to Lie Markov models (Sumner et al., 2012; Fernández-Sánchez et al., 2015). These models comprise inhomogeneous continuous-time Markov processes whose endpoint can be expressed as the result of a time-homogeneous process (essentially the time-resolved average of the inhomogeneous process). These processes permit the instantaneous rate matrix to vary over time (and along different branches in a phylogeny) and are useful for identifying the root position of a phylogeny without specifying a molecular clock (Hannaford et al., 2020).

2.3.2.2 Evolutionary covariates and the curse of dimensionality

Phylogenetic models are certainly not immune from the curse of dimensionality. This phenomenon is particularly acute in phylogeographic models where the number of discrete locations can be quite large. Assuming a continuous-time Markov process along the phylogeny with L discrete states and infinitesimal rate matrix $\mathbf{Q} = \{q_{\ell m}\}$, the number of free parameters in $\mathbf{Q}$ scales $\mathcal{O}(L^2)$. While there is no theoretical prohibition on inferring more parameters than there are observations, it becomes increasingly difficult to extract meaningful information in these settings.

This challenge is also an opportunity, as one can reduce the size of the parameter space by assuming the $\mathcal{O}(L^2)$ transition rates are functions of some low-dimensional process parameterized by scientifically relevant covariates. Lemey et al. (2014) and Zhao et al. (2016) develop a generalized linear model (GLM) that assumes the log-transition rates are a linear function of relevant covariates (e.g. pairwise air traffic between two locations, local temperature) with the number of parameters scaling linearly with the number of covariates. To further penalize over-parameterization within the GLM, Lemey et al. (2014) also assume *a priori* that some unspecified number of covariates have no influence on the transition rates as follows. Let $\mathbf{Z} = \{z_{\ell m,i}\}$ be the covariate observations associated with all ordered pairs $\ell, m \in \{1, \ldots, L\}^2, \ell \neq m$ and covariates $i = 1, \ldots, R$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_R)^t$ be a vector of regression coefficients and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_R)^t$ be a vector of indicator variables such that $\log q_{\ell m} = \sum_{i=1}^{R} \delta_i \beta_i z_{\ell m,i}$. Inference of the indicators i can be achieved via Bayesian stochastic search variable selection (Kuo and Mallick, 1998; Chipman et al., 2001). To sample efficiently from a posterior with high correlation between regression coefficients $\boldsymbol{\beta}$, Lemey et al. (2014) rely on a Markov chain transition kernel that draws the proposal $\boldsymbol{\beta}^* \sim \mathcal{N}(\boldsymbol{\beta}, \alpha \mathbf{Z}^t \mathbf{Z})$, where α is a tunable scaling factor. This kernel accounts for the prior expectation that coefficients associated with correlated covariates will also be correlated. Zhao et al. (2016), as an alternative, develop an HMC sampler for the regression coefficients. These GLM approaches are applicable beyond phylogenetics and facilitate inference of the rate matrix of any discrete-state continuous-time Markov process.

2.3.2.3 Piece-wise deterministic, non-reversible Markov processes

Bouchard-Côté et al. (2018) introduce the bouncy particle sampler (BPS) as a non-reversible, rejection-free alternative to reversible Metropolis-Hastings and HMC samplers. While they evaluate the BPS as a way to efficiently sample from the phylogenetic rate matrix $\mathbf{Q}$, it has broad utility beyond statistical phylogenetics. Inspired by the physics literature (Peters and de With, 2012), the BPS relies on piece-wise linear trajectories of a particle (the parameters) through a potential field (the negative log-posterior). Bouchard-Côté et al. (2018) generalize this sampler and develop methods to exactly simulate the parameter trajectories. The BPS relies on finding the parameter value along a line that maximizes the posterior density. Bouchard-Côté et al. (2018) use gradient calculations from the HMC sampler of Zhao et al. (2016) to identify these maxima and sample efficiently from a high-dimensional evolutionary rate matrix. See Section 2.3.3.2 for additional applications of piece-wise deterministic, non-reversible Markov processes.

2.3.3 Gaussian processes on a tree

While discrete-trait models discussed above are typically based on the same model of molecular sequences introduced in Section 2.1.1, continuous data integration requires new statistical models. Due to their computational tractability, Gaussian processes form the backbone of most continuous trait analyses. The simplest such model is one where correlated traits evolve according to a P-dimensional multivariate Brownian diffusion (MBD) process (Edwards and Cavalli-Sforza, 1964; Felsenstein, 1985b). Using the notation of Section 2.3.1, we have

$$\mathbf{x}_i \,\big|\, \mathbf{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\big(\mathbf{x}_{\mathrm{pa}(i)}, t_i \boldsymbol{\Sigma}\big) \quad \text{and} \quad \mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{x}_i. \tag{2.7}$$

Marginalizing the latent traits (except the root traits $\mathbf{x}_{2N-1}$) results in the likelihood

$$\mathrm{vec}(\mathbf{Y}) \,|\, \mathcal{F}, \mathbf{x}_{2N-1}, \boldsymbol{\Sigma} \sim \mathcal{N}\big(\mathrm{vec}\big(\mathbf{1}_N \mathbf{x}_{2N-1}^t\big), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}\big), \tag{2.8}$$

where $\otimes$ is the Kronecker product and $\boldsymbol{\Psi}$ is a deterministic function of the phylogenetic tree $\mathcal{F}$ capturing the phylogenetically-induced covariance between taxa.

Likelihood-based inference frequently requires repeated evaluation of the likelihood function $p(\mathbf{Y} \,|\, \mathcal{F}, \mathbf{x}_{2N-1}, \boldsymbol{\Sigma})$, which naively scales $\mathcal{O}(N^3 P^3)$. Exploiting the Kronecker product to invert the variance reduces this complexity to $\mathcal{O}(N^3 + P^3)$. As both N and P can be large, even this greatly simplified calculation can be intractable. Freckleton (2012) (based on Felsenstein (1973b)), Pybus et al. (2012) and Ho and Ané (2014) develop strategies for computing this likelihood in $\mathcal{O}(NP^2 + P^3)$ using approaches conceptually similar to Felsenstein's pruning algorithm for computing the sequence-based likelihood (Felsenstein, 1973a). The Ho and Ané (2014) approach uses the tree structure to efficiently compute

$$\left(\mathbf{Y} - \mathbf{1}_N \mathbf{x}_{2N-1}^t\right)^t \boldsymbol{\Psi}^{-1} \left(\mathbf{Y} - \mathbf{1}_N \mathbf{x}_{2N-1}^t\right) \tag{2.9}$$

in $\mathcal{O}(NP^2)$ for any matrix $\boldsymbol{\Psi}$ that satisfies what they dub the 3-point structure. Specifically, any matrix $\boldsymbol{\Psi}$ has a 3-point structure if for all i, j, k the two smallest covariances of ψ_{ij}, ψ_{ik}, ψ_{jk} are equal to each other. Ho and Ané (2014) generalize this to allow negative covariances in $\boldsymbol{\Psi}$ under certain conditions. More recently, Bastide et al. (2021) develop an HMC-based approach that can calculate gradients for nearly all relevant parameters in these hierarchical Gaussian models in linear time.

2.3.3.1 Gaussian processes and Matrix-Normal likelihoods with missing data

Unfortunately, the previous methods for computing the likelihood fail with partially missing data. Cybis et al. (2015) address missing data within a tip in these hierarchical Gaussian process models via data augmentation. Let $\mathbf{y}_i^{\text{mis}}$ and $\mathbf{y}_i^{\text{obs}}$ be the missing and observed data, respectively, associated with tip node ν_i. Cybis et al. (2015) develop a procedure that can sample from $\mathbf{y}_i^{\text{mis}} \,\middle|\, \mathbf{Y}^{\text{obs}}, \mathcal{F}, \boldsymbol{\Sigma}$ for $i = 1, \ldots, N$. Each sample requires $\mathcal{O}(NP^2)$ computations for $\mathcal{O}(N^2P^2)$ complexity to sample from all N tips.

Bastide et al. (2018); Mitov et al. (2020) and Hassler et al. (2020, Chapter 3) develop an alternative approach that analytically integrates out missing observations rather than relying on data augmentation. This approach assumes that

$$\mathbf{y}_i \,|\, \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}_i, \mathbf{R}_i^t \begin{pmatrix} \infty\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}_i\right) \tag{2.10}$$

where $\mathbf{R}_i$ is a permutation matrix that arranges the ∞ values to correspond to the indices of $\mathbf{y}_i^{\text{mis}}$ and the 0 values to correspond to the indices of $\mathbf{y}_i^{\text{obs}}$. This specification of missingness gives rise to a series of non-standard operations involving square matrices with 0 or ∞ diagonal elements. For example, the special inverse of some arbitrary matrix

$$\left[\mathbf{R}_i^t \begin{pmatrix} \infty\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}_i\right]^{-} = \mathbf{R}_i^t \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \infty\mathbf{I} \end{pmatrix} \mathbf{R}_i. \tag{2.11}$$

Propagating missing information up the tree via singular precision matrices allows marginal likelihood calculations of the observed data only in $\mathcal{O}(NP^3)$.

This algorithm applies to a much broader range of statistical models than MBD on a tree and helps solve the longstanding statistical challenge of efficiently calculating multivariate normal likelihoods with missing data. Specifically, it applies to any multivariate normal likelihood with a 3-point structured covariance matrix discussed above (Ho and Ané, 2014). This structure is common in hierarchical Gaussian models. While Allen and Tibshirani (2010) and Glanz and Carvalho (2018) use the expectation-maximization algorithm to perform maximum likelihood imputation, the Bastide et al. (2018)/Mitov et al. (2020)/Hassler et al. (2020, Chapter 3) approach permits inference relying on only the observed-data likelihood. For situations where imputation is desired, this approach allows one to sample from the full conditional distribution of all missing observations simultaneously in $\mathcal{O}(NP^3)$ time as well.

2.3.3.2 Multivariate probit models and sampling from high-dimensional truncated Gaussian distributions

Bayesian phylogenetics has also served as the motivation for many novel methods in multivariate probit models. Cybis et al. (2015) develop a phylogenetically informed multivariate probit model with correlations between both traits and taxa. Under this model, the data are a mix of continuous and discrete traits. Underlying all traits is an MBD process on the tree. Here, the mapping $f(\mathbf{x}_i) = (f_1(x_{i1}), \dots, f_P(x_{iP}))^t$ between the continuous latent traits $\mathbf{x}_i$ and mixed continuous/discrete observed data $\mathbf{y}_i$ is not the simple identity function. For a binary trait j, we have $y_{ij} = f_j(x_{ij}) = 1_{\{x_{ij}>0\}}$ (see Cybis et al. (2015) for mappings to ordinal or categorical traits). For continuous traits k, the link function remains $f_k(x_{ij}) = x_{ij}$.

Let $\mathbf{x}_i^{\text{obs}}$ be the components of $\mathbf{x}_i$ associated with the continuous phenotypes and let $\mathbf{x}_i^{\text{lat}}$ be the latent components informing the discrete traits. Efficient inference under this model requires data augmentation of $\mathbf{x}_i^{\text{lat}}$ for $i = 1, \dots, N$. As mentioned in Section 2.3.3.1, this procedure relies on sampling from $\mathbf{x}_i^{\text{lat}} \,\big|\, \mathbf{y}_i, \mathbf{X}_{\setminus i}, \mathcal{F}, \boldsymbol{\Sigma}$ for $i = 1, \dots, N$, where $\mathbf{X}_{\setminus i} = \{\mathbf{x}_j; j \neq i\}$. This full conditional posterior is a (potentially high-dimensional) truncated Gaussian distribution due to the constraints in the stochastic link function. While Cybis et al. (2015) rely on a multiple-try rejection sampler, this sampler can be prohibitively slow for high-dimensional truncated Gaussian distributions. Zhang et al. (2021), however, employ a novel approach, the BPS (Bouchard-Côté et al., 2018, see Section 2.3.2.3), to more efficiently sample from this challenging distribution. As noted previously, the BPS requires calculating the gradient of the log-posterior density with respect to the latent parameters $\mathbf{x}_i^{\text{lat}}$ for $i = 1, \dots, N$, which Zhang et al. (2021) achieve in linear time with a post-order tree traversal similar to that employed by Pybus et al. (2012). This Zhang et al. (2021) sampler essentially bounces off the truncations of the full conditional posterior. As the truncations are defined on a univariate basis, evaluating when these boundary events occur is trivial, and Zhang et al. (2021) observe increases in computational efficiency over rejection sampling approaching two orders of magnitude.

Seeking improvement on the BPS, Zhang et al. (2022) develop a zigzag Hamiltonian Monte Carlo sampler (Nishimura et al., 2020, zigzag-HMC) to further address the challenge of sampling from a high-dimensional truncated Gaussian distribution in the phylogenetic context. Zigzag-HMC differs from traditional HMC as it posits a Laplace momentum which imparts the unusual property that the Hamiltonian trajectory may only have slopes in $\{\pm 1\}^d$ where d is the dimensionality of the parameter space (i.e. the element-wise slopes may be 1 or -1 only). As the velocity restricted to $\{\pm 1\}^d$ only depends on the sign of the momentum, the particle moves with a constant velocity until one momentum component changes its sign, at which point the particle updates its velocity and moves along a new linear trajectory. See Figure 2.4 for a simple example. For Gaussian distributions, one can analytically simulate the zigzag Hamiltonian dynamics by calculating when these sign changes occur, eliminating the need for an accept/reject step. Zigzag-HMC handles truncations in the same way as the BPS and it also takes advantage of the linear time log-posterior gradient evaluations. Besides being more efficient than BPS on a truncated Gaussian, zigzag-HMC also enables a joint update of latent parameters and the across-trait correlation, further improving the sampling efficiency. Importantly, this Zhang et al. (2022) method is able to learn the conditional dependence between any two traits in large problems where BPS fails.

2.3.3.3 Highly structured, high dimensional data and latent factor models

Up to this point, we have primarily discussed the computational challenges associated with big-N problems. Big-P data sets are increasingly common in phylogenetic problems, and the methods discussed previously scale at best quadratically in P. Bayesian latent factor models (Press and Shigemasu, 1989; Lopes and West, 2004) are a common approach to reduce both computational and model complexity. These models assume that the P-dimensional observed data $\mathbf{y}_i$ arise from $K < P$ dimensional latent processes $\mathbf{x}_i$. Specifically, $\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{L}^t\mathbf{x}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{L}$ is a $K \times P$ estimable matrix and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \text{diag}[\boldsymbol{\sigma}])$. The standard (non-phylogenetic) model assumes the prior distribution $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, but this specification

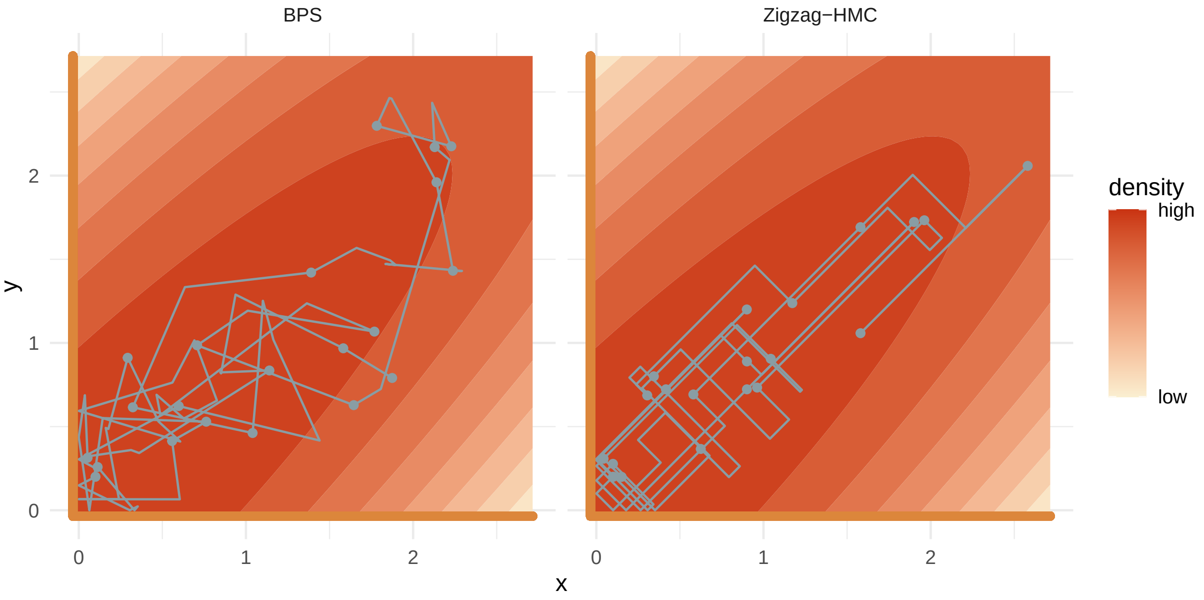


Figure 2.4: Sampling from a 2-dimensional truncated Gaussian distribution using both the BPS (left) and zigzag-HMC (right) samplers. Orange lines represent the truncations. Grey lines represent the particle trajectories, while grey dots represent samples from the posterior.

precludes the requisite correlation between the latent factors that the phylogeny induces. As such, Tolkoff et al. (2018) introduce phylogenetic factor analysis, where the $\mathbf{x}_i$ evolve along the phylogenetic tree via MBD. Standard procedures for sampling from the full conditional posterior of the loadings matrix $\mathbf{L}$ require conditioning on the latent traits $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^t$, and Tolkoff et al. (2018) rely on the procedure outlined in Cybis et al. (2015) to sample from $\mathbf{x}_i \,\big|\, \mathbf{y}_i, \mathbf{X}_{\backslash i}, \mathcal{F}, \boldsymbol{\sigma}$ for $i = 1, \ldots, N$ with overall complexity $\mathcal{O}(N^2 P K^2)$. Hassler et al. (2022, Chapter 4) apply the likelihood calculation and data augmentation algorithms of Hassler et al. (2020, Section 3.2.2.1) to sample from $\mathbf{X} \,|\, \mathbf{Y}, \mathcal{F}, \mathbf{L}, \boldsymbol{\sigma}$ in $\mathcal{O}(NPK^3)$. As K is by design small, the cubic scaling in K is preferable to the quadratic scaling in N.

Hassler et al. (2022, Section 4.3.1.2) also develop a novel HMC approach to efficiently sample directly from $\mathbf{L} \,|\, \mathbf{Y}, \mathcal{F}, \boldsymbol{\sigma}$ without conditioning on the latent factors $\mathbf{X}$ that applies to latent factor models generally. Hassler et al. (2022) show that one can calculate the gradient $\nabla_{\mathbf{L}} \log p(\mathbf{L} \,|\, \mathbf{Y}, \mathcal{F}, \boldsymbol{\sigma})$ required for HMC as a function of the full conditional mean and variance of each $\mathbf{x}_i$, but not the values of $\mathbf{x}_i$ explicitly. In the phylogenetic context, Hassler et al. (2022) use methods previously developed by Bastide et al. (2018) and Fisher et al. (2021) to calculate these gradients in $\mathcal{O}(NPK^3)$. This approach is easily transferable to non-phylogenetic latent factor models.

2.3.3.4 Beyond MBD

While the continuous trait models discussed above rely on MBD, we emphasize work on other models of continuous evolution. The closely related Ornstein–Uhlenbeck process (Uhlenbeck and Ornstein, 1930) is a Gaussian process where traits tend to revert to some mean value (i.e. some evolutionary optimum). Recent work has focused on inferring the points along the phylogeny at which these optima change, known as adaptive shifts (Uyeda and Harmon, 2014). Bastide et al. (2018) develop efficient likelihood calculations under a special case of this model. Other models include diffusion on a sphere (Bouckaert, 2016) and within a latent space arising from a multidimensional scaling (Holbrook et al., 2021) when only pair-wise distances between traits are observed.

2.3.4 Preferential sampling and bias

Phylogenetic analyses typically study biological populations evolving in the real world and are inherently observational. As such, data ascertainment is an important factor in any phylogenetic study, with preferential sampling possibly biasing results (Karcher et al., 2016). Phylogeographic models that capture spatiotemporal evolution are particularly susceptible to non-uniform sampling across both space and time (Guindon and De Maio, 2021; Kalkauskas et al., 2021). In infectious disease phylogeography, data ascertainment typically requires sequencing the viral genome associated with an individual infection. Unsurprisingly, there are numerous disparities that lead to preferential sampling across both time and space. Both testing and sequencing can be expensive, and resource-rich regions tend to sequence a higher proportion of actual infections (Brito et al., 2021). In the extreme case there may be no sequences available from a location with high levels of known transmission. In addition to sub-sampling to create more representative data sets, researchers have developed several strategies to address bias induced by preferential sampling.

2.3.4.1 Directly modeling ascertainment

The coalescent tree priors mentioned in Section 2.1.2 enable inference of (possibly time-varying) effective population size (EPS). Unsurprisingly, estimates of time-varying EPS are particularly sensitive to preferential sampling in time. While standard models (often inappropriately) assume that sequence ascertainment does not depend on EPS, Karcher et al. (2016) explicitly model ascertainment as an inhomogeneous Poisson process with intensity a function of EPS. They demonstrate via simulation that this approach reduces bias in EPS estimates when sequence ascertainment is proportional to EPS, a common scenario in epidemiological studies.

2.3.4.2 Sequence-free observations

When the spatiotemporal distribution of an epidemic can be estimated *a priori*, one can partially correct for preferential sampling by introducing sequence-free samples into the phylogenetic trait reconstruction. Up to this point we have taken for granted that all tip nodes in the phylogeny correspond to an associated molecular sequence as the sequences are the primary source of information for inferring the phylogeny itself. As there are situations where one has access to information about the spatiotemporal distribution of an epidemic (e.g. regional case counts) but relatively few sequences from certain locations, Lemey et al. (2020) and Kalkauskas et al. (2021) propose introducing sequence-free nodes to the phylogenetic tree and demonstrate that this approach can reduce bias induced by extremely biased sampling. Of course, this approach requires prior knowledge of the true spatiotemporal distribution of the process of interest.

2.3.4.3 Structured coalescent

An alternative model of discrete phylogeographic migration is the structured coalescent (Notohara, 1990), which posits a backward-in-time process where lineages converge and migrate between sub-populations. Where the previously-discussed discrete-trait model assumes the tree is *a priori* independent of the location data, the structured coalescent explicitly models dependence of the tree on the locations, which can reduce bias in both ancestral state reconstructions and rates of migration between locations. As the population demographics are explicit model parameters, they can in turn be informed by other sources of data, further avoiding some biases introduced by preferential sampling of individuals in some states (De Maio et al., 2015). The primary challenge to inference under these structured coalescent models is that there is no analog to Felsenstein's pruning algorithm (Felsenstein, 1973a, 1981, see Section 2.1.1) that analytically integrates out the migration events. As such, inference under these models requires numerically marginalizing the migration history, typically via MCMC (Vaughan et al., 2014).

De Maio et al. (2015) develop an approximation to the standard structured coalescent model that does allow analytic integration of the migration histories, avoiding laborious numerical integration. Volz (2012) and Müller et al. (2017) also develop efficient numerical approximations of the structured coalescent likelihood. Existing implementations of structured coalescent models, however, still compare poorly computationally with the simpler discrete trait models and are intractable for large-scale problems. Improving computational efficiency in these models is an active area of research.

2.4 Case study

Phylogenetics has increasingly played a role in studying viral epidemic dynamics, sometimes in real time (Dellicour et al., 2021; Hodcroft et al., 2021). Researchers can integrate information about the spatiotemporal spread of a virus into phylogenetic models to identify an epidemic's origin (Plantier et al., 2009; Liu et al., 2013; Worobey et al., 2016) and transmission dynamics (Ehichioya et al., 2011; Dudas et al., 2017; Du Plessis et al., 2021). In these phylodynamic analyses, the sampling time and location of a genetic sequence are critical data that allow researchers to reconstruct how a virus spreads through populations.

Here, we consider a case study arising out of the paper by Lemey et al. (2020) on early SARS-CoV-2 international transmission. In addition to viral genetic sequences, sample dates and sample locations, Lemey et al. (2020) incorporate information on individual travel history, global air traffic patterns, local outbreak intensity and within-host infection dynamics. The authors seek to identify the paths along which SARS-CoV-2 traveled as it escaped Hubei province, China, and spread globally. As discussed in Section 2.3.4, phylogeographic analyses are susceptible to ascertainment bias, which is often unavoidable as viral transmission does not respect administrative boundaries with consistent sequencing and reporting. To address this challenge, Lemey et al. (2020) integrate both individual-level travel history and location-specific estimated case counts into their phylogeographic analysis.

Lemey et al. (2020) collect 282 early SARS-CoV-2 sequences from around the world. Roughly 20% of these sequences were associated with recorded international travel. As they consider 44 discrete locations, they parameterize the transition rate matrix via a GLM with pairwise air traffic connectivity and geographic distance as covariates (see Section 2.3.2.2). To incorporate travel history, they introduce additional degree-2 internal nodes (i.e. nodes with a single parent and single child) into the phylogeny and assign the travel origins to those nodes. The dates of these nodes are fixed to the travel dates (when known) or inferred assuming a prior informed by the SARS-CoV-2 incubation time. The travel destinations remain assigned to tip nodes. Finally, Lemey et al. (2020) incorporate sequence-free observations from undersampled locations such as Italy and Iran.

Ultimately, incorporating these various sources of information into the discrete trait phylogeographic model resulted in more plausible transmission patterns and a statistical model with greater out-of-sample predictive performance (see Figure 2.5). The Bayesian approach allows seamless incorporation of prior knowledge in 1) SARS-CoV-2 case counts informing the locations and dates of sequence-free tip nodes and 2) SARS-CoV-2 within-host dynamics informing the prior on the time between the origin and destination nodes associated with specific travelers. These approaches also permitted accommodation of uncertainty in the phylogenetic tree itself, as the phylogenetic tree was inferred simultaneously with all transmission dynamics via MCMC simulation.

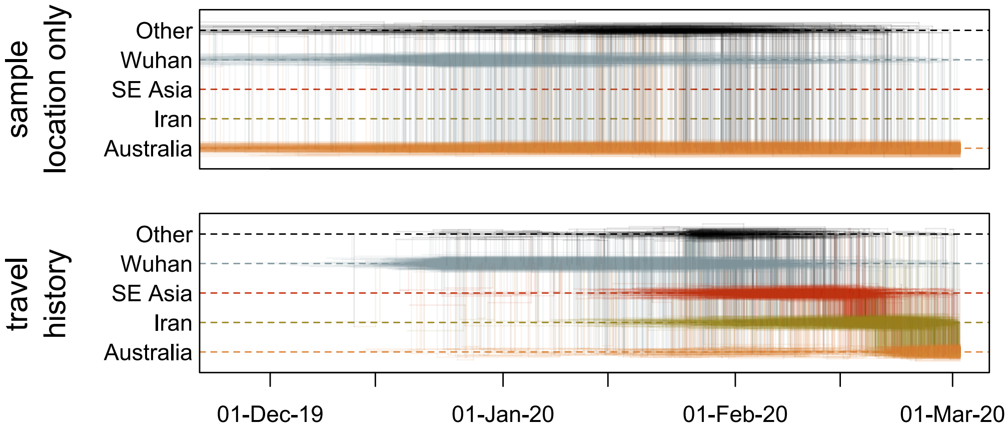


Figure 2.5: A toy example of the influence of travel history on discrete trait analyses. Horizontal lines represent persistent lineages within a location, while vertical lines represent transitions between locations in the Markov chain. We inferred a tree with 9 sequences (3 each from Wuhan, Australia, and Europe) where some of the infected individuals sampled in Australia had traveled from Iran or Southeast (SE) Asia. The analysis incorporating travel history captures more information in that the virus is present in all locations and there is less variance in the dates of transition events. This figure was modeled on the tutorial presented in the BEAST documentation. Please note that this is a toy analysis and should not be interpreted as providing insight into the early spread of SARS-CoV-2.

2.5 Discussion

Phylogenetics has motivated numerous theoretical, methodological, and computational advances in the statistics of Bayesian networks, continuous-time Markov processes and Gaussian processes. The challenges of dealing with complex, hierarchical statistical models with combined continuous/discrete parameter spaces continue to spur creative statistical innovations. Many of the topics discussed are active areas of research.

The Bayesian approach is particularly useful in phylogenetics as the phylogeny itself is frequently a nuisance parameter. Analyses that condition on a single phylogeny do not properly account for the often high degree of uncertainty in the phylogenetic estimates. Numerically marginalizing over the phylogeny via MCMC or other approaches discussed in Section 2.2 conveniently addresses this uncertainty. Similarly, the Bayesian approach offers a intuitive way to account for uncertainty in the phylogeny. Beyond properly measuring uncertainty, there are cases where we do indeed have prior information about relevant parameter values such as the root date (e.g. the temporal origin of a pandemic) or branch lengths (e.g. rapidly growing populations tend to have shorter branch lengths near the root).

Despite the many advances, there are persistent challenges in both inferring the tree itself and data integration. The SARS-CoV-2 pandemic greatly accelerated previous gains in epidemic genomic surveillance. Bayesian methods are typically limited to several thousand taxa and currently require down-sampling when analyzing some pandemic-scale data sets. Recent work has focused on computationally efficient implementations of simpler models (https://beast.community/thorney_beast) or approximate likelihoods (De Maio et al., 2022). Additionally, as discussed in Section 2.3.4, common phylogeographic models exhibit a tradeoff between computational efficiency and robustness to sampling bias.

Finally, while we focus here on the statistical implications related to data integration in Bayesian phylogenetics, we direct the reader to Baele et al. (2017) for a thorough discussion of data integration from a more biological perspective with more specific examples.

2.6 Acknowledgments

This work was supported through US National Institutes of Health grants HG006139, AI153044, AI154824 and AI162611; the European Research Council under the European Union's Horizon 2020 grant agreement no. 725422-ReservoirDOCS; and Wellcome Trust project 206298/Z/17/Z (Artic Network). GB acknowledges support from the Internal Funds KU Leuven (Grant No. C14/18/094) and the Research Foundation - Flanders ("Fonds voor Wetenschappelijk Onderzoek – Vlaanderen," G0E1420N, G098321N). PL acknowledges support by the Research Foundation – Flanders ("Fonds voor Wetenschappelijk Onderzoek – Vlaanderen," G0D5117N and G051322N).

CHAPTER 3

Inferring phenotypic trait evolution on large trees with many incomplete measurements

3.1 Introduction

Phylogenetic comparative methods explore the relationships between different biological phenotypes across sets of organisms. To properly understand these phenotypic trait relationships, methods must adjust for the shared evolutionary history of the taxa (Felsenstein, 1985b). Molecular sequences from emerging sequencing technology and high-throughput biological experimentation enable such phylogenetic adjustment for rapidly growing numbers of taxa and increasing numbers of trait measurements. Comparative studies incorporating dense taxonomic sampling create the potential for new research into general patterns in phenotypic evolution, key differences between subgroups and the relationship between phenotypic and genetic evolutionary dynamics. Unfortunately, many phylogenetic comparative methods remain poorly equipped to handle these research questions at scale.

Popular methods often assume an underlying Brownian diffusion process acts along each branch of a phylogenetic tree, such that the traits are multivariate normally distributed. Revell (2012) and Adams (2014b), for example, parameterize this distribution in terms of a highly-structured variance-covariance matrix that characterizes the tree and trait covariation. Computational work to invert this matrix to evaluate the multivariate normal likelihood scales cubically with the number of taxa. This work stands even more troublesome when the phylogenetic tree remains unknown and requires joint inference with the trait process, necessitating repeated inversion. Freckleton (2012), Pybus et al. (2012), and Ho and Ané (2014) all independently develop algorithms that take advantage of the matrix-normal structure of the data under the MBD model to evaluate the likelihood. Using the tree structure, these algorithms then scale linearly with the number of taxa with complete data, but this ideal run-time currently stumbles when trait measurements are missing.

As the number of taxa grows large, measuring a complete suite of traits for all taxa becomes increasingly challenging. While stripping any rows of data with missing values may create a "complete" data set, this procedure both reduces statistical power and can introduce bias (Nakagawa and Freckleton, 2008). Recent solutions to this problem that take advantage of all available data include those by Goolsby (2017), Tolkoff et al. (2018), Bastide et al. (2018), and Mitov et al. (2020). Tolkoff et al. (2018), for example, treat the missing data points as unknown model parameters and integrate them out via Markov Chain Monte Carlo (MCMC). This method, however, requires iterative manipulation of the likelihood function on a per-taxon basis and remains computationally prohibitive for large trees. Alternatively, Goolsby (2017), Bastide et al. (2018), and Mitov et al. (2020) take a different approach and develop algorithms that can compute the likelihood of the observed data only in linear time with respect to the number of taxa. However, the inference strategy of all three groups (implemented in Rphylopars (Goolsby et al., 2017), PCMFit (Mitov et al., 2019), and PhylogeneticEM (Bastide et al., 2018) respectively) rely on maximum likelihood estimation (MLE) regimes that assume the phylogenetic tree is known *a priori*. While this assumption may be appropriate when the phylogenetic tree is known with a high degree of certainty, this is not the case for many practical problems. If there is any uncertainty in the tree, these methods will likely be both biased and over-confident in their estimates.

In this paper, we reformulate evaluation of the data likelihood function under a Brownian diffusion process on a tree such that we achieve the marginalized likelihood of the observed trait measurements only. This innovation arises from thinking about observed tip traits as multivariate normally distributed with infinite precision in their sampling, while missing traits have zero precision, and appropriately propagating these precisions up the tree through dynamic programming involving an unusual matrix pseudo-inverse definition. This pseudo-inverse finds similar use, but independent discovery, in Bastide et al. (2018). Unlike previous approaches, the integration avoids EM iteration making simultaneous inference with the phylogeny practical and enables researchers to analyze all available measurements when inferring the trait relationships. Surprisingly, we can still evaluate the observed-data likelihood in linear time with respect to the number of taxa. The price to be paid is that computation now scales cubically, rather than quadratically, in the number of traits. This remains a small price since the number of taxa is often orders-of-magnitude larger than the number of traits. It is also notable that this method has applications beyond phylogenetic comparative methods and can be used more generally in a special class of matrix-normal and multivariate normal distributions with missing data. This has been a long standing problem in statistics since at least the 1930's (Wilks, 1932), with more recent work by Dominici et al. (2000); Cantet et al. (2004); Allen and Tibshirani (2010); and Glanz and Carvalho (2018). One important limitation to our approach is that it assumes data are missing at random (Little and Rubin, 1987) which is inappropriate for many data sets.

We also demonstrate how this framework can be easily extended to incorporate residual variance in the MBD model, which is only one of many possible model extensions. Our strategy of analytically marginalizing the observed data likelihood extends seamlessly to this and other model extensions and allows for efficient inference on these models while maintaining likelihood computations that scale linearly with the number of taxa. These extensions open up lines of inquiry not available in the simple MBD model. In particular, including residual variance in the model enables inference of phylogenetic heritability.

We demonstrate the broad utility of our algorithm to compute the marginalized likelihood through three examples. First, we examine covariation in mammalian life history traits using data on 3649 taxa from the PanTHERIA ecological database (Jones et al., 2009). Second, we use our new efficient algorithm to simultaneously evaluate several theories regarding prokaryotic evolutionary theory. We use data from NCBI Genome and a recent study by Goberna and Verdú (2016), along with matching 16S sequences from the ARB Silva Database (Ludwig et al., 2004), to jointly infer both the phylogenetic tree and evolutionary correlation between several prokaryotic genotypic and phenotypic traits. Finally, we apply our multivariate residual variance model extension to data presented by Blanquart et al. (2017) concerning HIV virulence to evaluate the heritability of HIV viral load and CD4 T-cell decline. We compare the computation speed of our analytical integration method against current best-practice methods and observed increases in speed that top two orders-of-magnitude.

3.2 Phenotypic diffusion on trees

Consider a data-complete collection $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^t$ where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iP})^t$ of P real-valued phenotypic traits measured across N biological taxa. Relating the taxa stands a known and fixed or unknown and random phylogeny $\mathscr{F}$ that is a bifurcating, directed acyclic graph whose $2N-1$ vertices originate with a degree-2 root node ν_{2N-1} and terminate with degree-1 tip nodes $(\nu_1, \ldots, \nu_N)$ that correspond to the N taxa. Linking vertices are edge weights or branch lengths $(t_1, \ldots, t_{2N-2})$. Let $\mathbf{x}_k = (x_{k1}, \ldots, x_{kP})$ be latent values of the traits at node ν_k on the tree for $k = 1, \ldots, 2N-1$. For tip nodes $i = 1, \ldots, N$, we posit a stochastic link $p\left(\mathbf{y}_i \,|\mathbf{x}_i\right)$ where $\mathbf{y}_i$ is drawn from some distribution parameterized by $\mathbf{x}_i$ and other hyperparameters (see Figure 3.1). Comparative methods standardly assume that the density $p\left(\mathbf{y}_i \,|\mathbf{x}_i\right)$ is degenerate at $\mathbf{x}_i$ (i.e. $\mathbf{y}_i = \mathbf{x}_i$ with probability 1), but we relax this assumption in future sections.

The most common phenotypic model of evolution (Felsenstein, 1985b) assumes a multivariate Brownian diffusion process acts conditionally independently along each branch generating a multivariate normal (MVN) increment,

$$\mathbf{x}_k \sim \text{MVN}\left(\mathbf{x}_{\text{pa}(k)}, t_k \mathbf{\Sigma}\right) \text{ for } k = 1, \ldots, 2N-2, \tag{3.1}$$

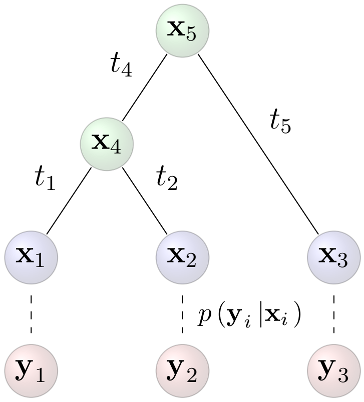


Figure 3.1: Schematic of diffusion model with stochastic link function. The data $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)^t$ arise from latent values $\mathbf{x}_i$ at the tips of the tree via the stochastic link function $p(\mathbf{y}_i \mid \mathbf{x}_i)$ for $i = 1, \ldots, N$.

centered around the realized value $\mathbf{x}_{\text{pa}(k)}$ at its parent node and variance proportional to an estimable $P \times P$ positive-definite matrix $\mathbf{\Sigma}$. Since the trait values at the root are also unknown, Pybus et al. (2012) suggest further assuming $\mathbf{x}_{2N-1} \sim \text{MVN}\left(\boldsymbol{\mu}_0, \kappa_0^{-1}\mathbf{\Sigma}\right)$ with fixed prior mean $\boldsymbol{\mu}_0$ and sample-size κ_0.

3.2.1 Computation of observed data likelihood

When there are no missing data and under our standard assumption that $p(\mathbf{y}_i \mid \mathbf{x}_i)$ is degenerate, integrating out unobserved internal and root node traits leads to a seemingly simple expression for the data likelihood $p(\mathbf{Y} \mid \mathbf{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0)$ (Freckleton, 2012; Vrancken et al., 2015). Namely, $\mathbf{Y}$ is matrix-normal (MN) distributed around mean $\mathbf{1}_N \boldsymbol{\mu}_0^t$, with across-row variance $\mathbf{\Upsilon} + \kappa_0^{-1}\mathbf{J}_N$ and across-column variance $\mathbf{\Sigma}$, where $\mathbf{1}_N$ is a vector of length N populated by ones, $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^t$, and $\mathbf{\Upsilon}$ is a deterministic function of $\mathscr{F}$. Specifically, element $\Upsilon_{ii'}$ measures shared evolutionary history and equals the sum of the branch lengths from the root to the most recent common ancestral node of taxa i and i' when $i \neq i'$ or the sum of the branch lengths from the root to taxon i otherwise. For example, in Figure 3.1, $\Upsilon_{12} = t_4$ and $\Upsilon_{11} = t_1 + t_4$. One can evaluate this highly structured matrix-normal likelihood function with computational complexity $\mathcal{O}(NP^2)$ given the acyclic nature of $\mathscr{F}$. When some data points are missing, however, the observed-data likelihood is no longer matrix-normal and new approaches are needed. This becomes increasingly urgent as the prevalence of missing observations grows with the size of trait data sets. In this context we wish to compute

$$p(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) = \int p(\mathbf{Y}^{\text{obs}}, \,|\, \mathbf{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0)\mathrm{d}\mathbf{Y}^{\text{mis}}, \tag{3.2}$$

where $\mathbf{Y}^{\text{obs}}$ and $\mathbf{Y}^{\text{mis}}$ contain the observed and missing trait values, respectively.

The two simplest strategies for calculating the observed-data likelihood are, unfortunately, computationally prohibitive for most large problems. One such solution forfeits the MN structure of the data in favor a simple expression of the observed-data likelihood. This strategy uses the fact that the matrix-normal distribution of $\mathbf{Y}$ can also be expressed as

$$\text{vec}\left[\mathbf{Y} \,|\, \mathbf{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0\right] \sim \text{MVN}\left(\text{vec}\left[\mathbf{1}_N \boldsymbol{\mu}_0^t\right], \mathbf{\Sigma} \otimes \left(\mathbf{\Upsilon} + \kappa_0^{-1}\mathbf{J}_N\right)\right), \tag{3.3}$$

using the Kronecker product $\otimes$. Assuming data are missing at random (Little and Rubin, 1987), one can simply remove the rows and columns of $\text{vec}\left[\mathbf{1}_N \boldsymbol{\mu}_0^t\right]$ and $\mathbf{\Sigma} \otimes \left(\mathbf{\Upsilon} + \kappa_0^{-1}\mathbf{J}_N\right)$ corresponding to the missing data and compute the likelihood for this $NP - M'$ dimensional MVN distribution, where M' is the number of missing measurements. This likelihood calculation carries the onerous computational complexity $\mathcal{O}\big((NP - M')^3\big)$. Alternatively, from a Bayesian perspective, one could numerically integrate out the missing data by treating each missing data point as an unknown model parameter and employing MCMC to sample each value. This strategy restores the matrix-normal structure, but requires the likelihood be evaluated each time one samples a missing data point. This results in computation complexity of at least $\mathcal{O}(NP^2M)$, where M is the number of taxa with missing measurements. Because M often scales with N, this method remains prohibitively slow for many data sets with large N. Our goal is to integrate out these missing values analytically using a dynamic programming algorithm in order to bring run time down to a much more manageable

$\mathcal{O}(NP^3)$.

3.2.1.1 Missing data definitions and operations

To develop our algorithm, we first introduce some useful abstractions and notation. At each tip in $\mathscr{F}$, information about each of the P traits comes in one of three forms: a trait value may be directly observed, latent, or completely missing. When directly observed, we posit without loss of generality that the value arises from a normal distribution centered at the observed value with infinite precision. We assume that trait data that arise from latent values are jointly multivariate normally distributed about the unknown latent values with known or estimable precision. Finally, a completely missing value arises also without loss of generality from a normal distribution centered at 0 with zero precision. To formalize this, for tip $i = 1, \ldots, N$, we construct a permutation matrix $\mathbf{C}_i$ that groups traits in directly observed, latent, and completely missing order and populate a pseudo-precision matrix

$$\mathbf{P}_i = \mathbf{C}_i \operatorname{diag}\left[\infty\mathbf{I}, \mathbf{R}_i, 0\mathbf{I}\right] \mathbf{C}_i^t, \tag{3.4}$$

where $\operatorname{diag}[\cdot]$ is a function that arranges its constituent elements into block-diagonal form and $\mathbf{R}_i$ is the latent block precision. Note that any block may be 0-dimensional. This construction arbitrarily forces off-diagonal elements of $\mathbf{P}_i$ involving directly observed and completely missing traits to equal 0 and plays an important role in simplifying computations.

We additionally define a series of operations that we will find useful for defining this algorithm. We define the pseudo-inverse

$$\mathbf{P}_i^- = \mathbf{C}_i \operatorname{diag}\left[0\mathbf{I}, \mathbf{R}_i^{-1}, \infty\mathbf{I}\right] \mathbf{C}_i^t. \tag{3.5}$$

We define the pseudo-determinant $\hat{\det}()$ as the product of the non-zero singular values. We also define the matrix $\boldsymbol{\delta}_i = \operatorname{diag}\left[\delta_{i1}, \ldots, \delta_{iP}\right]$ for $i = 1, \ldots, N$, where δ_{ij} is an indicator variable which takes a value of 1 if trait Y_{ij} is observed or latent and 0 if it is missing. Lastly, we define the possibly degenerate multivariate normal density function

$$\log \hat{\phi}(\mathbf{z}; \boldsymbol{\mu}, \mathbf{P}) = \frac{1}{2} \log \hat{\det}(\mathbf{P}) - \frac{\operatorname{rank}(\mathbf{P})}{2} \log 2\pi - \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^t \, \mathbf{P} \, (\mathbf{z} - \boldsymbol{\mu}),$$

for some argument $\mathbf{z}$, mean $\boldsymbol{\mu}$ and precision $\mathbf{P}$ of appropriate dimensions.

3.2.1.2 Post-order observed data likelihood algorithm

Our goal is to efficiently compute the likelihood $p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0)$. Following from Pybus et al. (2012), we perform a post-order traversal where we calculate the observed-data partial likelihood $p(\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor} \,|\, \mathbf{x}_k, \boldsymbol{\Sigma}, \mathscr{F})$ at each node ν_k where $\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor}$ is the observed data restricted to all descendants of node k on the tree. For example, in Figure 3.1, $\mathbf{Y}^{\text{obs}}_{\lfloor 4 \rfloor} = \{\mathbf{y}^{\text{obs}}_1, \mathbf{y}^{\text{obs}}_2\}$.

We posit that, given an appropriate stochastic link function $p\,(\mathbf{y}_i \,|\mathbf{x}_i\,)$, we can express the observed-data partial likelihood as

$$p(\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor} \,|\, \mathbf{x}_k, \boldsymbol{\Sigma}, \mathscr{F}) = r_k \hat{\phi}(\mathbf{x}_k; \mathbf{m}_k, \mathbf{P}_k)\,, \tag{3.6}$$

for all nodes $k = 1, \ldots, 2N - 1$ and some remainder r_k, mean $\mathbf{m}_k$, and precision $\mathbf{P}_k$. Given a parent node ℓ with children j and k, let us assume we can express the observed-data likelihood of $\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}$ and $\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor}$ as in Equation 3.6. Conditioning on $\mathbf{x}_\ell$, we can compute

$$p(\mathbf{Y}^{\text{obs}}_{\lfloor \ell \rfloor} \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) = p(\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor} \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor} \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) \tag{3.7}$$

as $\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}$ and $\mathbf{Y}^{\text{obs}}_{\lfloor k \rfloor}$ are conditionally independent given $\mathbf{x}_\ell$. Using Equations 3.1 and 3.6, we form

$$p(\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor} \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) = \int p(\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor} \,|\, \mathbf{x}_j, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{x}_j \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) \mathrm{d}\mathbf{x}_j = r_j \hat{\phi}\left(\mathbf{x}_\ell; \mathbf{m}_j, \mathbf{P}^{\star}_j\right), \tag{3.8}$$

where the branch-deflated pseudo-precision $\mathbf{P}_j^{\star} = \big(\mathbf{P}_j^{-} + t_j \boldsymbol{\delta}_j \boldsymbol{\Sigma} \boldsymbol{\delta}_j\big)^{-}$. See Section 3.9.1 for details on computing this pseudo-inverse. We use the results of Equation 3.8 in Equation 3.7 to compute the partial log-likelihood

$$\begin{aligned} \log p(\mathbf{Y}^{\text{obs}}_{\lfloor \ell \rfloor} \,|\, \mathbf{x}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) &= \log r_j + \log r_k + \log \hat{\phi}\big(\mathbf{x}_\ell; \mathbf{m}_j, \mathbf{P}_j^{\star}\big) + \log \hat{\phi}(\mathbf{x}_\ell; \mathbf{m}_k, \mathbf{P}_k^{\star}) \\ &= \log r_\ell + \log \hat{\phi}(\mathbf{x}_\ell; \mathbf{m}_\ell, \mathbf{P}_\ell)\,, \end{aligned} \tag{3.9}$$

where $\mathbf{P}_\ell = \mathbf{P}_j^{\star} + \mathbf{P}_k^{\star}$, $\mathbf{m}_\ell$ is a solution to $\mathbf{P}_\ell \mathbf{m}_\ell = \mathbf{P}_j^{\star} \mathbf{m}_j + \mathbf{P}_k^{\star} \mathbf{m}_k$, and

$$\begin{aligned} \log r_\ell = \log r_j + \log r_k + \frac{1}{2} \log \widehat{\det}\big(\mathbf{P}_j^{\star}\big) + \frac{1}{2} \log \widehat{\det}(\mathbf{P}_k^{\star}) - \frac{\Delta_{jk\ell}}{2} \log 2\pi \\ - \frac{1}{2} \log \widehat{\det}(\mathbf{P}_\ell) - \frac{1}{2} \left(\mathbf{m}_j^t \mathbf{P}_j^{\star} \mathbf{m}_j + \mathbf{m}_k^t \mathbf{P}_k^{\star} \mathbf{m}_k - \mathbf{m}_\ell^t \mathbf{P}_\ell \mathbf{m}_\ell\right). \end{aligned} \tag{3.10}$$

Note that the change of informative dimensions $\Delta_{jk\ell} = \text{rank}\big(\mathbf{P}_j^{\star}\big) + \text{rank}(\mathbf{P}_k^{\star}) - \text{rank}(\mathbf{P}_\ell)$. We update $\boldsymbol{\delta}_\ell = \boldsymbol{\delta}_j \vee \boldsymbol{\delta}_k$, where $\vee$ is the element-wise "logical or" operation.

Our algorithm initializes r_i, $\mathbf{m}_i$, and $\mathbf{P}_i$ such that $p(\mathbf{y}_i^{\text{obs}} \,|\, \mathbf{x}_i) = r_i \hat{\phi}(\mathbf{x}_i; \mathbf{m}_i, \mathbf{P}_i)$ at the tips of the tree. For the standard assumption that $\mathbf{y}_i = \mathbf{x}_i$, we have $r_i = 1$, $\mathbf{m}_i = \mathbf{C}_i \left[\mathbf{y}_i^{\text{obs}}, \mathbf{0}\right]$, and $\mathbf{P}_i = \mathbf{C}_i \operatorname{diag}\left[\infty \mathbf{I}, 0\mathbf{I}\right] \mathbf{C}_i^t$. We perform a post-order traversal of the tree computing $\mathbf{m}_\ell, \mathbf{P}_\ell$, and r_ℓ for internal nodes $\ell = N+1, \ldots, 2N-2$ using the already-computed node remainders, means, and precisions for their respective child nodes. At the root, $\mathbf{Y}^{\text{obs}}_{\lfloor 2N-1 \rfloor} = \mathbf{Y}^{\text{obs}}$ and we return the observed-data log-likelihood

$$\begin{aligned} p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) &= \int p(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{x}_{2N-1}, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{x}_{2N-1} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0) \mathrm{d}\mathbf{x}_{2N-1} \\ &= \int r_{2N-1} \hat{\phi}(\mathbf{x}_{2N-1}; \mathbf{m}_{2N-1}, \mathbf{P}_{2N-1})\, \hat{\phi}\left(\mathbf{x}_{2N-1}; \boldsymbol{\mu}_0, \kappa_0 \boldsymbol{\Sigma}^{-}\right) \mathrm{d}\mathbf{x}_{2N-1} \\ &= r_{\text{full}} \int \hat{\phi}(\mathbf{x}_{2N-1}; \mathbf{m}_{\text{full}}, \mathbf{P}_{\text{full}})\, \mathrm{d}\mathbf{x}_{2N-1}, \end{aligned} \tag{3.11}$$

where $\mathbf{P}_{\text{full}} = \mathbf{P}_{2N-1} + \kappa_0 \boldsymbol{\Sigma}^{-1}$ and $\mathbf{m}_{\text{full}} = \mathbf{P}_{\text{full}}^{-1} \left(\mathbf{P}_{2N-1} \mathbf{m}_{2N-1} + \kappa_0 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0\right)$. The integral evaluates to one, leaving the observed-data log-likelihood

$$\begin{aligned}\log p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) &= \log r_{\text{full}} \\ &= \log r_{2N-1} - \frac{\operatorname{rank}(\mathbf{P}_{2N-1})}{2} \log 2\pi \\ &\quad + \frac{1}{2} \log \hat{\det}(\mathbf{P}_{2N-1}) + \frac{1}{2} \log \hat{\det}\left(\kappa_0 \boldsymbol{\Sigma}^{-1}\right) - \frac{1}{2} \log \hat{\det}(\mathbf{P}_{\text{full}}) \\ &\quad - \frac{1}{2} \left(\mathbf{m}_{2N-1}^t \mathbf{P}_{2N-1} \mathbf{m}_{2N-1} + \kappa_0 \boldsymbol{\mu}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \mathbf{m}_{\text{full}}^t \mathbf{P}_{\text{full}} \mathbf{m}_{\text{full}}\right). \end{aligned} \tag{3.12}$$

This tree traversal visits each node in $\mathscr{F}$ exactly once and inverts a $P \times P$ matrix each time, resulting in an overall computational complexity of $\mathcal{O}(NP^3)$ for each likelihood evaluation.

3.2.2 Inference

The primary parameter of scientific interest is the diffusion variance $\boldsymbol{\Sigma}$. We are also often interested in additional hyper-parameters $\boldsymbol{\theta}$ related to the stochastic link function $p\left(\mathbf{y}_i \,|\mathbf{x}_i\right)$. In cases where the tree structure is unknown, we use sequence data $\mathbf{S}$ to simultaneously infer $\mathscr{F}$. As such, from a Bayesian perspective, we are interested in approximating

$$p(\boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta} \,|\mathbf{Y}^{\text{obs}}, \mathbf{S}) \propto p(\mathbf{Y}^{\text{obs}} \,|\boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta})\, p(\mathscr{F}, \mathbf{S}) p(\boldsymbol{\Sigma}) p(\boldsymbol{\theta}), \tag{3.13}$$

for inference. We place a Wishart$_P\left(\boldsymbol{\Lambda}_0, \nu\right)$ prior on $\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Lambda}_0$ is a $P \times P$ rate matrix. The prior on $\boldsymbol{\theta}$ depends the problem of interest, and there are many ways to specify $p(\mathscr{F}, \mathbf{S})$ (see Suchard et al., 2018). To approximate the posterior distributions via MCMC simulation, we apply a random scan Metropolis-within-Gibbs (Liu et al., 1995) approach by which we sample parameter blocks one at a time at random from their full conditional distribution.

Let $\mathbf{X} = \left(\mathbf{x}_1, \cdots, \mathbf{x}_N\right)^t$ be the latent trait values at the tips of the phylogeny. The conjugate $\text{Wishart}_P\left(\boldsymbol{\Lambda}_0, \nu\right)$ prior on $\boldsymbol{\Sigma}^{-1}$ implies that

$$\boldsymbol{\Sigma}^{-1} \mid \mathbf{X}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0 \sim \text{Wishart}_P\left[\boldsymbol{\Lambda}_0 + \left(\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t\right)^t \left(\boldsymbol{\Upsilon} + \frac{1}{\kappa_0}\mathbf{J}_N\right)^{-1} \left(\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t\right), \nu + N\right]. \tag{3.14}$$

We apply the post-order computation method proposed by Ho and Ané (2014) to compute $\left(\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t\right)^t \left(\boldsymbol{\Upsilon} + \frac{1}{\kappa_0}\mathbf{J}\right)^{-1} \left(\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t\right)$, which has computational complexity $\mathcal{O}(NP^2)$. When $\mathbf{X}$ are known (i.e. when there are no missing values and $p\left(\mathbf{y}_i \mid \mathbf{x}_i\right)$ is degenerate at $\mathbf{x}_i$), we can sample from the distribution in Equation 3.14 immediately without any additional steps. However, if either assumption is violated, we must first draw from the full conditional distribution of $\mathbf{X}$ via the data augmentation algorithm described below. This algorithm is similar to the 'E' step of the EM algorithm developed by Bastide et al. (2018) to compute the moments of each $\mathbf{x}_i$. In our case, we sample from the joint posterior of all $\mathbf{x}_i$ simultaneously rather than computing the conditional moments of each $\mathbf{x}_i$ individually.

3.2.2.1 Pre-order missing data augmentation algorithm

To sample jointly from the full conditional of $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)^t$, we draw on the calculations made in Section 3.2.1.2 and perform a pre-order traversal of the tree. Note that we omit explicit dependence on the parameters $\boldsymbol{\Sigma}, \mathscr{F}$, and $\boldsymbol{\theta}$ in all calculations below for clarity. Starting at the root, $\mathbf{x}_{2N-1}$, we draw from $\mathbf{x}_{2N-1} \mid \mathbf{Y}^{\text{obs}}, \boldsymbol{\mu}_0, \kappa_0$. Using Bayes' rule and Equation 3.11, we see that

$$\begin{aligned} p(\mathbf{x}_{2N-1} \mid \mathbf{Y}^{\text{obs}}, \boldsymbol{\mu}_0, \kappa_0) &\propto p(\mathbf{Y}^{\text{obs}} \mid \mathbf{x}_{2N-1}) p(\mathbf{x}_{2N-1} \mid \boldsymbol{\mu}_0, \kappa_0) \\ &\propto \hat{\phi}\left(\mathbf{x}_{2N-1}; \mathbf{m}_{\text{full}}, \mathbf{P}_{\text{full}}\right), \text{ which implies that} \\ \mathbf{x}_{2N-1} \mid \mathbf{Y}^{\text{obs}}, \boldsymbol{\mu}_0, \kappa_0 &\sim \text{MVN}\left(\mathbf{m}_{\text{full}}, \mathbf{P}_{\text{full}}\right). \end{aligned} \tag{3.15}$$

After sampling the root traits from their full conditional, we continue the traversal of the tree where we sample each node $\mathbf{x}_j$ conditional on its (previously sampled) parent node $\mathbf{x}_{\text{pa}(j)}$ and the observed data below node j $\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}$ for $j = 1, \ldots, 2N-2$. For the internal nodes, we compute $p(\mathbf{x}_j \,|\, \mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}, \mathbf{x}_{\text{pa}(j)})$ as follows:

$$\begin{aligned} p(\mathbf{x}_j \,|\, \mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}, \mathbf{x}_{\text{pa}(j)}) &\propto p(\mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor} \,|\, \mathbf{x}_j) p(\mathbf{x}_j \,|\, \mathbf{x}_{\text{pa}(j)}) \\ &\propto \hat{\phi}(\mathbf{x}_j; \mathbf{m}_j, \mathbf{P}_j)\, \hat{\phi}\left(\mathbf{x}_j; \mathbf{x}_{\text{pa}(j)}, (t_j \boldsymbol{\Sigma})^{-1}\right) \\ &\propto \hat{\phi}(\mathbf{x}_j; \mathbf{n}_j, \mathbf{Q}_j) \end{aligned} \tag{3.16}$$

where $\mathbf{Q}_j = \mathbf{P}_j + (t_j \boldsymbol{\Sigma})^{-1}$ and $\mathbf{n}_j = \mathbf{Q}_j^{-1} \left(\mathbf{P}_j \mathbf{m}_j + (t_j \boldsymbol{\Sigma})^{-1} \mathbf{x}_{\text{pa}(j)} \right)$. This implies $\mathbf{x}_j \,\Big|\, \mathbf{Y}^{\text{obs}}_{\lfloor j \rfloor}, \mathbf{x}_{\text{pa}(j)} \sim \text{MVN}(\mathbf{n}_j, \mathbf{Q}_j)$, and we sample $\mathbf{x}_j$ from this distribution.

At the tips, we employ one of two techniques depending on the specific model. Under our standard assumption (i.e. $\mathbf{x}_i = \mathbf{y}_i$ with probability 1), we partition the precision $\boldsymbol{\Sigma}^{-1}$ and trait values $\mathbf{x}_i$ and $\mathbf{x}_{\text{pa}(i)}$ such that

$$\boldsymbol{\Sigma}^{-1} = \mathbf{C}_i \begin{pmatrix} \mathbf{S}_i^{\text{obs}} & \mathbf{S}_i^{\text{om}} \\ \mathbf{S}_i^{\text{mo}} & \mathbf{S}_i^{\text{mis}} \end{pmatrix} \mathbf{C}_i^t, \quad \mathbf{x}_i = \mathbf{C}_i \begin{pmatrix} \mathbf{x}_i^{\text{obs}} \\ \mathbf{x}_i^{\text{mis}} \end{pmatrix}, \text{ and } \quad \mathbf{x}_{\text{pa}(i)} = \mathbf{C}_i \begin{pmatrix} \mathbf{x}_{\text{pa}(i)}^{\text{obs}} \\ \mathbf{x}_{\text{pa}(i)}^{\text{mis}} \end{pmatrix} \tag{3.17}$$

and draw from $\mathbf{x}_i^{\text{mis}} \,\big|\, \mathbf{y}_i^{\text{obs}}, \mathbf{x}_{\text{pa}(i)} \sim \text{MVN}\left(\mathbf{x}_{\text{pa}(i)}^{\text{mis}} + \mathbf{S}_i^{\text{mis}-1} \mathbf{S}_i^{\text{mo}} \left(\mathbf{x}_{\text{pa}(i)}^{\text{obs}} - \mathbf{x}_i^{\text{obs}} \right), \frac{1}{t_i} \mathbf{S}_i^{\text{mis}} \right)$ for $i = 1, \ldots, N$. For cases where $p\left(\mathbf{y}_i \,|\mathbf{x}_i\right)$ is non-degenerate, we simply use Equation 3.16 to sample from $\mathbf{x}_i \,\big|\, \mathbf{y}_i^{\text{obs}}, \mathbf{x}_{\text{pa}(i)}$. Once we have sampled $\mathbf{X} \,|\mathbf{Y}^{\text{obs}}, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta}$, we can draw from the full conditional distribution of $\boldsymbol{\Sigma}^{-1}$ via Equation 3.14. This pre-order data augmentation procedure requires a single $P \times P$ matrix inversion at each of the $2N-1$ nodes in the tree, resulting in overall computational complexity $\mathcal{O}(NP^3)$.

3.3 Model extension: residual variance

We extend the MBD model of phenotypic evolution to include multivariate normal residual variance at each of the tips. Under this model, we assume

$$p(\mathbf{y}_i \,|\mathbf{x}_i) = \hat{\phi}(\mathbf{y}_i; \mathbf{x}_i, \mathbf{\Gamma}) \text{ for } i = 1, \dots, N, \tag{3.18}$$

where $\mathbf{\Gamma}$ is a $P \times P$ precision matrix. Under this model, the vectorization of $\mathbf{Y}$ is MVN-distributed with $NP \times NP$ variance-covariance matrix $\mathbf{\Sigma} \otimes \left(\mathbf{\Upsilon} + \kappa_0^{-1}\mathbf{J}_N\right) + \mathbf{\Gamma}^{-1} \otimes \mathbf{I}_N$ where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Unlike the case where $\mathbf{y}_i = \mathbf{x}_i$, $\mathbf{Y}$ cannot be expressed as matrix-normal even in the data-complete case because the variance cannot be expressed as the Kronecker product of two matrices. As such, our post-order likelihood computation algorithm is useful for this extended model, even when there are no missing data points.

3.3.1 Inference of residual variance

Similar to our inference of $\mathbf{\Sigma}$ in the diffusion process, we place a conjugate $\text{Wishart}_P\left(\mathbf{\Lambda}_s, \nu_s\right)$ prior on $\mathbf{\Gamma}$ using the rate parameterization. This yields the full conditional distribution

$$\mathbf{\Gamma}\,|\mathbf{Y}, \mathbf{X} \sim \text{Wishart}_P\left(\mathbf{\Lambda}_s + (\mathbf{Y} - \mathbf{X})^t\,(\mathbf{Y} - \mathbf{X})\,, \nu_s + N\right). \tag{3.19}$$

Because $\mathbf{X}$ is latent in this model, each time we update $\mathbf{\Gamma}$ we first draw from the full conditional posterior of $\mathbf{X}$ using the algorithm described in Section 3.2.2.1. For cases where $\mathbf{Y}$ is not completely observed, we must perform an additional data augmentation step where we draw from $\mathbf{Y}^{\text{mis}}\,|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{\Gamma}$. To do this, we decompose the sampling precision matrix into blocks such that

$$\mathbf{\Gamma} = \mathbf{C}_i \begin{pmatrix} \mathbf{\Gamma}_i^{\text{obs}} & \mathbf{\Gamma}_i^{\text{mo}t} \\ \mathbf{\Gamma}_i^{\text{mo}} & \mathbf{\Gamma}_i^{\text{mis}} \end{pmatrix} \mathbf{C}_i^t \text{ for } i = 1, \dots, N. \tag{3.20}$$

From Equation 3.18, we see that

$$p\left(\mathbf{y}_i^{\text{mis}} \left| \mathbf{y}_i^{\text{obs}}, \mathbf{x}_i, \boldsymbol{\Gamma}\right.\right) = \hat{\phi}\left(\mathbf{y}_i^{\text{mis}}; \mathbf{x}_i^{\text{mis}} + \boldsymbol{\Gamma}_i^{\text{mis}-1}\boldsymbol{\Gamma}_i^{\text{mo}}\left(\mathbf{x}_i^{\text{obs}} - \mathbf{y}_i^{\text{obs}}\right), \boldsymbol{\Gamma}_i^{\text{mis}}\right). \tag{3.21}$$

As such, we can directly sample $\mathbf{y}_i^{\text{mis}}$ from its full conditional above and update $\mathbf{y}_i = \mathbf{C}_i\left[\mathbf{y}_i^{\text{obs}}, \mathbf{y}_i^{\text{mis}}\right]^t$ for $i = 1, \ldots, N$. This process also has computational complexity $\mathcal{O}(NP^3)$.

Note that we can draw from the joint full conditional of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ by performing a single pre-order data augmentation where we draw from $p(\mathbf{X}, \mathbf{Y}^{\text{mis}} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\Gamma})$ and subsequently draw from $p(\boldsymbol{\Sigma}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{Y}) = p(\boldsymbol{\Sigma} \,|\, \mathbf{X})p(\boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{Y})$. These distributions are conditionally independent due to the fact that $\mathbf{X}$ and $\mathbf{X} - \mathbf{Y}$ are independent by construction. This procedure effectively halves the computation time as we only need to perform a single post-order likelihood computation/pre-order data augmentation step to sample both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$, rather than each time we sample one.

3.3.2 Heritability statistic

The residual variance extension enables us to estimate phenotypic heritability over evolutionary time. We use a definition analogous to the broad-sense heritability in statistical genetics (see Visscher et al., 2008). Namely, we seek to quantify the proportion of variance in a trait attributable to the Brownian diffusion process on the phylogeny (as opposed to the residual variance). Note that we are primarily interested in heritability in the HIV example below, for which we use data from a recent paper by Blanquart et al. (2017). As such, we use a multivariate generalization of the heritability statistic from that paper. Specifically, we estimate phylogenetic heritability by taking the expectation of the empirical sample variance under our extended model. We define the $P \times P$ empirical covariance matrix as

$$\mathbf{S}^2(\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)^t = \frac{1}{N}\left(\mathbf{Y} - \bar{\mathbf{Y}}\right)^t\left(\mathbf{Y} - \bar{\mathbf{Y}}\right), \tag{3.22}$$

where $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i = \frac{1}{N}\mathbf{Y}^t\mathbf{1}_N$ and $\bar{\mathbf{Y}} = \mathbf{1}_N\bar{\mathbf{y}}^t = \frac{1}{N}\mathbf{J}_N\mathbf{Y}$. The expectation of this quantity reduces to the following expression (see Section 3.9.2 for details):

$$\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] = \frac{N-1}{N}\mathbf{\Gamma}^{-1} + \left(\frac{1}{N}\text{tr}\left[\mathbf{\Upsilon}\right] - \frac{1}{N^2}\mathbf{1}_N^t\mathbf{\Upsilon}\mathbf{1}_N\right)\mathbf{\Sigma}. \tag{3.23}$$

Because $\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right]$ is a linear combination of $\mathbf{\Sigma}$ and $\mathbf{\Gamma}^{-1}$, we propose the $P \times P$ heritability matrix $\mathbf{H} = \{h_{kl}\}$ with entries

$$h_{kl} = \frac{c_\sigma \Sigma_{kl}}{\sqrt{\left(c_\sigma\Sigma_{kk} + c_\gamma\Gamma_{kk}^{-1}\right)\left(c_\sigma\Sigma_{ll} + c_\gamma\Gamma_{ll}^{-1}\right)}}, \tag{3.24}$$

where $c_\sigma = \frac{1}{N}\text{tr}\left[\mathbf{\Upsilon}\right] - \frac{1}{N^2}\mathbf{1}_N^t\mathbf{\Upsilon}\mathbf{1}_N$ and $c_\gamma = \frac{N-1}{N}$. Each diagonal entry $h_{kk} = h_k^2$ represents the marginal phylogenetic heritability of that trait, and each off-diagonal entry represents the pair-wise co-heritability (Falconer, 1960, chap. 19) between traits.

Note that naive computation of $c_\sigma = \frac{1}{N}\text{tr}\left[\mathbf{\Upsilon}\right] + \frac{1}{N^2}\mathbf{1}_N^t\mathbf{\Upsilon}\mathbf{1}_N$ in Equation 3.23 would require constructing the $N \times N$ matrix $\mathbf{\Upsilon}$ and summing over all its elements, which has computation complexity of at least $\mathcal{O}(N^2)$. For cases where $\mathscr{F}$ is random and changes throughout the MCMC simulation, this quantity must be re-computed each time we compute the statistic. To avoid this issue, we implement an algorithm that avoids constructing $\mathbf{\Upsilon}$ in its entirety and simply calculates both $\text{tr}\left[\mathbf{\Upsilon}\right]$ and $\mathbf{1}_N^t\mathbf{\Upsilon}\mathbf{1}_N$ in $\mathcal{O}(N)$ time. The algorithm performs a post-order traversal of the tree where at each internal node ν_ℓ we compute $N_{\lfloor\ell\rfloor}$ (the number of tips below ν_ℓ), $s_{\lfloor\ell\rfloor}$ (the sum of all elements in $\mathbf{\Upsilon}_{\lfloor\ell\rfloor}$), and $d_{\lfloor\ell\rfloor}$ (the sum of the diagonal elements in $\mathbf{\Upsilon}_{\lfloor\ell\rfloor}$). We define $\mathbf{\Upsilon}_{\lfloor\ell\rfloor}$ as the tree variance-covariance matrix constructed from the sub-tree $\mathscr{F}_{\lfloor\ell\rfloor}$ that is simply the tree that contains only the nodes below ν_ℓ with node ν_ℓ

as its root. For internal nodes ν_ℓ with child nodes ν_j and ν_k, we accumulate

$$\begin{aligned} N_{\lfloor \ell \rfloor} &= N_{\lfloor j \rfloor} + N_{\lfloor k \rfloor} + 1, \\ s_{\lfloor \ell \rfloor} &= s_{\lfloor j \rfloor} + s_{\lfloor k \rfloor} + t_j N_{\lfloor j \rfloor}^2 + t_k N_{\lfloor k \rfloor}^2, \text{ and} \\ d_{\lfloor \ell \rfloor} &= d_{\lfloor j \rfloor} + d_{\lfloor k \rfloor} + t_j N_{\lfloor j \rfloor} + t_k N_{\lfloor k \rfloor}. \end{aligned} \tag{3.25}$$

At the tips, we initialize with $s_{\lfloor i \rfloor} = d_{\lfloor i \rfloor} = 0$ and $N_{\lfloor i \rfloor} = 1$. At the root, $s_{\lfloor 2N-1 \rfloor} = \mathbf{1}_N^t \boldsymbol{\Upsilon} \mathbf{1}_N$ and $d_{\lfloor 2N-1 \rfloor} = \operatorname{tr}[\boldsymbol{\Upsilon}]$. This algorithm visits each node in $\mathscr{F}$ exactly once and has run time $\mathcal{O}(N)$.

While the breadth of research in heritability is extensive across both statistical genetics and phylogenetics (see in particular the recent paper by Mitov and Stadler, 2018), we choose the same heritability statistic as used by Blanquart et al. (2017) for direct comparison with their analysis. That being said, our methods could be readily adapted to approximate the posterior distribution of several of the alternative heritability statistics presented in Mitov and Stadler (2018). Additionally, our pre-order data augmentation procedure allows us to generate samples directly from the posterior of the latent trip traits $\mathbf{X}$, from which we can directly compute the genetic covariance $\mathbf{S}^2(\mathbf{X})$ rather than relying on expectations.

3.4 Research materials

We have implemented these methods in the development version of BEAST (Suchard et al., 2018). The data files, scripts, and instructions necessary for running the following analyses are available at `https://github.com/suchard-group/incomplete_measurements`.

3.5 Computational efficiency

Our method dramatically increases computational efficiency over the current best-practice method. This latter procedure, developed by Cybis et al. (2015), treats the missing and latent values of $\mathbf{X}$ as unknown parameters and numerically integrates them out by placing a Gibbs sampler on each tip $\mathbf{x}_i$ that draws from its full conditional distribution $p\left(\mathbf{x}_i \left| \mathbf{y}_i, \mathbf{X}_{\lceil i \rceil}\right.\right)$ for $i = 1, \ldots, N$ where $\mathbf{X}_{\lceil i \rceil} = \mathbf{X} \backslash \mathbf{x}_i$. Because the full conditional distribution of $\mathbf{x}_i$ relies on the other missing and latent values in $\mathbf{X}$, we sample each tip individually. The advantage of this is that the likelihood calculation, the Gibbs sampler of the diffusion variance $\mathbf{\Sigma}$, and the data augmentation procedure for each tip all have complexity $\mathcal{O}(NP^2)$ rather than our $\mathcal{O}(NP^3)$. As such, this numerical integration procedure has overall complexity $\mathcal{O}(MNP^2)$ where M is the number of tips with missing or latent values. For any extended model where $p\left(\mathbf{y}_i \left| \mathbf{x}_i\right.\right)$ is not degenerate at $\mathbf{x}_i$, all values of $\mathbf{X}$ are latent and $M = N$.

We formalize our comparison by computing the median and minimum effective sample size (ESS) per hour for all parameters of interest under both our analytical integration method and the sampling method discussed above. Typically researchers run MCMC chains until the ESS for all parameters reach some minimum value, so the minimum ESS per hour is most reflective of actual computation time. We also compute the ESS per sample and samples per hour to understand how our improved method influences both the autocorrelation between MCMC samples and the amount of computational work required to generate a single draw from the posterior. Higher ESS per sample indicates lower autocorrelation, while higher samples per hour indicates less computational work per sample. We define the number of samples as the number of states in which the MCMC simulation updates the parameters of interest (as opposed to missing trait values). Note that for the numerical sampling strategy, we tested a range of sampling ratios between the parameters of interest and the missing trait values and chose the ratios with the best performance for each dataset/model combination.

Table 3.1 presents the results of our efficiency comparisons. We compare computation time under both models (only Brownian diffusion or Brownian diffusion with residual variance) for both the mammalian and HIV data set. We omit the prokaryote data set from this analysis as simultaneous inference of the tree made the “sampling” technique prohibitively slow. For each of the four scenarios, we performed 10 MCMC runs and compute the average

Table 3.1: Algorithmic improvement. We report MCMC sampling efficiency through effective sample size (ESS) that shows both a decrease in autocorrelation (as shows by ESS / Sample) and in the overall work required per sample (as shown by Samples / Hour).

Data set	Model	Integration method	ESS/hour		ESS/sample		Samples /hour
			minimum	median	minimum	median	
Mammals	Diffusion only	Analytic	1,200	3,600	0.043	0.13	27,000
		Sampling	3.0	9.8	0.0043	0.014	700
		Speed-up	**400**×	**370**×	**10**×	**9.5**×	**39**×
	Diffusion with residual	Analytic	140	320	0.0062	0.015	22,000
		Sampling	0.38	3.0	2.5e-5	0.00019	16,000
		Speed-up	**350**×	**110**×	**250**×	**76**×	**1.4**×
HIV	Diffusion only	Analytic	100,000	220,000	0.31	0.66	320,000
		Sampling	1,500	8,500	0.01	0.057	150,000
		Speed-up	**65**×	**25**×	**30**×	**12**×	**2.2**×
	Diffusion with residual	Analytic	1,600	2,500	0.0061	0.0096	260,000
		Sampling	5.1	8.7	5.1e-5	8.7e-5	100,000
		Speed-up	**320**×	**290**×	**120**×	**110**×	**2.6**×

ESS for each parameter, using the minimum and median of the averaged parameter ESSs in the table. We also report the speedup (analytic divided by sampling) for all values of interest in each analysis. Note that we only report up to two significant figures for clarity.

3.6 Simulation study

To understand the behavior of our inference techniques, we conduct a simulation study based on the empirical examples we discuss in Section 3.7. While these simulation studies cannot confirm that these models are appropriate for these real-world data sets, they do demonstrate the theoretical properties of our inference on these specific data sets assuming the model is appropriate. We use the mammals ($N = 3649$, $P = 8$), prokaryote ($N = 705$, $P = 7$), and HIV ($N = 1536$, $P = 3$) data sets. For each empirical example, we select the posterior mean diffusion variance $\mathbf{\Sigma}$ and residual variance $\mathbf{\Gamma}^{-1}$ to simulate traits. We also sub-sample the phylogenies from each example to vary the number of taxa. Note that for the prokaryotes example, we simulate conditional on the maximum clade credibility tree inferred from our analysis in Section 3.7.2. We keep the number of traits fixed within each empirical data set. Additionally, we randomly remove 0%, 25%, 50%, and (if possible) 75% of the data from each set of simulated values. We require that at least one observation from each taxon remain observed, so it is not possible to remove 75% of the data from the HIV example where $P = 3$.

For each unique combination of example, number of taxa, and percent of missing values, we simulate ten replicate data sets. Note that for each repetition we sub-sample a different set of taxa from the original phylogeny. We approximate the posterior of the diffusion correlation and residual correlation (i.e. the correlation derived from $\mathbf{\Sigma}$ and $\mathbf{\Gamma}^{-1}$ respectively) as well as the diagonals of the heritability matrix $\mathbf{H}$. These are the statistics that are of most scientific interest in our empirical analyses, and these model parameters remain invariant if the data are re-scaled while covariances do not. Across repetitions, we estimate the posterior bias and log mean squared error (logMSE) from the “true” values used for simulation. Figure 3.2 presents the posterior logMSE of all three parameters of interest for all example analyses. As expected the logMSE decreases with increasing taxa and decreasing missing values for all parameters of interest. Also, note that the HIV logMSE in the diffusion correlation is relatively higher when compared to the mammals and prokaryote examples for equivalent numbers of taxa and amounts of missing data. This is likely due to the fact that we infer relatively low heritability for the HIV traits (see Section 3.7.3) and use these values for simulation. Low heritability indicates less phylogenetic signal, that suggests more data would be needed to understand the evolutionary relationships between the different traits. For the same reason, we see the opposite pattern with the residual correlation, with lower error observed for the HIV example. See Section 3.9.4 for further simulation study results.

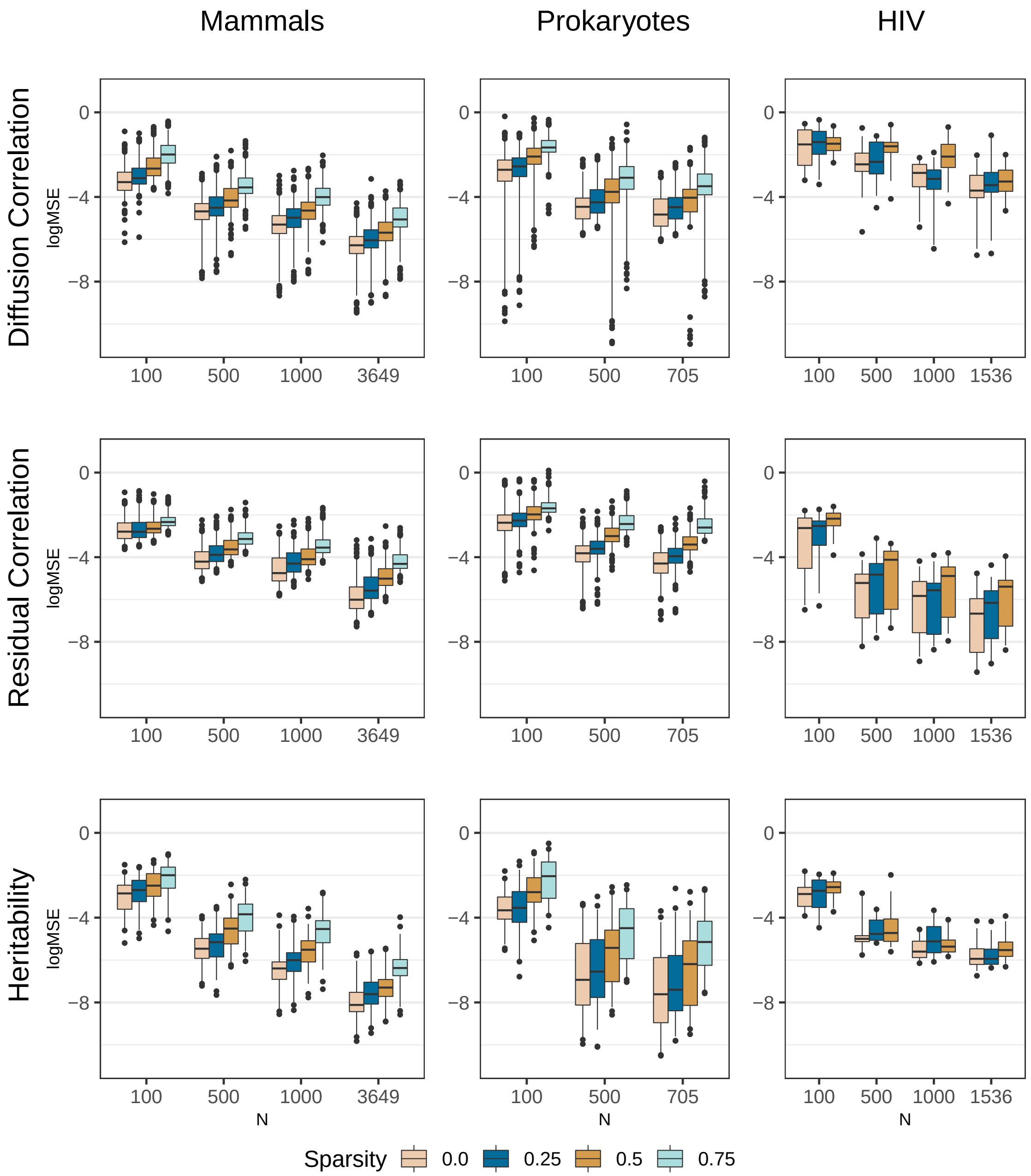


Figure 3.2: Posterior log mean squared-error of the diffusion correlation, residual correlation, and heritability over ten simulated replicates based on three empirical examples. The boxes extend from the 25th to the 75th posterior percentiles with the middle bar representing the median. The lines extend from the 2.5th through the 97.5th percentiles, with outliers depicted as dots. The sparsity depicted by different colors represents different percentages of randomly removed data.

3.7 Applications

3.7.1 Mammalian life history

A major task for life history theory is to understand the ecological and evolutionary significance of correlation between life history traits such as age at sexual maturity, the number of offspring per reproductive event, and reproductive lifespan (Roff, 2002). Establishing patterns of such correlation grants insight into whether life history variation between individuals, populations or species is consistent with pace-of-life theory (Reynolds, 2003; Réale et al., 2010). This theory predicts that 'fast' traits such as early maturity, large broods, small offspring, frequent reproduction and a short lifespan are positively associated with each other as a consequence of organisms pursuing strategies that prioritize either current or future reproduction. Existing approaches using comparative life history data to investigate fast-slow trait covariation patterns (e.g. mammals: Bielby et al., 2007; hymenoptera: Blackburn, 1991; lizards: Clobert et al., 1998; birds: Sæther and Bakke, 2000; plants: Salguro-Gómez, 2017; fish: Wiedmann et al., 2014) generally support the fast-slow hypothesis; however, results are rarely consistent across taxa. This may reflect important taxonomic differences in life history evolution, but there is concern that differences are an artifact of different methodologies (Jeschke and Kokko, 2009).

One key limitation is that previous methods have required complete data for each species. As complete measurements across a rich suite of varied life history traits are not yet available for most species, this means that researchers must choose to either reduce the number of traits or reduce the number of species included in analyses. By integrating out missing traits, we resolve this issue and analyze the life history dataset used in Capellini et al. (2015), which is based largely on the final PanTHERIA dataset (Jones et al., 2009), supplemented with measurements from Ernest (2003) and additional sources. Our analysis includes all the variables analyzed by Bielby et al. (2007) (gestation length, weaning age, neonatal body mass, litter size, litter frequency, and age at first birth) plus reproductive lifespan (maximum lifespan minus age at first birth). We include female body mass as a trait rather than analyze size-corrected residuals and log-transform and standardize all traits prior to analysis. The analysis assumes the phylogeny of Fritz et al. (2009) that remains the most complete phylogeny for mammals. In total, 3649 species in the phylogeny have measurement of at least one trait and are included. Table 3.2 reports the number of species with measurements for each trait. Only 136 species have complete data on all 8 traits; thus the ability to include species with partially missing traits enables inclusion of 932% more measurements.

Table 3.2: Missing data summary for all three examples.

Data set	Trait	Number observed	Percent missing
Mammals $N = 3649$	Body mass	3467	5.0%
	Litter size	2477	32.1%
	Gestation length	1359	62.8%
	Weaning age	1161	68.2%
	Litter frequency	888	75.7%
	Neonatal body mass	1083	70.3%
	Age at first birth	444	87.8%
	Reproductive lifespan	348	90.5%
	Total	**11227**	**61.5%**
Prokaryotes $N = 705$	Cell diameter	690	2.1%
	Cell length	657	6.8%
	Genome length	563	20.1%
	GC content	563	20.1%
	Coding sequence length	558	20.9%
	Optimal temperature	548	22.3%
	Optimal pH	487	30.9%
	Total	**4066**	**17.6%**
HIV $N = 1536$	GSVL	1536	0.0%
	SPVL	1536	0.0%
	CD4 slope	1102	28.3%
	Total	**4174**	**9.4%**

To estimate the correlation between these traits throughout mammalian evolution, we jointly model them with an MBD process on the tree with residual variance. In this analysis, we are primarily interested in the correlation between traits during the MBD process on the tree and estimate trait correlations from the marginal posterior of $\mathbf{\Sigma}$. Figure 3.3 summarizes these findings. Our results are clearly consistent with the fast-slow trait covariation patterns

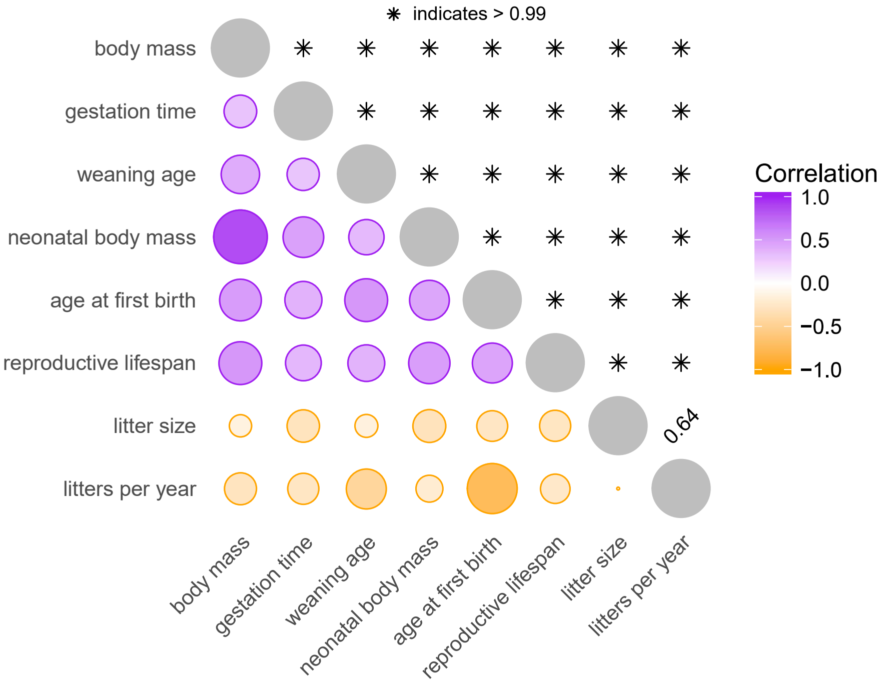


Figure 3.3: Correlation among mammalian life-history traits. The circles below the diagonal summarize the posterior mean correlation between each pair of traits. Purple represents a positive correlation while orange represents a negative correlation. Circle size and color intensity both represent the absolute value of the correlation. The numbers above the diagonal report the posterior probability that the correlation is of the same sign as its mean.

that pace-of-life theory predicts. The 'slow' life history traits (longer gestation, later weaning, larger neonatal body mass, later age at first birth, and longer reproductive lifespan) are all positively correlated with each other and negatively correlated with the two 'fast' life history traits (greater litter size and more frequent litters). All correlations are significant (determined by $< 5\%$ posterior tail probability) with the notable exception of that between litter size and litter frequency. This apparent lack of correlation may be due to the opposing effects of their joint positive correlation with body mass combined with a trade-off between these two traits that life history theory predicts (Stearns, 1989). Nevertheless, our results demonstrate that larger animals tend to have slower life history traits, confirming known patterns and reflecting the central role of body size in life history evolution.

We compare the computational efficiency of our method against that of the sampling method using the MBD model both with and without residual variance. Table 3.1 shows an increase in overall computational efficiency of two orders-of-magnitude as indicated by the change in ESS per hour. Additionally, we see that our method succeeds at reducing both the amount of computational work per MCMC sample (as indicated by the increase in samples per hour) and autocorrelation (as indicated by the increase in ESS per sample).

3.7.2 Prokaryote evolution

Comparative genomics has greatly assisted in the formulation of prokaryote evolutionary theories. Several such theories have been inspired by and tested through measuring correlation among different phenotypic and genomic traits. For example, the thermal adaptation hypothesis posits that higher GC content is involved in adaptation to high temperatures because it may offer thermostability to genetic material (Bernardi and Bernardi, 1986). The genome streamlining hypothesis attempts to explain the compactness of prokaryotic genomes through natural selection favoring small genomes (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Giovannoni et al., 2014). Sabath et al. (2013) argue that lower cell volume is an adaptive response to high temperature. The field is well-aware of the need to account for phylogenetic relationships when measuring correlation, but statistical analyses generally rely on fixed, poorly resolved trees and simple models of trait evolution.

Here, we estimate correlation among a set of genotypic and phenotypic traits while simultaneously accounting for phylogenetic uncertainty and accommodating complexity in the trait evolutionary process. We construct our data set from a study by Goberna and Verdú (2016), who collated cell diameter, cell length, optimum temperature and pH measurements for a large set of prokaryotes. Prior experience in resolving large, unknown trees suggests that we limit our analysis to less than ∼750 taxa. As such, we include all taxa with three or more measurements and a selection of the taxa with only two measurements in our analysis. For our selection of 705 taxa, we obtain data on genome length, the number of coding sequences, and GC content from the prokaryotes table in NCBI Genome. Table 3.2 presents the number of measurements for each trait. We log-transform and standardize all traits (except for GC content which we logit-transform and standardize). To infer the phylogeny, we obtain matching 16S sequences via the ARB software package (Ludwig et al., 2004) that we then align using the SINA Alignment Service (Pruesse et al., 2012) and manually edit.

Through MCMC simulation, we simultaneously infer the sequence and trait evolutionary process. We model the sequence evolutionary process using a general time-reversible model (Tavaré, 1986) with gamma-distributed rate variation among sites (Yang, 1994). We use an uncorrelated lognormal relaxed clock to model rate variation among branches (Drummond et al., 2006) and specify a Yule birth prior process on the unknown tree (Gernhard, 2008). For the trait evolutionary process, we assume an MBD model with residual variance.

Figure 3.4 displays our estimated maximum clade credibility phylogeny with associated trait measurements, and Figure 3.5 presents the phylogenetic correlation between those traits. One notable result is the positive correlation between optimal temperature and GC content (0.22 posterior mean, $[0.08, 0.37]$ 95% highest posterior density interval) that the thermal adaptation hypothesis predicts (Bernardi and Bernardi, 1986). Researchers have discussed this hypothesis for years with mixed support (Hurst and Merchant, 2001; Musto et al., 2004; Wang et al., 2006; Wu et al., 2012; Sabath et al., 2013; Aptekmann and Nadra, 2018). Our analysis, however, includes 435 taxa with measurements for both GC content and optimal growth temperature, making it the largest study we are aware of that accounts for phylogenetic relationships. Interestingly, while cell diameter and cell length are not significantly correlated, they are both positively correlated with genome length. Smaller cells have been associated with smaller genomes in both prokaryotes and unicellular eukaryotes (Shuter et al., 1983; Lynch, 2007), but the reasons for this are not fully understood (Dill et al., 2011).

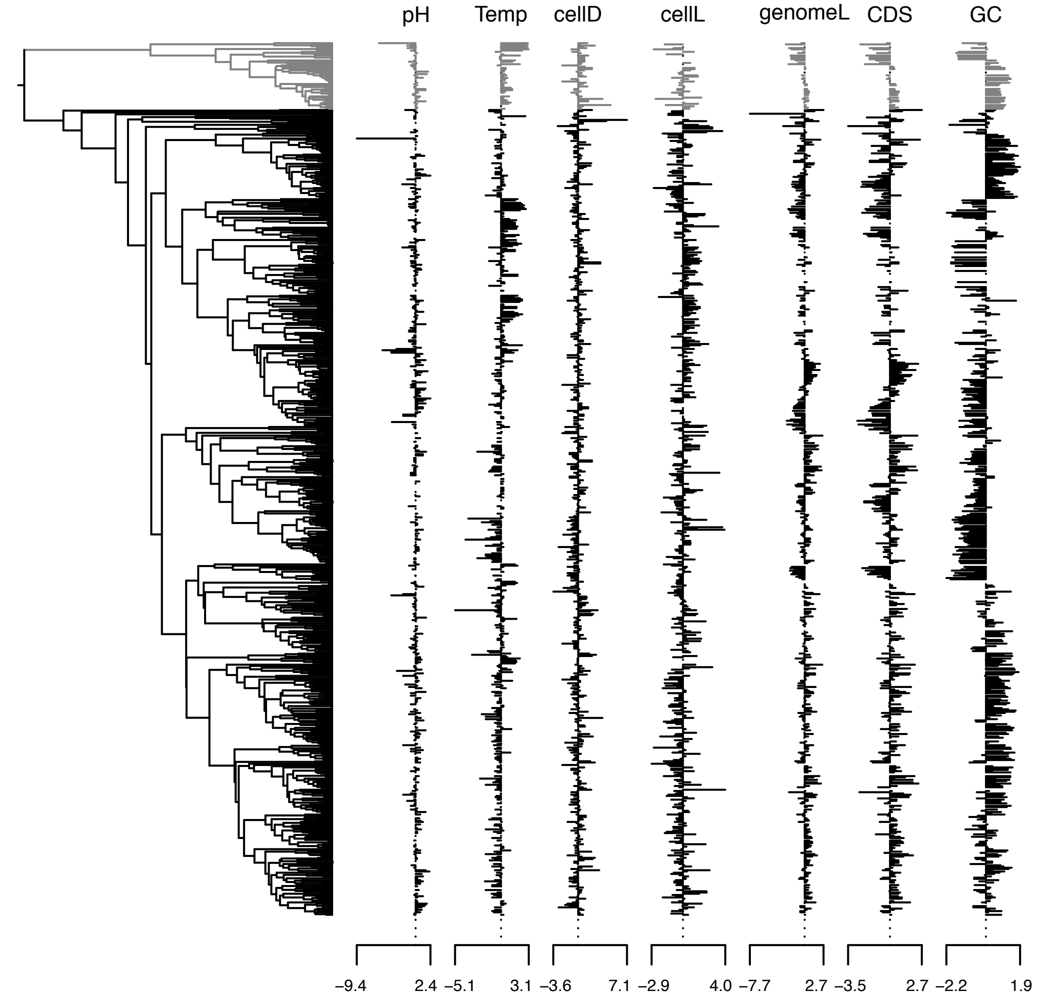


Figure 3.4: Prokaryote phylogeny and traits. The phylogeny depicts the inferred maximum clade credibility tree. The archaea clade ($N = 54$) and the associated trait measurements are depicted in grey.

We also estimate a relatively strong negative correlation between genome length and optimal temperature (-0.52 $[-0.67, -0.37]$), supporting the genomic streamlining hypothesis during thermal adaptation. Note that we do not compare computation times here, as simultaneous inference of the phylogenetic tree makes the sampling method prohibitively slow.

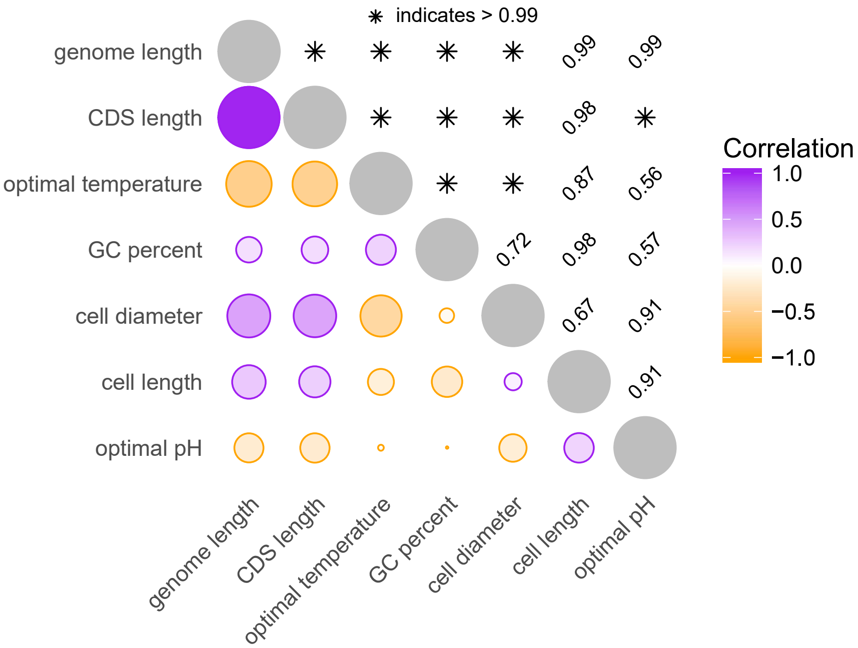


Figure 3.5: Correlation among prokaryotic growth properties. See Figure 3.3 caption.

3.7.3 HIV-1 virulence

Recent years have witnessed a strong interest in using phylogenetic comparative methods to study the heritability of HIV-1 virulence. Initially, Alizon et al. (2010) employed Pagel's λ (Pagel, 1999) to measure the extent to which HIV-1 set-point viral load reflects viral shared evolutionary history in the Swiss HIV Cohort Study (Swiss HIV Cohort Study et al., 2009) patients. A relatively high heritability estimate of set-point viral load, a predictive measure of clinical outcome, motivated others to examine to what extent the viral genotype can control for this trait (e.g. Hodcroft et al., 2014; Vrancken et al., 2015). These efforts have resulted in widely varying estimates, from 6% to 59%, prompting a discussion on the methods used to estimate the heritability of virulence (see Mitov and Stadler, 2018; Bertels et al., 2018). Here, we revisit the most comprehensive data set recently analyzed (Blanquart et al., 2017) to determine the extent to which variability in HIV-1 virulence is attributable to viral genetic variation. We focus on the dataset of subtype B viruses from Blanquart et al.

(2017) that encompasses 1581 taxa with associated measures of set-point viral load and CD4 cell count decline. We rely on the maximum likelihood phylogeny inferred for this data set, but convert it to a time-measured tree with dated tips using a heuristic procedure (To et al., 2016). A prior examination of the correlation between sampling time and root-to-tip divergence using TempEst (Rambaut et al., 2016) indicated the presence of outliers, most of which can be attributed to a basal lineage in the phylogeny. As the subtyping of the taxa in this basal lineage also was ambiguous (Blanquart, personal communication), we remove this lineage (36 taxa) together with 9 other outlier taxa. We note that this resulted in a time to the most recent common ancestor (TMRCA) estimate of about 1960 that is much more in line with a recent subtype B TMRCA estimate (1967, 95% Bayesian credibility interval of 1963–1970; Worobey et al., 2016) than the estimate including the basal lineage (∼1930).

Two measures of set-point viral load are available for all remaining taxa: (i) one based on a standardized choice of assay on a single sample taken between 6 and 24 months after infection and before the initiation of antiretroviral therapy ("gold standard viral load", GSVL) and (ii) a more classical measure of set-point viral load (SPVL) based on the mean of all log viral loads measured between 6 and 24 months after infection. Figure 3.6 presents the phylogeny and associated trait values. To estimate heritability of both set-point viral load measures and CD4 slope, we model all three measurements as a multivariate trait in our MBD model with residual variance and approximate the posterior distribution of the heritability statistic $\mathbf{H}$ via MCMC. Our estimated heritabilities are 0.21 $[0.11, 0.3]$ for GSVL, 0.18 $[0.1, 0.26]$ for SPVL, and 0.16 $[0.07, 0.25]$ for CD4 cell decline. These estimates are consistent with similar estimates reported by Blanquart et al. (2017).

We further asses model fit by assessing predictive performance of GSVL on SPVL. We omit CD4 slope from our analysis as it is measured concurrently with SPVL. We randomly remove 5% of the SPVL measurements from the data set and consider four different models. We consider both a bivariate case where we assume a multivariate process and a univariate case where we analyze SPVL alone. For both the bivariate and univariate cases, we use the

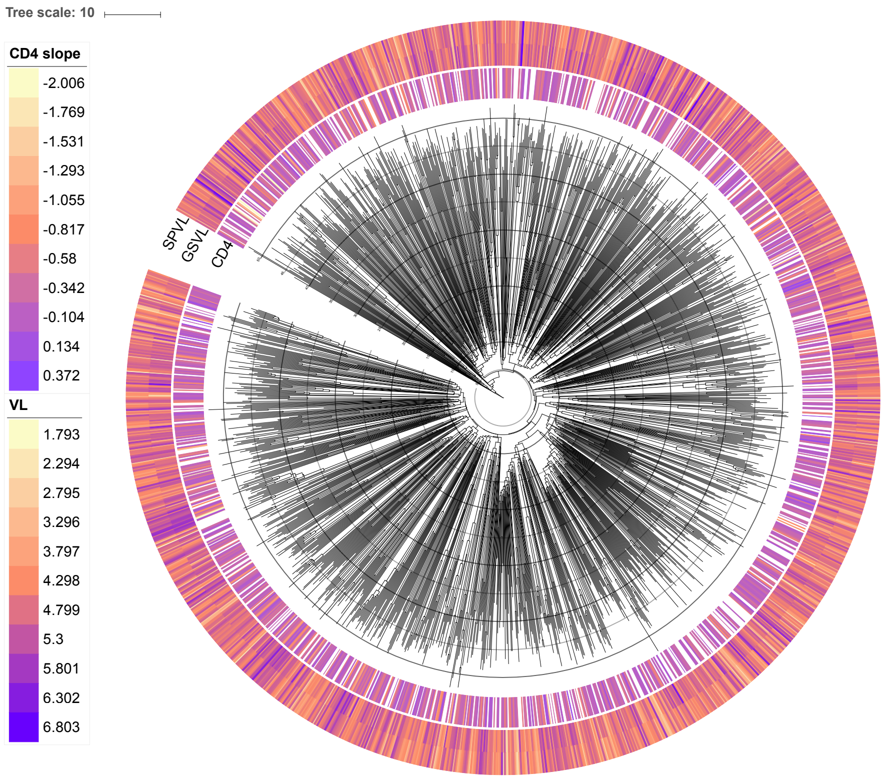


Figure 3.6: HIV-1 phylogeny with associated CD4 slope, SPVL, and GSVL values for each viral host.

MBD model both with and without the residual variance extension. For each removed SPVL measurement, we compute the mean squared error (MSE) between the predicted and true values. We repeat each analysis 50 times and report the cumulative results in Figure 3.7, from which two results emerge. First, the MSE of prediction in the bivariate cases are lower than those in the univariate cases. This is unsurprising given the strong correlation between SPVL and GSVL. Second, addition of residual variance to the model results in modestly better prediction of SPVL in both the bivariate and univariate cases. This emphasizes the importance of including model extensions like residual variance in these analyses.

We again demonstrate improvements in computational efficiency (see Table 3.1). While less dramatic than the mammals example, we still see an order-of-magnitude increase in

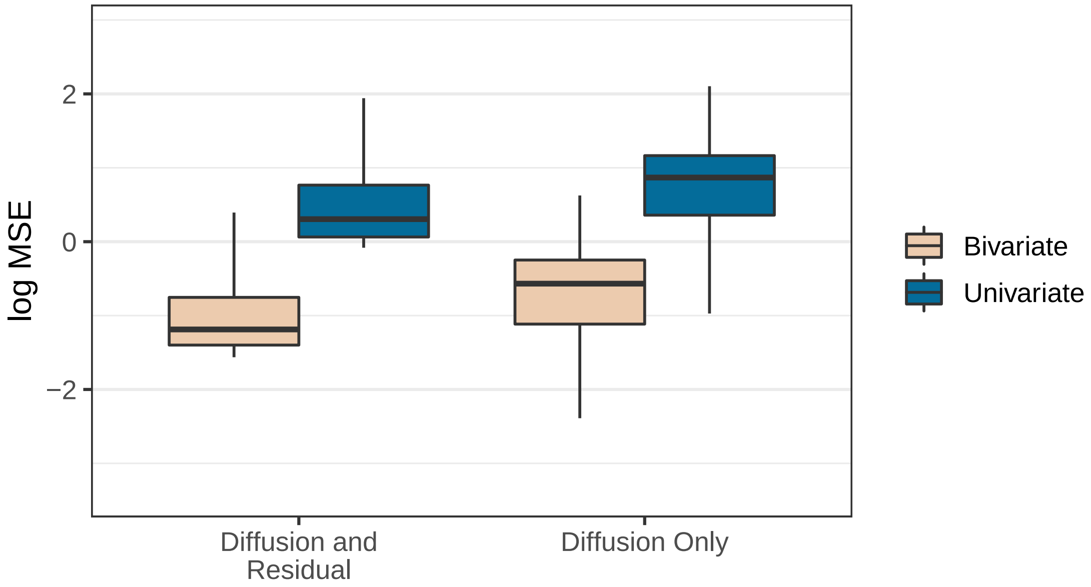


Figure 3.7: Model predictive performance of HIV set-point viral load. Each box-and-whisker plot depicts the posterior mean-squared-error of prediction under a different model. The boxes represent the interquartile range, while the lines extend to include the 2.5^{th} through 97.5^{th} percentiles. Outliers are omitted.

effective sample size per hour in the MBD model without residual variance. This attenuation is to be expected, as there are far fewer missing measurements in the HIV data set than the mammal data set. Nevertheless, our method still outperforms the sampling method in the simple MBD model even when only 9.4% of data points are missing. For the model with residual variance, our method outperforms the sampling method by two orders-of-magnitude.

3.8 Discussion

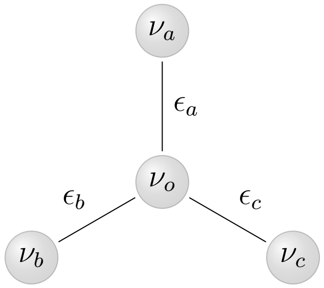
Oftentimes comparative biologists are interested in phylogenetically adjusted methods for assessing relationships between traits of organisms. However, frequently when the number of taxa grows large the level of missing data increases, making inference challenging. Here, we have developed a method for evaluating the likelihood of observed traits given a tree while integrating out missing values analytically that dramatically outperforms current best-practice methods. In the mammalian data set, with $N = 3649$ and 61.5% missing data, we achieve a minimum effective sample size per hour 400× greater than previous methods. This increase in speed brings computation times down from more than a week to less than an hour. Even in the more tractable HIV data set, with $N = 1536$ and 9.4% missing data, we increase the minimum ESS per hour by a factor of 65. Both increases in speed are due to an overall decrease in both autocorrelation between MCMC samples and the amount of computational work required per sample. Importantly, this increase in computational efficiency allows for previously intractable analyses on large trees. Specifically, we incorporate residual variance into the model and (in the prokaryotes example) simultaneously infer $\mathbf{\Sigma}$, $\mathbf{\Gamma}$, and an unknown phylogeny $\mathscr{F}$. Further, the residual variance extension is only one of several potential extensions. Other possible extensions could incorporate data sets with repeated measurements at the tips of the tree and factor analyses (Tolkoff et al., 2018).

Additionally, our strategy could be used in a more diverse array of phylogenetic models than the fixed-rate MBD process. Recently, Fisher et al. (2021) have used our method in a scale-mixture of multivariate normals diffusion model where there is a different evolutionary rate on each of the tree branches. This model assumes that the rate of evolution changes over time and across taxa. Moreover, these methods also easily translate to multi-optima Ornstein–Uhlenbeck (OU) diffusions, where there is some (potentially changing) optimum trait value that traits tend to evolve toward. Following from Bastide et al. (2018), a modified version of our method has already been implemented for the OU process in BEAST.

We also note that our pre-order missing data augmentation algorithm presented in Section 3.2.2.1 has far broader utility than computing the conjugate Wishart statistics. Notably, it allows for joint sampling of all missing values in linear-time. As such, this data augmentation procedure serves as a bridge between any data set with missing data and statistical methods that require complete data. Such cases occur, for example, in computing the residual sum of squares in phylogenetic mixed models (Lynch, 1991) as well as the gradient of the log likelihood with respect to the model parameters.

An important limitation of our and previous methods is that they assume an ignorable missing data mechanism (i.e. that the data are missing at random and that the prior on any model parameters is independent of the missing data mechanism). Note that this is assumption is not as restrictive as it seems as we only require that the data are *missing and random* and not necessarily *missing completely at random* (Little and Rubin, 1987). While these conditions may hold in some comparative biology examples, possible violations abound. Any solution to this problem would necessarily depend on the specific missing data mechanism. One commonly used missing data mechanism is the thresholding model where data above or below some limit are omitted from the analysis. This could occur, for example, when there is some minimum detection limit below which a value cannot be measured. To explicitly account for these omissions, we could modify our model to assume the observed data at the tips are drawn from a truncated multivariate-normal distribution rather than a full multivariate normal distribution. Under this model, the observed data likelihood would remain the same up to a normalizing constant and indicator function. As the distribution of the internal nodes would remain un-truncated and the Gaussian kernel on all nodes would remain unchanged, our likelihood calculation algorithm would remain largely unchanged. For the likelihood computation, the normalizing constants and indicator functions would simply be propagated up the tree in the same way as the integration remainders r_i. One challenge of this approach would be to compute the normalizing constants for all taxa with missing data. This may be particularly challenging as, depending on the specific missing data mechanism, these constants may depend on the latent trait values immediately internal to the tip nodes. An additional challenge to this approach would be to formalize the distribution of the missing data so that we could appropriately apply our pre-order data augmentation algorithm. We may simply be able to draw each missing value from their un-truncated full conditional distribution, but more work would be necessary to determine whether this augmentation regime is appropriate. We leave these challenges as future work.

Finally, and perhaps most importantly, we propose our method as a special case solution



Sample covariance matrices

$$\mathbf{M}_1 = \begin{pmatrix} \epsilon_a & 0 & 0 \\ 0 & \epsilon_b & 0 \\ 0 & 0 & \epsilon_c \end{pmatrix} \quad \mathbf{M}_2 = \begin{pmatrix} \epsilon_a & \epsilon_a & \epsilon_a \\ \epsilon_a & \epsilon_b + \epsilon_a & \epsilon_a \\ \epsilon_a & \epsilon_a & \epsilon_c + \epsilon_a \end{pmatrix}$$

Figure 3.8: An acyclic graph with nodes $\{\nu_o, \nu_a, \nu_b, \nu_c\}$ and edge weights $\{\epsilon_a, \epsilon_b, \epsilon_c\}$. The covariance matrix $\mathbf{\Lambda} = \{\Lambda_{ij}\}$ is additive on an acyclic graph if each Λ_{ij} is equal to the sum of the shared non-negative edge-weights in the paths from ν_i and ν_j to some origin node. For example, the matrix $\mathbf{M}_1$ is additive for nodes $(\nu_a, \nu_b, \nu_c)^t$ with ν_o at the origin, while the matrix $\mathbf{M}_2$ is additive for nodes $(\nu_o, \nu_b, \nu_c)^t$ with ν_a at the origin.

to the long-standing statistical problem involving multivariate normal distributions with missing data. Specifically, our method applies to any MVN distribution with a three-point structured covariance matrix (see Ho and Ané, 2014). Intuitively, this condition arises in covariance matrices generated from processes that are additive on an acyclic graph (see Figure 3.8). This restriction, however, is not overly limiting and applies to a broad range of normal models including multilevel hierarchical models and matrix-normal distributions such as the one we use here. Additionally, our pre-order data augmentation procedure enables $\mathcal{O}(N)$ imputation in these highly structured models. While Allen and Tibshirani (2010) and Glanz and Carvalho (2018) have utilized the EM algorithm (Dempster et al., 1977) to efficiently perform maximum likelihood imputation in similar problems, our method could serve as an alternative for approaches that base inference on the observed-data likelihood.

3.9 Appendix

3.9.1 Matrix inversion computations

To evaluate the observed data likelihood, we must compute branch-deflated precisions $\mathbf{P}_i^\star = \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \mathbf{\Sigma} \boldsymbol{\delta}_i\right)^-$ for $i = 1, \ldots, 2N-2$. We demonstrate below that this matrix exists and is well-defined under the definition of our pseudo-inverse. Using the permutation matrix $\mathbf{C}_i$

from Section 3.2.1.1, we decompose the diffusion variance $\mathbf{\Sigma}$ and node precision $\mathbf{P}_i$ such that

$$\begin{aligned}
\mathbf{\Sigma} &= \mathbf{C}_i \begin{pmatrix} \mathbf{\Sigma}_i^{\text{obs}} & \mathbf{\Sigma}_i^{\text{ol}} & \mathbf{\Sigma}_i^{\text{om}} \\ - & \mathbf{\Sigma}_i^{\text{lat}} & \mathbf{\Sigma}_i^{\text{lm}} \\ - & - & \mathbf{\Sigma}_i^{\text{mis}} \end{pmatrix} \mathbf{C}_i^t \text{ and} \\
\mathbf{P}_i &= \mathbf{C}_i \text{diag}\left[\infty\mathbf{I}, \tilde{\mathbf{P}}_i, 0\mathbf{I}\right] \mathbf{C}_i^t,
\end{aligned}$$

for $i = 1, \ldots, 2N-2$. We use this decomposition to identify that:

$$\begin{aligned}
\mathbf{P}_i^\star &= \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \mathbf{\Sigma} \boldsymbol{\delta}_i\right)^- \\
&= \mathbf{C}_i \left(\left(\text{diag}\left[\infty\mathbf{I}, \tilde{\mathbf{P}}_i, 0\mathbf{I}\right]\right)^- + \text{diag}\left[t_i \begin{pmatrix} \mathbf{\Sigma}_i^{\text{obs}} & \mathbf{\Sigma}_i^{\text{ol}} \\ - & \mathbf{\Sigma}_i^{\text{lat}} \end{pmatrix}, 0\mathbf{I} \right] \right)^- \mathbf{C}_i^t \\
&= \mathbf{C}_i \left(\text{diag}\left[0\mathbf{I}, \tilde{\mathbf{P}}_i^{-1}, \infty\mathbf{I}\right] + \text{diag}\left[t_i \begin{pmatrix} \mathbf{\Sigma}_i^{\text{obs}} & \mathbf{\Sigma}_i^{\text{ol}} \\ - & \mathbf{\Sigma}_i^{\text{lat}} \end{pmatrix}, 0\mathbf{I} \right] \right)^- \mathbf{C}_i^t \\
&= \mathbf{C}_i \left(\text{diag}\left[\mathbf{T}, \infty\mathbf{I}\right]\right)^- \mathbf{C}_i^t \\
&= \mathbf{C}_i \, \text{diag}\left[\mathbf{T}^{-1}, 0\mathbf{I}\right] \mathbf{C}_i^t,
\end{aligned} \tag{3.26}$$

where

$$\mathbf{T} = \text{diag}\left[0\mathbf{I}, \tilde{\mathbf{P}}_i^{-1}\right] + t_i \begin{pmatrix} \mathbf{\Sigma}_i^{\text{obs}} & \mathbf{\Sigma}_i^{\text{ol}} \\ - & \mathbf{\Sigma}_i^{\text{lat}} \end{pmatrix} = \begin{pmatrix} t_i \mathbf{\Sigma}_i^{\text{obs}} & t_i \mathbf{\Sigma}_i^{\text{ol}} \\ - & \tilde{\mathbf{P}}_i^{-1} + t_i \mathbf{\Sigma}_i^{\text{lat}} \end{pmatrix}. \tag{3.27}$$

The matrix $\mathbf{T}$ is the sum of a positive-definite matrix and positive-semidefinite matrix and is therefore invertible.

3.9.2 Heritability statistic

We compute the expectation of the empirical variance $\mathbb{E}[\mathbf{S}^2(\mathbf{Y})]$ under the MBD model with residual variance as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] &= \mathbb{E}\left[\frac{1}{N}\left(\mathbf{Y}-\bar{\mathbf{Y}}\right)^t\left(\mathbf{Y}-\bar{\mathbf{Y}}\right)\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{2}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}+\frac{1}{N^2}\mathbf{Y}^t\mathbf{J}_N\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{2}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}+\frac{1}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{1}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{y}_i\mathbf{y}_i^t\right]-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\mathbf{y}_i\mathbf{y}_j^t\right].
\end{aligned} \tag{3.28}$$

The multivariate normal distribution of $\operatorname{vec}[\mathbf{Y}]$ implies $\operatorname{Cov}(Y_{ik},Y_{jl}) = \Sigma_{kl}\Upsilon_{ij}+\Gamma_{kl}^{-1}1_{\{i\}}j$ where $1_{\{i\}}j$ is an indicator function. Using this information in Equation 3.28,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] &= \frac{1}{N}\sum_{i=1}^{N}\left(\Upsilon_{ii}\mathbf{\Sigma}+\mathbf{\Gamma}^{-1}+\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}[\mathbf{y}_i]^t\right) \\
&\quad -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\Upsilon_{ij}\mathbf{\Sigma}+\mathbf{\Gamma}^{-1}1_{\{i\}}j+\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}\left[\mathbf{y}_j\right]^t\right) \\
&= \frac{1}{N}\operatorname{tr}[\mathbf{\Upsilon}]\,\mathbf{\Sigma}+\mathbf{\Gamma}^{-1}-\left(\frac{1}{N^2}\mathbf{1}_N^t\mathbf{\Upsilon}\mathbf{1}_N\right)\mathbf{\Sigma}-\frac{1}{N}\mathbf{\Gamma}^{-1} \\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}[\mathbf{y}_i]^t-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}\left[\mathbf{y}_j\right]^t.
\end{aligned} \tag{3.29}$$

Note that $\mathbb{E}[\mathbf{y}_i]=\mathbf{y}_{2N-1}$ for $i=1\ldots N$, which implies

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}[\mathbf{y}_i]^t-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\mathbf{y}_i]\,\mathbb{E}\left[\mathbf{y}_j\right]^t=0. \tag{3.30}$$

Table 3.3: Likelihood calculation speed comparison between BEAST and PCMBaseCpp. Each data set was run 10 times for 1,000 likelihood evaluations each. We report the median likelihood evaluations per second and speed-up over the 10 runs.

Data set	N	P	Likelihood evaluations/sec		Speed-up
			BEAST	PCMBaseCpp	
Prokaryotes	705	7	240	40	6.0×
HIV	1536	3	490	67	7.2×
Mammals	3649	8	60	12	5.1×

As such, our expression for the expected empirical variance reduces to the following:

$$\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] = \frac{N-1}{N}\boldsymbol{\Gamma}^{-1} + \left(\frac{1}{N}\text{tr}\left[\boldsymbol{\Upsilon}\right] - \frac{1}{N^2}\mathbf{1}_N^t\boldsymbol{\Upsilon}\mathbf{1}_N\right)\boldsymbol{\Sigma}. \tag{3.31}$$

3.9.3 Comparison with PCMBaseCpp

As our algorithm for efficiently computing the likelihood with incomplete trait measurements relies on a similar strategy as that presented by Mitov et al. (2020), we compare the likelihood computation speed of our BEAST (Suchard et al., 2018) implementation against and the PCMBaseCpp implementation. We record the time it takes to evaluate the likelihood 1,000 times using the data and trees from all three examples we discuss in the text, and repeat this ten times for each example. We report the median likelihoods per second in Table 3.3. We also perform the same comparisons with simulated trees and data sets, and report these results in Table 3.4.

Note that while we do show consistently faster likelihood evaluations than PCMBase, we do not believe that our implementation is necessarily "better" than that of Mitov et al. (2020). The primary difficulty in comparing the speed of the two software packages is that we implement our software in different languages (BEAST in Java and PCMBase in R and C++), and the specific Java and C++ compilers used could influence their speed. It is

Table 3.4: Likelihood calculation speed comparison between BEAST and PCMBaseCpp on simulated data. For each N, P combination, data was simulated 10 times under random conditions and run for 1,000 likelihood evaluations each. We report the median likelihood evaluations per second and speed-up over the 10 runs.

N	P	Likelihood evaluations/sec		Speed-up
		BEAST	PCMBaseCpp	
100	2	3300	1300	2.6×
100	10	690	180	3.8×
100	20	220	26	8.3×
1,000	2	780	170	4.5×
1,000	10	100	13	7.9×
1,000	20	25	2.8	8.8×
10,000	2	82	16	5.1×
10,000	10	11	1.7	6.4×
10,000	20	2.5	0.29	8.7×

difficult to determine the exact sources of the differences in speed without testing both implementations on a wide range of computer architectures and compilers.

Nevertheless, the PCMBase / PCMFit packages and BEAST are fundamentally different in that PCMFit relies on maximum likelihood estimation (MLE) while BEAST performs Bayesian inference. The MLE framework is certainly appropriate when the phylogenetic tree is known with a high degree of certainty, but poses problems when the phylogenetic tree is unknown and must be jointly inferred with the trait evolutionary process. Specifically, MLE will likely produce biased results and has difficulty constructing confidence intervals that take into account the uncertainty of the tree. From the Bayesian perspective, however, we can simply integrate out the tree via Markov Chain Monte Carlo, that results in posterior estimates of the trait evolution parameters that accurately reflect the uncertainty of the tree.

3.9.4 Simulation study

The setup of our simulation study is described in Section 3.6. Figures 3.9, 3.10, and 3.11 present the results of our simulation study. In general, results indicate that our inference machinery is sufficiently well-powered to accurately and precisely recapture the parameters used to simulate the data. All parameters of interest achieve low posterior mean squared error (MSE) when all available taxa are included. Additionally, there is no apparent bias in our parameter estimation with the notable exception of the diagonal heritabilities. Note that despite the fact that there is some bias in the heritability estimates, they also achieve low logMSE and are indeed close to their "true" values. We believed the induced prior on the diagonal heritabilities may be responsible for this bias, but have not fully explored this phenomenon. Regardless, these results suggest that (conditioning on the model being appropriate) our results accurately reflect biological reality.

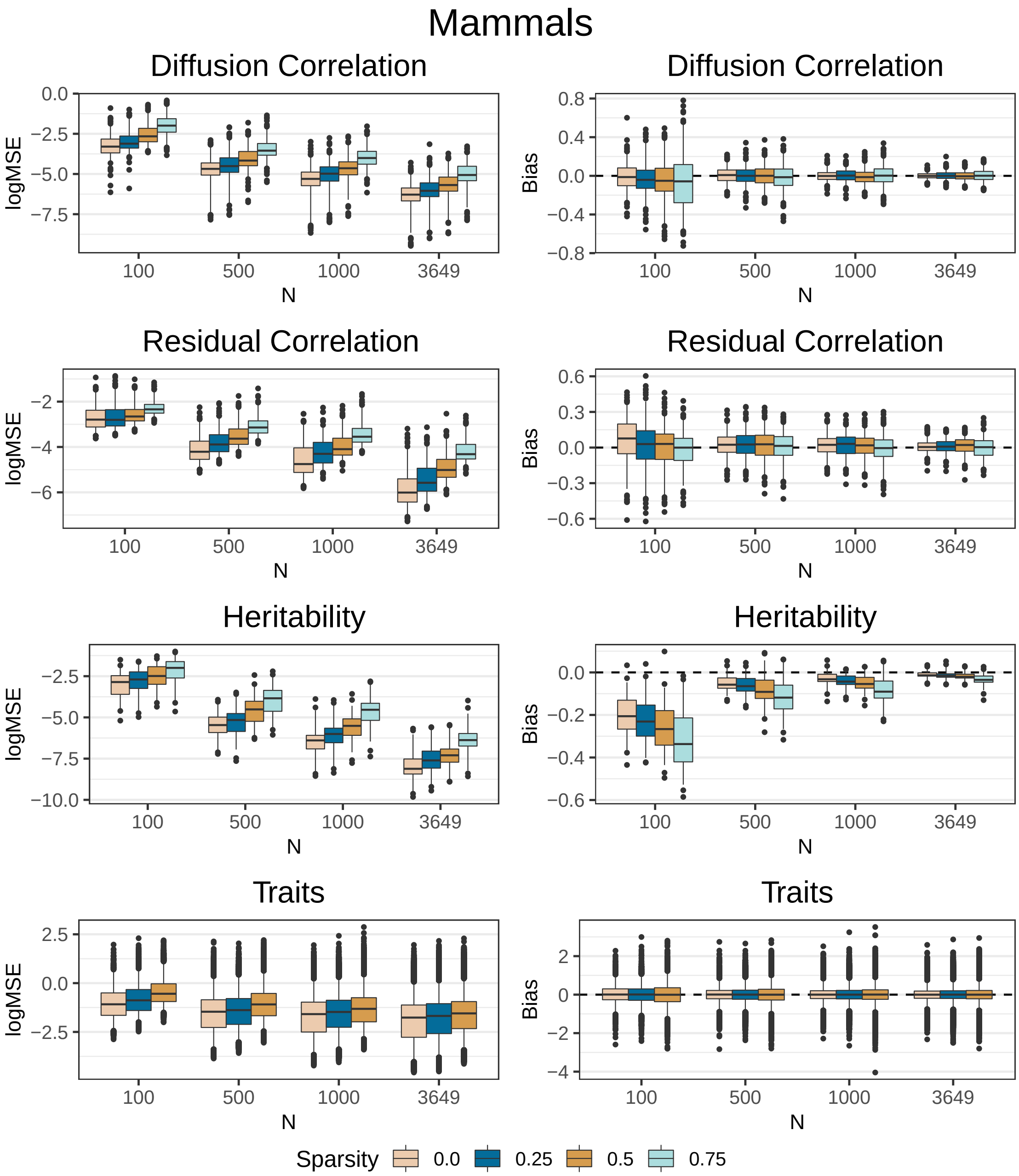


Figure 3.9: Mammals simulation study. Posterior log mean squared-error and bias of the parameters of interest over ten simulated replicates. The boxes extend from the 25th to the 75th posterior percentiles with the middle bar representing the median. The lines extend from the 2.5th through the 97.5th percentiles, with outliers depicted as dots. The sparsity depicted by different colors represents different percentages of randomly removed data.

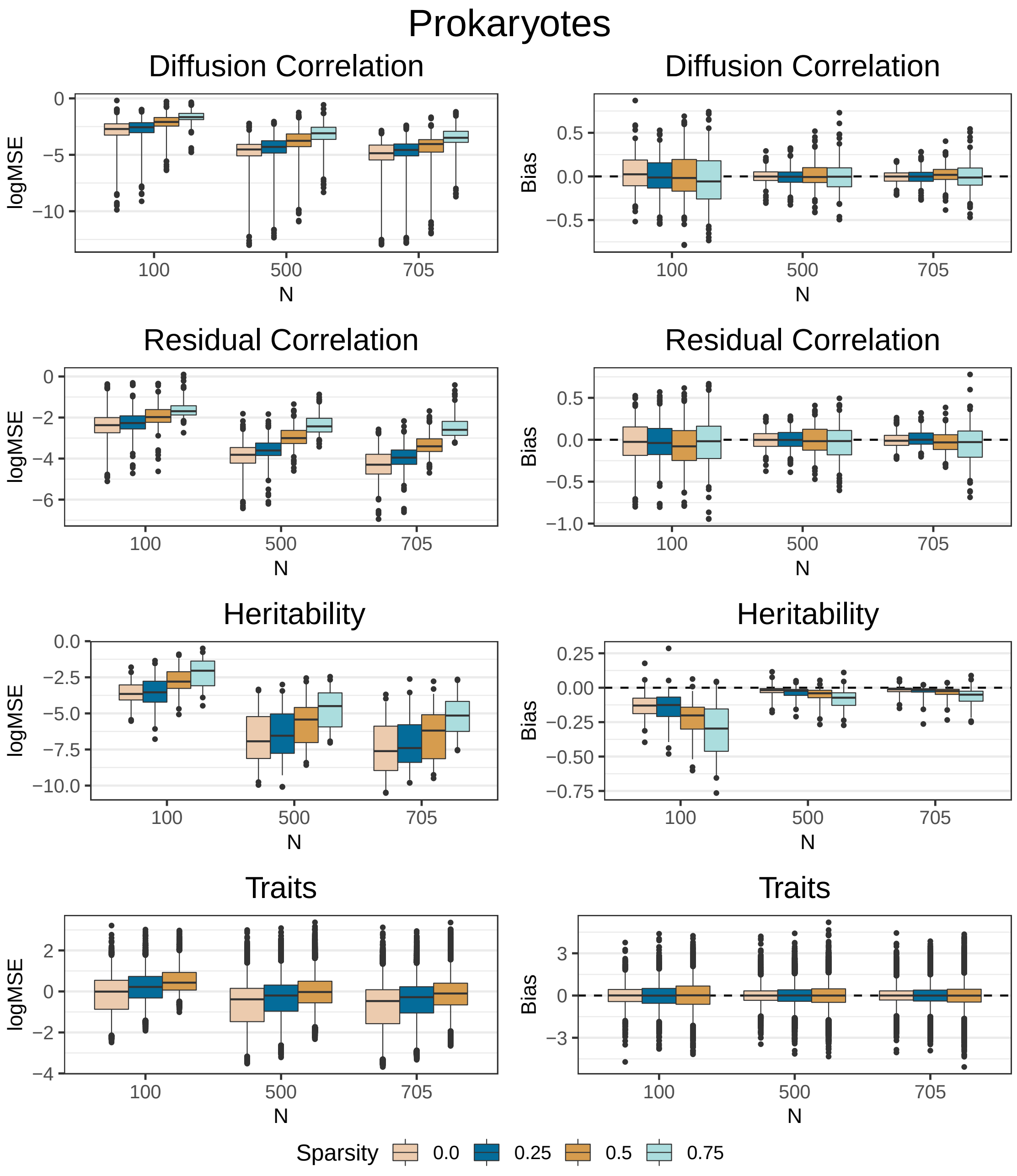


Figure 3.10: Prokaryote simulation study results. See Figure 3.9 for description.

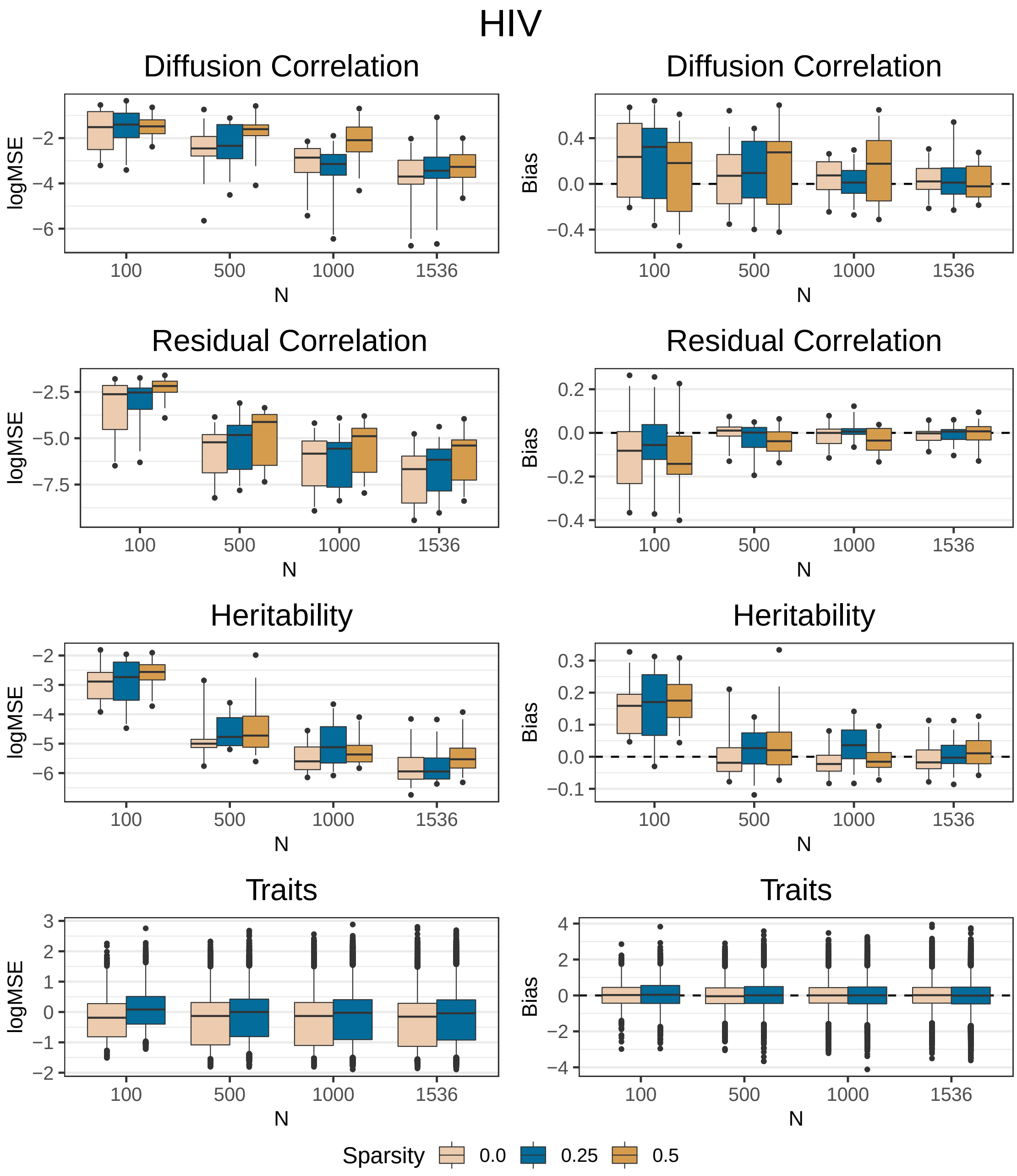


Figure 3.11: HIV-1 simulation study results. See Figure 3.9 for description.

CHAPTER 4

Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis

4.1 Introduction

Biological phenotypes are the result of numerous evolutionary forces acting in complex and often conflicting ways throughout an organism's evolutionary history. Phylogenetic comparative methods seek to untangle this web of selective pressures and elucidate the forces that have shaped organisms over time. As implied by their name, these methods compare phenotypes across numerous biological taxa connected by a phylogenetic tree that captures their shared evolutionary history. Accounting for shared evolutionary history via the phylogeny is necessary to avoid biased inference, as this shared history implies phenotypes are non-independent across taxa. Statistical models that inappropriately ignore this dependence can identify spurious associations between phenotypes (Felsenstein, 1985b). However, accounting for these relationships between taxa poses challenges to statistical inference.

Starting with Felsenstein (1985b), there has been much work developing computationally efficient phylogenetic comparative methods (see Rohlf, 2001; Revell and Harmon, 2008; Pybus et al., 2012; Ho and Ané, 2014). While methods development has typically focused on scaling inference to large trees, these methods struggle to accommodate data with a large number of traits or high-dimensional phenotypes. Most approaches scale quadratically or cubically with the number of traits, making inference intractable as the number of traits increases. Additionally, methods that estimate the evolutionary correlation structure between traits are difficult to interpret for data sets with high-dimensional phenotypes, as the number of pairwise correlations requiring interpretation scales quadratically with the number of traits.

4.1.1 Why phylogenetic factor analysis?

Phylogenetic factor analysis (PFA, Tolkoff et al., 2018) provides an all-in-one approach to high-dimensional comparative analyses that simultaneously simplifies complex data via dimension reduction, similar to phylogenetic principal component analysis (pPCA, Revell, 2009), and statistically evaluates evolutionary correlations between groups of phenotypes, as with phylogenetic independent contrasts (Felsenstein, 1985b). In Section 4.6.1, for example, we use PFA to understand the relationship between 11 floral phenotypes and pollinator species in columbines. We identify two axes along which floral phenotypes evolve: a first differentiating hummingbird pollination from hawk moth pollination and a second capturing phenotypes differentiating bumblebee pollination from the latter two pollination strategies. Similarly, in Section 4.6.2, we explore evolutionary relationships between 82 phenotypes of industrial yeast: growth rates under 62 different stress conditions, production of 16 metabolites and 4 metrics related to reproduction. In this example, we identify a group of phenotypes characterizing the early domestication of beer yeast. Additionally, PFA allows for flexible model specifications. For example, in Section 4.6.3 we study the evolution of life history strategies in mammals. We structure the PFA model to isolate the influence of a particular trait (body size) so that we can infer size-independent patterns of life history evolution. Finally, as with pPCA, researchers can employ PFA as a descriptive technique useful for identifying and visualizing low-dimensional structure in high-dimensional data (see Section 4.6.4 for an example of this with New World monkey brain shape). Unlike pPCA, however, Bayesian PFA incorporates uncertainty into the loadings (the analogs of the pPCA weights) and factors (the analogs of the pPCA scores).

4.1.2 Statistical developments in high-dimensional trait analyses

As the primary motivation of PFA is analyzing high-dimensional trait data, we briefly discuss existing methods that deal with the computational and interpretive burden of high-dimensional phenotypes. As mentioned above, pPCA (Revell, 2009) is one such solution that constructs a low-dimensional, phylogenetically-informed summary of the relationships between traits. More recently, several distance-based methods have been developed by Adams (2014a,b,c) to study phylogenetic signal, high-dimensional phylogenetic regression and evolutionary rates, respectively.While these methods are statistically efficient for high-dimensional phenotypes, they rely on operations that scale cubically with the number of taxa and may struggle computationally with very large trees or in cases where they must be applied over many large trees. Additionally, existing implementations of pPCA and the Adams (2014a,b,c) distance-based methods do not readily accommodate missing data, a common scourge in many relevant data sets. PFA (Tolkoff et al., 2018) adapts the Bayesian latent factor model of Aguilar and West (2000) to the phylogenetic context.Like pPCA, PFA is a linear dimension reduction approach that assumes the P-dimensional data arise from K latent factors that evolve independently along a phylogenetic tree. Unlike pPCA, PFA readily accommodates missing data without data imputation or augmentation. Additionally, PFA fits seamlessly into Bayesian phylogenetic inference and estimates the uncertainty of the influence of a particular factor on a particular trait. However, the inference regime proposed by Tolkoff et al. (2018) scales quadratically with the number of taxa and is intractable for large trees.

Finally, Clavel et al. (2019) propose a penalized likelihood framework for studying high-dimensional phenotypes. While this procedure involves an operation that scales quadratically in number of taxa, the rate-limiting calculations scale linearly in the number of taxa but cubically in the number of traits. Nevertheless, Clavel et al. (2019) demonstrate success handling data sets with more than a thousand traits. While PFA reduces the size of the parameter space by assuming the between-trait covariance is low-rank, the penalized likelihood

Table 4.1: Example of how the ordering of three hypothetical traits (A, B and C) influences results in a simple two-factor model under the assumptions made by Tolkoff et al. (2018).

	trait order 1: A, B, C	trait order 2: B, A, C
first factor	captures relationships of trait A with traits B and C	captures relationships of trait B with traits A and C
second factor	captures relationships between traits B and C independent of A	captures relationships between traits A and C independent of B

approach of Clavel et al. (2019) achieves a similar goal by assuming *a priori* that relatively few of the between-trait covariances are non-zero. The specific implementations also differ in that Clavel et al. (2019) rely on maximum likelihood inference while our work here and Tolkoff et al. (2018) approach PFA from a Bayesian perspective.

4.1.3 A new approach to PFA

We propose two new PFA inference regimes that each scale linearly with both the number of traits P and the number of taxa N. While Tolkoff et al. (2018) rely on data augmentation, our new methods rely on a novel likelihood-calculation algorithm that analytically integrates out the latent factors. We also address two other shortcomings of PFA and latent factor models generally. First, Tolkoff et al. (2018) constrain the factor loadings matrix to be upper triangular, which induces an implicit ordering to the phenotypes. Specifically, the first trait is influenced only by the first factor, the second trait is influenced only by the first two factors, etc. until the K^{th} trait and beyond which are influenced by all K factors (see Table 4.1 for an example). As justifying a specific ordering of the phenotypes *a priori* can be difficult, we extend an alternative constraint proposed by Holbrook et al. (2016) that eliminates such ordering. Second, a common challenge in exploratory factor analysis generally is determining an appropriate number of factors. As such, we implement a cross-validation model selection procedure that identifies the number of factors that confers the best predictive performance.

To facilitate use among researchers seeking to employ these methods, we develop an analysis plan with practical guidance on the most significant modeling and inference decisions. We codify this plan in the Julia package PhylogeneticFactorAnalysis.jl, which uses relatively simple instructions to automatically perform model selection and run more complex analyses in the Bayesian phylogenetic inference software BEAST (Suchard et al., 2018).

For clarity, we emphasize which methods below are completely new statistical innovations and which are novel applications of previously developed statistical practices. The calculations in Sections 4.3.1.2 and 4.3.2.1 that allow inference of the loadings without conditioning on the latent factors are novel, and we are unaware of any similar work in the statistics literature. The fast likelihood calculations in Section 4.2.1.1 are based on earlier work by Hassler et al. (2020, Section 3.2.1.2) but require non-trivial adjustment for application to this context (see Section 4.8.1). Finally, the modeling decisions described in Section 4.2.2 and inference techniques described in Sections 4.3.1.1, 4.3.1.3 and 4.3.2 are previously developed statistical procedures that find novel application to phylogenetic comparative methods here.

4.1.4 Brief overview

PFA allows researchers to identify high-dimensional patterns of trait variation using a model that reduces the computational and interpretive burden of high-dimensional analyses. We begin by specifying the technical details of the PFA model in Section 4.2. Intuitively, PFA assumes that the evolution of high-dimensional trait data can be approximated by the evolution of some small number of latent (unobserved) factors, with each of these latent factors influencing the observed traits in some estimable way. In Section 4.3 we present the technical details of several approaches to statistical inference under this model, and in Section 4.4 we compare the computational efficiency of these various approaches. As we recognize that researchers seeking to use these methods face an array of technical modeling and inference decisions, we devote Section 4.5 to practical guidance on how to make these decisions. Finally, in Section 4.6 we demonstrate the utility of PFA on 4 real-world examples.

4.2 Phylogenetic latent factor model

We approach inference from a Bayesian perspective and propose two statistical models which share a likelihood but have distinct priors. As we discuss below, each model has advantages under different circumstances, and allowing researchers to choose a model (with our guidance) offers maximum flexibility while keeping modeling decisions to a minimum.

4.2.1 Likelihood

Both statistical models share the same latent factor likelihood introduced by Tolkoff et al. (2018). This likelihood assumes the $N \times P$ trait data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^t$ arise from $N \times K$ latent factors $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^t$ via the linear transformation $\mathbf{Y} = \mathbf{FL} + \boldsymbol{\epsilon}$, where $\mathbf{L}$ is a $K \times P$ loadings matrix that must be inferred and $\boldsymbol{\epsilon} \sim \text{MN}\big(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\Lambda}^{-1}\big)$ is matrix-normally distributed with mean $\mathbf{0}$, between row variance $\mathbf{I}_N$ and diagonal between column precision $\boldsymbol{\Lambda} = \text{diag}[\lambda_1, \ldots, \lambda_P]$. The latent factors $\mathbf{F}$ arise from K independent Brownian diffusion processes on the phylogenetic tree $\mathcal{F}$.The tree $\mathcal{F}$ is rooted and bifurcating with degree-two root node ν_{2N-1}, degree-three internal nodes $\{\nu_{N+1}, \ldots, \nu_{2N-2}\}$ and degree-one leaf nodes $\{\nu_1, \ldots, \nu_N\}$. Under the Brownian diffusion model, all internal and tip factors are normally distributed as $\mathbf{f}_j \sim \mathcal{N}\big(\mathbf{f}_{\text{pa}(j)}, t_j \mathbf{I}_K\big)$, where $\mathbf{f}_{\text{pa}(j)}$ are the factors of the parent of node ν_j and t_j is the distance (time) between nodes $\nu_{\text{pa}(j)}$ and ν_j. Following from Pybus et al. (2012), we assume the ancestral root traits $\mathbf{f}_{2N-1} \sim \mathcal{N}\Big(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\mathbf{I}_K\Big)$, where κ_0 is some (typically small) predetermined prior sample size. This construction implies the tip factors are jointly matrix-normally distributed as $\mathbf{F} \sim \text{MN}\Big(\mathbf{1}_N \boldsymbol{\mu}_0^t, \boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}_N, \mathbf{I}_K\Big)$, where $\mathbf{1}_N$ is an N-vector of ones, $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^t$ and $\boldsymbol{\Psi}$ is the standard variance-covariance (VCV) representation of the phylogeny $\mathcal{F}$. Specifically, the diagonal elements Ψ_{ii} are the sum of the edge lengths connecting ν_i to the root ν_{2N-1}. The off-diagonal elements Ψ_{ij} are the total amount of shared evolutionary history or time from the most recent common ancestor of ν_i and ν_j to the root node ν_{2N-1}.

Given this model, the vectorized data vec($\mathbf{Y}$) are multivariate normally distributed as

$$\text{vec}(\mathbf{Y}) \,|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F} \sim \mathcal{N}\left(\text{vec}\big(\mathbf{1}_N \boldsymbol{\mu}_0^t\big), \mathbf{L}^t\mathbf{L} \otimes \left[\boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}_N\right] + \boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_N\right), \tag{4.1}$$

where $\otimes$ is the Kronecker product operator. Computing the likelihood in this form, however, requires inverting the $NP \times NP$ dimensional variance matrix, which has computational complexity $\mathcal{O}(N^3P^3)$. Tolkoff et al. (2018) avoid this by treating the latent factors $\mathbf{F}$ as model parameters that they integrate out via Markov chain Monte Carlo (MCMC) simulation.This augmented likelihood $p(\mathbf{Y}, \mathbf{F} \,|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}) = p(\mathbf{Y} \,|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathbf{F})p(\mathbf{F} \,|\, \mathcal{F})$ is far easier to compute, but sampling from the full conditional distribution of $\mathbf{F}$ (i.e. the posterior distribution of $\mathbf{F}$ conditional on the data and all other model parameters) as proposed by Tolkoff et al. (2018) scales quadratically with the size of the phylogenetic tree and is intractable for big-N.

4.2.1.1 Fast likelihood calculation

To avoid costly data augmentation, we adapt the likelihood-computation algorithm independently developed by Bastide et al. (2018), Mitov et al. (2020) and Hassler et al. (2020, Section 3.2.1.2). This algorithm analytically integrates out latent traits (in our case factors) and missing data to compute the likelihood $p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big)$ of the observed data $\mathbf{Y}^{\text{obs}}$ in $\mathcal{O}(NPK^2 + NK^3)$ via a post-order traversal of the tree (i.e. computations start at the tips and are carried up the tree to the root).This procedure naturally accommodates missing data assuming an ignorable missing data mechanism (Rubin, 1976). We also utilize a more numerically stable modification of this post-order algorithm proposed by Bastide et al. (2021). We detail these calculations in Section 4.8.1.

4.2.1.2 Loadings identifiability

A major challenge in latent factor models generally is the non-identifiability of the loadings matrix $\mathbf{L}$ (see Shapiro, 1985). In statistical models, non-identifiability occurs when there are multiple parameter values that result in the same probability density over the data. In these cases, inference procedures cannot distinguish between the equally valid parameter values.This lack of identifiability in PFA stems from the fact that the likelihood as defined in Equation 4.1 depends only on $\mathbf{L}^t\mathbf{L}$ rather than $\mathbf{L}$ itself. As such, for any $K \times K$ orthonormal matrix $\mathbf{Q}$ (i.e. $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_K$), $p(\mathbf{Y} \,|\, \mathbf{L}, \dots) = p(\mathbf{Y} \,|\, \mathbf{QL}, \dots)$ because $(\mathbf{QL})^t\,(\mathbf{QL}) = \mathbf{L}^t\mathbf{L}$. This identifiability problem inspires our choice of priors below.

4.2.2 Priors

We assume the diagonal precisions $\lambda_j \sim \text{Gamma}(a_{\boldsymbol{\Lambda}}, b_{\boldsymbol{\Lambda}})$ for $j = 1, \dots, P$ (shape/rate parameterization). For the loadings $\mathbf{L} = \{\ell_{kj}\}$, we propose two different priors. Each prior on $\mathbf{L}$ admits a different inference regime for sampling from $\mathbf{L}$ which in turn have their own strengths and weaknesses that we discuss in Section 4.3.

4.2.2.1 Independent Gaussian priors on the loadings L

The standard assumption in Bayesian latent factor models is that each element of the loadings $\ell_{kj} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, where typically $\sigma^2 = 1$. As this prior is also invariant with respect to orthogonal rotations, additional constraints are required for posterior identifiability.One solution is to assume certain elements of the loadings matrix $\mathbf{L}$ (typically those below the diagonal) are fixed at zero (Geweke and Zhou, 1996; Aguilar and West, 2000). This approach solves the identifiability problem, but it induces an implicit ordering to the data (see Table 4.1). While this ordering may be well-informed in some cases, there is typically no principled way to choose such an ordering *a priori*.

An alternative to the sparsity constraint is to assume that the loadings matrix has rows that 1) are orthogonal and 2) have decreasing norms (Holbrook et al., 2016). This constraint does not require any *a priori* ordering of the traits. However, it does require sampling from the space of orthogonal matrices, which is a notoriously challenging problem (see Hoff, 2009;

Byrne and Girolami, 2013; Jauch et al., 2021; Pourzanjani et al., 2021). We address this challenge via post-processing in Section 4.3.1.3.

4.2.2.2 Orthogonal shrinkage prior

While post-processing to orthogonality is often sufficient, we find in practice that the loadings may be only loosely identifiable with this procedure in small-N problems. As such, we seek an alternative prior that enforces the orthogonality constraint directly. Following from Holbrook et al. (2017), we decompose the loadings $\mathbf{L} = \mathbf{\Sigma}\mathbf{V}$ where $\mathbf{\Sigma} = \text{diag}[\boldsymbol{\sigma}]$ is a $K \times K$ diagonal matrix whose diagonals $\boldsymbol{\sigma}$ have descending absolute values and $\mathbf{V}$ is a $K \times P$ orthonormal matrix (i.e. $\mathbf{V}\mathbf{V}^t = \mathbf{I}_K$). We assume $\mathbf{V}^t$ is uniformly distributed over the Stiefel manifold $\mathcal{V}_K\big(\mathbb{R}^P\big)$ (i.e. the space of $P \times K$ orthonormal matrices). For the scale component $\mathbf{\Sigma} = \text{diag}[\sigma_1, \ldots, \sigma_K]$ we assume a multiplicative gamma prior inspired by Bhattacharya and Dunson (2011):

$$\begin{aligned}
\sigma_k &\sim \mathcal{N}\big(0, \tau_k^{-1}\big) \text{ for } k = 1, \ldots, K, \text{ where} \\
\tau_k &= \prod_1^k \nu_\ell \text{ and} \\
\nu_\ell &\sim \text{Gamma}(a_\ell, b_\ell) \text{ for } \ell = 1, \ldots, K.
\end{aligned} \tag{4.2}$$

For $\ell > 1$, we constrain the prior shape a_ℓ and rate b_ℓ such that $a_\ell > b_\ell$ (i.e. $\mathbb{E}[\nu_\ell] > 1$).This constraint implies that the τ_k are (stochastically) increasing with k, which results in scale parameters σ_k with (stochastically) decreasing magnitudes.

This prior induces posterior identifiability, as it is not invariant under rotations of the loadings. However, in some cases we find that this prior does not induce sufficient identifiability in practice, particularly when K is relatively large (i.e. > 5). For these cases, we multiply the joint prior on $\mathbf{\Sigma}$ by an indicator function $1\{|\sigma_k| < \alpha\,|\sigma_{k-1}| \text{ for } k = 2, \ldots, K\}$. Setting $\alpha < 1$ forces spacing between the diagonals of $\mathbf{\Sigma}$, which results in more identifiable posteriors.

4.3 Inference

Our Bayesian inference regime seeks to approximate the posterior distribution of the parameters of scientific interest via MCMC simulation. We typically use molecular sequence data $\mathbf{S}$ to simultaneously infer the factor model parameters and phylogenetic tree by approximating

$$p\big(\mathbf{L}, \mathbf{\Lambda}, \mathcal{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{S}\big) \propto p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}\big) p(\mathcal{F}, \mathbf{S}) p(\mathbf{L}) p(\mathbf{\Lambda}), \tag{4.3}$$

]["]SeqModelwhere the model of sequence evolution $p(\mathcal{F}, \mathbf{S})$ is developed elsewhere (see Suchard et al., 2018). For cases where we lack sequence data or $\mathcal{F}$ is too large to infer efficiently, we simply fix the tree $\mathcal{F}$.

4.3.1 Loadings under the i.i.d. Gaussian prior

We propose two different samplers to draw from the full conditional distribution of the loadings $\mathbf{L}$ under the i.i.d. Gaussian prior from Section 4.2.2.1. The first relies on the Gibbs sampler used by Tolkoff et al. (2018), where we sample from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{\Lambda}$. The second avoids data augmentation and can sample directly from the full conditional distribution $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{\Lambda}, \mathcal{F}$ without conditioning on the latent factors $\mathbf{F}$.

4.3.1.1 Gibbs sampler with data augmentation

Tolkoff et al. (2018) use the conjugate Gibbs sampler of Lopes and West (2004) to sample from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{\Lambda}$. As this sampler conditions on the latent factors $\mathbf{F}$, Tolkoff et al. (2018) simultaneously infer the factors by sequentially drawing from $\mathbf{f}_i \,\big|\, \mathbf{F}_{/i}, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}$ for $i = 1, \ldots, N$, where $\mathbf{F}_{/i}$ represents all factors except $\mathbf{f}_i$. As sampling $\mathbf{f}_i$ for all N taxa requires $\mathcal{O}(N^2K^2)$ work, this procedure quickly becomes intractable with increasing taxa.

Rather than relying on this per-taxon sampling scheme, we employ the pre-order data augmentation algorithm of Hassler et al. (2020, 3.2.2.1) that uses statistics from the post-order likelihood computation to draw jointly from $\mathbf{F} \,\middle|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}$ in $\mathcal{O}(NK^3)$ via a single pre-order traversal of the tree (see Section 4.8.2.1 for details). After sampling from $\mathbf{F} \,\middle|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}$, we can draw directly from $\mathbf{L} \,\middle|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \boldsymbol{\Lambda}$ using the procedure developed by Lopes and West (2004) with computational complexity $\mathcal{O}(NPK^2)$ (see Section 4.8.2.2 for details).

4.3.1.2 Hamiltonian Monte Carlo sampler

We also propose an alternative Hamiltonian Monte Carlo (HMC; Neal, 2010) sampler for the loadings that does not require data augmentation. Intuitively, HMC (a form of MCMC) treats parameter values as the position of a particle in a landscape informed by the posterior distribution. Parameter proposals are the end-point of a trajectory initiated by "kicking" the particle and allowing it to traverse this landscape according to Hamiltonian dynamics for a pre-determined amount of time. As the parameter trajectories are informed by the geometry of the posterior, HMC tends to propose parameter updates that are both relatively far away from the current position and have high acceptance probabilities.

While we cannot compute these continuous trajectories analytically, we can approximate them numerically.Each trajectory approximation, however, requires numerous gradient calculations, and we must efficiently compute the gradient $\nabla_{\mathbf{L}} \log p\big(\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda}, \mathcal{F}\big) = \nabla_{\mathbf{L}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big) + \nabla_{\mathbf{L}} \log p(\mathbf{L})$ to effectively employ HMC to update the loadings $\mathbf{L}$. As we assume each element of the loadings are *a priori* i.i.d. $\mathcal{N}(0,1)$, the gradient of the log-prior $\nabla_{\mathbf{L}} \log p(\mathbf{L})$ can be computed simply as $\frac{\partial}{\partial \ell_{kj}} \log p(\mathbf{L}) = -\ell_{kj}$ for $j = 1, \ldots, P$, $k = 1, \ldots, K$.

As computing $\nabla_{\mathbf{L}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big)$ directly via Equation 4.1 scales $\mathcal{O}(N^3P^3)$ and is intractable for most problems, we use the highly structured nature of the phylogeny to compute this gradient in $\mathcal{O}(NPK^2 + NK^3)$. We calculate the gradient of the likelihood with respect to each column of the loadings $\boldsymbol{\ell}_j$ individually to accommodate variation in the missing data structure across traits.

$$\nabla_{\boldsymbol{\ell}_j} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big) = \lambda_j \mathbb{E}\big[\mathbf{F}^t \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big] \boldsymbol{\delta}'_j \mathbf{y}_j^{\text{obs}'} - \lambda_j \mathbb{E}\big[\mathbf{F}^t \boldsymbol{\delta}'_j \mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big] \boldsymbol{\ell}_j, \tag{4.4}$$

where $\mathbf{y}_j^{\text{obs}'}$ is the j^{th} column of $\mathbf{Y}^{\text{obs}}$ and $\boldsymbol{\delta}'_j = \text{diag}[\delta_{1j}, \ldots, \delta_{Nj}]$ is a diagonal matrix of observed-data indicators (i.e. $\delta_{ij} = 1$ if y_{ij} is observed and 0 otherwise). Note that these calculations rely only on the conditional mean and variance of the factors, not the factors themselves. We compute the expectations using statistics from the post-order likelihood calculation (see Section 4.8.1) in a pre-order tree traversal (Bastide et al., 2018; Fisher et al., 2021) that takes $\mathcal{O}(NK^3)$ additional time. See Section 4.8.3 for detailed calculations.

4.3.1.3 Orthogonality constraint and post-processing

While both the Gibbs and HMC samplers above can enforce the structured sparsity constraint, neither can enforce the orthogonality constraint directly. However, as both the likelihood and i.i.d. prior are invariant with respect to orthonormal rotations of $\mathbf{L}$, applying such a rotation to all posterior samples via post-processing results in a valid posterior. We can easily rotate the loadings to have orthogonal rows with descending norms via singular value decomposition (see Section 4.8.4 for details).

4.3.2 Loadings under the orthogonal shrinkage prior

Both samplers above are incompatible with the orthogonal shrinkage prior from Section 4.2.2.2 as 1) they cannot enforce the orthogonality constraint directly and 2) post-processing is invalid because the prior is not rotationally invariant. Therefore, we sample directly from the full conditional distributions of both $\boldsymbol{\Sigma}$ and $\mathbf{V}$ rather than their product $\mathbf{L}$.

4.3.2.1 Geodesic HMC sampler on the orthonormal component $\mathbf{V}$

Requiring $\mathbf{V}^t$ to be orthonormal allows us to employ existing techniques for sampling from the Stiefel manifold (i.e. the space of orthonormal matrices). Geodesic HMC (Byrne and Girolami, 2013) uses the same fundamental principles of standard HMC, but progresses parameters along geodesics on manifolds (e.g. an arc on a sphere) rather than through Euclidean space.This procedure also relies on the gradient of the log-posterior with respect to the parameter of interest. As such, to efficiently employ geodesic HMC to update the orthonormal matrix $\mathbf{V}$, we must efficiently compute the gradient

$$\nabla_{\mathbf{V}} \log p\big(\mathbf{V} \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \mathcal{F}\big) = \nabla_{\mathbf{V}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{V}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \mathcal{F}\big) + \nabla_{\mathbf{V}} \log p(\mathbf{V}). \tag{4.5}$$

As noted in Section 4.2.2.2, we place a uniform prior on $\mathbf{V}$ and can therefore ignore $\nabla_{\mathbf{V}} \log p(\mathbf{V})$. Using our calculations for $\nabla_{\mathbf{L}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big)$ from Section 4.3.1.2, the chain rule provides a simple formula for the gradient of the likelihood with respect to $\mathbf{V}$ as $\mathbf{L} = \boldsymbol{\Sigma}\mathbf{V}$:

$$\nabla_{\mathbf{V}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{V}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \mathcal{F}\big) = \boldsymbol{\Sigma}\nabla_{\mathbf{L}} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big). \tag{4.6}$$

We then use this gradient in the geodesic HMC algorithm of Holbrook et al. (2016) to sample from the full conditional distribution of $\mathbf{V}$.

4.3.2.2 Gibbs sampler on the diagonal scale component $\boldsymbol{\Sigma}$

While we can employ HMC to sample from $\boldsymbol{\Sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{V}, \boldsymbol{\Lambda}, \mathcal{F}$, our implementation did not mix well in practice. We develop a Gibbs sampler to draw from $\boldsymbol{\Sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{V}, \boldsymbol{\Lambda}, \mathbf{F}$ as an efficient alternative that relies on the data augmentation of $\mathbf{F}$ in Section 4.8.2.1. See Section 4.8.5 for details.

4.3.2.3 Gibbs sampler on the precision multipliers

We must also sample from the shrinkage multipliers $\nu_1, \ldots, \nu_K$ when using the shrinkage prior on the loadings. Bhattacharya and Dunson (2011, Section 3.1, Step 5) develop a conjugate Gibbs sampler for these multipliers that we apply directly to this model.

4.3.3 Sign constraint on the loadings

Regardless of which prior (i.i.d. vs. orthogonal shrinkage) or constraint (sparsity vs. orthogonality) we choose, we must enforce a sign constraint on a single element in each row of $\mathbf{L}$ for full identifiability (see Section 4.8.6 for details).

4.3.4 Gibbs sampler on the error precisions $\boldsymbol{\Lambda}$

We sample from $\boldsymbol{\Lambda} \,\big|\, \mathbf{F}, \mathbf{Y}^{\text{obs}}, \mathbf{L}$ using the same procedure as Tolkoff et al. (2018) in conjunction with the data augmentation algorithm in Section 4.8.2.1 (see Section 4.8.7 for details).

4.4 Computational efficiency

We compare the computational efficiency of the inference regimes discussed in Sections 4.3.1.1, 4.3.1.2 and 4.3.2 with that of Tolkoff et al. (2018). To understand performance across a wide range of situations, we simulate three unique data sets for all 36 combinations of $N \in \{50, 100, 500, 1000\}$, $P \in \{10, 100, 1000\}$ and $K \in \{1, 2, 4\}$ (see Section 4.8.8.1 for simulation details). To understand the relative performance of each inference regime, we compare the effective sample size (ESS) per second of the loadings across all four samplers (see Section 4.8.8.2 for details) and report our results in Figure 4.1.

Compared against the conditional Gibbs sampler of Tolkoff et al. (2018), both our joint Gibbs and HMC samplers under the i.i.d. prior consistently yield efficiency gains of an order of magnitude in small-N data sets and two orders of magnitude in big-N data sets. While

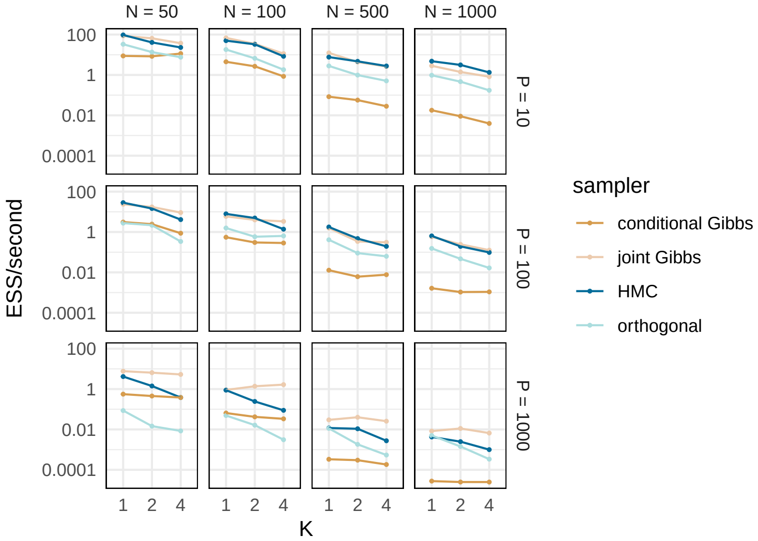


Figure 4.1: Timing comparison between inference regimes. We run three MCMC chain simulations for each combination of N (the number of taxa), P (the number of traits), K (the number of factors) and sampler and present the average minimum ESS per second for each. The "conditional Gibbs" sampler refers to the methods used by Tolkoff et al. (2018). The "joint Gibbs", "HMC" and "orthogonal" samplers refer to the methods presented in Sections 4.3.1.1, 4.3.1.2 and 4.3.2 respectively. Our joint Gibbs and HMC samplers are an order of magnitude faster than the conditional Gibbs sampler with relatively few taxa ($N = 50$) but more than two orders of magnitude faster with many taxa ($N = 1000$). The orthogonal sampler is slower than the joint Gibbs and HMC samplers (and even the conditional Gibbs in the case of small-N, big-P) but scales well to large trees. Values are available in Table 4.2.

the sampling regime under the orthogonal shrinkage prior is slower than either the joint Gibbs or HMC sampler (and even the conditional Gibbs sampler for small-N, big-P), it has clear advantages over the others that we discuss in Section 4.5.2.

4.5 Principled analysis plan

The modeling decisions required for Bayesian factor analysis can be daunting. In addition to the priors, identifiability constraints and sampling procedures discussed above, researchers must also choose an appropriate number of factors K. Making such choices in a principled manner is challenging, and experimenting with different combinations to determine which "work best" is time consuming and opens the door to modeling decisions based on publication concerns. We propose a generalizable analysis plan to guide researchers through this process. To aid researchers seeking to employ phylogenetic factor analysis specifically, we also develop software tools that codify this plan and automate core procedures.

4.5.1 Choosing the loadings constraint

The decision to apply the sparsity constraint versus the orthogonality constraint depends on the biological question of interest. While the sparsity constraint induces ordering onto the traits, this ordering can be desirable under certain circumstances. For example, if one is trying to isolate the effects of a particular set of traits, placing those traits first in conjunction with the upper triangular constraint ensures that they will load only onto the first few factors and all subsequent factors will be independent of their influence. If one does not want to apply such an ordering, the orthogonality constraint may be a better alternative. We emphasize, however, that the orthogonality constraint is no less restrictive than the sparsity constraint; rather, it replaces a series of potentially arbitrary modeling decisions (i.e. the ordering of the first K traits) with a single, perhaps equally arbitrary, constraint.

Researchers can also apply a hybrid approach where one or more traits load only onto a certain factor(s) while the remaining traits are free to load onto all factors. If the specific sparsity structure is not sufficient to induce identifiability, then any unconstrained sub-matrices of the loadings would require rotation to orthogonality. We present a simple example of this in Section 4.6.3, where the the first trait (body mass) loads only onto the first factor and the remaining traits load onto all K factors. In this case, the first row of the loadings is identifiable and captures mass-dependent relationships, while the sub-matrix composed of rows $2, \ldots, K$ and columns $2, \ldots, P$ is rotated to orthogonality via post-processing.

4.5.2 Choosing the loadings prior

Those choosing the sparsity (or hybrid) constraint must use the i.i.d. prior on the loadings, as orthogonality is implicit in our definition of the shrinkage prior. For those opting for the orthogonality constraint, we recommend choosing a prior based on the characteristics of the specific application. For big-N data sets ($N > 1000$) the geodesic HMC sampler on $\mathbf{V}$ under the shrinkage prior may be prohibitively slow (particularly when combined with big-P), and we suggest using the i.i.d. prior with post-processing.

One serious limitation of the post-processing regime, however, is the potential for label switching (Celeux, 1998). This phenomenon occurs when the posterior distributions of certain scale parameters $\boldsymbol{\sigma}$ overlap enough that a given factor switches its ordering. When this occurs, the resulting estimated factor (e.g. factor 1) may actually be a mixture of factors that shuffle in order during MCMC and post-processing. Figure 4.2 provides an example of this phenomenon and shows how the orthogonal shrinkage prior can address it. Examining the MCMC trace plots (i.e. plots of parameter values over each sample from the MCMC chain) in software such as the CODA R package (Plummer et al., 2006) or Tracer (Rambaut et al., 2018) is the best way to check for label switching. If the trace plot of the scale parameters $\boldsymbol{\sigma}$ appear to be touching (as in the top, left panel of Figure 4.2), then label switching is likely occurring. See Section 4.8.9 for a more thorough discussion of identifying label switching in the context of PFA.

Conveniently, label switching does not typically occur in big-N analyses, so we recommend the more computationally efficient i.i.d. prior with post-processing in these situations. For small- or moderate-N analyses, we still suggest attempting the i.i.d. sampler with post-processing, but we caution users to look for evidence of label switching. If such evidence exists, we recommend using the shrinkage prior with forced ordering and separation.

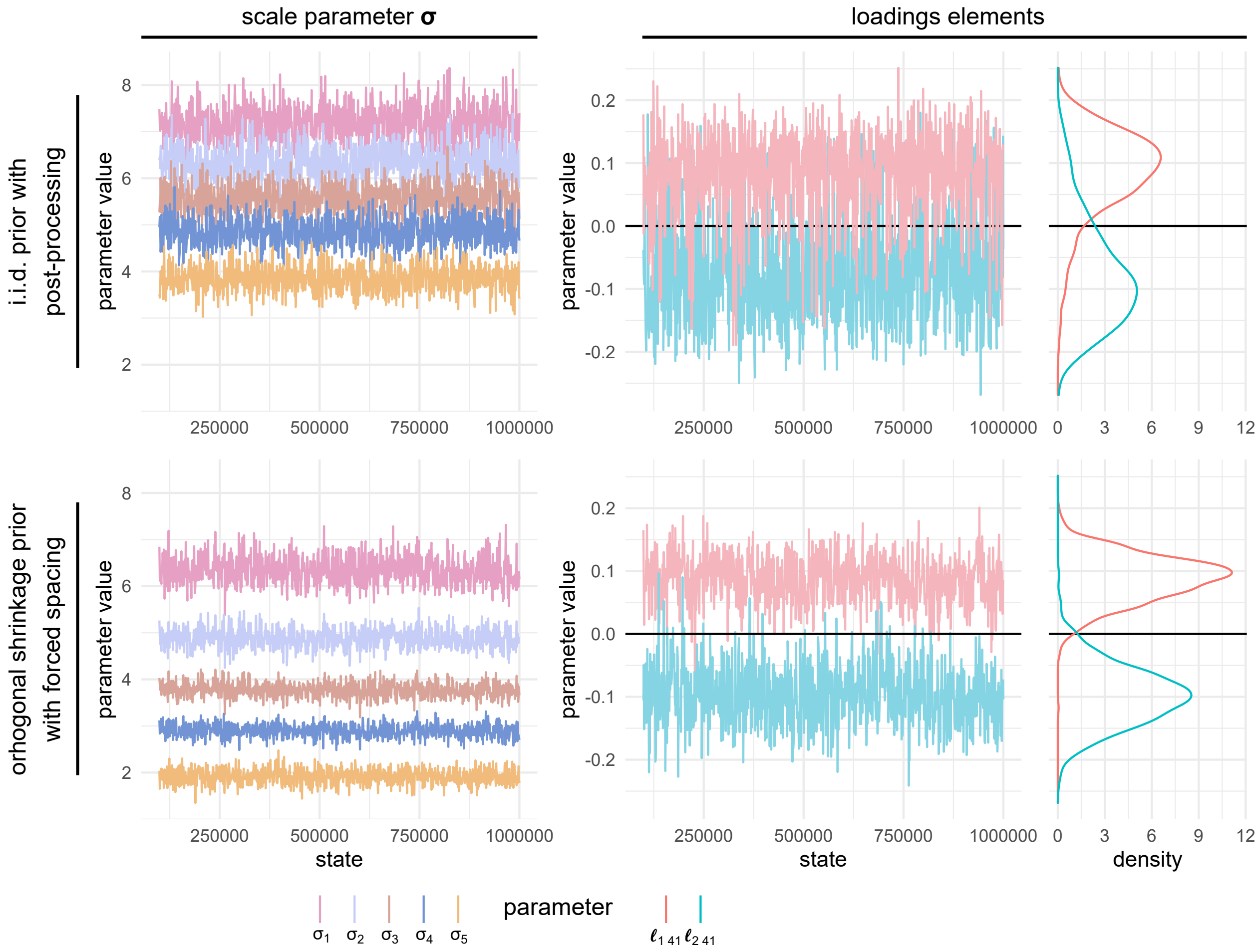


Figure 4.2: Trace plots of relevant parameters from analysis in Section 4.6.2. Estimates under the i.i.d. Gaussian prior are characteristic of poorly-identifiable conditions (the scales $\boldsymbol{\sigma}$ are overlapping resulting in label switching / row-wise convolution of the loadings). The shrinkage prior with forced spacing ($\alpha = 0.8$) largely eliminates this problem.

4.5.3 Constraining the number of factors

We propose cross-validation for identifying the number of factors with optimal predictive performance. In the case of the i.i.d. prior, this procedure compares models with different number of factors directly, while in the case of the orthogonal shrinkage prior it tunes the strength of the shrinkage on the loadings scales. See Section 4.8.10 for details.

We fully recognize that complex evolutionary processes do not, in reality, conform exactly to the phylogenetic latent factor model (or any tractable statistical model) and caution against seeking to identify the “true” number of underlying evolutionary processes driving the phenotypes of interest, as such ground truth likely does not exist. Rather, we encourage researchers to use this model selection procedure to identify the limitations of the information available in a particular data set and the model’s ability to extract it. For example, if model selection determines that a four factor model provides optimal predictive performance, one should be wary of interpreting results from a model with greater than four factors as it is likely some of the perceived signal is an artifact of noise in the data.

Prior to model selection, one must choose some maximum number of factors K_{max} that balances model interpretability, flexibility, identifiability and tractability. Models with more factors are inherently more flexible and can potentially capture more information about underlying biological phenomena. However, interpretation becomes challenging as the number of factors increases. While the model with optimal predictive performance may have $K < K_{\text{max}}$, one should be open to interpreting a model where $K = K_{\text{max}}$. Limiting K_{max} provides additional benefits, as 1) the identifiability challenges discussed in Section 4.5.2 intensify with increasing K and 2) inference scales cubically with K and some big-K models may be intractable. In practice, we settle on $K_{\text{max}} = 5$ for most examples below, as we find that the computation time and identifiability issues are typically manageable at $K = 5$ and feel most researchers would rarely need to interpret more than five factors.

4.5.4 Software implementation

We implement all inference procedures in Section 4.3 in the Bayesian phylogenetic inference software BEAST (Suchard et al., 2018). While BEAST is an extraordinarily flexible tool, this flexibility can result in a user experience that is overwhelming for the uninitiated.

We develop the Julia package PhylogeneticFactorAnalysis.jl to both simplify the BEAST user experience (in the context of PFA) and automate model selection, post-processing, diagnostics and plotting. Users must input the trait data, a phylogenetic tree, the identifiability constraint on the loadings and the prior on the loadings. Users may also optionally specify other modeling decisions such as whether to standardize the trait data (which we recommend) and the model selection meta-parameters as well as a BEAST input file with instructions for inferring the phylogenetic tree from sequence data.

After receiving appropriate input, PhylogeneticFactorAnalysis.jl automatically performs model selection and outputs a series of files including the sub-sampled MCMC realizations and plots of both the loadings (see Figures 4.3B, 4.4A and 4.5A) and factors on the tree (see Figures 4.4B, 4.5B and 4.6B) using the ggplot2 (Wickham, 2016) and ggtree (Yu et al., 2017) plotting libraries.PhylogeneticFactorAnalysis.jl is registered under the Julia General registry. Source code and documentation can be accessed at:

`https://github.com/gabehassler/PhylogeneticFactorAnalysis.jl`

4.6 Example analyses

We demonstrate the utility of these methods in the four examples below. Unless otherwise noted, all data are standardized on a per-trait basis (i.e. subtracting the trait mean and dividing the by the trait standard deviation) prior to analysis.

4.6.1 Pollinator-flower co-evolution in *Aquilegia*

The intimate relationship between plants and their pollinators has played a defining role in the evolution of angiosperms (see Kay and Sargent, 2009; Van der Niet and Johnson, 2012). Here we re-evaluate the relationship between floral phenotypes and pollinators in the genus *Aquilegia* (columbines). Whittall and Hodges (2007) identify three primary *Aquilegia* "pollination syndromes" associated with bumblebees, hummingbirds and hawk moths respectively. Tolkoff et al. (2018) apply phylogenetic factor analysis to study the relationship between 11 floral phenotypes and these pollination syndromes in *Aquilegia* and identify two factors, only one of which is associated with pollinator type.

We re-evaluate this previous work for two reasons. First, Tolkoff et al. (2018) assume the upper-triangular constraint on the loadings which requires that the vertical angle of the flower loads only onto the first factor. Our orthogonality constraint eliminates arbitrarily singling out this phenotype. Additionally, we compare our cross-validation model selection procedure with the marginal likelihood-based approach of Tolkoff et al. (2018), which identifies a two-factor model as having greatest posterior support.

As four of the traits (anthocyanin production and the three pollination syndromes) are binary, we follow Tolkoff et al. (2018) in adapting the latent-liability model of Cybis et al. (2015) to the latent factor model (see Section 4.8.11). We use the i.i.d. prior with orthogonality constraint, and our model selection procedure, indeed, identifies two factors. We present our results in Figure 4.3. The first factor captures patterns differentiating hummingbird-pollinated plants from hawk moth-pollinated plants, while the second factor appears to separate the bumblebee pollinated flowers from the other two pollination syndromes. Note that in Figure 4.3A, the first factor falls along a relatively uniform continuum, while the second factor has a clear out-group consisting of the bumblebee-pollinated plants. While only two taxa are coded as being pollinated by both hummingbirds and hawk moths, this suggests that non-bumblebee *Aquilegia* pollination strategies may lie on a continuum rather than strict a hawk moth/hummingbird dichotomy, and it is possible that many of the plants listed as having a single pollinator in reality attract both hummingbirds and hawk moths.

4.6.2 Yeast domestication

The brewer's yeast *Saccharomyces cerevisiae* is essential to a variety of industrial applications due to its ability to convert sugars into ethanol, carbon dioxide and aroma compounds. In addition to its well-known role in the production of fermented food and beverages, it also plays a key role in the production of of bio-fuels and serves as model organism for basic biological research. Industrial strains within this species adapted to thrive within specialized environments and can withstand stress conditions often suited to the specific industrial niche

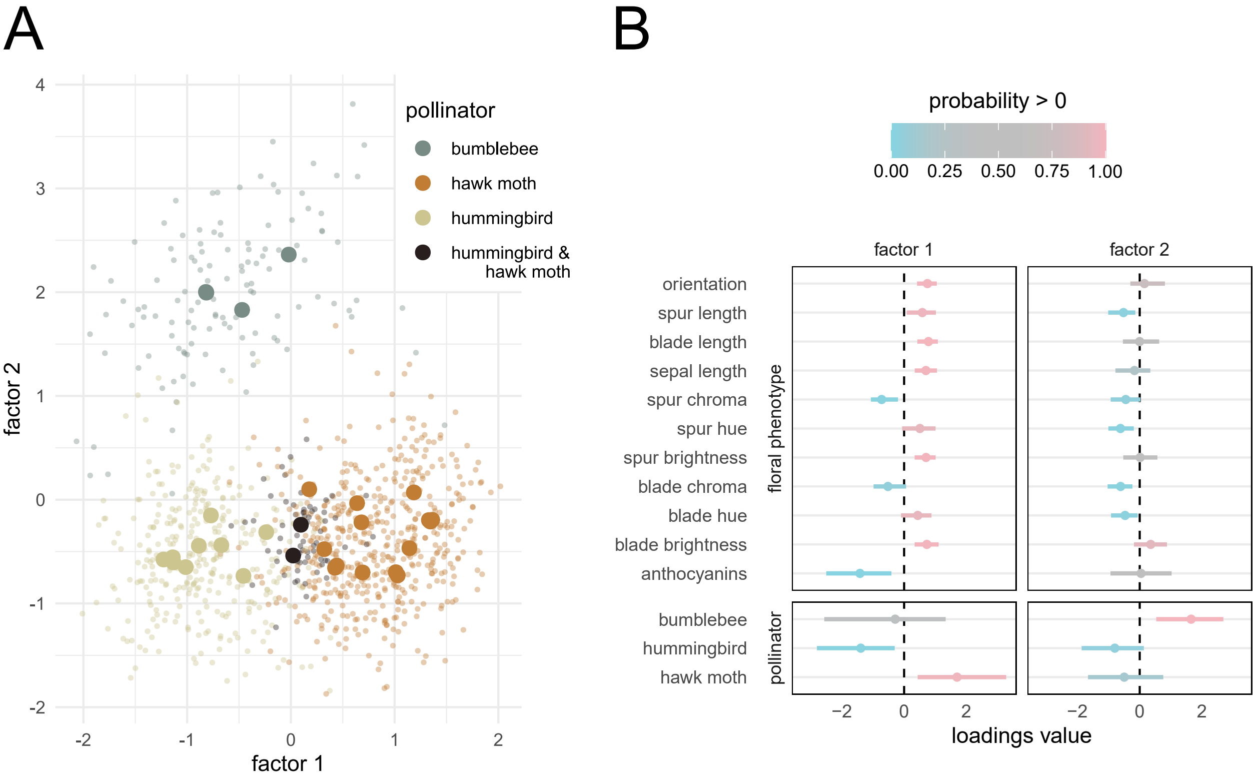


Figure 4.3: *Aquilegia* results. **A)** Factor values colored by pollinator(s) for each species of *Aquilegia*. Large, solid points represent posterior means for each species. Small, transparent points represent a random sample from the posterior distribution of the factors. **B)** Posterior summary of the loadings matrix. Dots represent posterior means while bars cover the 95% highest posterior density (HPD) interval. Colors represent the posterior probability that the parameter is greater than 0. While the second factor clearly separates the bumblebee-pollinated plants from the others, the first factor captures a more gradual transition from hummingbird pollination to hawk moth pollination.

they evolved in, such as ethanol, osmotic, acidic and temperature stresses.

Recent work by Gallone et al. (2016) and Gallone et al. (2019) uses phylogenetic methods to study the domestication of *S. cerevisiae* within industrial environments. To elucidate the effects of domestication on yeast phenotypes, Gallone et al. (2016) sequence and phenotype 154 strains of industrial and wild *S. cerevisiae*. The 82 phenotypes include numerous measurements of growth rates under varying environmental and nutrient stresses, the levels of production of various metabolites and the ability to reproduce sexually.

Domestication in plants and animals is typically characterized by limited reproduction outside of domestic contexts, increased yield and decreased tolerance to rare or novel environmental stressors (Doebley et al., 2006; Larson and Fuller, 2014). Gallone et al. (2016) observe these same patterns in the yeast strains they study, with additional niche-specific patterns of covariation. While their analysis examines the specific hypotheses above, they do not employ a data-generative model of phenotypic evolution capable of studying broad changes across all measured phenotypes.

The phylogenetic latent factor model, however, is ideally suited for such a task. We first infer a phylogenetic tree for the 154 phenotyped strains using the 2.8 megabase DNA sequence alignment of Gallone et al. (2016) (see Section 4.8.12.1). We fix this tree during model selection due to the computational costs of inferring the phylogeny. Based on the principles discussed in Section 4.5, we opt for the orthogonality constraint, the orthogonal shrinkage prior with forced spacing ($\alpha = 0.8$) and $K_{\text{max}} = 5$. Our model selection procedure yields a final model with five significant factors. For the final analysis we infer the tree jointly with factor model parameters using the same tree model in Section 4.8.12.1. As the number of significant factors K is equal to the maximum K_{max}, we are confident any signal is biologically relevant but recognize we have not completely captured the full phenotypic covariance structure. That being said, the final factor captures only 7% (5%-9% HPD interval) of the heritable variance and 3% (2%-4%) of the total variance, suggesting that adding additional factors will yield diminishing returns at the expense of exacerbating identifiability challenges.

We plot the loadings associated with the first factor and the first factor on the tree in Figure 4.4 (see Figures 4.8 and 4.9 for the full results). For the first factor that accounts for 44% (33%-52%) of the heritable variance, we observe a clear separation between strains in the Beer 1 clade and strains isolated from other fermentation processes and from the wild. Notably, the domestication of beer strains in this clade led to an impaired sexual cycle as observed in the reduced sporulation efficiency and spore viability. This loss of a functional sexual cycle is paired with the additional loss of tolerance to environment and nutrient stresses generally. These stresses are not encountered during continuous growth in the nutrient-rich wort medium. The higher tolerance to high temperature outside of Beer 1 might reflect other more cryptic specializations of non-Beer clade 1 strains selected for different industrial processes (e.g. bioethanol or cocoa fermentation). Beyond these general patterns, we also note specific traits selected for in the Beer 1 clade. For example: strains within this clade do not produce 4-vinyl guaiacol (4-VG), a renown off-flavor in beer that is less relevant to other industrial niches. Additionally, the first factor in this clade is associated with efficient utilization of maltotriose, an important carbon source in beer wort but rarely found in high concentrations in natural environments. These results overall recapitulate one of the main findings of Gallone et al. (2016): the transition from complex and variable natural niches to the stable, nutrient-rich, beer medium favored certain adaptations (e.g. efficient utilization of maltotriose) and accentuation of certain traits (lost of beer off-flavours) at the cost of becoming sub-optimal for survival in the wild.

We emphasize that in this dataset there are different domestication trajectories targeted to very diverse industrial processes, and the life histories of the different clades took separate paths that the additional factors likely capture.

4.6.3 Mammalian life history

Life history strategies vary greatly across the tree of life. Generally speaking, organisms exist along a spectrum between fast-reproducing species that produce many offspring with little investment into any single child and slow-reproducing species that invest relatively great time and energy into each of their (comparatively fewer) offspring (Pianka, 1970). While allometric (size-dependent) constraints clearly influence these life history strategies (Boukal et al., 2014), pace-of-life theory predicts size-independent life-history variation as a major driver of phenotypic covariation (Reynolds, 2003; Réale et al., 2010). Much work has been done evaluating these hypotheses across numerous taxonomic groups (see Blackburn, 1991; Bielby et al., 2007; Salguro-Gómez, 2017), but most studies are limited by methodologies

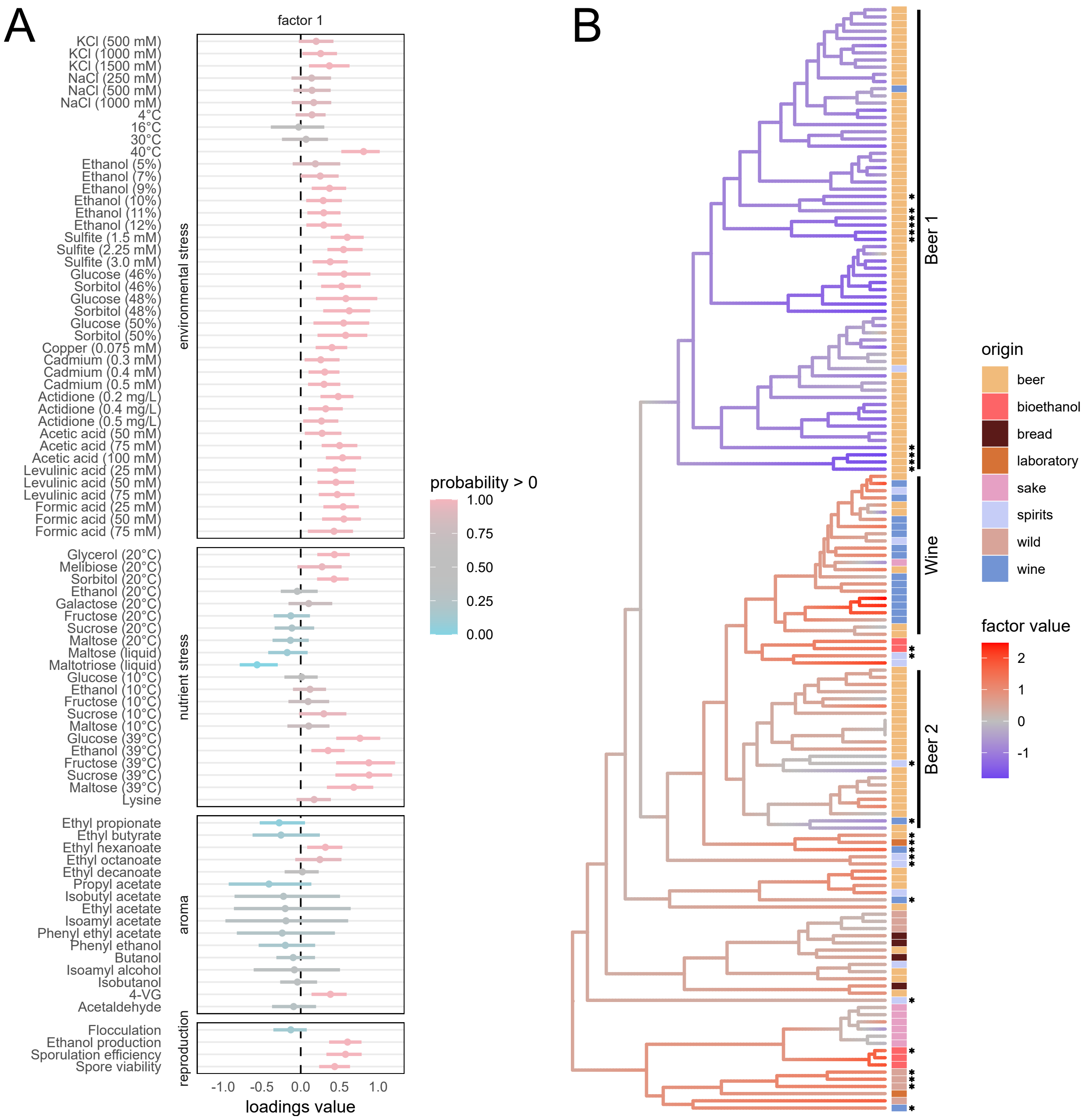


Figure 4.4: Results associated with first factor in yeast analysis. **A)** Posterior summary of first row of the loadings of 5-factor PFA on yeast data set. This first factor primarily captures differences associated with tolerance to environment and nutrient stress as well as reproductive ability. See Figure 4.3B for description of plot elements. **B)** The first factor plotted on yeast phylogeny with strain origin. Stars at the tips indicate mosaic strains as identified by Gallone et al. (2016). *(caption continues on next page)*

4.4 *(previous page)*: Low factor values in the Beer 1 clade indicate poor tolerance of environmental and nutrient stress generally and a lower capacity to reproduce sexually, all of which are signs of domestication. The Beer 1 clade includes strains from Belgium, Germany, Britain and the United States, and Gallone et al. (2016) estimate its origin ca. 1590 AD that coincides with the transition from home-brewing to large-scale beer production across Europe.

that require complete data and scale poorly to very large trees and many traits.

We explore the evolution of mammalian life history using the PanTHERIA ecological database (Jones et al., 2009). We select a sub-set of this data including body mass and 10 life history traits for the 3,691 species with at least one non-missing observation. While Hassler et al. (2020, Section 3.7.1) explore a similar subset of the PanTHERIA data using a multivariate Brownian diffusion (MBD) model, the MBD model cannot partition the covariance structure into size-dependent and size-independent components.

PFA, however, is ideally suited to this task as we can structure the loadings matrix *a priori* to reveal these relationships. Specifically, we apply the hybrid constraint introduced in Section 4.5.1 where elements $\ell_{21}, \dots, \ell_{K1}$ are fixed to zero, forcing body mass to load only onto the first factor. To avoid ordering the other life-history traits, we assume that the sub-matrix consisting of rows $2, \dots, K$ and columns $2, \dots, P$ is orthogonal (which we enforce via post-processing). We use the fixed tree of Fritz et al. (2009), which we prune to include only the 3,691 taxa for which we have trait data. We perform model selection assuming $K_{\max} = 5$, with the optimal model having $K = 5$. However, the first three factors explain 85% of the heritable variance (with the last factor explaining only 4%), suggesting that $K = 5$ is sufficient to capture the major patterns of variation in mammalian life-history evolution. We plot our results in Figure 4.5.

Consistent with the Hassler et al. (2020, Section 3.7.1) analysis, body size is clearly associated with the “slow” life history strategy (i.e. smaller and less frequent litters, longer lives). Notably, this allometric factor is not the dominant factor and explains only 16%

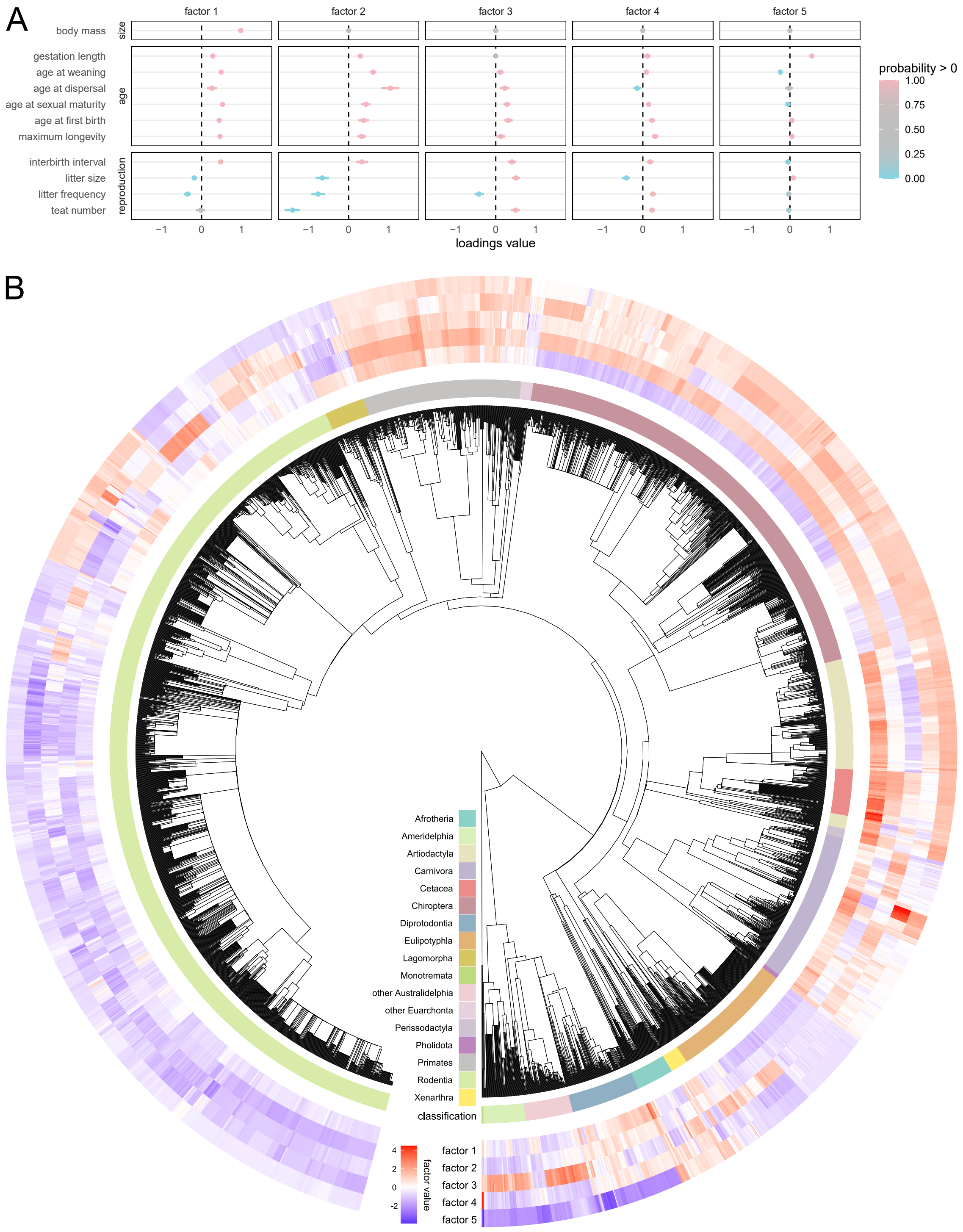


Figure 4.5: Mammalian life history results. *(caption on next page)*

4.5 *(previous page)*: **A)** Posterior summary of the loadings. Loadings of body size onto factors 2-5 is set to 0 *a priori*. See Figure 4.3B for detailed description of figure elements. The first factor captures allometric relationships (by design) and explains only 16% of the heritable variance, while the remaining factors capture size-independent relationships. The second factor, accounting for the plurality (46%) of the heritable variance, captures a fast-slow life history axis. Remaining factors capture more specific strategies (e.g. factors three and four appear to support the energy trade-off between litter size and litter frequency). This suggests that body size is not the main driver of life history evolution and that natural selection primarily acts on life history directly. **B)** Evolution of factors along the mammalian phylogeny. Most factors are strongly phylogenetically conserved throughout the tree, with large clades sharing similar factor values. There is relatively little correlation between the the first and second factors, with clades of small, slow species (e.g. bats) and large, fast species (e.g. lagomorphs).

(14%-18%) of the heritable variance. The second factor, however, explains 46% (42%-51%) of this variance and clearly captures a size-independent fast-slow life history axis, suggesting that size-independent life-history strategies play a major role in mammalian evolution. As evident in Figure 4.5, this primary life-history axis (factor 2) varies independently of the allometric one (factor 1) with examples of large/slow (cetaceans), large/fast (lagomorphs), small/slow (bats) and small/fast (rodents) taxonomic groups. This primary life-history factor is well-conserved across the phylogenetic tree, with large taxonomic groups sharing life-history strategies.

Factors 3, 4 and 5 explain comparatively less of the heritable variance (23%, 11% and 4% respectively). Factors 3 and 4 appear to capture trade-offs between litter size and litter frequency, while the 5th factor primarily captures a negative relationship between weaning age and gestation length and is strongly expressed in monotremes and marsupials that employ different reproductive strategies than placental mammals.

4.6.4 New World monkey cranial morphology

While much effort has been devoted to studying the evolution of primate brain size, relatively few studies have focused on understanding diversity in brain morphology or shape. Notable exceptions to this trend include Aristide et al. (2016) and Sansalone et al. (2020). Here we re-analyze the data presented in Aristide et al. (2016), that consist of 399 endocranial landmarks in 3-dimensional Euclidean space (standardized by generalized Procrustes analysis) for 48 species of New World monkey (NWM). While Aristide et al. (2016) perform principal component analysis on the Procrustes coordinates and use the principal component scores as traits in a larger evolutionary analysis, this procedure lacks a complete data-generative statistical model that explicitly accounts for uncertainty or noise in the shape data.

We simultaneously infer the phylogeny with the PFA parameters using DNA sequence alignments from Aristide et al. (2015) (see Section 4.8.12.2 for details). Preliminary results suggest 1) optimal predictive performance requires a very large number of factors (> 20), which is unsurprising given the complexity of this data set, and 2) identifiability poses an unusually great challenge due to the "small-N big-P" nature of the data. As such, we settle on a 3-factor model with orthogonal shrinkage prior and strong shrinkage to maximize identifiability. To maintain differences in scale between traits, we do not re-scale on a per-trait basis but rather divide all traits by the maximum per-trait standard deviation.

We plot the influence of each factor on brain shape and the evolution of these factors on the tree in Figure 4.6. These three factors capture similar patterns of variation as the first three principal components in Aristide et al. (2016), who identify several ecological processes associated with the evolution of these principal components. As the latent factor model can capture uncertainty that PCA cannot, we are eager to re-evaluate these relationships via a more structured latent factor model that directly models the relationship between the brain shape factors and ecological phenotypes such as social structure or diet. While preliminary results suggest that the first factor is correlated with relative brain volume (i.e. brain volume divided by body mass) and social group size and that the second factor is correlated with body mass and absolute brain volume, we leave this more structured analysis as future work.

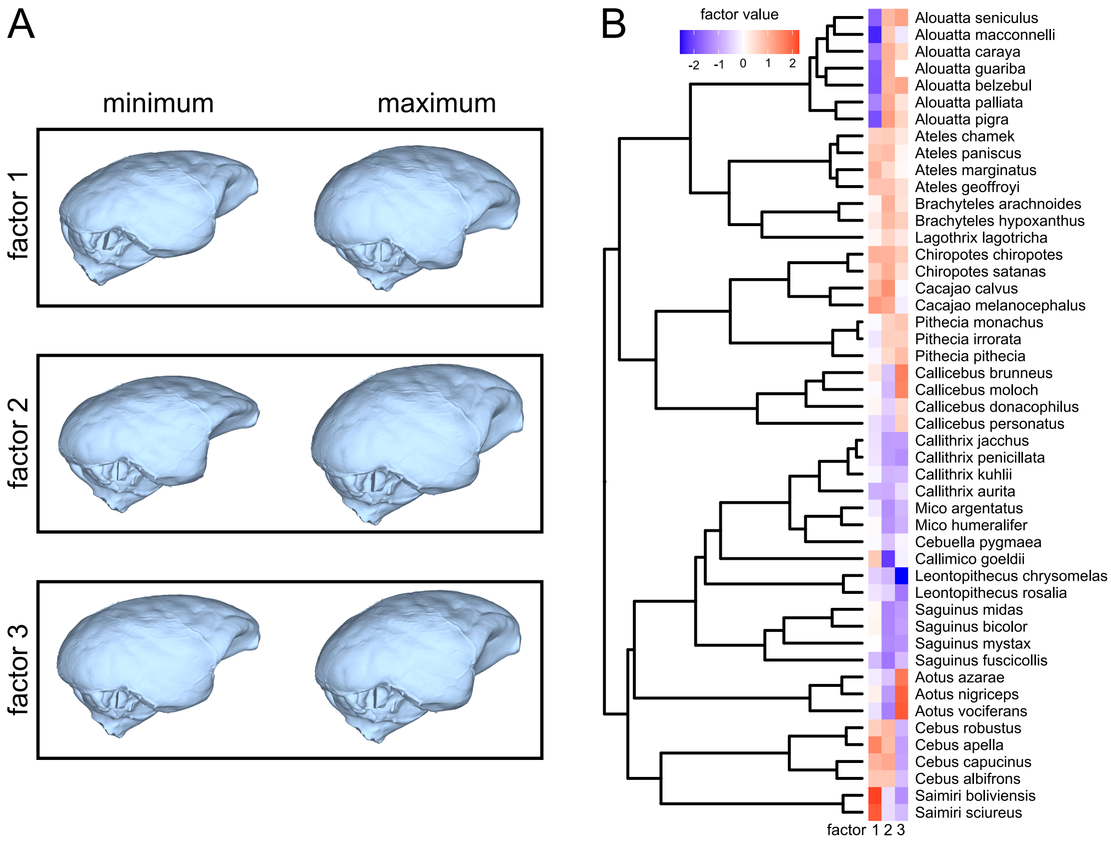


Figure 4.6: **A)** Influence of each factor on New World monkey brain shape. **B)** Brain shape factors plotted along New World monkey phylogeny. The coefficients of the first three principal components (PCs) from Aristide et al. (2016) are highly correlated with the corresponding rows of the loadings matrix. While we do not explore such an analysis here, Aristide et al. (2016) provide evidence of association of PC1 (strongly correlated with our first factor) with relative brain size and PC2 (strongly correlated with our second factor) with diet.

4.7 Discussion

We develop a practical and scalable analysis plan requiring minimal user decisions enabled by computationally innovative inference procedures. Previously, researchers performing phylogenetic factor analysis were limited by computational constraints and had to determine *a priori* the ordering of the traits and optimal number of factors. These computational and modeling advances are not independent but rather complement each other. Our default model selection procedure requires 26 individual MCMC chain simulations (5-fold cross validation with 5 sets of meta-parameters plus the final run). Such an analysis would be intractable for all but the smallest data sets using existing inference techniques. However, our new inference procedures take only a few hours to run all 26 simulations for even the largest data sets we analyze. Additionally, we have made these tools both flexible and accessible with the Julia package PhylogeneticFactorAnalysis.jl, which assembles and runs all BEAST input files, automatically performs model selection, plots the results and performs basic quality control. Our implementation allows researchers to focus on big-picture modeling decisions and leave low-level implementation details to the software.

Limitations of this work that we plan to address in the future include the following. First, while we can accommodate discrete phenotypes through the latent probit model of Cybis et al. (2015) (see Section 4.8.11), we notice both in our analysis and Tolkoff et al. (2018) that the discrete parameters tend to have a far higher influence than their continuous counterparts (i.e. the loadings entries associated with the discrete traits have greater magnitude than those associated with continuous traits). This is likely due to the fact that we control the variance of the latent liabilities indirectly by fixing the discrete trait precisions $\mathbf{\Lambda}$ to a constant as do Tolkoff et al. (2018). It is possible that the (potentially) inflated significance of these discrete traits can influence the loadings structure in unexpected ways, and we seek an alternative solution that places the continuous and discrete traits on more equal footing.

Second, there may be cases where label switching persists despite our efforts to induce identifiability. Additional post-processing procedures developed for Bayesian mixture models (Rodríguez and Walker, 2014) or multidimensional scaling (Okada and Mayekawa, 2018) may serve as solutions to these unusually convolved posteriors. While preliminary work suggests that these methods can efficiently identify and deconvolve individual modes of multi-modal posteriors, we are concerned about their potential to identify non-existent signal in the data and believe a careful analysis of their properties is warranted.

Additionally, as proposed in Section 4.6.4, this work can be readily extended to incorporate parallel evolutionary models for different suites of traits. In this framework, we could simultaneously perform factor analysis on a high-dimensional trait (e.g. brain shape) and infer the evolutionary correlation between the latent factors and other phenotypes of interest (e.g. brain size, diet, group size) using an MBD model. Note that we could study relationships between multiple, distinct high-dimensional phenotypes as well from structural equation modeling paradigm (Lee and Song, 2012). While likelihood calculations under such models are straightforward given this and previous work, inferring the joint evolutionary covariance matrix requires additional inference machinery that we leave as future work.

Finally, while we focus on the multivariate Brownian diffusion model of phenotypic evolution for simplicity, all inference machinery can be readily adapted to other Gaussian processes, such as the multivariate Ornstein–Uhlenbeck (OU) process (Hansen, 1997). Indeed, the OU model and inference procedure of Bastide et al. (2018) have already been implemented in BEAST and are easily integrated with the methods presented in this paper.

4.8 Appendix

4.8.1 Post-order traversal likelihood calculations

We seek to compute $p(\mathbf{Y} \,|\, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F})$ in $\mathcal{O}(NPK^2 + NK^3)$ by adapting the methods developed by Bastide et al. (2018), Mitov et al. (2020) and Hassler et al. (2020). Let $\mathbf{Y}^{\text{obs}} = \left(\mathbf{y}_1^{\text{obs}}, \ldots, \mathbf{y}_N^{\text{obs}}\right)^t$ be the $N \times P$ matrix of observed data, where all missing measurements in $\mathbf{Y}$ have been replaced with 0's. This post-order algorithm requires that one can compute the partial mean $\mathbf{m}_i$, precision $\mathbf{P}_i$ and remainder r_i such that

$$\begin{aligned} &p\left(\mathbf{y}_i^{\text{obs}} \,\middle|\, \mathbf{f}_i, \mathbf{L}, \mathbf{\Lambda}\right) = r_i \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \text{ where} \\ &\qquad \hat{\theta}(x; \boldsymbol{\mu}, \mathbf{P}) = (2\pi)^{-\text{rank}(\mathbf{P})/2}\, \hat{\det}(\mathbf{P})^{1/2} \exp\left(-\frac{1}{2}\left(x - \boldsymbol{\mu}\right)^t \mathbf{P}\left(x - \boldsymbol{\mu}\right)\right), \end{aligned} \tag{4.7}$$

rank($\mathbf{P}$) is the number of non-zero singular values of $\mathbf{P}$ and $\hat{\det}(\mathbf{P})$ is the product of the non-zero singular values of $\mathbf{P}$. We also define the indicator matrices $\boldsymbol{\delta}_i = \text{diag}[\delta_{i1}, \ldots, \delta_{iP}]$

where $\delta_{ij} = 1$ if y_{ij} is observed and $\delta_{ij} = 0$ if it is missing. Finally, we define $P_i^{\text{obs}} = \sum_{j=1}^{P} \delta_{ij}$ as the number of observed traits for taxon i.

In the context of PFA, we calculate

$$\begin{aligned}\log p\big(\mathbf{y}_i^{\text{obs}} \,\big|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}\big) &= -\frac{\operatorname{rank}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i)}{2} \log 2\pi + \frac{1}{2} \log \hat{\det}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i) \\ &\qquad - \frac{1}{2} \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right)^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right) \\ &= \log r_i + \log \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \text{ where}\end{aligned} \tag{4.8}$$

the precision $\mathbf{P}_i = \mathbf{L} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{L}^t$, the mean $\mathbf{m}_i$ is a solution to $\mathbf{P}_i \mathbf{m}_i = \mathbf{L}^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}}$ and

$$\begin{aligned}\log r_i = -\frac{P_i^{\text{obs}} - \operatorname{rank}(\mathbf{P}_i)}{2} \log 2\pi + \frac{1}{2} \left(\sum_{j=1}^{P} \delta_{ij} \log \lambda_j - \log \hat{\det}(\mathbf{P}_i) \right) \\ - \frac{1}{2} \left[\mathbf{y}_i^{\text{obs}^t} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t \mathbf{P}_i \mathbf{m}_i \right].\end{aligned} \tag{4.9}$$

See Section 4.8.1.1 for detailed calculations. As $\boldsymbol{\Lambda}$ is diagonal, computing all $\mathbf{P}_i$ has complexity $\mathcal{O}(NPK^2)$, which dominates the computation time for these operations.

After computing $\mathbf{m}_i$, $\mathbf{P}_i$ and r_i, the Hassler et al. (2020, Section 3.2.1.2) algorithm requires minor modification to compute the likelihood $p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big)$ in $\mathcal{O}(NK^3)$ additional time. Specifically, $\mathbf{P}_i$ may not be invertible via the special inverse defined in Hassler et al. (2020, Section 3.2.1.1). Section 4.8.1.2 offers an alternative approach that avoids this inversion via the continuously rediscovered identity $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \left(\mathbf{I} + \mathbf{B}\mathbf{A}^{-1}\right)^{-1} \mathbf{B}\mathbf{A}^{-1}$ for conformable square matrices $\mathbf{A}$ and $\mathbf{B}$ (Henderson et al., 1959; Henderson and Searle, 1981). We also utilize a more numerically stable modification of this post-order algorithm proposed by Bastide et al. (2021).

4.8.1.1 Partial likelihood calculations under the latent factor model

We present the detailed calculations from Equation 4.8.

$$
\begin{aligned}
\log p\big(\mathbf{y}_i^{\text{obs}} \,\big|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}\big) &= -\frac{\text{rank}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i)}{2} \log 2\pi + \frac{1}{2} \log \hat{\det}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i) \\
&\qquad - \frac{1}{2} \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right)^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right) \\
&= -\frac{P_i^{\text{obs}}}{2} \log 2\pi + \frac{1}{2} \sum_{j=1}^{P} \delta_{ij} \log \lambda_j \\
&\qquad - \frac{1}{2} \left[\mathbf{f}_i^t \mathbf{L}^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{L}^t \mathbf{f}_i - 2\mathbf{f}_i^t \mathbf{L}^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}} + \mathbf{y}_i^{\text{obs}^t} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}}\right] \\
&= -\frac{P_i^{\text{obs}}}{2} \log 2\pi + \frac{1}{2} \sum_{j=1}^{P} \delta_{ij} \log \lambda_j \\
&\qquad - \frac{1}{2} \left[(\mathbf{f}_i - \mathbf{m}_i)^t \, \mathbf{P}_i \, (\mathbf{f}_i - \mathbf{m}_i)\right] - \frac{1}{2} \left[\mathbf{y}_i^{\text{obs}^t} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t \mathbf{P}_i \mathbf{m}_i\right] \\
&= \log r_i - \frac{\text{rank}(\mathbf{P}_i)}{2} \log 2\pi + \frac{1}{2} \log \hat{\det}(\mathbf{P}_i) \\
&\qquad - \frac{1}{2} \left[(\mathbf{f}_i - \mathbf{m}_i)^t \, \mathbf{P}_i \, (\mathbf{f}_i - \mathbf{m}_i)\right] \\
&= \log r_i + \log \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \ \text{ where}
\end{aligned}
\tag{4.10}
$$

the partial precision $\mathbf{P}_i = \mathbf{L} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{L}^t$, the partial mean $\mathbf{m}_i$ is a (not necessarily unique) solution to $\mathbf{P}_i \mathbf{m}_i = \mathbf{L}^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}}$ and the remainder

$$
\begin{aligned}
\log r_i = -\frac{P_i^{\text{obs}} - \text{rank}(\mathbf{P}_i)}{2} \log 2\pi + \frac{1}{2} \left(\sum_{j=1}^{P} \delta_{ij} \log \lambda_j - \log \hat{\det}(\mathbf{P}_i) \right) \\
- \frac{1}{2} \left[\mathbf{y}_i^{\text{obs}^t} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t \mathbf{P}_i \mathbf{m}_i\right].
\end{aligned}
\tag{4.11}
$$

4.8.1.2 Special inverse calculations

One challenge that the PFA model poses to this approach is that the partial precisions at the tips $\mathbf{P}_i$ for $i = 1, \ldots, N$ may not be invertible via the pseudoinverse used by Hassler et al. (2020, Section 3.2.1.1). The post-order traversal algorithm requires that for each internal node ν_i for $i = N+1, \ldots, 2N-1$ in $\mathcal{F}$, we must compute $\mathbf{P}_i^*$ such that $p\big(\mathbf{Y}_{\lfloor i \rfloor} \,\big|\, \mathbf{f}_{\text{pa}(i)}\big) = r_i \hat{\theta}\big(\mathbf{f}_{\text{pa}(i)}; \mathbf{m}_i, \mathbf{P}_i^*\big)$, where $\mathbf{Y}_{\lfloor i \rfloor}$ represents the trait values of all terminal descendants of node ν_i. In the PFA model, this results in $\mathbf{P}_i^* = \big(\mathbf{P}_i^{-1} + t_i \mathbf{I}_K\big)^{-1}$. However, it is possible that the initial partial precisions $\mathbf{P}_i$ at the tip nodes $\nu_1, \ldots, \nu_N$ may be rank-deficient. This situation arises, for example, when the number of non-missing traits P_i^{obs} at taxon i is less than the number of factors K. To avoid this inversion, we use an algebraic slight-of-hand to compute $\mathbf{P}_i^*$ in terms of $\mathbf{P}_i$ directly (rather than its non-existing inverse). Specifically we use an identity for the inverse of the sum of two square matrices that has been discovered and forgotten several times (see, for example, Henderson et al., 1959; Henderson and Searle, 1981)

$$(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \left(\mathbf{I} + \mathbf{B}\mathbf{A}^{-1}\right)^{-1} \mathbf{B}\mathbf{A}^{-1}. \tag{4.12}$$

Applying this to our equation for $\mathbf{P}_i^*$, we get

$$\mathbf{P}_i^* = \mathbf{P}_i - t_i \mathbf{P}_i \left(\mathbf{I}_k + t_i \mathbf{P}_i\right)^{-1} \mathbf{P}_i. \tag{4.13}$$

Note that the matrix $\mathbf{I}_K + t_i \mathbf{P}_i$ is the sum of the positive semi-definite matrix $t_i \mathbf{P}_i$ with the positive definite matrix $\mathbf{I}_K$ and is therefore invertible. As such, computing $\mathbf{P}_i^*$ is indeed possible and the Hassler et al. (2020, Section 3.2.1.2) algorithm can proceed to compute the likelihood.

4.8.2 Sampling from the loadings L via data augmentation

To employ the Gibbs sampler of Tolkoff et al. (2018) to sample from the loading $\mathbf{L}$, we follow the procedure below:

1. Sample from $\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}$ via the pre-order algorithm of Hassler et al. (2020, Section 3.2.2.1)

2. Sample from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{\Lambda}$ via the methods discussed in Lopes and West (2004)

4.8.2.1 Pre-order data augmentation algorithm

We seek to sample from $\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}$ via the pre-order algorithm of Hassler et al. (2020, Section 3.2.2.1). This procedure relies on first computing the statistics $\mathbf{m}_i$ and $\mathbf{P}_i$ such that

$$p\big(\mathbf{Y}^{\text{obs}}_{\lfloor i \rfloor} \,\big|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\big) \propto \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i) \tag{4.14}$$

for $i = 1, \ldots, 2N-1$ (i.e. all nodes in the tree), where $\mathbf{Y}^{\text{obs}}_{\lfloor i \rfloor}$ is the subset of $\mathbf{Y}^{\text{obs}}$ restricted to the descendants of node ν_i. We compute these statistics at the tips as described in Section 4.2.1.1 and at internal nodes as described in Section 3.2.1.2.

Once we have computed these statistics, we draw the factors at the root from their full conditional distribution $\mathbf{f}_{2N-1} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}, \boldsymbol{\mu}_0, \kappa_0$ as described by Equation 3.15 in Hassler et al. (2020). After sampling the factors $\mathbf{f}_{2N-1}$ at the root node ν_{2N-1} from their full conditional distribution, we perform a pre-order traversal of the tree sampling from $\mathbf{f}_i \,\Big|\, \mathbf{f}_{\text{pa}(i)}, \mathbf{Y}^{\text{obs}}_{\lfloor i \rfloor}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}$ for $j = 1, \ldots, 2N-2$ as described in Section 3.2.2.1. After we have completed this pre-order traversal, we have sampled from the full conditional distribution of $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^t$.

4.8.2.2 Conjugate Gibbs sampler on the loadings L

Here we describe our procedure for sampling from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \boldsymbol{\Lambda}$ via the conjugate Gibbs sampler developed by Lopes and West (2004) and Tolkoff et al. (2018). Let us first introduce notation related to both structured sparsity in the loadings and missing data. Let the K-dimensional vector $\boldsymbol{\ell}_j$ and N-dimensional vector $\mathbf{y}'_j$ be the j^{th} column of $\mathbf{L}$ and $\mathbf{Y}$ respectively for $j = 1, \ldots, P$. Let $\mathbf{x}_j \subseteq \{1, \ldots, K\}$ be the indices corresponding to the unconstrained elements of $\boldsymbol{\ell}_j$ (i.e. those that are not fixed at 0), and let $\mathbf{z}_j \subseteq \{1, \ldots, N\}$ be the indices of the observed (non-missing) elements of $\mathbf{y}'_j$. Finally let the sub-vectors $\boldsymbol{\ell}_{j,\mathbf{x}_j}$ and $\mathbf{f}_{i,\mathbf{x}_j}$ be the elements of $\boldsymbol{\ell}_j$ and $\mathbf{f}_i$, respectively, restricted to the indices in $\mathbf{x}_j$, and let $\mathbf{y}'_{j,\mathbf{z}_j}$ be the elements of $\mathbf{y}'_j$ restricted to the elements in $\mathbf{z}_j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, P$. Note that conditional on the latent factors, the full conditional distributions of each column of the loadings are independent. Additionally, the full conditional of $\boldsymbol{\ell}_j$ depends only on $\mathbf{y}'_j$, and does not depend on the other columns of the data matrix $\mathbf{Y}$ (Lopes and West, 2004). As such, we draw from $\boldsymbol{\ell}_{j,\mathbf{x}_j} \Big| \mathbf{F}, \mathbf{y}'_{j,\mathbf{z}_j}, \boldsymbol{\Lambda}$ as follows:

$$\begin{aligned} p\Big(\boldsymbol{\ell}_{j,\mathbf{x}_j} \Big| \mathbf{y}'_{j,\mathbf{z}_j}, \mathbf{F}, \lambda_j\Big) &\propto p\Big(\mathbf{y}'_{j,\mathbf{z}_j} \Big| \boldsymbol{\ell}_{j,\mathbf{x}_j}, \mathbf{L}, \lambda_j\Big) p\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}\big) \\ &= \prod_{i \in \mathbf{z}_j} p\big(y_{ij} \,\big|\, \mathbf{f}_i, \boldsymbol{\ell}_{j,\mathbf{x}_j}, \lambda_j\big) p\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}\big) \\ &= \prod_{i \in \mathbf{z}_j} \theta\Big(y_{ij}; \boldsymbol{\ell}^t_{j,\mathbf{x}_j} \mathbf{f}_{i,\mathbf{x}_j}, \lambda_j\Big)\, \theta\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}; \mathbf{0}, \boldsymbol{\Lambda}_j\big) \\ &= \theta\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}; \boldsymbol{\eta}_j, \boldsymbol{\Gamma}_j\big) \end{aligned} \tag{4.15}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_j &= \frac{1}{\sigma^2} \mathbf{I}_{|\mathbf{x}_j|}, \\ \boldsymbol{\Gamma}_j &= \boldsymbol{\Lambda}_j + \lambda_j \sum_{i \in \mathbf{z}_j} \mathbf{f}_{i,\mathbf{x}_j} \mathbf{f}^t_{i,\mathbf{x}_j}, \\ \boldsymbol{\eta}_j &= \boldsymbol{\Gamma}_j^{-1} \left(\boldsymbol{\Lambda}_j \mathbf{0} + \lambda_j \sum_{i \in \mathbf{z}_j} y_{ij} \mathbf{f}_{i,\mathbf{x}_j} \right) \end{aligned} \tag{4.16}$$

and $\theta(\mathbf{x}; \boldsymbol{\mu}, \mathbf{P})$ is the multivariate normal density function with argument $\mathbf{x}$, mean $\boldsymbol{\mu}$ and precision $\mathbf{P}$.

Computing $\boldsymbol{\Gamma}_j$ has computational complexity $\mathcal{O}(NK^2)$, so computing all P precisions has overall complexity $\mathcal{O}(NPK^2)$. Once the precisions have been computed, computing the means has complexity $\mathcal{O}(NPK + PK^3)$, which contributes relatively little to overall computation time as $N >> K$ for most problems. Note that if the data are completely observed and there is no structured sparsity in the loadings, then $\boldsymbol{\Gamma}_j = \boldsymbol{\Lambda}_j + \lambda_j \mathbf{F}^t\mathbf{F}$. In that case, we only need to compute $\mathbf{F}^t\mathbf{F}$ once (not P times), which brings the overall complexity down to $\mathcal{O}(NPK)$ (as we still need to compute the means for al P columns of $\mathbf{L}$). Drawing all $\boldsymbol{\ell}_j$ for $j = 1, \ldots, P$ results in a complete sample from the full conditional distribution of $\mathbf{L}$.

4.8.3 Loadings gradient calculation

We calculate the gradient of the likelihood with respect to each column of the loadings $\boldsymbol{\ell}_j$ individually to accommodate variation in the missing data structure across traits. Note that in the calculations below, we omit explicit dependence on the residual precision $\mathbf{\Lambda}$ and tree $\mathcal{F}$ in the interest of notational simplicity.

$$\begin{aligned}
\nabla_{\boldsymbol{\ell}_j} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}\big) &= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}\big) \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \nabla_{\boldsymbol{\ell}_j} \left[\int p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) p(\mathbf{F}) \mathrm{d}\mathbf{F} \right] \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \int p(\mathbf{F}) \nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \mathrm{d}\mathbf{F}.
\end{aligned} \tag{4.17}$$

Based on the fact that the elements of $\mathbf{Y}^{\text{obs}}$ are independent conditional on the loadings and factors, we have:

$$\begin{aligned}
p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) &= \prod_{i=1}^{N} \prod_{k=1}^{P} p\big(y_{ij} \,|\, \mathbf{f}_i, \mathbf{L}\big)^{\delta_{ik}} \\
&= \prod_{i=1}^{N} \prod_{k=1}^{P} (2\pi\lambda_k)^{-\delta_{ik}/2} \exp\left(-\frac{1}{2} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2 \right) \\
&= c \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2 \right),
\end{aligned} \tag{4.18}$$

where δ_{ij} is an indicator that equals 1 if y_{ij} is observed and 0 if it is missing, and c is a normalization constant that does not depend on the loadings $\mathbf{L}$. Therefore,

$$\begin{aligned}
\nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) &= \nabla_{\boldsymbol{\ell}_j}\left[c\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{P}\lambda_k\delta_{ik}\big(y_{ik} - \mathbf{f}_i^t\boldsymbol{\ell}_k\big)^2\right)\right] \\
&= c\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{P}\lambda_k\delta_{ik}\big(y_{ik} - \mathbf{f}_i^t\boldsymbol{\ell}_k\big)^2\right) \\
&\qquad \times \nabla_{\boldsymbol{\ell}_j}\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{P}\lambda_k\delta_{ik}\big(y_{ik} - \mathbf{f}_i^t\boldsymbol{\ell}_k\big)^2\right] \\
&= p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \times \nabla_{\boldsymbol{\ell}_j}\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{P}\lambda_k\delta_{ik}\big(y_{ik} - \mathbf{f}_i^t\boldsymbol{\ell}_k\big)^2\right] \\
&= p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \times -\frac{1}{2}\lambda_j\sum_{i=1}^{N}\delta_{ij}\nabla_{\boldsymbol{\ell}_j}\left[\big(y_{ij} - \mathbf{f}_i^t\boldsymbol{\ell}_j\big)^2\right] \\
&= p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big)\lambda_j\sum_{i=1}^{N}\delta_{ij}\mathbf{f}_i\big(y_{ij} - \mathbf{f}_i^t\boldsymbol{\ell}_j\big) \\
&= p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big)\lambda_j\left(\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\mathrm{obs}'} - \mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\ell}_j\right)
\end{aligned} \tag{4.19}$$

where $\mathbf{y}_j^{\mathrm{obs}'}$ is the j^{th} column of $\mathbf{Y}^{\mathrm{obs}}$ and $\boldsymbol{\delta}'_j = \mathrm{diag}[\delta_{1j}, \ldots, \delta_{Nj}]$ is a diagonal matrix of observed-data indicators. Using this result in Equation 4.17, we calculate

$$\begin{aligned}
\nabla_{\boldsymbol{\ell}_j} \log p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{L}\big) &= \int \frac{p(\mathbf{F})p\big(\mathbf{Y}^{\mathrm{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big)}{p\big(\mathbf{Y}^{\mathrm{obs}} \,|\, \mathbf{L}\big)}\lambda_j\left(\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\mathrm{obs}'} - \mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\ell}_j\right)\mathrm{d}\mathbf{F} \\
&= \int p\big(\mathbf{F} \,\big|\, \mathbf{Y}^{\mathrm{obs}}, \mathbf{L}\big)\lambda_j\left(\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\mathrm{obs}'} - \mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\ell}_j\right)\mathrm{d}\mathbf{F} \\
&= \mathbb{E}\left[\lambda_j\left(\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\mathrm{obs}'} - \mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\ell}_j\right) \,\middle|\, \mathbf{Y}^{\mathrm{obs}}, \mathbf{L}\right] \\
&= \lambda_j\mathbb{E}\big[\mathbf{F}^t \,\big|\, \mathbf{Y}^{\mathrm{obs}}, \mathbf{L}\big]\boldsymbol{\delta}'_j\mathbf{y}_j^{\mathrm{obs}'} - \lambda_j\mathbb{E}\big[\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F} \,\big|\, \mathbf{Y}^{\mathrm{obs}}, \mathbf{L}\big]\boldsymbol{\ell}_j.
\end{aligned} \tag{4.20}$$

Note that

$$\begin{aligned}\mathbb{E}\big[\mathbf{F}^t\boldsymbol{\delta}_j'\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] &= \sum_{i=1}^{N} \delta_{ij}\mathbb{E}\big[\mathbf{f}_i\mathbf{f}_i^t \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] \\ &= \sum_{i=1}^{N} \delta_{ij}\mathbb{V}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] + \delta_{ij}\mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]\mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]^t.\end{aligned} \tag{4.21}$$

We compute $\mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]$ and $\mathbb{V}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]$ for $i = 1, \ldots, N$ in $\mathcal{O}(NPK^2 + NK^3)$ via a post-order likelihood calculation algorithm (see Section 4.8.1) followed by the pre-order algorithms independently developed by Bastide et al. (2018) and Fisher et al. (2021).

For the case where there is no missing data, we can simplify Equation 4.19 to be

$$\nabla_{\mathbf{L}} p(\,\mathbf{Y} \,|\, \mathbf{F}, \mathbf{L}\,) = p(\,\mathbf{Y} \,|\, \mathbf{F}, \mathbf{L}\,)\left[\mathbf{F}^t\mathbf{Y}\boldsymbol{\Lambda} - \mathbf{F}^t\mathbf{F}\mathbf{L}\boldsymbol{\Lambda}\right]. \tag{4.22}$$

4.8.4 Post-processing procedure

We employ singular value decomposition (SVD) to enforce the orthogonality constraint on the loadings via post-processing. In practice, we sample from the orthogonally-constrained loadings as follows. Let $\mathbf{L}^{(d)}$ be a sample from the posterior distribution $\mathbf{L} \,|\, \mathbf{Y}$ at the d^{th} state in the MCMC chain. For each $\mathbf{L}^{(d)}$, we compute the SVD $\mathbf{L}^{(d)} = \mathbf{U}^{(d)}\boldsymbol{\Sigma}^{(d)}\mathbf{V}^{(d)}$ where $\mathbf{U}^{(d)}$ is a $K \times K$ orthonormal matrix and $\boldsymbol{\Sigma}^{(d)}$ and $\mathbf{V}^{(d)}$ retain their constraints from Section 4.2.2.2 (i.e. $\boldsymbol{\Sigma}^{(d)}$ is diagonal with descending positive entries and $\mathbf{V}^{(d)}{\mathbf{V}^{(d)}}^t = \mathbf{I}_K$). While the parameter $\mathbf{U}$ is not identifiable, $\boldsymbol{\Sigma}$ and $\mathbf{V}$ are (Holbrook et al., 2016). As such, we then treat ${\mathbf{L}^{\perp}}^{(d)} = \boldsymbol{\Sigma}^{(d)}\mathbf{V}^{(d)}$ as (now identifiable) samples from the posterior of the loadings. If we also sample the factors $\mathbf{F}$, we rotate the factors to sample from ${\mathbf{F}^{\perp}}^{(d)} = \mathbf{F}^{(d)}\mathbf{U}^{(d)}$ to ensure that ${\mathbf{F}^{\perp}}^{(d)}{\mathbf{L}^{\perp}}^{(d)} = \mathbf{F}^{(d)}\mathbf{U}^{(d)}\boldsymbol{\Sigma}^{(d)}\mathbf{V}^{(d)} = \mathbf{F}^{(d)}\mathbf{L}^{(d)}$.

4.8.5 Sampling from $\boldsymbol{\Sigma}$

We define the K-vector $\boldsymbol{\sigma}$ such that $\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\sigma}]$ and sample $\boldsymbol{\sigma}$ as follows (see Section 4.8.5.1 for derivation):

$$\begin{aligned} \boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda} \sim \text{MVN}\big(\boldsymbol{\mu}_{\sigma}, \mathbf{P}_{\sigma}^{-1}\big), \text{ where} & \\ \mathbf{P}_{\sigma} &= \text{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P} \lambda_j \,\text{diag}[\mathbf{v}_j]\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F}\,\text{diag}[\mathbf{v}_j], \\ \boldsymbol{\mu}_{\sigma} &= \mathbf{P}_{\sigma}^{-1} \left(\sum_{j=1}^{P} \lambda_j \,\text{diag}[\mathbf{v}_j]\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} \right), \end{aligned} \tag{4.23}$$

$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ and $\mathbf{v}_j$ is the j^{th} column of $\mathbf{V}$.

While the prior encourages the elements of $\boldsymbol{\sigma}$ to have descending absolute value, it does not enforce this constraint strictly. As discussed in Section 4.2.2.2, for some problems a strict ordering with forced spacing may be necessary in practice for full identifiability. In these cases we employ a rejection sampler where we draw from the full conditional distribution of $\boldsymbol{\sigma}$ using the unrestricted multivariate normal distribution and reject any samples that do not conform to the particular constraint. As the unconstrained prior already induces a soft ordering, we find that this rejection sampler typically has high acceptance probability.

4.8.5.1 Loadings scale full conditional distribution

We detail our derivation of Equation 4.23 below. Recall that we define the K-vector $\boldsymbol{\sigma}$ such that $\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\sigma}]$, and note that all proportional symbols imply log-proportional:

$$
\begin{aligned}
&\log p\big(\boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) \\
&\qquad \propto \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \boldsymbol{\sigma}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) + \log p(\boldsymbol{\sigma}) \\
&\qquad = \sum_{j=1}^{P} \log p\Big(\mathbf{y}_j^{\text{obs}'} \,\Big|\, \boldsymbol{\sigma}, \mathbf{F}, \mathbf{v}_j, \lambda_j\Big) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - \mathbf{y}_j^{\text{obs}'}\right)^t \boldsymbol{\delta}'_j \left(\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - \mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\mathbf{v}_j^t\boldsymbol{\Sigma}\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - 2\mathbf{v}_j^t\boldsymbol{\Sigma}\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\boldsymbol{\sigma}^t\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\boldsymbol{\sigma} - 2\boldsymbol{\sigma}^t\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\right)\boldsymbol{\sigma} \\
&\qquad\qquad\quad - \boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\boldsymbol{\sigma}^t\left(\text{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\right)\boldsymbol{\sigma} \\
&\qquad\qquad\quad - \boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) \\
&\qquad \propto -\frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\mu}_\sigma\right)^t\mathbf{P}_\sigma\left(\boldsymbol{\sigma} - \boldsymbol{\mu}_\sigma\right),
\end{aligned}
\tag{4.24}
$$

where

$$
\begin{aligned}
\mathbf{P}_\sigma &= \operatorname{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P} \lambda_j \operatorname{diag}[\mathbf{v}_j] \mathbf{F}^t \boldsymbol{\delta}'_j \mathbf{F} \operatorname{diag}[\mathbf{v}_j] \text{ and} \\
\boldsymbol{\mu}_\sigma &= \mathbf{P}_\sigma^{-1} \left(\sum_{j=1}^{P} \lambda_j \operatorname{diag}[\mathbf{v}_j] \mathbf{F}^t \boldsymbol{\delta}'_j \mathbf{y}_j^{\text{obs}\prime} \right)
\end{aligned}
\tag{4.25}
$$

This implies

$$
\log p\big(\boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) = \theta\big(\boldsymbol{\sigma}; \boldsymbol{\mu}_\sigma, \mathbf{P}_\sigma\big) . \tag{4.26}
$$

4.8.6 Sign constraint on the loadings

Regardless of which prior (i.i.d. vs shrinkage) or constraint (sparsity vs orthogonality) we choose, we must enforce a sign constraint on a single element in each row of $\mathbf{L}$ for full identifiability. Let $\gamma_k \in \{1, \ldots, P\}$ be the index of the K^{th} row of $\mathbf{L}$ with the sign constraint (i.e. require $\ell_{\gamma_k k} \geq 0$). If the sample $\ell_{k\gamma_k}^{(d)} < 0$, then we simply multiply row k of $\mathbf{L}^{(d)}$ by -1 to ensure $\ell_{k\gamma_k}^{(d)} \geq 0$. These K sign-constrained elements are not required to be in the same row of $\mathbf{L}$, and we choose these rows in a way that maximizes the posterior identifiability of $\mathbf{L}$. In practice, we apply a simple heuristic where for $k = 1, \ldots, K$

$$
\gamma_k = \underset{j \in 1, \ldots, P}{\arg\max} \left(\frac{\bar{\ell}_{jk}^{\text{abs}}}{\sqrt{\sum_{d=1}^{D} \left(\left| \ell_{jk}^{(d)} \right| - \bar{\ell}_{jk}^{\text{abs}} \right)^2}} \right) \quad \text{and} \quad \bar{\ell}_{jk}^{\text{abs}} = \frac{1}{D} \sum_{d=1}^{D} \left| \ell_{jk}^{(d)} \right| . \tag{4.27}
$$

In the absence of sign constraints, the marginal posteriors of many elements of $\mathbf{L}$ are bimodal and symmetric across zero. Our heuristic aims to find an index in each column of $\mathbf{L}$ with low mass near 0 and simply chose the positive mode.

4.8.7 Sampling from $\boldsymbol{\Lambda}$

Regardless of the prior on the loadings, we sample from $\boldsymbol{\Lambda} \,\big|\, \mathbf{F}, \mathbf{Y}^{\text{obs}}, \mathbf{L}$ using the same conjugate Gibbs sampler as Tolkoff et al. (2018) in conjunction with the data augmentation algorithm from Section 4.3.1.1. The Gamma$(a_{\boldsymbol{\Lambda}}, b_{\boldsymbol{\Lambda}})$ (shape, rate parameterization) prior on the diagonal elements of $\boldsymbol{\Lambda}$ results in a simple expression for the full conditional distribution of λ_j for $j = 1, \ldots, P$ conditional on the factors $\mathbf{F}$. Specifically, each λ_j is distributed as

$$\lambda_j \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{L} \sim \text{Gamma}\left(a_{\boldsymbol{\Lambda}} + \frac{N_j^{\text{obs}}}{2}, b_{\boldsymbol{\Lambda}} + \frac{1}{2} \sum_{i=1}^{N} \delta_{ij} \left(y_{ij} - \boldsymbol{\ell}_j^t \mathbf{f}_i \right)^2 \right). \tag{4.28}$$

This computation only requires run time $\mathcal{O}(NPK)$ and, in our experience, time spent estimating $\boldsymbol{\Lambda}$ does not contribute significantly to the overall run time of the MCMC chain.

Note that as with the loadings in Section 4.3.1.2, we also derive a strategy for sampling from these precisions without conditioning on $\mathbf{F}$ via HMC. As we are satisfied with the Tolkoff et al. (2018) procedure, we have not implemented this strategy, but the derivation can be found below. Naturally, this HMC sampler requires we compute the gradient of the likelihood with respect to the loadings as follows:

$$\begin{aligned}
\frac{\partial \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \boldsymbol{\Lambda} \big)}{\partial \lambda_j} &= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Lambda} \big)} \int p(\mathbf{F}) \frac{\partial p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \boldsymbol{\Lambda} \big)}{\partial \lambda_j} \mathrm{d}\mathbf{F} \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Lambda} \big)} \int p(\mathbf{F}) p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \boldsymbol{\Lambda} \big) \\
&\qquad\qquad \times \left(\frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right)^t \boldsymbol{\delta}_j^{\prime} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right) \right) \mathrm{d}\mathbf{F} \\
&= \mathbb{E}\left[\frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right)^t \boldsymbol{\delta}_j^{\prime} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right) \,\middle|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda} \right] \\
&= \frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \boldsymbol{\ell}_j^t \mathbb{E}\big[\mathbf{F}^t \boldsymbol{\delta}_j^{\prime} \mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda} \big] \boldsymbol{\ell}_j + \boldsymbol{\ell}_j^t \mathbb{E}\big[\mathbf{F}^t \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda} \big] \boldsymbol{\delta}_j^{\prime} \mathbf{y}_j^{\text{obs}\prime} \\
&\quad - \frac{1}{2} \mathbf{y}_j^{\text{obs}\prime t} \boldsymbol{\delta}_j^{\prime} \mathbf{y}_j^{\text{obs}\prime}
\end{aligned} \tag{4.29}$$

The conditional expectations of the latent factors are the same as in Section 4.3.1.2. Note that we restrict $\boldsymbol{\Lambda}$ to be diagonal, so we only consider the diagonal elements of the gradient. Once we have computed this gradient, we employ it in standard HMC to sample from the full conditional of $\boldsymbol{\Lambda}$.

4.8.8 Timing

4.8.8.1 Simulation details

To simulate each data set for the timing comparison, we generate a random coalescent tree with N tips (Kingman, 1982). We then simulate the factors $\mathbf{F}$ according to K independent Brownian diffusion processes on the tree and subsequently re-scale the factors so that each column has unit variance. We draw $\mathbf{V}$ from a uniform distribution on the Stiefel manifold. To avoid identifiability challenges associated with values of $\boldsymbol{\Sigma}$ having similar magnitudes, we set $\sigma_k = 2^{-k}\sqrt{P}$ for $k = 1, \ldots, K$. Note that we multiply by $\sqrt{P}$ so that the expectations of $\ell_{kj}^2 = \sigma_k^2 v_{kj}^2$ remain the same regardless of P. We sample the residual variances λ_j^{-1} independently from Gamma$(2, 4)$ for $j = 1, \ldots, P$, which keeps the contribution of the residual variance to the total variance similar to that of the latent factors. Finally, we draw $\boldsymbol{\epsilon} \sim \text{MN}\big(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\Lambda}^{-1}\big)$ and compute $\mathbf{Y} = \mathbf{F}\boldsymbol{\Sigma}\mathbf{V} + \boldsymbol{\epsilon}$. As all methods rely on the same principles for handling missing data, we do not remove any observations from the simulated data sets.

When performing inference, we assume the tree is fixed to its true value used to simulate the factors $\mathbf{F}$. We use the orthogonality constraint on the loadings and employ the post-processing regime discussed in Section 4.3.1.3 to rotate results from each sampler (except the one associated with the orthogonal shrinkage prior) to enforce this constraint. For the model with the orthogonal shrinkage prior, we assume both forced ordering and spacing ($\alpha = 0.9$).

4.8.8.2 Effective sample size calculations

To understand the relative performance of each inference regime, we compare the effective sample size (ESS) per second of the loadings across all four samplers. Draws from an MCMC simulation are often auto-correlated, and the total number of steps in the chain is rarely a direct proxy for our confidence in the posterior estimates. ESS approximates the number of *independent* samples from the chain. As researchers typically set a minimum ESS threshold to determine the length of MCMC simulations, we compare the minimum ESS per unit time.

Let $\mathrm{ESS}_{kj}^{(m)}$ be the effective sample size for ℓ_{kj} in replicate m and $\mathrm{ESS}_{\min}^{(m)} = \min_{k,j} \mathrm{ESS}_{kj}^{(m)}$ for $m = 1, \ldots, 3$. We compute $\overline{\mathrm{ESS}}_{\min} = \frac{1}{3}\sum_{m=1}^{3} \mathrm{ESS}_{\min}^{(m)}/t^{(m)}$ for all models, where $t^{(m)}$ is the time required for the m^{th} MCMC simulation. Actual ESS values were calculated using the Julia package MCMCDiagnostics.jl. We compare these values in Figure 4.1 and Table 4.2.

Table 4.2: Comparison of computational efficiency. Effective sample size computed using the Julia package MCMCDiagnostics.jl.

N	P	K	minimum ESS per minute				speed increase over sampled		
			Sampled	Gibbs	HMC	orthogonal	Gibbs	HMC	orthogonal
50	10	1	530	5100	5700	2000	9.8×	11.0×	3.8×
		2	500	3900	2500	810	7.8×	4.9×	1.6×
		4	680	2200	1400	450	3.3×	2.0×	0.7×
	100	1	190	1400	1700	170	7.6×	9.1×	0.89×
		2	150	1000	870	130	7.1×	5.9×	0.89×
		4	52	550	250	20	11×	4.7×	0.39×
	1000	1	34	460	250	5.2	14×	7.4×	0.15×
		2	27	390	85	0.87	14×	3.1×	0.032×
		4	23	320	23	0.51	14×	1.0×	0.022×
100	10	1	270	4100	3000	1100	15×	11×	4.0×
		2	160	2100	2000	400	13×	12×	2.5×
		4	51	680	500	110	13×	9.9×	2.1×
	100	1	33	360	480	94	11×	14×	2.9×
		2	18	240	290	35	13×	16×	1.9×
		4	17	200	83	38	12×	4.8×	2.2×
	1000	1	3.9	54	53	2.9	14×	14×	0.75×
		2	2.5	82	15	0.98	33×	5.8×	0.39×
		4	2.0	99	5.3	0.19	49×	2.6×	0.092×
500	10	1	5.0	740	460	170	150×	92×	33×
		2	3.4	260	280	59	77×	83×	17×
		4	1.7	160	170	30	93×	98×	18×
	100	1	0.77	95	110	25	120×	140×	32×
		2	0.37	20	28	5.4	56×	77×	15×
		4	0.46	18	12	3.7	40×	25×	8.1×
	1000	1	0.02	1.8	0.71	0.68	90×	35×	34×
		2	0.018	2.4	0.65	0.11	130×	36×	6.1×
		4	0.011	1.5	0.16	0.032	140×	15×	2.9×

(table continued on next page)

N	P	K	minimum ESS per minute				speed increase over sampled		
			Sampled	Gibbs	HMC	orthogonal	Gibbs	HMC	orthogonal
1000	10	1	1.1	170	290	58	160×	270×	54×
		2	0.54	84	190	28	160×	350×	52×
		4	0.24	49	80	10	210×	340×	44×
	100	1	0.098	35	38	9.2	350×	390×	94×
		2	0.064	15	12	2.8	230×	180×	44×
		4	0.065	7.6	5.8	1.0	120×	90×	15×
	1000	1	0.0017	0.5	0.25	0.3	300×	150×	180×
		2	0.0015	0.67	0.15	0.085	450×	100×	57×
		4	0.0015	0.4	0.06	0.02	270×	40×	14×

4.8.9 Identifying label switching

As discussed in Section 4.5.2, the post-processing algorithm used to induce orthogonality when sampling under the i.i.d. prior can result in label switching. This phenomenon occurs when elements of the scale parameter $\boldsymbol{\sigma}$ have significantly overlapping posterior distributions. As the post-processing algorithm orders the factors based on the magnitude of the scale parameters, it will swap two factors when their scales switch in order. Because of this, the estimated posterior distribution of the loadings and factor values associated with two factors undergoing label-switching will be a mixture of some (unknown) underlying distributions that we are trying to estimate. This mixing can obscure signals in our data.

Consider Figure 4.7 as a toy example where we know the true underlying distributions. In practice, we do not know these distributions (if we did we wouldn't need to infer them). We assume a 3-factor model where the posterior of scales σ_1 and σ_2 are slightly overlapping. If we then order the rows of the loadings according to the the scales $\boldsymbol{\sigma}$, the estimated rows of the loadings clearly switch at the places where the true $\sigma_1 < \sigma_2$. We see evidence of this occurring in the plot of the loadings where samples from the posterior of ℓ_{11}, which are normally greater than 1, occasionally have unusually low values near 0. At the same points in the chain samples from ℓ_{21}, which are normally near 0, have unusually high values near 1. It appears that the estimated samples from the posterior of ℓ_{11} and ℓ_{21} are occasionally switching between the two.

Label switching is not always as obvious as the simple example depicted here in Figure 4.7. In Figure 4.2, all elements of $\boldsymbol{\sigma}$ appear close to each other and there is likely a higher degree of overlap between pairs of factors. Rather than obvious switching, the posteriors of the loadings under the i.i.d. prior appear to blend into each other. While it is possible that the posteriors of the loadings really are overlapping, the apparently overlapping scale parameters and skewed tails of the each of loadings posterior densities toward the mean of the other distribution suggests label switching. Repeating the analysis with the orthogonal shrinkage prior reveals distinct posterior distributions in the relevant parameters of the loadings, confirming that label switching is occurring under first analysis (i.i.d. prior with post processing).

4.8.10 Cross validation

Our model selection strategy seeks to identify the shrinkage strength (when using the shrinkage prior) or number of factors (when using the i.i.d. prior) that provides optimal predictive performance via cross-validation. To this end, we posit M sub-models characterized by the meta-parameters $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M$. Under the i.i.d. prior, $\mathbf{\Omega}_i = K^{[i]}$ is the number of factors in model i. For example, our default for the i.i.d. prior assumes $K_{\max} = 5$ and $M = 5$ models with $(K^{[1]}, \ldots, K^{[M]}) = (1, 2, 3, 4, 5)$. Under the shrinkage prior, let $\mathbf{\Omega}_i = \{\mathbf{a}^{[i]}, \mathbf{b}^{[i]}\}$ be the shapes and rates, respectively, of the gamma priors on the shrinkage multipliers $\nu_1, \ldots, \nu_K$ for model i. We typically retain $K_{\max} = 5$ and define the 5 sub-models as $\mathbf{a}^{[i]} = 10^{(i+1)/2}\mathbf{1}_{K_{\max}}$ and $\mathbf{b}^{[i]} = \mathbf{1}_{K_{\max}}$ for $i = 1, \ldots, 5$.

We evaluate the predictive performance of each model on R replicate data sets via R-fold

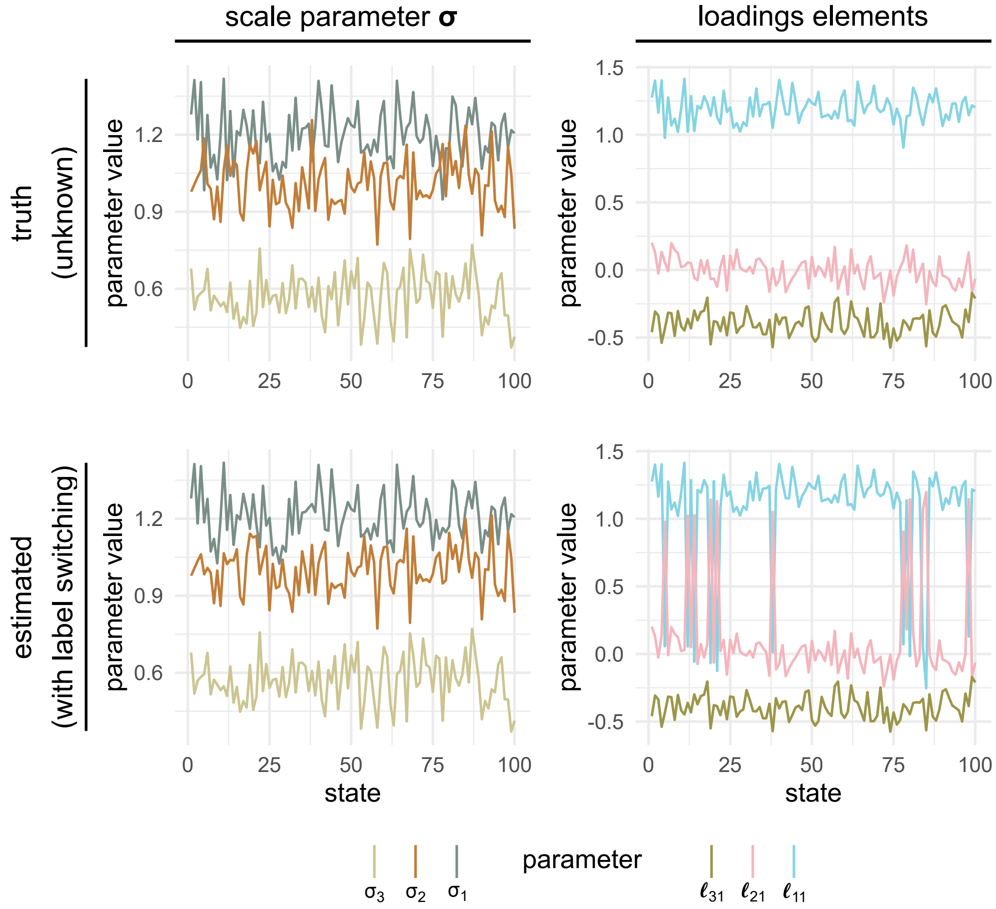


Figure 4.7: Example of label switching. The top trace plots are samples from a known distribution. Note that in practice, we to not know the true underlying distribution. The bottom plot demonstrates how ordering the scale parameters can induce label switching between rows of the loadings. Here, there is label switching between the first two factors, but not the third. The switching in the estimated parameters occurs at the MCMC states where the "true" $\sigma_1 < \sigma_2$ (normally the reverse is true).

cross-validation. For each replicate $j = 1, \ldots, R$, we randomly partition the observed data $\mathbf{Y}^{\text{obs}}$ into a training set $\mathbf{Y}_j^{\text{tr}}$ containing $(100 - \frac{100}{R})\%$ of the data and a validation set $\mathbf{Y}_j^{\text{val}}$ with the remaining $\frac{100}{R}\%$ such that each observation occurs in exactly one validation set.

Let $\boldsymbol{\Theta} = \{\mathbf{L}, \boldsymbol{\Lambda}\}$ be the model parameters relevant to the likelihood. We first approximate $p\big(\boldsymbol{\Theta} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \boldsymbol{\Omega}_i\big)$ for $i = 1, \ldots, M$, $j = 1, \ldots, R$ via MCMC simulation as described in Section 4.3. We then compute the expected log predictive density (Gelman et al., 2013) $\pi_{ij} = \mathbb{E}\big[\log p\big(\mathbf{Y}_j^{\text{val}} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \boldsymbol{\Theta}_{ij}\big)\big]$ for $i = 1, \ldots, M$, $j = 1, \ldots, R$, where $\boldsymbol{\Theta}_{ij}$ is a random variable with density $p\big(\boldsymbol{\Theta} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \boldsymbol{\Omega}_i\big)$. We select $\boldsymbol{\Omega}_m$, where $m = \arg\max_i \frac{1}{R}\sum_j \pi_{ij}$, as the optimal model and approximate $p\big(\mathbf{L}, \boldsymbol{\Lambda} \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Omega}_m\big)$ as the final step in the analysis plan.

4.8.11 Phylogenetic latent liability model

In the case of binary traits, we assume the latent liability model of Cybis et al. (2015). Specifically, rather than assuming the observations $\mathbf{Y} = \mathbf{FL} + \boldsymbol{\epsilon}$, we introduce an additional latent variable $\mathbf{Z} = \{z_{ij}\}$ for $i = 1, \ldots, N$, $j = 1, \ldots, P$ and assume $\mathbf{Z} = \mathbf{FL} + \boldsymbol{\epsilon}$. These latent liabilities z_{ij} are connected to the observations y_{ij} via the link function $y_{ij} = g_j(z_{ij})$ where $g_j(x) = x$ if trait j is continuous, $g_j(x) = 1\{x \leq 0\}$ if j is binary.

Under this model, the full conditional distributions of the latent liabilities are independent truncated Gaussian distributions with densities

$$p\big(z_{ij} \,\big|\, y_{ij}, \mathbf{f}_i, \boldsymbol{\ell}_j, \lambda_j, \mathbf{t}_j\big) \sim \theta\big(z_{ij}; \mathbf{f}_i^t \boldsymbol{\ell}_j, \lambda_j\big)\, 1\{g_j(z_{ij}) = y_{ij}\}\,. \tag{4.30}$$

As these full conditional distributions are independent, we can sample from them efficiently via a simple rejection sampler. Specifically, we first draw from $\mathbf{F} \,|\, \mathbf{Z}, \boldsymbol{\Lambda}, \mathcal{F}$ as in Section 4.3.1.1. We then sample the proposal $z_{ij} \sim \mathcal{N}\big(\mathbf{f}_i^t \boldsymbol{\ell}_j, 1/\lambda_j\big)$ that we accept if $g_j(z_{ij}) = y_{ij}$ and reject otherwise. Note that for each discrete trait j, we must also fix $\lambda_j = 1$ to ensure the variance of the latent traits j are identifiable (see Tolkoff et al., 2018).

4.8.12 Phylogenetic tree inference

4.8.12.1 Yeast

For they yeast analysis, we first infer a phylogenetic tree for the 154 phenotyped strains using the 2.8 megabase DNA sequence alignment of Gallone et al. (2016) (see subsection *Phylogenetic Tree for the Sequenced Collection* in *Methods* of Gallone et al. (2016) for details). Our phylogenetic tree model includes an uncorrelated relaxed clock model (Drummond et al., 2006), an HKY+G substitution model (Hasegawa et al., 1985; Yang, 1994) and a constant-population coalescent prior on the tree (Kingman, 1982).

We perform MCMC simulation via BEAST (Suchard et al., 2018) to approximate the posterior distribution of the phylogenetic tree. We run the MCMC chain for 10 million states, sampling the tree and related parameters every thousand states and the factor related parameters every 10 thousand states. Inspection of relevant trace plots indicated the the MCMC chain had achieved stationarity by 1 million states, and we exclude the first million states as burn-in. We compute the maximum clade credibility (MCC) tree as a point estimate of the phylogenetic tree using TreeAnnotator (Rambaut and Drummond, 2015).

4.8.12.2 New World monkeys

We simultaneously infer the NWM tree structure with the latent factor model using DNA sequence alignments of Aristide et al. (2015). To infer the tree structure, we partition the taxa into four monophyletic clades consisting of the 1) *Atelidae*, 2) *Aotidae* and *Callitrichidae*, 3) *Cebidae* and 4) *Pitheciidae* respectively and place zero prior probability on tree topologies that do not maintain these clades. Otherwise, we use the same phylogenetic tree model and inference procedure as described in Section 4.8.12.1.

4.8.13 Additional results

We present the full results of our yeast analysis below.

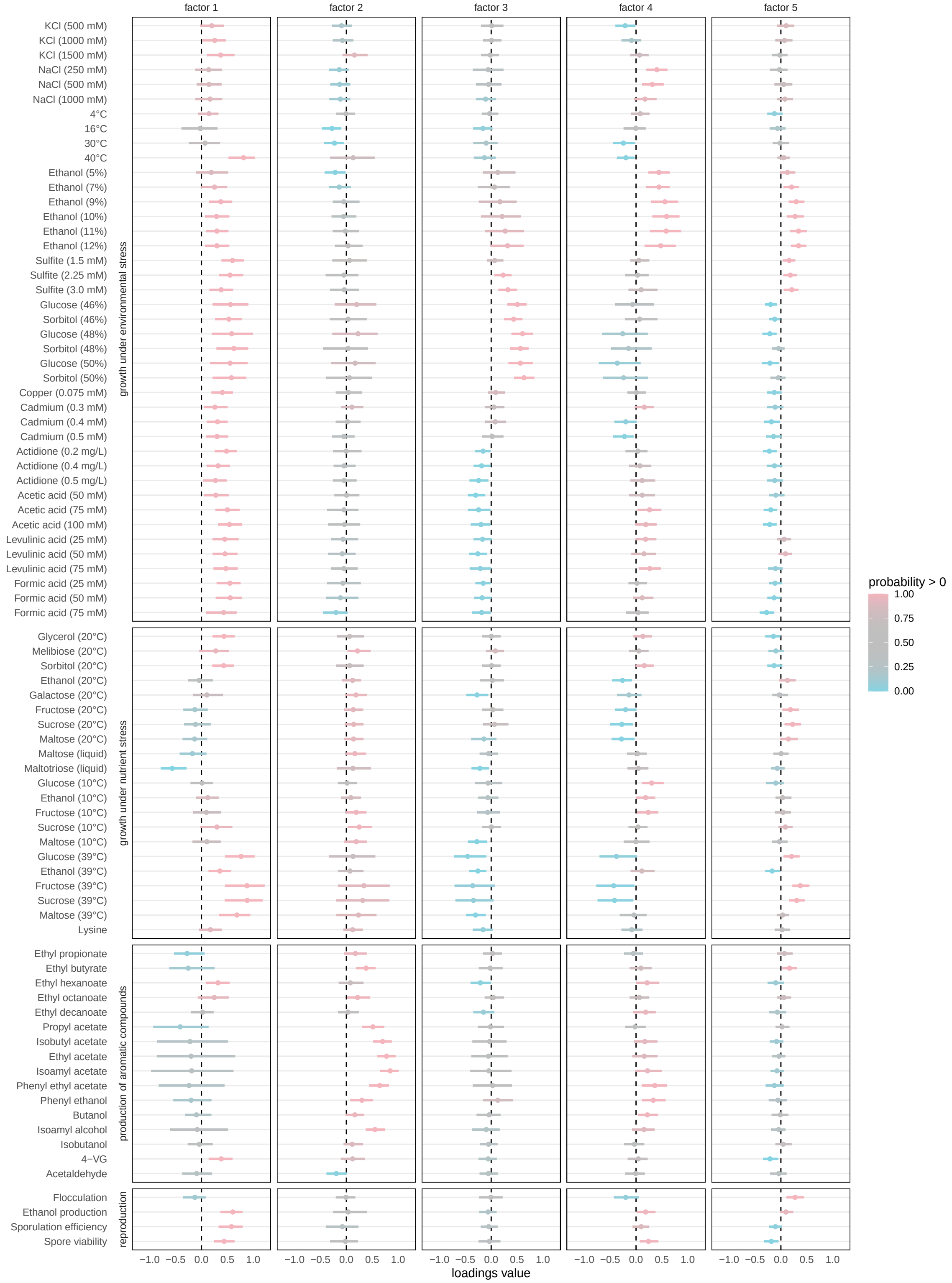


Figure 4.8: Posterior summary of loadings of 5-factor PFA on yeast data set. *(caption on next page)*

4.8 *(previous page)*: The first factor primarily captures differences associated with tolerance to environment and nutrient stress as well as reproductive ability. Dots represent posterior means while bars cover the 95% highest posterior density (HPD) interval. Colors represent the posterior probability that the parameter is greater than 0.

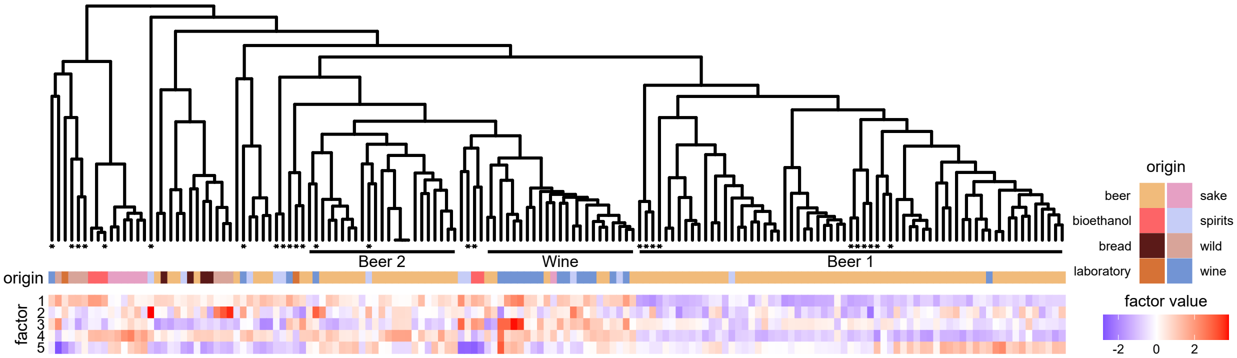


Figure 4.9: All five factors plotted on yeast phylogeny with strain origin. Stars at the tips indicate mosaic strains as identified by Gallone et al. (2016). The first factor separates the Beer 1 clade from the remaining strains.

CHAPTER 5

Phylogenetic structural equation modeling

5.1 Introduction

Phylogenetic comparative methods seek to untangle the complex relationships between phenotypes over a group of organisms' evolutionary history. Until recently, Bayesian comparative methods have been limited computationally in the number of unique traits or phenotypes they can simultaneously examine. Recent work by Tolkoff et al. (2018) and Hassler et al. (2022, Chapter 4) has expanded the realm of computationally feasible Bayesian analyses from those examining dozens of traits to those examining thousands. These approaches, however, rely on a phylogenetically-informed latent factor model that is agnostic to the structure of the data.

We generalize phylogenetic factor analysis to phylogenetic structural equation modeling that explicitly accounts for the underlying structure in the data. By structure, we mean that the overall data set can be naturally partitioned into two or more groups of traits that may be related to each other but are likely to have different sets of selective pressures acting on them. For example, in Section 5.4.3 we study the domestication of beer yeast with data from Gallone et al. (2016). We analyze a data set with 154 strains of yeast and 82 observations per strain. These 82 traits are divided as follows: 62 traits describing yeast growth rates under varying conditions, 16 traits related to the production of various aromatic compounds and four other traits primarily associated with reproductive ability. When Hassler et al. (2022, Section 4.6.2) analyze this data set using phylogenetic factor analysis, they include all 82 traits in a single latent factor model. However, it is possible, and indeed likely, that the selective forces driving growth under varying stress conditions are distinct from those driving the production of aromatic compounds. Mapping all 82 traits to the same low-dimensional latent factors ignores this.

As an alternative, we develop a more structured model that can map different sets of traits to an evolutionary process on a phylogenetic tree in different ways. Some groups of traits may be mapped to the tree using their own phylogenetic factor model, where the high-dimensional traits arise from the evolution of low-dimensional factors on tree. Other traits may be mapped directly to the tree without dimension reduction. Importantly, the model permits correlated between different groups of latent factors / traits. In the yeast domestication analysis, for example, we measure the evolutionary correlation between the traits related to reproductive ability and both sets of latent factors associated with growth under stress and production of aromatic compounds, respectively.

Phylogenetic structural equation modeling inherits many of the advantages and challenges of the less-structured models on which it is based. Namely, assuming the model extensions all meet certain criteria, one can compute likelihoods and perform inference in linear time with respect to both the number of taxa and traits. Conversely, the lack of identifiability that plagues latent factor models is compounded by this more-structured model. We develop new inference procedures to address these challenges, including a novel procedure for sampling from the space of unusually structured correlation matrices.

We demonstrate the utility of this new modeling framework in four real-world applications, including measuring the relationships between 1) HIV immune escape mutations and clinical outcomes, 2) floral phenotypes and pollinator species, 3) yeast stress tolerance and reproductive ability and 4) SARS-CoV-2 ACE2 binding affinity (related to infectiousness) and the virus' ability to evade human immune response. We describe the model and inference details in the sections below.

5.2 Phylogenetic structural equation model

Assume that we have observed a set of P traits across N taxa. We partition those traits into M non-overlapping subsets and form the $N \times P_m$ matrices $\mathbf{Y}_m$ for $m = 1, \ldots, M$. Let $\mathbf{y}_{mi}$ be the traits for taxon i in partition m. We assume that

$$\mathbf{y}_{mi} \,|\, \mathbf{f}_{mi}, \mathbf{L}_m, \boldsymbol{\Lambda}_m \overset{\text{ind}}{\sim} \mathcal{N}\big(\mathbf{L}_m^t \mathbf{f}_{mi}, \boldsymbol{\Lambda}_m^{-1}\big) \text{ for } i = 1, \ldots, N, m = 1, \ldots, M \tag{5.1}$$

where $\mathbf{f}_{mi}$ is a K_m-vector of latent factors, $\mathbf{L}_m$ is a $K_m \times P_m$ ($K_m \leq P_m$) loadings matrix and $\boldsymbol{\Lambda}_m$ is a $P_m \times P_m$ precision matrix.

Let $\mathbf{f}_i = \big(\mathbf{f}_{1i}^t, \ldots, \mathbf{f}_{Mi}^t\big)^t$ be the K-vector ($K = \sum_{m=1}^M K_m$) of latent factors for taxon i and all M trait subsets. We assume that these latent factors arise from a multivariate Brownian diffusion (MBD) process on a phylogenetic tree. The phylogeny $\mathcal{F}$ is a directed acyclic graph with N degree-one tip nodes $\nu_1, \ldots, \nu_N$, $N-2$ degree-three internal nodes $\nu_{N+1}, \ldots, \nu_{2N-2}$ and one degree-two root node ν_{2N-1}. With the exception of the root, there is an edge connecting each node ν_i to its parent $\nu_{\text{pa}(i)}$ with length t_i corresponding to the amount of evolutionary time separating the two nodes. The MBD process implies that for $i = 1, \ldots, 2N-2$ (i.e. all non-root nodes), the latent factors $\mathbf{f}_i$ associated with each node ν_i are drawn from a normal distribution $\mathbf{f}_i \sim \mathcal{N}\big(\mathbf{f}_{\text{pa}(i)}, t_i \boldsymbol{\Sigma}\big)$, where $\boldsymbol{\Sigma}$ is an estimable $K \times K$ between-factor covariance matrix. This results in the latent factors at the tips $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^t$ being matrix-normally (MN) distributed as follows:

$$\mathbf{F} \sim \text{MN}\left(\mathbf{1}_N \boldsymbol{\mu}^t, \boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}_N, \boldsymbol{\Sigma}\right) \tag{5.2}$$

where $\boldsymbol{\mu}$ is latent factor value at the root of the tree, $\mathbf{1}_N$ is a N-vector of ones, $\mathbf{J}_N$ is a $N \times N$ matrix of ones, κ_0 is the prior sample size of the root mean and $\boldsymbol{\Psi} = \{\Psi_{ij}\}$ is the between-taxon covariance matrix that is a deterministic function of the phylogeny. Specifically, each diagonal element Ψ_{ii} of $\boldsymbol{\Psi}$ is the sum of the branch lengths from the root node ν_{2N-1} of the phylogeny $\mathcal{F}$ to tip node ν_i, and each off-diagonal element Ψ_{ij} is the sum from the root to the most recent common ancestor of nodes ν_i and ν_j.

This model construction results in the data $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_M)$ being distributed as

$$\text{vec}(\mathbf{Y}) \sim \mathcal{N}\Big(\text{vec}\big(\mathbf{1}_N \boldsymbol{\mu}^t \mathbf{L}\big), \mathbf{L}^t \boldsymbol{\Sigma} \mathbf{L} \otimes \left(\boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}_N\right) + \boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_N\Big) \tag{5.3}$$

where vec(.) is the column-wise vectorization operator,

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{L}_M \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Lambda}_M \end{pmatrix}. \tag{5.4}$$

Qualitatively, this model is structured so that each partition of the data $\mathbf{Y}_m$ for $m = 1, \ldots, M$ has its own sub-model characterized by parameters $\mathbf{L}_m$ and $\boldsymbol{\Lambda}_m$ that map the data to the tips of the phylogenetic tree. Conditional on the factors $\mathbf{F}$ at the tips of the tree, the data associated with each sub-model are independent from each other. However, the latent factors $\mathbf{F}$ evolve along the phylogeny $\mathcal{F}$ where they may be correlated with each other via the evolutionary covariance matrix $\boldsymbol{\Sigma}$. As such, information is shared between these various sub-models through the evolutionary process on the tree. This model allows researchers to learn the low-dimensional structure of different groups of traits independently while also measuring the evolutionary correlation between groups that may or may not be mapped to the tree via dimension reduction.

5.2.1 Fast likelihood computation

Computing the likelihood naively from the likelihood in Equation 5.3 scales $\mathcal{O}(N^3P^3)$ and is computationally prohibitive for all but the smallest problems. However, straightforward application of the algorithms developed by Hassler et al. (2020, Section 3.2.1.2) and Hassler et al. (2022, Section 4.2.1.1) allow for likelihood calculation in order $\mathcal{O}(NPK^2 + NK^3)$. Specifically, these approaches allow fast likelihood calculation when one can compute some r_i, $\mathbf{m}_i$ and $\mathbf{P}_i$ for $i = 1, \ldots, N$ such that

$$p(\mathbf{y}_i \,|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}) = r_i \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \tag{5.5}$$

where

$$\log\hat{\theta}(\mathbf{z}; \boldsymbol{\mu}, \mathbf{P}) = \frac{1}{2}\log\hat{\det}(\mathbf{P}) - \frac{\text{rank}(\mathbf{P})}{2}\log 2\pi - \frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}\right)^t \mathbf{P}\left(\mathbf{z} - \boldsymbol{\mu}\right) \tag{5.6}$$

and $\hat{\det}(\mathbf{P})$ is the product of the non-zero singular values of $\mathbf{P}$. Given that $p(\mathbf{y}_{mi} \,|\, \mathbf{f}_{mi}, \mathbf{L}_m, \boldsymbol{\Lambda}_m) = r_{mi}\hat{\theta}(\mathbf{f}_{mi}; \mathbf{m}_{mi}, \mathbf{P}_{mi})$ for $m = 1, \ldots, M$ (Hassler et al., 2022, Section 4.8.1) and that the tip observations $\mathbf{y}_{mi}$ are independent conditional on the latent factors $\mathbf{f}_i$, we have

$$\begin{aligned}
\log p(\mathbf{y}_i \,|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}) &= \sum_{m=1}^{M} \log p(\mathbf{y}_{mi} \,|\, \mathbf{f}_{mi}, \mathbf{L}_m, \boldsymbol{\Lambda}_m) \\
&= \sum_{m=1}^{M} \log r_{mi} + \log\hat{\theta}(\mathbf{f}_{mi}; \mathbf{m}_{mi}, \mathbf{P}_{mi}) \\
&= \log r_i + \log\hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i), \text{ where} \\
r_i &= \prod_{m=1}^{M} r_{mi}, \\
\mathbf{m}_i &= \left(\mathbf{m}_{1i}^t, \ldots, \mathbf{m}_{Mi}\right)^t \text{ and} \\
\mathbf{P}_i &= \begin{pmatrix} \mathbf{P}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{2i} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{Mi} \end{pmatrix}.
\end{aligned} \tag{5.7}$$

As such, the computational burden of computing these partials in this more structured model is simply the sum of the computational costs of each of the sub-models. Once these partials have been calculated, we can complete the likelihood calculation in $\mathcal{O}(NK^3)$ using the Hassler et al. (2020, Section 3.2.1.2) algorithm.

5.2.2 Common model extensions

While there are numerous possible model extensions that place constraints on $\mathbf{L}_m$ and $\mathbf{\Lambda}_m$, we focus on three that have been previously explored. We summarize these various extensions in Table 5.1.

Sub-model	Description	Dimensions	$\mathbf{L}_m$	$\mathbf{\Lambda}_m$
un-extended	trait values are directly observed at the tips of the tree	$K_m = P_m$	$\mathbf{L}_m = \mathbf{I}_{P_m}$	$\mathbf{\Lambda}_m = \infty \mathbf{I}_{P_m}$
residual variance	trait values are observed with some uncertainty	$K_m = P_m$	$\mathbf{L}_m = \mathbf{I}_{P_m}$	positive-definite
latent factor	high-dimensional trait values are generated via some low-dimensional process	$K_m < P_m$	unrestricted*	diagonal

Table 5.1: Three possible Gaussian model extensions for the phylogenetic structural equation model. *Restrictions may be necessary in practice for identifiability during inference.

5.3 Inference

We approach inference from a Bayesian perspective as it allows for both 1) simpler estimation of uncertainty in model parameters and 2) simultaneous inference of the phylogenetic tree to account for uncertainty in the evolutionary history. From this perspective, we seek to sample from the posterior distribution

$$p(\mathbf{\Sigma}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F} \,|\, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \mathbf{\Sigma}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}) p(\mathbf{S} \,|\, \mathcal{F}) p(\mathbf{\Sigma}) p(\mathbf{L}) p(\mathbf{\Lambda}) p(\mathcal{F}), \tag{5.8}$$

where $\mathbf{S}$ is genetic sequence data (e.g. DNA or RNA). See Lemey et al. (2009) for a thorough discussion of estimating the tree structure $\mathcal{F}$.

Relevant model parameters that must be inferred are the loadings $\mathbf{L}_m$ associated with any latent factor model, residual precision $\mathbf{\Lambda}_m$ associated with any residual variance or latent factor model, and joint evolutionary covariance matrix $\mathbf{\Sigma}$. To exploit existing approaches for sampling from correlation matrices, we decompose $\mathbf{\Sigma}$ into a diagonal scale component $\mathbf{D} = \text{diag}[\mathbf{d}]$ and correlation component $\mathbf{C}$ such that $\mathbf{\Sigma} = \mathbf{DCD}$. We partition $\mathbf{\Sigma}$ by sub-model such that

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \cdots & \mathbf{\Sigma}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{1M}^t & \cdots & \mathbf{\Sigma}_{MM} \end{pmatrix} = \begin{pmatrix} \mathbf{D}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{D}_M \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{1M}^t & \cdots & \mathbf{C}_{MM} \end{pmatrix} \begin{pmatrix} \mathbf{D}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{D}_M \end{pmatrix} \tag{5.9}$$

We consider two different kinds of sub-models. We define low-rank sub-models as those where $K_m < P_m$, and full-rank sub-models as those where $K_m = P_m$. For any low-rank sub-model m, we impose the identifiability constraint $\mathbf{\Sigma}_{mm} = \mathbf{D}_m = \mathbf{C}_{mm} = \mathbf{I}_{K_m}$. Note that we do not place any restrictions on off-diagonal blocks $\mathbf{C}_{mn}$ for $m \neq n$ besides the global constraint that $\mathbf{C}$ be positive-definite. This means that latent factors may not be correlated with other factors within the same sub-model, but may be correlated with factors from other sub-models.

We explore the procedures for sampling from the posterior distributions of these model parameters in the sections below.

5.3.1 Inferring the residual precisions $\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_M$

For un-extended sub-models n, the residual precision $\mathbf{\Lambda}_n$ is fixed and need not be inferred. For residual variance or latent factor sub-models m, we utilize the conditional independence of each partition to sample independently from $\mathbf{\Lambda}_m \,|\, \mathbf{Y}_m, \mathbf{F}_m, \mathbf{L}_m$. Of course, $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_M)$ is unobserved, so we first draw from $\mathbf{F} \,|\, \mathbf{Y}, \mathbf{\Sigma}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}$ via a full conditional Gibbs sample using methods discussed in Hassler et al. (2022, Section 4.8.2.1).

For sub-models where $\mathbf{\Lambda}_m$ is diagonal, we place conjugate gamma priors on each diagonal element of λ_j and sample them according to the procedure in Hassler et al. (2022, Section 4.3.4). For sub-models where $\mathbf{\Lambda}_m$ is not necessarily diagonal, we place a Wishart prior on $\mathbf{\Lambda}_m$ and sample it according to the procedure in Hassler et al. (2020, Section 3.3.1).

5.3.2 Inferring the loadings $\mathbf{L}_1, \ldots, \mathbf{L}_M$

For latent factor sub-models, we must infer the loadings matrices $\mathbf{L}_m$. As with the error precisions $\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_M$, the sampling methods of Hassler et al. (2022, Section 4.3.1) apply in this context with only trivial modification. Preliminary experiments with inference under this more structured model, however, suggested that the methods developed by Hassler et al. (2022) to induce posterior identifiability and prevent label switching may be inadequate in this new context. As such, we explore alternative procedures for inducing posterior identifiability below.

5.3.2.1 Procrustes post-processing

While all the procedures for sampling from the loadings discussed in Tolkoff et al. (2018) and Hassler et al. (2022, Section 4.3) apply in this partitioned model, we are still faced with the same identifiability challenges that plague latent factor models generally. While much work has been done to address these challenges (see Holbrook et al., 2016; Jauch et al., 2021; Papastamoulis and Ntzoufras, 2022), Hassler et al. (2022, Section 4.3.1.3) attempt to address this issue by post-processing the posterior samples so that the loadings matrix has orthogonal rows with decreasing norms. This post-processing induces identifiability up to sign changes and row-wise permutations of the loadings. In some cases, however, the lack of identifiability with respect to row-wise permutations of the loadings results in label switching in the posterior.

To combat the label switching problem, Hassler et al. (2022, Section 4.3.2.1) also implement a geodesic Hamiltonian Monte Carlo (HMC) sampler modeled on work by Holbrook et al. (2016) that can sample directly from the space or orthogonal matrices. The geodesic HMC sampler, however, is relatively slow compared to alternatives that sample from unrestricted space, particularly for big-P data sets. It also sometimes requires artificial identifiability constraints that may potentially bias the relative magnitudes of each row of the loadings matrix.

We borrow from the mixture model and multidimensional scaling literature to find an alternative solution to the label switching problem. Specifically, we follow Okada and Mayekawa (2018) and use Procrustes analysis to find the rotation for each sample from the posterior that minimizes the distance to some iteratively updated value. We adapt this post-processing procedure to the case of factor analysis in Algorithm 5.1. This approach is substantially similar to the procedure independently developed in the economics literature by Aßmann et al. (2016) to induce identifiability in latent factor models.

Algorithm 5.1 Generalized Procrustes rotations to induce posterior identifiability in Bayesian latent factor analysis

```
$\mathbf{L}_m^{\text{ref}} \leftarrow \frac{1}{D}\sum_{d=1}^{D} \mathbf{L}_m^{(d)}$
$\epsilon \leftarrow \infty$
while $\epsilon > \text{tol}$ do
    $\bar{\mathbf{L}}_m \leftarrow \mathbf{0}$
    for $d \leftarrow 1$ to $D$ do
        $\mathbf{W} \leftarrow \mathbf{L}_m^{\text{ref}} \mathbf{L}_m^{(d)^t}$
        Compute orthonormal $\mathbf{U}, \mathbf{V}$ and diagonal $\mathbf{\Sigma}$ s.t. $\mathbf{U\Sigma V}^t = \mathbf{W}$
                via singular value decomposition
        $\mathbf{Q} \leftarrow \mathbf{UV}^t$
        $\mathbf{L}_m^{(d)} \leftarrow \mathbf{QL}_m^{(d)}$        ▷ $\mathbf{F}$ and $\mathbf{C}$, if sampled, must also be rotated by $\mathbf{Q}$ or $\mathbf{Q}^t$
        $\bar{\mathbf{L}}_m \leftarrow \bar{\mathbf{L}}_m + \mathbf{L}_m^{(d)}/D$
    end for
    $\epsilon \leftarrow ||\mathbf{L}_m^{\text{ref}} - \bar{\mathbf{L}}_m||$
    $\mathbf{L}_m^{\text{ref}} \leftarrow \bar{\mathbf{L}}_m$
end while
```

This generalized Procrustes procedure differs from alternative post-processing techniques in that, rather than restricting the loadings to some sub-space (e.g. the space of orthogonal matrices), it seeks to minimize posterior the variance of the loadings. This means that even if the initial reference matrix $\mathbf{L}_m^{\mathrm{ref}}$ is restricted to some space, there is no guarantee that any other samples $\mathbf{L}_m^{(d)}$ will be rotated to that space. However, we do not find this particularly restricting as the constraints on the loadings are typically not motivated by biological insights but rather the necessities of inducing identifiability.

In Figure 5.1, we demonstrate the ability of this approach to dealing with the label switching problem in the posterior of the loadings matrix from the New World monkey brain shape analysis of Hassler et al. (2022, Section 4.6.4). Label switching was particularly present in the Hassler et al. (2022) analysis, and they rely on Geodesic HMC to address this. Rather than relying on geodesic HMC, we use the faster unrestricted Gibbs sampler of Hassler et al. (2022, Section 4.3.1.1) coupled with Procrustes post-processing. Notably, Procrustes post-processing results in substantially similar results as the geodesic HMC sampler. However, the geodesic HMC approach requires about 80 minutes to achieve a median effective sample size (ESS) of 200, whereas sampling from un-restricted space using the Gibbs sampler and Procrustes post-processing requires only 30 seconds to reach equivalent ESS. This is a speedup of roughly 160 times for this particular $N = 48$, $P = 1197$, $K = 3$ data set.

5.3.2.2 Optimal rotation of the posterior

One advantage of the Procrustes post-processing approach is that it does not require arbitrarily imposing an identifiability constraint on the loadings. As such, we are free to find a particular rotation of the loadings that is informed by biological data and the problem of interest. As the motivating goal of the phylogenetic structural equation model introduced in this paper is to measure associations between latent factors and some number of additional traits, we propose an (optional) additional step after post-processing that orients the loadings and factors in a way that is most biologically informative (i.e. that maximizes the posterior correlation between a factor and the trait(s) of interest). Assume trait j is the trait of interest. Let the K_m-vector $\mathbf{c}_{mj}$ be the correlations between the latent factors in

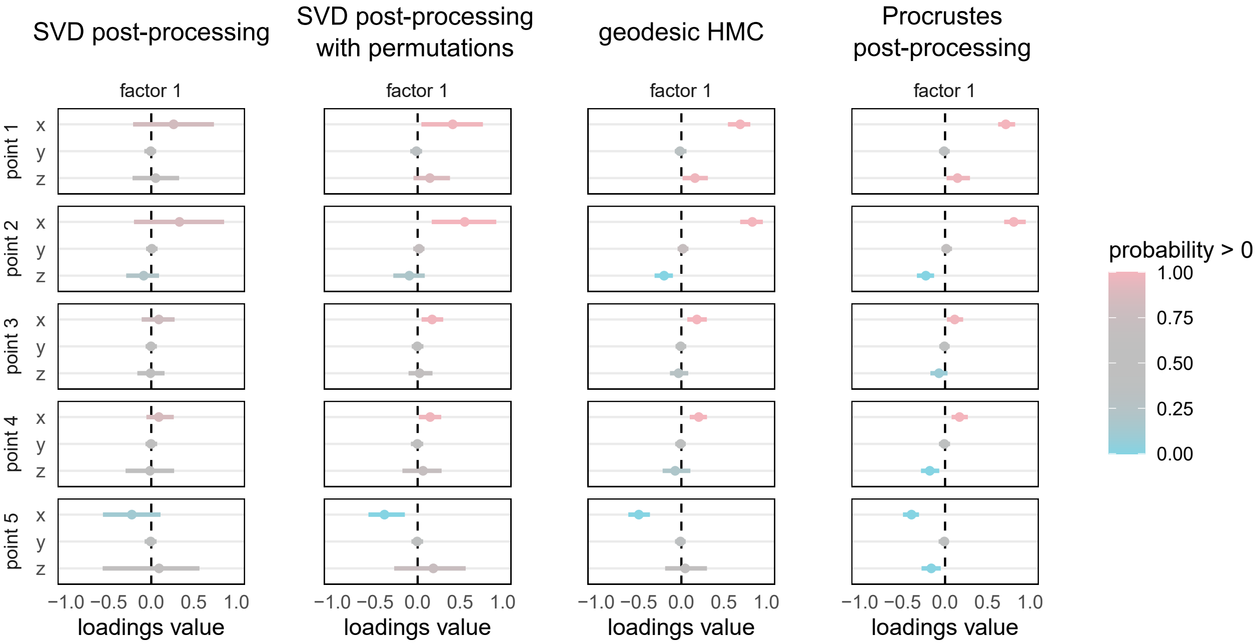


Figure 5.1: Comparison of sampling and post-processing methods for inducing identifiability in the loadings matrix. Data consists of 3-dimensional coordinates for 399 endocranial landmarks ($P = 1197$) in 48 species of New World monkeys (Aristide et al., 2016). We fit a 3-factor model using an unrestricted Gibbs sampler to sample from the full conditional posterior of the loadings (with the exception of the "geodesic HMC" run where we use geodesic HMC (Holbrook et al., 2016) to sample directly from the space of orthogonal matrices). For the unrestricted run, we post-process the output using 1) SVD post-processing as described in Hassler et al. (2022, Section 4.3.1.3), 2) SVD post-processing followed by iteratively permuting the rows of the loadings to minimize posterior variance and 3) Procrustes post-processing described in Algorithm 5.1. We show the marginal posterior distributions for elements $\ell_{1,1}, \ldots, \ell_{1,15}$ of the loadings matrix corresponding to coordinates for the first five endocranial landmarks (points). The SVD post-processing regime results in a largely non-identifiable posterior, characterized by high posterior variance on the elements of the loadings matrix. Procrustes post-processing and geodesic HMC avoid this problem. SVD post-processing with permutations does decrease posterior variance, but not to the extent of Procrustes post-processing or geodesic HMC.

sub-model m and trait j (not in sub-model m). Any rotation of the loadings $\mathbf{QL}_m$ must be accompanied by an equivalent rotation of the correlations $\mathbf{Qc}_{mj}$ so that post-processing leaves the posterior density invariant.

Let

$$R_m = \{r_{m1}, \ldots, r_{mK_m}\}, r_{mk} = k + \sum_{n=1}^{m-1} K_n \tag{5.10}$$

be the indices of the latent space associated with sub-model m. Without loss of generality, we seek to find the rotation matrix $\mathbf{Q}^*$ that maximizes the posterior mean of the element $c_{r_{m1}j}$ in the correlation matrix $\mathbf{C}$ (i.e. the first element of $\mathbf{c}_{mj}$). If we let $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{K_m})^t$ and $\mathbf{c}^*_{mj} = \mathbf{Q}\mathbf{c}_{mj}$, then $c^*_{r_{m1}j} = \mathbf{q}_1^t \mathbf{c}_{mj}$. As such, we seek to identify some $\mathbf{q}^*_1$ such that

$$\begin{aligned} \mathbf{q}^*_1 &= \underset{\mathbf{q}_1;\mathbf{q}_1^t\mathbf{q}_1=1}{\arg\max}\left(\frac{1}{D}\sum_{d=1}^{D} \mathbf{q}_1^t \mathbf{c}_{mj}^{(d)}\right) \\ &= \underset{\mathbf{q}_1;\mathbf{q}_1^t\mathbf{q}_1=1}{\arg\max}\big(\mathbf{q}_1^t \bar{\mathbf{c}}_{mj}\big), \text{ where} \\ \bar{\mathbf{c}}_{mj} &= \frac{1}{D}\sum_{d=1}^{D} \mathbf{c}_{mj}^{(d)}. \end{aligned} \tag{5.11}$$

Therefore, $\mathbf{q}^*_1 = \bar{\mathbf{c}}_{mj} / \left|\left|\bar{\mathbf{c}}_{mj}\right|\right|$ is just the normalized posterior mean of $\mathbf{c}_{mj}$. The remaining $\mathbf{q}^*_2, \ldots, \mathbf{q}^*_{K_m}$ can then be set to an arbitrary basis for the null space of $\mathbf{q}^*_1$. Alternatively, one can repeat this procedure for additional factors in sub-model m with additional traits k. In this case, you would perform the same calculations to calculate $\mathbf{q}^*_2, \ldots, \mathbf{q}^*_{K_m}$. At each step in the process however, one would project the result onto the null-space of the previously optimized rows. In this way, one can choose a rotation of the loadings that maximizes the association of the latent factors with specific traits of interest, rather than some arbitrary rotation.

We emphasize that the rotations to optimize the posterior correlation between a latent factor and a trait of interest are fundamentally different from the Procrustes rotations for inducing identifiability. The Procrustes post-processing procedure starts with a posterior distribution over unidentifiable or only weakly identifiable parameters and finds *many rotations* (one for each sample from the posterior) that minimize the distance between each sample and an iteratively updated reference matrix. The optimal rotation suggested in this section, however, starts with an identifiable posterior distribution and finds a *single rotation* that, when applied to all samples from the posterior together, maximizes the posterior mean of the correlation between one of the latent factors and a trait of scientific interest.

Finally, if one cannot easily choose a single trait with which to orient the posterior, then it is possible to optimize some linear combination of the correlations between a factor and several traits. For example, in Section 5.4.3, rather than rotating the posterior to maximize correlation between the first factor and either yeast spore viability or sporulation efficiency, we maximize the sum of the correlations between the first factor and spore viability and sporulation efficiency.

5.3.3 Inferring the evolutionary scales d

Recall that we decompose the evolutionary covariance matrix $\mathbf{\Sigma} = \mathbf{DCD}$ where $\mathbf{D} = \text{diag}[\mathbf{d}]$ and $\mathbf{C}$ is a correlation matrix with ones on the diagonal. As such, the vector $\mathbf{d}$ describes the rate at which individual traits evolve along the tree.

One can sample from the scales $\mathbf{d}$ with HMC using the method proposed by Bastide et al. (2021). HMC uses the gradient of the log-posterior

$$\nabla_{\mathbf{d}} \log p(\,\mathbf{d}, \mathbf{C}, \mathbf{L}, \mathbf{\Lambda} \,|\, \mathbf{Y}, \mathcal{F}\,) = \nabla_{\mathbf{d}} \log p(\,\mathbf{Y} \,|\, \mathbf{d}, \mathbf{C}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}\,) + \nabla_{\mathbf{d}} \log p(\mathbf{d}) \tag{5.12}$$

to more efficiently traverse high-dimensional parameter space than the Metropolis-Hastings algorithm. The Bastide et al. (2021) approach works for any model where we can calculate the partial mean $\mathbf{m}_i$ and precision $\mathbf{P}_i$ for $i = 1, \ldots, N$ as described in Section 5.2.1. We modify this approach slightly so that we only update the elements of $\mathbf{d}$ that are not fixed to 1 due to identifiability constraints. Otherwise, we use this approach with no additional modification.

5.3.4 Inferring the evolutionary correlation C

While inference procedures for other model parameters remain largely unchanged from previous work, the phylogenetic structural equation model requires new approaches to infer the evolutionary correlation matrix $\mathbf{C}$. Recall that to induce posterior identifiability, we assume $\mathbf{C}_{mm} = \mathbf{I}$ for any latent factor sub-model m. However, we do not place any constraints on off-diagonal blocks $\mathbf{C}_{mn}, m \neq n$. This unusual structure where $\mathbf{C}$ has structural zeros in some diagonal blocks but is not block-diagonal poses challenges to inference. While a block-diagonal structure would solve these problems, the primary motivation for this model is to explore the correlations between latent factors across sub-models, which would be zero by construction in a block-diagonal $\mathbf{C}$. We separate our inference strategies into two cases: a special case with exactly one low-rank sub-model and a general case with two or more low-rank sub-models. Note that, without loss of generality, we assume that there is at most one full-rank sub-model, as multiple full-rank sub-models can be grouped together to form one large, full-rank sub-model.

5.3.4.1 LKJ prior with structural zeros

Regardless of which scenario applies above, we place a modified Lewandowski-Kurowicka-Joe (LKJ, Lewandowski et al., 2009) prior on the correlation matrix $\mathbf{C}$:

$$p(\mathbf{C}) \propto \det(\mathbf{C})^{\eta-1} \prod_{(k,\ell)\in S} 1\{c_{k\ell} = 0\} \tag{5.13}$$

where $S = \{(k, \ell) : 1 \leq k < \ell \leq K, \{k, \ell\} \subset R_m \text{ for some } m = 1, \ldots, M\}$ is the set of index pairs such that $c_{k\ell}$ is identically 0 and R_m (defined in Equation 5.10) is the set of indices corresponding to sub-model m. We show below that this LKJ distribution modified to have structural zeros arises naturally from the original derivation of the LKJ distribution in Joe (2006). We also show that the structural zeros do not influence the marginal distributions of the non-zero elements of $\mathbf{C}$.

Joe (2006) and Lewandowski et al. (2009) derive the LKJ distribution over the space of correlation matrices (i.e. positive-definite matrices with ones along the diagonal). The LKJ distribution has a appealingly simple density $p(\mathbf{C}) \propto \det(\mathbf{C})^{\eta-1}$, where high values ($> 1$) of the parameter η place higher density on smaller correlations and low values (< 1) do the reverse. Notably, $\eta = 1$ results in a uniform distribution across all correlation matrices.

While the LKJ distribution was developed for dense correlation matrices, it readily extends to our case with structural zeros. Joe (2006) derives this density by placing independent Beta distributions (shifted to have support (-1, 1)) on the partial correlations arising from a D-vine on the correlation matrix $\mathbf{C}$. Lewandowski et al. (2009) extend this derivation to any set of partial correlations that arise from any regular vine. While the positive-definite constraint on $\mathbf{C}$ does not permit one to sample $\mathbf{C}$ by independently sampling each of its off-diagonal elements from (-1, 1), the partial correlations, by construction, have no such constraint.

A vine (Bedford and Cooke, 2002) on K variables is a series of nested trees $\{T_1, \ldots, T_{K-1}\}$ where the edges tree T_i become the nodes of tree T_{i+1} (see Joe and Kurowicka (2011) for a detailed discussion of vines). The nodes in the base tree T_1 represent the indices of the correlation matrix $\{1, \ldots, K\}$ and the $K-1$ edges represent a subset of the pairwise correlations. Edges of the trees $T_2, \ldots, T_{K-1}$ capture the remaining $\binom{K}{2} - (K-1)$ pair-wise partial correlations. The D-vine is a vine where all nodes are at most degree-2. In this case, the base tree T_1 can be represented as a line of nodes connected by edges and can be easily represented by $\mathbf{v} = (v_1, \ldots, v_K) \in \mathcal{P}_K$ where $\mathcal{P}_K$ is the set of all permutations of $(1, \ldots, K)$. Joe (2006) assumes $\mathbf{v} = (1, \ldots, K)$, but we relax this assumption for the sake of demonstrating that structural zeros do not influence the marginal distributions of the free parameters.

We introduce the following notation before demonstrating invariance of the LKJ distribution with structural zeros. Recall that we partition the correlation matrix as follows:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{1M}^t & \cdot & \mathbf{C}_{MM} \end{pmatrix}, \tag{5.14}$$

where each diagonal block is a square matrix. Assume some $U \subset \{1, \ldots, M\}$ such that $m \in U \Rightarrow \mathbf{C}_{mm} = \mathbf{I}$. Let $R_m = \left\{1 + \sum_{n=1}^{m-1} K_n, \ldots, \sum_{n=1}^{m} K_n\right\}$ be the indices associated with partition m, and let $\mathbf{v}_m = \left(v_1^m, \ldots, v_{K_m}^m\right)$ be a vector of length K_m containing all the elements of R_m in no particular order. We define the concatenation operator $(x_1, \ldots, x_n) \frown (y_1, \ldots, y_m) = (x_1, \ldots, x_n, y_1, \ldots, y_m)$. Let $\mathbf{P} = \{p_{k\ell}\}$ be the matrix of partial correlations defined by the D-vine on $\mathbf{C}$ with upper-triangular elements:

$$p_{v_k v_\ell} = \begin{cases} c_{v_k v_\ell} \text{ if } \ell - k \leq 1 \\ c_{v_k v_\ell : v_{k+1}, \ldots, v_{\ell-1}} \text{ if } \ell - k \geq 2 \end{cases} \tag{5.15}$$

where $c_{v_k v_\ell : v_{k+1}, \ldots, v_{\ell-1}}$ is the partial correlation between traits v_k and v_ℓ conditional on traits $v_{k+1}, \ldots, v_{\ell-1}$. Finally, let $f(.)$ be an invertible function such that $f(\mathbf{C}) = \mathbf{P}$.

Claim: Assume a D-vine with ordering $\mathbf{v} = (\mathbf{v}_{u_1} \frown \ldots \frown \mathbf{v}_{u_M})$, where $\mathbf{u} = (u_1, \ldots, u_M) \in \mathcal{P}_M$. The following then hold:

1. $\mathbf{P}_{mm} = \mathbf{I} \Rightarrow \mathbf{C}_{mm} = \mathbf{I}$

2. Sampling

$$\begin{aligned} &p_{v_k v_\ell} \sim \text{Beta}_{(\text{-}1,1)}(\alpha_{\ell-k}, \alpha_{\ell-k}) && \text{if } \{v_k, v_\ell\} \not\subset R_m \text{ for all } m = 1, \ldots, M \\ &p_{v_k v_\ell} = 0 && \text{if } \{v_k, v_\ell\} \subset R_m \text{ for any } m = 1, \ldots, M, \end{aligned} \tag{5.16}$$

where $\alpha_i = \eta + [K - 1 - i]/2$, imposes a the joint distribution on $\mathbf{C} = f^{-1}(\mathbf{P})$

$$p(\mathbf{C}) \propto \det(\mathbf{C})^{\eta-1} \prod_{(k,\ell) \in S} 1\{c_{k\ell} = 0\}\,. \tag{5.17}$$

3. A random correlation matrix distributed with kernel defined by Equation 5.17 will have identical (but not independent) marginal correlations

$$c_{k\ell} \sim \text{Beta}_{(\text{-}1,1)}(\eta + K/2 - 1, \eta + K/2 - 1). \tag{5.18}$$

Proof (part 1): Within a partition, we must show that $\mathbf{P}_{mm} = \mathbf{I} \Rightarrow \mathbf{C}_{mm} = \mathbf{I}$. Assume we have a partition defined by indices $\mathbf{v}_m = \left(v_1^m, \ldots, v_{K_m}^m\right)$. We can divide this partition into a series of overlapping sub-partitions $\mathbf{v}_m^{1|j}, \ldots, \mathbf{v}_m^{K_m - j|j}$ for $j = 1, \ldots, K_m - 1$ where $\mathbf{v}_m^{i|j} = \left(v_i^m, \ldots, v_{i+j}^m\right)$. Finally, let $\mathbf{C}\{(x_1, , \ldots, , x_n)\}$ be the diagonal block of $\mathbf{C}$ associated with indices $(x_1, , \ldots, , x_n)$.

We demonstrate $\mathbf{P}_{mm} = \mathbf{I} \Rightarrow \mathbf{C}_{mm} = \mathbf{I}$ by induction by noting that:

$$\begin{array}{l} \mathbf{C}\left\{\mathbf{v}_m^{i|j}\right\} = \mathbf{I} \text{ for } i = 1, \ldots, K_m - j \text{ and} \\ p_{v_i^m v_{i+j+1}^m} = 0 \text{ for } i = 1, \ldots, K_m - j - 1 \end{array} \Rightarrow \mathbf{C}\left\{\mathbf{v}_m^{i|j+1}\right\} = \mathbf{I} \text{ for } i = 1, \ldots, K_m - (j+1) \tag{5.19}$$

By the construction of the D-vine, $p_{v_i^m v_{i+1}^m} = c_{v_i^m v_{i+1}^m}$ for $i = 1, \ldots, K_m - 1$. Therefore, $p_{v_i^m v_{i+1}^m} = 0$ for $i = 1, \ldots, K_m - 1$ implies $\mathbf{C}\left\{\mathbf{v}_m^{i|1}\right\} = \mathbf{I}$ for $i = 1, \ldots, K_m - 1$.

To finish the proof we follow Lewandowski et al. (2009) and use the following formula for calculating partial correlations. Given random variables $X_1, \ldots, X_n$ with correlation matrix $\mathbf{R}$, the partial correlation between X_1 and X_n with all other variables held constant is

$$c_{1n:2,\ldots,n-1} = \frac{\text{cof}(\mathbf{R}; 1, n)}{\sqrt{\text{cof}(\mathbf{R}; 1, 1)\,\text{cof}(\mathbf{R}; n, n)}} \tag{5.20}$$

where $\text{cof}(\mathbf{R}; i, j)$ is the (i, j) cofactor of $\mathbf{R}$ (i.e. $(-1)^{i+j}$ times the determinant of the matrix formed by removing row i and column j of $\mathbf{R}$). In the context of the recursion, the becomes:

$$p_{v_i^m v_{i+j+1}^m} = 0 = \frac{\text{cof}\left(\mathbf{C}\left\{\mathbf{v}_m^{i|j+1}\right\}; 1, j+2\right)}{\sqrt{\text{cof}\left(\mathbf{C}\left\{\mathbf{v}_m^{i|j+1}\right\}; 1, 1\right)\,\text{cof}\left(\mathbf{C}\left\{\mathbf{v}_m^{i|j+1}\right\}; j+2, j+2\right)}} \tag{5.21}$$

Note that $\operatorname{cof}\Big(\mathbf{C}\Big\{\mathbf{v}_m^{i|j+1}\Big\};1,1\Big) = \det\Big(\mathbf{C}\Big\{\mathbf{v}_m^{i+1|j}\Big\}\Big)$ and $\operatorname{cof}\Big(\mathbf{C}\Big\{\mathbf{v}_m^{i|j+1}\Big\};j+2,j+2\Big) = \det\Big(\mathbf{C}\Big\{\mathbf{v}_m^{i|j}\Big\}\Big)$. By the inductive assumption $\mathbf{C}\Big\{\mathbf{v}_m^{i|j}\Big\} = \mathbf{I}$ for $i = 1,\ldots,K_m - j$, we have

$$\operatorname{cof}\big(\mathbf{C}\{\mathbf{v}_m^{i|j+1}\};1,j+2\big) = 0. \tag{5.22}$$

To express this cofactor in terms of the correlation $c_{v_i^m v_{i+j+1}^m}$, note that the inductive assumption results in

$$\mathbf{C}\{\mathbf{v}_m^{i|j+1}\} = \begin{pmatrix} 1 & \mathbf{0} & c_{v_i^m v_{i+j+1}^m} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ c_{v_i^m v_{i+j+1}^m} & \mathbf{0} & 1 \end{pmatrix} \tag{5.23}$$

Therefore:

$$\begin{aligned} \operatorname{cof}\big(\mathbf{C}\{\mathbf{v}_m^{i|j+1}\};1,j+2\big) &= (-1)^{j+3}\det\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ c_{v_i^m v_{i+j+1}^m} & \mathbf{0} \end{pmatrix} \\ &= -c_{v_i^m v_{i+j+1}^m} \end{aligned} \tag{5.24}$$

Equations 5.22 and 5.24 imply that $c_{v_i^m v_{i+j+1}^m} = 0$ and $\mathbf{C}\Big\{\mathbf{v}_m^{i|j+1}\Big\} = \mathbf{I}$. We conclude the proof by noting that $\mathbf{C}\Big\{\mathbf{v}_m^{1|K_m-1}\Big\} = \mathbf{C}_{mm}$. ■

Proof (part 2): We use the standard change of variable formula:

$$p_{\mathbf{C}}(\mathbf{C}) = p_{\mathbf{P}}(\mathbf{P})\,|\det(\mathbf{J})|\,, \tag{5.25}$$

where $f(\mathbf{C}) = \mathbf{P}$ and $\mathbf{J} = \frac{\partial f(\mathbf{C})}{\partial \mathbf{C}}$. The following proof follows directly from Joe (2006). While Joe (2006) relies on a specific D-vine with ordering $\mathbf{v} = (1,\ldots,K)$, symmetry implies that all results hold for any D-vine with arbitrary ordering. As such, we replace all indices k,ℓ in Joe (2006) results with v_k, v_ℓ.

The only challenge that prohibits immediate application of the Joe (2006) result to the situation with structural zeros is the fact that the number of free parameters in $\mathbf{P}$ from $\binom{K}{2}$ to $\binom{K}{2} - \sum_{m\in U}\binom{K_m}{2}$. Conveniently, the proof above demonstrates that $\mathbf{P}$ and $\mathbf{C} = f^{-1}(\mathbf{P})$ have the same structure. Let $R = \{(k, \ell) : 1 \leq k < \ell \leq K, \{k, \ell\} \not\subset R_m$ for all $m = 1, \ldots, M\}$ be the set of index pairs such that $p_{k\ell}$ is not identically 0. We must compute the Jacobian $\mathbf{J}$ that maps from $\{p_{v_k v_\ell} : (v_k, v_\ell) \in R\} \rightarrow \{c_{v_k v_\ell} : (v_k, v_\ell) \in R\}$.

Joe (2006) demonstrate that the fully-parameterized Jacobian is lower-triangular. For the case without structural zeros, Lemma 3 and Theorem 4 in Joe (2006) rely on this lower-triangular structure of the Jacobian matrix $\mathbf{J}$ to calculate.

$$\begin{aligned} |\det(\mathbf{J})| &= \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \frac{\partial p_{v_k v_\ell}}{\partial c_{v_k v_\ell}} \\ &= \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \prod_{m=1}^{\ell-k-1} \left[(1 - p^2_{v_k v_{k+m}})(1 - p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} \\ &= \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \left(1 - p^2_{v_k v_\ell}\right)^{[1+(\ell-k)-K]/2}. \end{aligned} \tag{5.26}$$

As the columns corresponding to the zero values of $\mathbf{C}$ are the same as the rows corresponding to the zero values of $\mathbf{P}$, removing the necessary rows and columns retains the same lower-triangular structure. In the case with structural zeros, Equation 5.26 becomes

$$\begin{aligned}
|\det(\mathbf{J})| &= \prod_{(v_k,v_\ell)\in R} \frac{\partial p_{v_k v_\ell}}{\partial c_{v_k v_\ell}} \\
&= \prod_{(v_k,v_\ell)\in R} \prod_{m=1}^{\ell-k-1} \left[(1-p^2_{v_k v_{k+m}})(1-p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} \\
&= \prod_{(v_k,v_\ell)\in R} \prod_{m=1}^{\ell-k-1} \left[(1-p^2_{v_k v_{k+m}})(1-p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} \\
&\qquad \times \prod_{(v_k,v_\ell)\notin R} \prod_{m=1}^{\ell-k-1} \left[(1-p^2_{v_k v_{k+m}})(1-p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} \qquad (5.27)\\
&= \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \prod_{m=1}^{\ell-k-1} \left[(1-p^2_{v_k v_{k+m}})(1-p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} \\
&= \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \left(1-p^2_{v_k v_\ell}\right)^{[1+(\ell-k)-K]/2} \\
&= \prod_{(v_k,v_\ell)\in R} \left(1-p^2_{v_k v_\ell}\right)^{[1+(\ell-k)-K]/2}
\end{aligned}$$

The third equal sign is the result of the fact that, in the case with structural zeros

$$\prod_{(v_k,v_\ell)\notin R} \prod_{m=1}^{\ell-k-1} \left[(1-p^2_{v_k v_{k+m}})(1-p^2_{v_{\ell-m} v_\ell})\right]^{-1/2} = 1 \qquad (5.28)$$

Theorem 1 in Joe (2006) states

$$\det(\mathbf{C}) = \prod_{k=1}^{K-1} \prod_{\ell=k+1}^{K} \left(1-p^2_{v_k v_\ell}\right). \qquad (5.29)$$

We can rewrite Equation 5.29 as

$$\begin{aligned}\det(\mathbf{C}) &= \left[\prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)\right]\left[\prod_{(v_k,v_\ell)\notin R}\left(1-p^2_{v_kv_\ell}\right)\right]\\ &= \left[\prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)\right]\left[\prod_{(v_k,v_\ell)\notin R}(1-0)^2\right]\\ &= \prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)\end{aligned} \tag{5.30}$$

Finally, the shifted Beta distribution $\text{Beta}_{(\text{-}1,1)}(\alpha,\beta)$ has density:

$$p(x) \propto (1+x)^{\alpha-1}(1-x)^{\beta-1} \tag{5.31}$$

If $\alpha=\beta$, this becomes:

$$p(x) \propto \left(1-x^2\right)^{\alpha-1} \tag{5.32}$$

Therefore, drawing all partial correlations as in Equation 5.16 results in a density (over the partial correlations) of:

$$p(\mathbf{P}) = \prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)^{\eta+[K-1-(\ell-k)]/2-1} \tag{5.33}$$

Applying Equations 5.27 and 5.33 in the change-of-variable formula, we get

$$\begin{aligned}p_{\mathbf{C}}(\mathbf{C}) &= p_{\mathbf{P}}(\mathbf{P})\,|\det(\mathbf{J})|\\ &= \left[\prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)^{\eta+[K-1-(\ell-k)]/2-1}\right]\left[\prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)^{[1+(\ell-k)-K]/2}\right]\\ &= \prod_{(v_k,v_\ell)\in R}\left(1-p^2_{v_kv_\ell}\right)^{\eta-1}\\ &= \det(\mathbf{C})^{\eta-1}.\end{aligned} \tag{5.34}$$

The last equality follows from Equation 5.30. ■

Proof (part 3): Finally, we demonstrate that sampling the partial correlations as in 5.16 induces identical marginal distributions on the correlations

$$c_{k\ell} \sim \text{Beta}_{(\text{-}1,1)}(\eta + K/2 - 1, \eta + K/2 - 1). \tag{5.35}$$

Given that the marginal density can be derived from the joint density alone, the marginal densities resulting from any particular D-vine with ordering $\mathbf{v}$ must hold regardless of choice of $\mathbf{v}$ as long as the locations of the structural zeros remains the same. By construction of the D-vine and assumptions in Equation 5.16,

$$c_{v_i v_{i+1}} = p_{v_i v_{i+1}} \sim \text{Beta}_{(\text{-}1,1)}(\eta + K/2 - 1, \eta + K/2 - 1) \quad \forall \quad i \text{ such that } (v_i, v_{i+1}) \in R \tag{5.36}$$

given any choice of $\mathbf{v}$ that respects the location of the structural zeros. For any pair $(k, \ell) \in R$ we can define some $\mathbf{v}$ such that $(k, \ell) = (v_i, v_{i+1})$ for some i. As such, for any arbitrary $c_{k\ell}$ such that $(k, \ell) \in R$, we can construct a $\mathbf{v}$ such that $c_{k\ell} \sim \text{Beta}_{(\text{-}1,1)}(\eta + K/2 - 1, \eta + K/2 - 1)$. ■

5.3.4.2 Special case: exactly one low-rank sub-model

We develop two different strategies for inferring the evolutionary correlation matrix $\mathbf{C}$ with structural zeros. The first, presented here, applies in the special case where there is exactly one low-rank sub-model. In this special case we modify the inference procedure for sampling from unconstrained $\mathbf{C}$ developed by Bastide et al. (2021). Bastide et al. (2021) decompose $\mathbf{C}$ into its Cholesky factors $\mathbf{C} = \mathbf{\Gamma}\mathbf{\Gamma}^t$ (i.e. $\mathbf{\Gamma}$ is lower-triangular with positive diagonals). If we arrange the sub-models such that the $m = 1$ corresponds to the single low-rank sub-model and $m = 2$ corresponds to the full-rank sub-model, $\mathbf{C}$ has the following structure:

$$\mathbf{C} = \begin{pmatrix} \mathbf{I} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^t & \mathbf{C}_{22} \end{pmatrix}. \tag{5.37}$$

Claim: A correlation matrix structured as in Equation 5.37 will have a Cholesky decomposition structured as

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{\Gamma}_{12}^t & \mathbf{\Gamma}_{22} \end{pmatrix}. \tag{5.38}$$

Proof: Let

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{11} & \mathbf{0} \\ \mathbf{\Gamma}_{12}^t & \mathbf{\Gamma}_{22} \end{pmatrix} \tag{5.39}$$

be the Cholesky decomposition of a matrix structured as in Equation 5.37. If $\mathbf{\Gamma}$ is lower-triangular with positive diagonals, then $\mathbf{\Gamma}_{11}$ must also be lower-triangular with positive diagonals. By definition

$$\mathbf{\Gamma}\mathbf{\Gamma}^t = \begin{pmatrix} \mathbf{\Gamma}_{11}\mathbf{\Gamma}_{11}^t & \mathbf{\Gamma}_{11}\mathbf{\Gamma}_{12}^t \\ \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{11}^t & \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{12}^t + \mathbf{\Gamma}_{22}\mathbf{\Gamma}_{22}^t \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^t & \mathbf{C}_{22} \end{pmatrix}. \tag{5.40}$$

This implies that $\mathbf{\Gamma}_{11}\mathbf{\Gamma}_{11}^t = \mathbf{I}$. As $\mathbf{\Gamma}_{11}$ is, by construction, a lower-triangular matrix with positive diagonals, it is itself a Cholesky decomposition. By the uniqueness of the Cholesky decomposition, $\mathbf{\Gamma}_{11}\mathbf{\Gamma}_{11}^t = \mathbf{I}$ implies that $\mathbf{\Gamma}_{11}$ is the Cholesky decomposition of the identity. As such $\mathbf{\Gamma}_{11} = \mathbf{I}$. ■

The transference of structural zeros in $\mathbf{C}$ to the same structural zeros in $\mathbf{\Gamma}$ permits a computationally convenient procedure for sampling from the posterior of $\mathbf{\Gamma}$ via MCMC. While the constraints on the Cholesky decomposition of a correlation matrix (i.e. lower-triangular with positive-diagonals and unit-norm rows) are substantially simpler to enforce than that of the raw correlation matrix, Bastide et al. (2021) develop a bijective map from the space of $K \times K$ Cholesky decompositions to $\mathbb{R}^{\binom{K}{2}}$ so that one can sample from unrestricted space. Let $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,k-1})$ be the first $k-1$ elements of the k^{th} row of $\boldsymbol{\Gamma}$. The vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ are sufficient to define $\boldsymbol{\Gamma}$ as $\gamma_{kk} = \sqrt{1 - \boldsymbol{\gamma}_k^t \boldsymbol{\gamma}_k}$ and $\gamma_{k,k+1} = \cdots = \gamma_{k,K} = 0$. Each $\boldsymbol{\gamma}_k$ is constrained to the Euclidean ball $\mathcal{B}_{k-1}$. For each row, Bastide et al. (2021) sample from the posterior of the transformed random variable $\mathbf{x}_k = \mathbf{f}^{-1}(\mathbf{g}^{-1}(\boldsymbol{\gamma}_k)) \in \mathbb{R}^{k-1}$ via HMC and map from $\mathbb{R}^{k-1} \rightarrow \mathcal{B}_{k-1}$ in two steps by first using the inverse Fisher z-transformation $\mathbf{f}: \mathbb{R}^{k-1} \rightarrow \mathcal{B}_{k-1}^{\infty}; f_\ell(y_1, \ldots, y_{k-1}) = \exp(2y_\ell) - 1/(\exp(2y_\ell) + 1)$ and then the function $\mathbf{g}: \mathcal{B}_{k-1}^{\infty} \rightarrow \mathcal{B}_{k-1}; g_\ell(z_1, \ldots, z_{k-1}) = z_\ell \prod_{i=1}^{\ell-1} \sqrt{1 - z_i^2}$, where $\mathcal{B}_{k-1}^{\infty}$ is the infinity-norm ball. Conveniently, this mapping preserves zeros, so maintaining the zeros while sampling in transformed space also maintains the zeros in Cholesky (and therefore correlation) space. In other words $\gamma_{k\ell} = g_\ell(\mathbf{f}(\mathbf{x}_k)) = 0 \iff x_{k\ell} = 0$. As such, one can adapt the Bastide et al. (2021) procedure for sampling from the posterior of the Cholesky decomposition $\boldsymbol{\Gamma}$ of the correlation matrix $\mathbf{C}$ by simply fixing to zero each $x_{k\ell} \in \mathbb{R}$ corresponding to the zeros in $\boldsymbol{\gamma}_k$.

5.3.4.3 General case: two or more low-rank sub-models

For the general case with two or more low-rank sub-models, zeros in correlation space do not necessarily translate to zeros in Cholesky space. Let $\boldsymbol{\Gamma}_m = \begin{pmatrix} \boldsymbol{\Gamma}_{m1} & \cdots & \boldsymbol{\Gamma}_{mM} \end{pmatrix}$ be the rows of $\boldsymbol{\Gamma}$ corresponding to sub-model m. Without loss of generality, assume that sub-models $1, \ldots, M-1$ are low-rank and sub-model M is full-rank. As demonstrated above, $\mathbf{C}_{11} = \mathbf{I} \iff \boldsymbol{\Gamma}_1 = \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix}$. For $m = 2, \ldots, M-1$, $\mathbf{C}_{mm} = \mathbf{I} \iff \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^t = \mathbf{I}$ (i.e. $\boldsymbol{\Gamma}_m^t$ inhabits the Stiefel manifold $\mathcal{V}_{K_m}(\mathbb{R}^K)$). However, the fact that $\boldsymbol{\Gamma}$ is lower-triangular means that there are structural zeros in $\boldsymbol{\Gamma}_m$, which existing methods for sampling from the Stiefel manifold (e.g. Holbrook et al., 2016) cannot easily accommodate.

To avoid the challenging task of sampling from this unusual sub-set of the Stiefel manifold, we do not re-parameterize via the Cholesky decomposition and instead sample directly from the full conditional posterior of $\mathbf{C}$ via HMC. The gradient $\nabla_{\mathbf{C}} \log p(\mathbf{Y} \mid \mathbf{C}, \mathbf{d}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F})$ is supplied by Bastide et al. (2021). However, the challenge of implementing HMC in correlation-space is that correlation-space is bounded with multivariate boundary $\det(\mathbf{C}) = 0$. Any HMC algorithm for sampling directly from correlation space must respect this boundary.

We first explain the fundamentals of HMC before discussing the specifics of dealing with this boundary. For a more thorough introduction, see Neal (2010). Like the Metropolis-Hasting algorithm, HMC samples from a distribution of interest by repeatedly proposing new parameter values conditional on the current parameters and subsequently accepting or rejecting those proposals. HMC, however, leverages the geometry of the posterior distribution to make proposals that are both farther away from the current position and have a high acceptance probability, allowing HMC to more efficiently traverse parameter space. Specifically, HMC posits a particle with position variable $\mathbf{q}$ corresponding to some set of model parameters and an auxiliary momentum variable $\mathbf{p}$. This particle traverses a potential energy landscape defined by $U(\mathbf{q})$ equal to the negative full conditional log-posterior of the parameter corresponding to $\mathbf{q}$. HMC generates parameter proposals $\mathbf{q}(t')$ for some t' where $\mathbf{q}(t)$ is the solution to the ordinary differential equation system

$$\begin{aligned}\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} &= \mathbf{M}^{-1}\mathbf{p}(t)\\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\nabla_{\mathbf{q}} U(\mathbf{q}(t))\end{aligned} \tag{5.41}$$

with initial conditions $\mathbf{q}(0)$ equal to the current parameter value and $\mathbf{p}(0)$ randomly drawn from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{M})$. Typically, the Hamiltonian equations do not have an analytical solution, and the trajectory of the parameter is approximated via a series of linear steps. We outline a common HMC algorithm in Algorithm 5.2.

Algorithm 5.2 assumes that the parameters reside in unrestricted Euclidean space. This algorithm, however, can easily be adapted to a bounded parameter space by allowing elastic collisions off boundaries. Without loss of generality, let the boundaries of the parameter space be defined by the curve $b(\mathbf{q}) = 0$. Following from Afshar and Domke (2015), we modify Algorithm 5.2 to handle reflections at the boundary (see Algorithm 5.3). In addition to

Algorithm 5.2 HMC using leapfrog approximation

```
$t \leftarrow 0$
Draw $\mathbf{p}(0) \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$
$\mathbf{q}(0) \leftarrow$ the current parameter value
for $h = 1, \ldots, H$ do
    $\mathbf{p}(t + \epsilon/2) \leftarrow \mathbf{p}(t) - (\epsilon/2)\nabla_{\mathbf{q}} U(\mathbf{q}(t))$
    $\mathbf{q}(t + \epsilon) \leftarrow \mathbf{q}(t) + \epsilon \mathbf{M}^{-1}\mathbf{p}(t + \epsilon/2)$
    $\mathbf{p}(t + \epsilon) \leftarrow \mathbf{p}(t + \epsilon/2) - (\epsilon/2)\nabla_{\mathbf{q}} U(\mathbf{q}(t + \epsilon))$
    $t \leftarrow t + \epsilon$
end for
$\mathbf{q}^* \leftarrow \mathbf{q}(t)$
$H \leftarrow \exp\big(-U(\mathbf{q}(t)) + U(\mathbf{q}(0)) - \mathbf{p}(t)^t\mathbf{M}^{-1}\mathbf{p}(t)/2 + \mathbf{p}(0)^t\mathbf{M}^{-1}\mathbf{p}(0)/2\big)$    ▷ Hastings ratio
$\alpha \leftarrow \min(1, H)$
Draw $u \sim \text{Uniform}(0, 1)$
if $u < \alpha$ then
    Accept proposal $\mathbf{q}^*$
else
    Reject proposal $\mathbf{q}^*$
end if
```

the gradient $\nabla_{\mathbf{q}} U(\mathbf{q}(t))$, reflective HMC also requires computing the time t at which the trajectory $\mathbf{q} + t\mathbf{v}$ hits the boundary and (assuming a collision occurs) the vector normal to the boundary at the collision point.

In the context of sampling from the space of correlation matrices, the boundary is the curve $\det(\mathbf{C}) = 0$. To implement reflective HMC, we must compute the time to collision $\epsilon_{\text{intersect}} = \min\{s : \det(\mathbf{C} + s\mathbf{V}) = 0, s > 0\}$ for some symmetric velocity matrix $\mathbf{V}$ with zero-diagonals. To find the solutions to $\det(\mathbf{C} + s\mathbf{V}) = 0$, recall that the eigenvalues of any matrix $\mathbf{A}$ are the roots of the characteristic polynomial $\det(\mathbf{A} - s\mathbf{I})$. For invertible $\mathbf{V}$, $\det(\mathbf{C} + s\mathbf{V}) = 0 \iff \det(-\mathbf{C}\mathbf{V}^{-1} - s\mathbf{I}) = 0$ which implies $\epsilon_{\text{intersect}}$ in Line 8 of Algorithm 5.3 is the smallest positive eigenvalue of $-\mathbf{C}\mathbf{V}^{-1}$. The velocity $\mathbf{V}$, however, is guaranteed to be singular in certain circumstances as the elements of $\mathbf{V}$ corresponding to the structural zeros in $\mathbf{C}$ must also be zero. To circumvent this obstacle, we note that for invertible $\mathbf{C}$, $\det(\mathbf{C} + s\mathbf{V}) = 0 \iff \det(-\mathbf{C}^{-1}\mathbf{V} - (1/s)\mathbf{I}) = 0$. As such, the solutions to $\det(\mathbf{C} + s\mathbf{V}) = 0$ are the reciprocal eigenvalues of $-\mathbf{C}^{-1}\mathbf{V}$. $\mathbf{C}$ is a correlation matrix and

Algorithm 5.3 HMC with reflections

```
⋮
for $h = 1, \ldots, H$ do
    ⋮
    $t_{\text{sub}} \leftarrow t$                ▷ this code block replaces Line 6 in Algorithm 5.2
    $\mathbf{p}_{\text{sub}} \leftarrow \mathbf{p}(t + \epsilon/2)$
    while $t_{\text{sub}} < t + \epsilon$ do
        $\mathbf{v} \leftarrow \mathbf{M}^{-1}\mathbf{p}_{\text{sub}}$
        $\epsilon_{\text{intersect}} \leftarrow \min\{s : b(\mathbf{q}(t_{\text{sub}}) + s\mathbf{v}) = 0, s > 0\}$
        if $t_{\text{sub}} + \epsilon_{\text{intersect}} \leq t + \epsilon$ then                ▷ reflection occurs
            $\mathbf{q}(t_{\text{sub}} + \epsilon_{\text{intersect}}) \leftarrow \mathbf{q}(t_{\text{sub}}) + \epsilon_{\text{intersect}}\mathbf{M}^{-1}\mathbf{p}_{\text{sub}}$
            $\mathbf{a} \leftarrow \nabla_{\mathbf{q}} b(\mathbf{q}(t_{\text{sub}} + \epsilon_{\text{intersect}}))$    ▷ vector normal to the boundary at intersection
            $\mathbf{p}_{\text{sub}} \leftarrow \mathbf{p}_{\text{sub}} - 2\frac{\mathbf{p}_{\text{sub}}^t\mathbf{a}}{\mathbf{a}^t\mathbf{a}}\mathbf{a}$                ▷ reflection off boundary
            $t_{\text{sub}} \leftarrow t_{\text{sub}} + \epsilon_{\text{intersect}}$
        else
            $\mathbf{q}(t + \epsilon) \leftarrow \mathbf{q}(t_{\text{sub}}) + (t + \epsilon - t_{\text{sub}})\mathbf{M}^{-1}\mathbf{p}_{\text{sub}}$
            $t_{\text{sub}} \leftarrow t + \epsilon$
        end if
    end while
    ⋮
end for
⋮
```

therefore invertible everywhere except along the boundary $\det(\mathbf{C}) = 0$. If $\mathbf{C}$ is in fact at the boundary and non-invertible, we can still find the solutions to $\det(\mathbf{C} + s\mathbf{V}) = 0$ by selecting some arbitrary offset u and noting that

$$\det(\mathbf{C} + s\mathbf{V}) = \det((\mathbf{C} + u\mathbf{V}) + (s - u)\mathbf{V}) \tag{5.42}$$

Even if $\mathbf{C}$ is singular, $\mathbf{C} + u\mathbf{V}$ is almost surely full rank (although not necessarily positive definite). If $\mathbf{e}$ are the reciprocal eigenvalues of $-(\mathbf{C} + u\mathbf{V})^{-1}\mathbf{V}$, then the solutions to $\det(\mathbf{C} + s\mathbf{V}) = 0$ are $\mathbf{e} + u\mathbf{1}$. As such, we can calculate the time to colliding with the boundary $\epsilon_{\text{intersect}} = \min\{s : \det(\mathbf{C} + s\mathbf{V}) = 0, s > 0\}$ even with singular $\mathbf{C}$ and $\mathbf{V}$. Each of these checks requires an inverse and eigendecomposition, resulting in complexity $\mathcal{O}(K^3)$. However, calculating the gradient $\nabla_{\mathbf{C}} p(\mathbf{Y} \,|\, \mathbf{C}, \ldots)$ requires at least $\mathcal{O}(NK^3)$ work, so the cost of computing $\epsilon_{\text{intersect}}$ does not significantly impact run time.

For HMC steps that require reflection off the boundary, we must also compute the vector $\mathbf{a}$ normal to the boundary, which is equal to the gradient of the boundary function at the point of interest. Therefore, we must compute the gradient $\nabla_{\mathbf{C}} \det(\mathbf{C}) = \text{adj}(\mathbf{C})^t$, where adj(.) is the adjugate operator. For invertible $\mathbf{C}$, $\text{adj}(\mathbf{C}) = \det(\mathbf{C})\mathbf{C}^{-1}$. However, at the boundaries where we must compute the gradient when $\det(\mathbf{C}) = 0$ and $\mathbf{C}$ is not invertible. While the adjugate may also be constructed using the cofactors of $\mathbf{C}$ (i.e. the matrix of determinants of $K - 1 \times K - 1$ sub-matrices formed by removing individual rows and columns of $\mathbf{C}$), this approach scales $\mathcal{O}(K^5)$ and we seek a more elegant solution. Stewart (1998) provides such a solution and demonstrates that one can compute the adjugate of singular matrices using the singular value decomposition with computational complexity $\mathcal{O}(K^3)$.

5.4 Example analyses

We demonstrate the utility of these methods in four examples below. All data are standardized on a per-traits basis unless otherwise noted.

5.4.1 HIV-1 virulence

As discussed in Section 3.7.3, clinical outcomes in virally infected individuals are at least partially attributable to viral genetics. Payne et al. (2014) examine the effects of specific human leukocyte antigen (HLA) escape mutations (i.e. mutations that help the virus evade the human immune system) on viral load and CD4 T-cell count in HIV-infected individuals as well as viral replicative capacity. While Payne et al. (2014) identify associations between clinically relevant viral phenotypes (viral load, CD4 T-cell count, replicative capacity) and specific HLA escape mutations, they do not account for the phylogenetic relationships between viruses in the study. Zhang et al. (2021, 2022) re-analyze this data set of $N = 535$ sequences with associated mutations and clinical measurements in the phylogenetic context using the multivariate probit model of Cybis et al. (2015). This multivariate probit model maps discrete traits on a tree to continuous space so that the correlation between discrete traits can be estimated using a Gaussian process on the tree (e.g. multivariate Brownian diffusion). Zhang et al. (2021, 2022) identify several mutations associated with viral load and replicative capacity and infer conditional dependency graphs based on these associations.

Here, we re-examine this work through the lens of phylogenetic structural equation modeling. Specifically, we assume two sub-models with sub-model one being a latent factor model capturing the low-dimensional structure of the HLA escape mutations ($P_1 = 20$) and the second sub-model capturing relationships between replicative capacity, viral load, and CD4 T-cell count ($P_2 = 3$). As the model selection strategy described in Chapter 4 underestimates the number of factors when a large proportion of traits are discrete, we arbitrarily choose $K_1 = 3$ factors for the first sub-model (the second sub-model does not rely on dimension reduction). We perform inference using a fixed phylogenetic tree inferred by Zhang et al. (2021). We post-process the results using Procrustes rotation (see Section 5.3.2) and subsequently rotate the posterior to maximize correlation between the first factor and a combination of replicative capacity, viral load, and (negative) CD4 T-cell count. We present the results in Figure 5.2.

The first factor has correlations of 0.27 (0.11 - 0.45 highest posterior density interval), 0.12 (-0.06 - 0.31) and -0.16 (-0.36 - 0.01) with replicative capacity, viral load, and CD4 T-Cell count, respectively. Notably, this first factor only loads positively onto two mutations, both with very small loadings magnitudes. There are, however, several mutations (e.g. V168I, Q182X, T186X, T190X) that this first factor loads negatively onto. This suggests that these mutations may interact and result in viruses with lower replicative capacity and infections with lower viral load and higher host CD4 T-cell counts. This recapitulates known biology to a certain extent, as T186X is a known immune escape mutation that dramatically lowers replicative capacity. However, when a T186X mutation is paired with a T190I mutation (a subset of T190X mutations), this drop in replicative capacity is partially offset (Wright et al.,

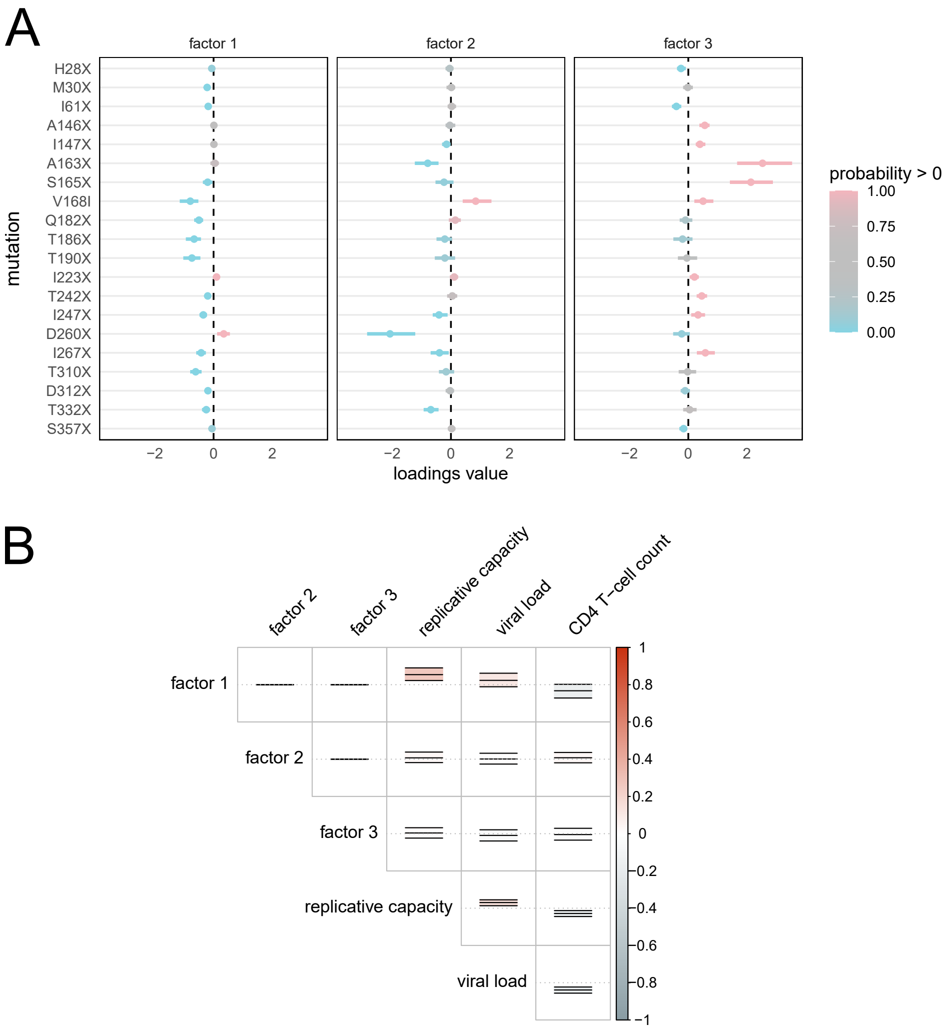


Figure 5.2: HIV mutation profile and correlation with clinically relevant phenotypes. **A**) Loadings matrix mapping latent factors on the phylogenetic tree to HIV immune escape mutations. The samples have been rotated to maximize the sum of the correlations between the first factor and replicative capacity, viral load, and (negative) CD4 T-cell count (see Section 5.3.2.2). **B**) Summary of posterior distribution of the correlation between the mutation factors and clinically relevant phenotypes. Color and middle lines represent the posterior means while lower and upper lines span the 95% HPD interval. Correlations between factors are fixed at zero by construction.

2012). Note that a similar relationship is known between A163X and S165X mutations. The interaction between those mutations is captured by the third factor, but this third factor does not appear to be significantly associated with any of the relevant clinical variables.

5.4.2 *Aquilegia* pollination

Plants in the genus *Aquilegia* (i.e. columbines) have a high degree of floral phenotypic diversity. This phenotypic diversity is accompanied by a diversity in the animals that pollinate these plants. Namely, *Aquilegia* species are pollinated by either bumblebees, hummingbirds, hawk moths or some combination thereof. Whittall and Hodges (2007) explore the relationship between floral phenotypes and pollinator and identify "pollination syndromes" (i.e. collections of floral phenotypes) associated with each pollinator. Tolkoff et al. (2018) and Hassler et al. (2022, Section 4.6.1) examine this problem using phylogenetic factor analysis to evaluate the extent to which the pollination syndromes of Whittall and Hodges (2007) correspond to the low-rank structure of floral phenotypes. One limitation of the Tolkoff et al. (2018) and Hassler et al. (2022) analyses, however, is that pollinator types are included as phenotypes in the latent factor model. Jointly inferring the low-dimensional structure of both the floral phenotypes and pollinator type makes the resulting factors difficult to interpret. Moreover, Hassler et al. (2022) note that discrete traits (e.g. pollinator type) seem to have an out-sized influence on the loadings matrix of a phylogenetic factor analysis, and inclusion of these discrete traits may influence the loadings associated with other traits.

Phylogenetic structural equation modeling solves these problems by allowing us to separate the dimension reduction over the floral phenotypes from the pollinator type. Rather than including all traits in a single latent factor model, we assume a latent factor model for the floral phenotypes only ($P_1 = 11$) and infer the association between these low-dimensional floral factors and pollinator type ($P_2 = 3$). Hassler et al. (2022) identify a two-factor model as having optimal predictive performance, so we opt for $K_1 = 2$. The second sub-model is a full-rank (i.e. $K_2 = P_2 = 3$) residual variance model with diagonal variance. To induce iden-

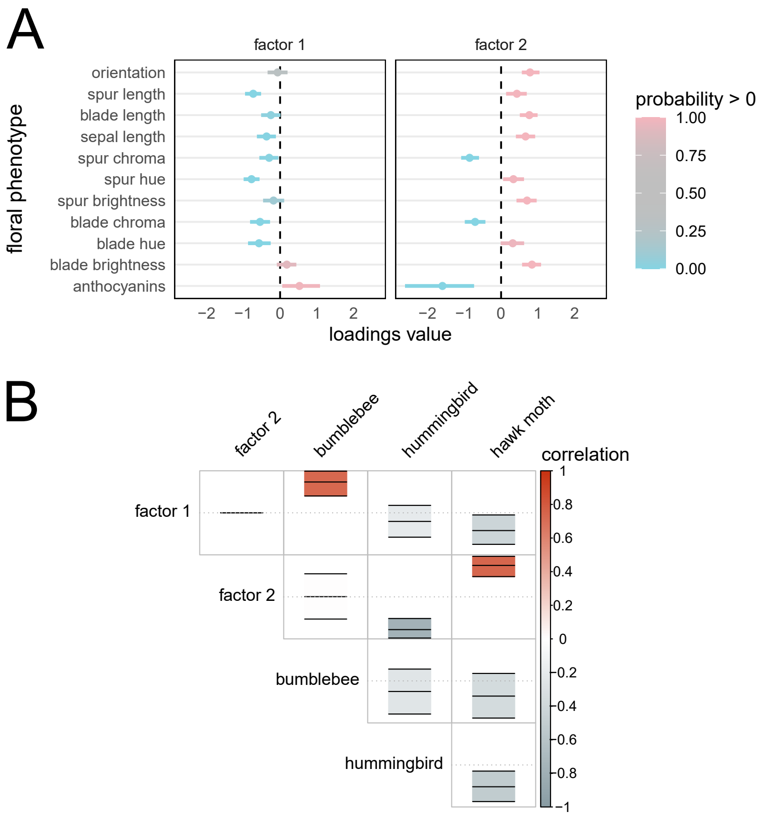


Figure 5.3: Relationship between floral phenotypes and pollinator type in *Aquilegia*. **A**) Loadings matrix mapping latent factors to floral phenotypes. The posterior of the loadings matrix has been rotated to maximize posterior correlation of the first factor with bumblebee pollination. **B**) Correlation between the latent factors and pollinators. See Figure 5.2 for figure details. The first factor is strongly associated with bumblebee pollination, while the second factor captures a hummingbird/hawk moth axis.

tifiability, we post-process the factor results using Procrustes rotation. Finally, we rotate the latent factors such that the correlation between the first factor and bumblebee pollination is maximized. We present the results in Figure 5.3.

The results are generally consistent with those of Hassler et al. (2022). Specifically, the first factor here, which we rotate to optimize association with bumblebee pollination versus hummingbird and hawk moth pollination, has the same general pattern as the second factor that Hassler et al. (2022) identify (see Figure 4.3). A major exception is that Hassler et al. (2022) did not infer association between bumblebee pollination and anthocyanin production, while we do here. Notably, the second factor here which is strongly negatively correlated with hummingbird pollination positively correlated with hawk moth pollination is nearly identical to the first factor identified by Hassler et al. (2022). Generally, the structural equation approach here confirms the earlier results of Hassler et al. (2022) and clearly identifies two axes along which *Aquilegia* flowers have evolved.

5.4.3 Yeast domestication

The yeast species *Saccharomyces cerevisiae* is a staple of human culinary and industrial activity and has been in the process of domestication for at least the last 400 years (Gallone et al., 2016). Gallone et al. (2016) and Gallone et al. (2019) study this domestication process by collecting numerous measurements of growth rates under varying stress conditions, the production of aromatic compounds, and the ability to reproduce outside of industrial settings for $N = 154$ strains of *S. cerevisiae*. Hassler et al. (2022, Section 4.6.2) seek to identify a "domestication phenotype" via phylogenetic factor analysis. However, this approach includes all phenotypes in a single latent factor model, which makes it challenging to explicitly study certain sub-sets of phenotypes.

Here, we reanalyze this $N = 154, P = 82$ data set using phylogenetic structural equation modeling. Specifically, we separate the 82 phenotypes into three partitions. The sub-model on the first partition with $P_1 = 62$ measurements of growth under stress conditions is a latent factor model with $K_1 = 2$. The second sub-model is also a latent factor model covering $P_2 = 16$ measurements of production of various aromatic compounds with $K_2 = 2$. Note that we arbitrarily choose $K = 2$ for both latent factor models. The final sub-model includes the four remaining traits without dimension reduction: flocculation (i.e. the tendency of yeast to clump together and fall out of solution), ethanol production, sporulation efficiency and spore viability. We assume a fixed tree inferred by Hassler et al. (2022). We post-process the results of both factor models using Procrustes rotation and rotate the resulting posterior to maximize the combined correlation with sporulation efficiency and spore viability. We present the results in Figures 5.4 and 5.5.

The first stress factor captures general patterns of tolerance to environmental and nutrient stress, with high factor values corresponding to faster growth under stress conditions. The notable exceptions are growth at 20°C (near room temperature) and in maltotriose growth medium (a common element of beer wort but rare outside of brewing contexts). This first factor has low values in a clade of pre-dominantly beer yeasts, suggesting that these yeast strains have adapted to survive primarily in beer wort and have lost tolerance to many stressors uncommon in industrial beer production. The first aroma factor (also rotated to have optimal correlation with sporulation efficiency and spore viability) paints a less clear picture, but concentrates posterior mass on several esters and 4-vinyl guaiacol (4-VG), known to cause off flavors in beer. As expected, yeast strains in the large beer clade have lower values for this factor suggesting that they produce less 4-VG. The first stress factor and first aroma factor show strong correlation, suggesting that the selective forces on both growth in stress conditions and production of aromatic compounds may have been driven by the same underlying process (i.e. adaptation of yeast to industrial beer production).

5.4.4 SARS-CoV-2 antigenic evolution

RNA viruses are some of the most fast-evolving organisms on the planet (Peck et al., 2018), and SARS-CoV-2 (the virus responsible for COVID-19) is no exception. While the selective pressures driving this evolution are numerous, two of the most important in the context of SARS-CoV-2 are the affinity of the SARS-CoV-2 spike protein for the human ACE2 receptor (Shang et al., 2020) and its ability to escape the human immune system (Harvey et al., 2021). Starr et al. (2020) and Greaney et al. (2021) perform deep mutational scanning to identify how individual mutations affect the binding affinities of the SARS-CoV-2 spike protein with the human ACE2 receptor and polyclonal antibodies, respectively. In both studies, researchers intentionally mutate the SARS-CoV-2 genome to produce all possible amino acid substitutions for a given codon. They then test the binding affinity between

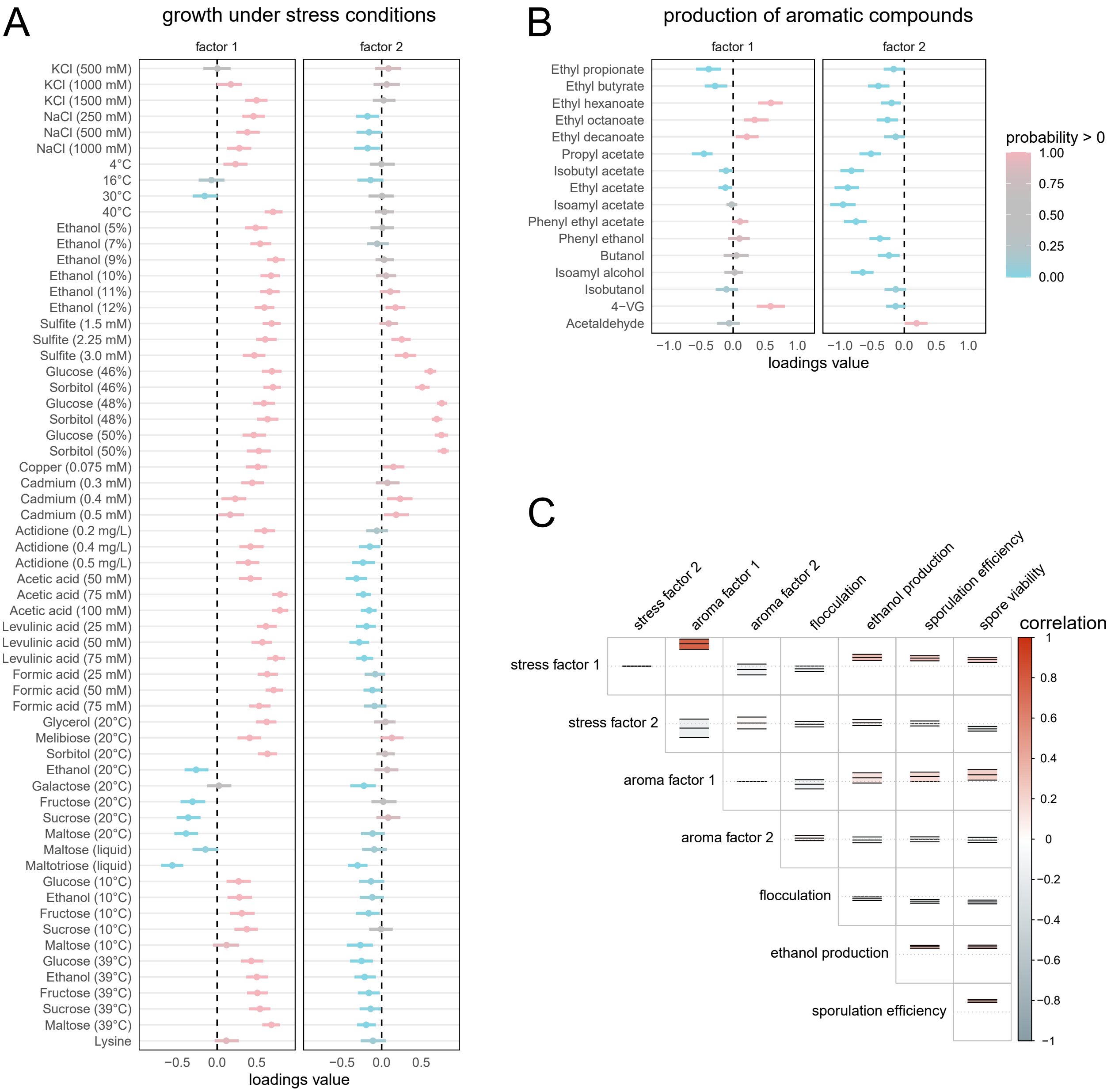


Figure 5.4: Relationship between yeast growth rates under stress conditions, production of aromatic compounds, ethanol production and reproductive ability. **A**) Loadings matrix for growth under stress conditions. The posterior has been rotated to maximize correlation of the first factor with sporulation efficiency and spore viability. This first factor characterizes tolerance to environmental stress generally, with the only negative values associated with either growth under normal temperatures or in environments similar to beer wort. **B**) Loadings matrix for production of aromatic compounds. 4-VG in particular causes off flavors in many beers. *(caption continues on next page)*

5.4 *(previous page)*: **C**) Correlation between both sets of latent factors and remaining phenotypes. The posterior distributions of the first stress factor and first aroma factor have both been rotated to have maximum correlation with sporulation efficiency and spore viability. See Figure 5.2 for a description of plot elements.

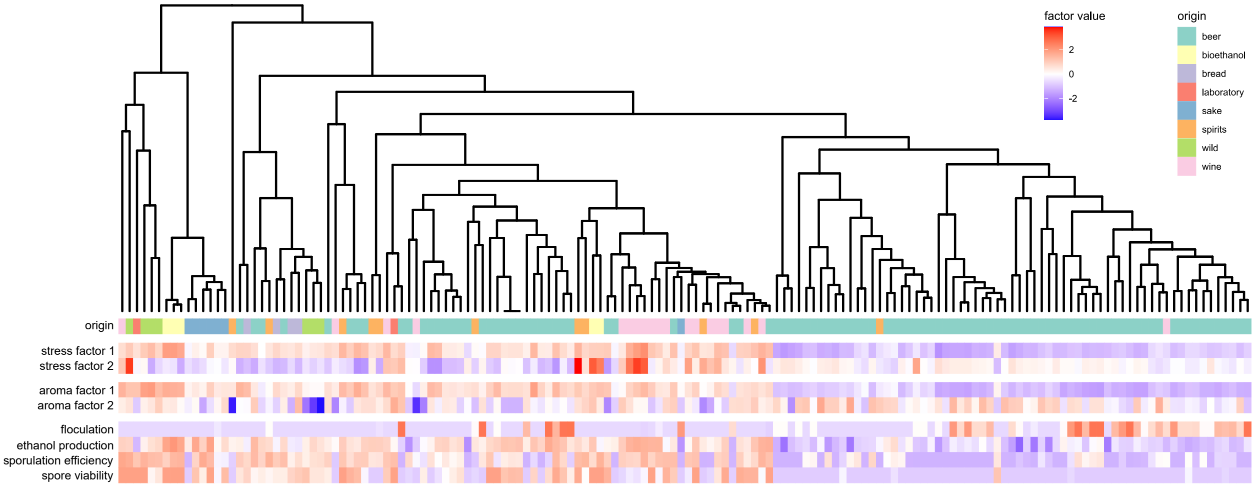


Figure 5.5: Latent factors and other yeast phenotypes plotted on a phylogenetic tree. The tree was inferred by Hassler et al. (2022, Section 4.6.2). The large clade consisting of predominantly beer yeast on the right show strong signs of domestication, with low values of the stress factor 1 (associated with tolerance to stress) and aroma factor 1 (associated with aromatic compounds known to produce off flavors). These factors co-evolved on the tree with lower ethanol production, sporulation efficiency, and spore viability.

the spike protein of the mutated virus and either the human ACE2 receptor or human polyclonal anti-SARS-CoV-2 antibodies. They repeat this for each amino acid in the spike protein. Note that this approach looks at the individual effects of mutations at a single codon. The researchers do not directly test how interactions between multiple mutations can influence binding. Regardless, these studies give valuable insight into which sites on the spike protein may be particularly good targets for natural selection.

Phylogenetic factor analysis and phylogenetic structural equation modeling offer tools for studying the evolution of SARS-CoV-2 through antigenic space similar to the HIV analysis in Section 5.4.1. We collect $N = 3210$ SARS-CoV-2 RNA sequences and build a phylogenetic tree. From the RNA sequences we determine per-site amino acid substitutions on the spike protein and identify the change in ACE2 and antibody binding affinity associated with each mutation. We structure the phylogenetic structural equation model with one latent factor sub-model ($P_1 = 199$, $K_1 = 2$) corresponding to the per-mutation ACE2 binding affinities and another ($P_2 = 172$, $K_2 = 2$) corresponding to the per-mutation polyclonal antibody binding affinities. We map the resulting factors onto the phylogenetic tree in Figure 5.6.

We first emphasize that the total proportion of variance explained by the factor sub-models for both ACE2 and antibody binding affinities are 3.95% (3.86%-4.04%) and 16.3% (15.5%-17.3%), respectively. The extremely low proportion of variance explained by the factor model for the ACE2 binding affinities suggests that either there is not some low-dimensional structure that well-describes the pattern of evolution of ACE2 binding affinity or that this low-dimensional process is not well captured by the assumptions the latent factor model on a phylogenetic tree. With that in mind, the first ACE2 factor captures general variation across the entire phylogeny, while the second has very low values for the Omicron variant. Similarly, both antibody factors capture essentially the same variation in the phylogenetic tree in that for a given taxon positive values in the first antibody factor are almost always accompanied by negative values in the second factor. The only notable exception to this pattern is the Omicron variant. While we cannot make broad claims about the evolution of SARS-CoV-2 in antigenic space due to the low proportion of variance explained by this model, one clear result is (unsurprisingly) that the Omicron variant has a highly unusual mutational profile with potential implications for both its ability to bind to human ACE2 receptors and evade pre-existing immunity. We do not explore these topics here, but they are an area of active research (see Han et al., 2022; Junker et al., 2022; Lupala et al., 2022; Planas et al., 2022; Wu et al., 2022).

Notably, there is also a high degree of correlation between the ACE2 and antibody factors, with correlations of some rotations of the factors (not shown) in excess of 0.99. This is not entirely surprising however, as both the ACE2 and antibody binding data sets rely on the same underlying mutational profile. In other words, if a particular sequence has a mutation

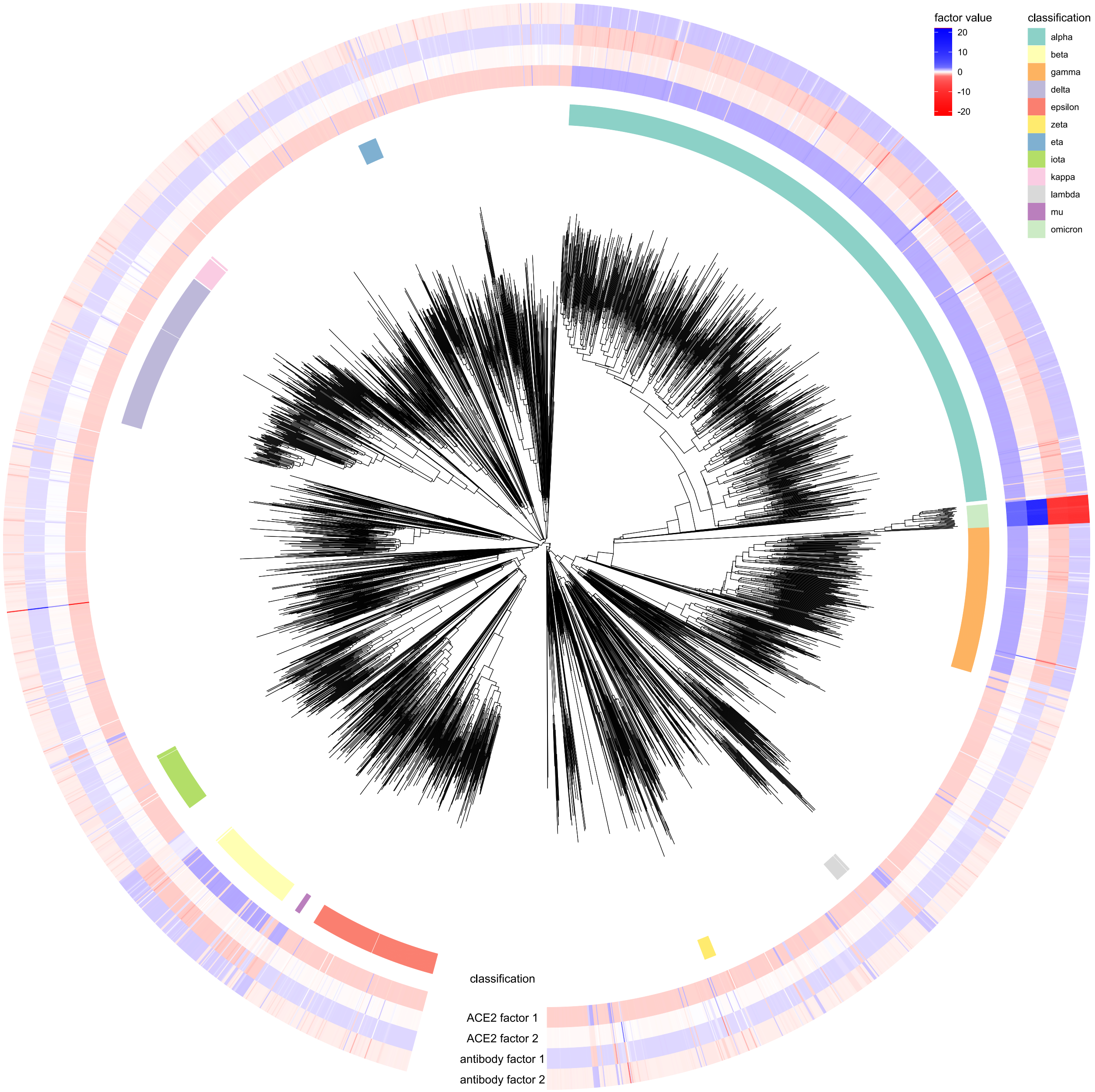


Figure 5.6: Posterior means of latent factors associated with binding between SARS-CoV-2 spike protein and human ACE2 receptors and polyclonal antibodies on a phylogenetic tree. The tree includes a small number of sequences belonging to the Omicron variant. Factors values associated with the Omicron sequences are outliers, which is consistent with existing research that the Omicron variant is antigenically highly distinct from other variants.

that changes a spike protein amino acid, the change in that amino acid will be registered in both the ACE2 and antibody data sets (although with different effects on their relative binding affinities). As such, one would expect extremely high correlation between these two sets of factors, which we indeed observe. This extremely high correlation between the factors means that there is a region of high posterior density near the boundary of the correlation space $\det(\mathbf{C}) = 0$. We demonstrate here that the reflective HMC sampler introduced in Section 5.3.4.3 functions well even in this extreme situation.

5.5 Discussion

We develop here an efficient and flexible set of models for studying the evolution of biological traits on phylogenetic trees. This approach synthesizes previous work (Tolkoff et al., 2018; Hassler et al., 2020, 2022) under a unifying framework that allows for highly structured yet computationally tractable models where information between sub-models is shared via some evolutionary process on a phylogenetic tree. We design these models so that all calculations for inference scale linearly in both the total number of species N and traits P despite complex dependencies between observations. These innovations allow Bayesian inference in phylogenetic trait models at unprecedentedly large scales. For example, in Section 5.4.4, we perform inference in a problem with $N = 3120$ sequences and $P = 371$ traits. Moreover, this analysis requires sampling from a nearly singular correlation matrix with non-trivial constraints.

In addition to the computational efficiency and ability to sample from challenging spaces, we adapt post-processing methods to the factor analytic context that dramatically reduce identifiability challenges. We also describe a simple approach to chose one particular rotation of the posterior based on the scientific question of interest rather than arbitrary identifiability constraints.

These approaches, however, have limitations that we would like to address before they can be broadly adopted. First, both the Procrustes post-processing and correlation-maximizing rotations have the potential to increase type 1 error. While both procedures create valid posterior samples, they do so in a way designed to minimize posterior variance in the loadings and maximize posterior correlation, respectively. While preliminary results from simulation studies suggest that these effects are both measurable and manageable, we believe further study is warranted.

Additionally, there appears to be a tendency in some data sets to infer nearly-singular correlation matrices when there are at least two latent factor sub-models, both of which have a high number of traits. While these observations may be the result of the idiosyncrasies of individual data sets (as with our example in Section 5.4.4), we believe this phenomenon should be more thoroughly explored. One advantage, however, of our approach to inferring the correlation matrix is that we can place non-traditional priors on different elements of the correlation matrix to correct for this tendency (if it does indeed exist). The commonly used LKJ prior induces exchangeability between elements of the correlation matrix, which hold even in the case where we require structural zeros (see Section 5.3.4.1). Alternative priors that shrink certain elements of the correlation (e.g. those corresponding to correlation between two latent factors in different high-dimensional sub-models) toward zero could help address this issue. While we have not fully explored this, it is possible that such priors could be motivated by partial-correlation vines similar to those used to construct the LKJ distribution (Joe, 2006).

Finally, while we have focused on the very specific case of multivariate Brownian diffusion on a phylogenetic tree, all methods discussed here apply more broadly. In the phylogenetic context, they apply to alternative Gaussian models such as Ornstein–Uhlenbeck (OU) processes commonly used to model directional selection. Outside of phylogenetics, these methods apply to any Gaussian model where the covariance between observations is additive on an acyclic graph (see Ho and Ané, 2014). This applies to any hierarchical Gaussian model where the dependence structure can be mapped to a tree.

CHAPTER 6

Conclusion and future directions

6.1 Methodological advances

The work I have presented here, culminating in Chapter 5, enables scalable inference in flexible models of continuous trait evolution on phylogenetic trees. These methods, however, are not unique to the phylogenetic context and are broadly applicable to any highly structured Gaussian hierarchical model. These methodological advances collectively permit efficient Bayesian inference with missing data (Chapter 3), model extensions with different between-trait covariance structures (Chapter 3), dimension reduction (Chapter 4) and multiple, correlated sub-models with unusual constraints on the correlation matrix (Chapter 5). The increases in computational efficiency are not simply a matter of convenience but often make the difference between analyses that are doable by a researcher with ordinary resources and those that are not. The methods in Chapters 3 and 4 achieve increases in computational efficiency of two orders of magnitude in large data sets, bringing computation times down from an order of months to an order of hours. While I have yet to formally benchmark the new statistical methods developed in Chapter 5, they rely on the same principles as those developed in the earlier chapters and I anticipate comparable computational performance.

6.2 Scientific contributions

While many of the empirical examples presented in Chapters 3, 4 and 5 were chose to illustrate the capabilities of the new methods, several have contributed to scientific knowledge in their own right. Specifically, Section 3.7.2 identifies a relationship between the GC content of bacterial genomes and their optimal growth temperatures. This relationship had been hypothesized, but evidence for it was inconclusive partly due to studies that either did not correct for phylogenetic relationships between bacterial lineages or had small sample size. Our analysis is the largest ($N = 435$) that examines this relationship while correcting for shared ancestry.

Similarly, in Sections 4.6.2 and 5.4.3 we identify a single latent factor contributing a plurality of the variance in yeast stress tolerance that separates industrial beer yeast from other strains of *S. cerevisiae*. While these traits had been examined on an individual basis, this was the first to quantitatively map a low-dimensional "domestication factor" capturing broad patterns of variation to the phylogenetic tree of industrial yeast.

In Section 4.6.3, we study the evolution of life history traits in the largest known phylogenetic analysis of its kind with $N = 3649$ species of mammals. This analysis partitions the variance in mammalian life history into that which is body-size dependent and that which is body-size independent. We identify that the primary source of variation on mammalian life history traits is body-size independent and that these size-independent patterns of variation correspond to predictions made by pace-of-life theory.

Finally, novel statistical methods are unhelpful to the broader scientific community without user-friendly software implementation. All of the core methods described in this dissertation have been implemented in the widely used Bayesian phylogenetic inference software BEAST (Suchard et al., 2018), which is well known in both the ecology and viral phylodynamics communities. I am in the process of including these methods the the BEAST graphical user interface to further aid community adoption. Additionally, I have written the Julia package PhylogeneticFactorAnalysis.jl (Hassler et al., 2022, Section 4.5.4) that wraps BEAST and includes all post-processing procedures developed in Chapters 4 and 5. The ultimate goal, however, is to include all PhylogeneticFactorAnalysis.jl functionality in BEAST and its graphical user interface, avoiding a two-language analysis pipeline.

6.3 Limitations and future directions

While statistical models discussed here are indeed flexible at modeling relationships between different traits, all impose a Gaussian likelihood over the data. Related to this assumption is that all dimension reduction relies on linear maps between high- and low-dimensional space. These assumptions are not appropriate for all data, particularly when some observations are discrete. While the multivariate probit model of Cybis et al. (2015) helps address this challenge for discrete traits, it requires sampling from a high-dimensional truncated Gaussian distribution that remains challenging, although substantially less so after recent work by Zhang et al. (2021) and Zhang et al. (2022). Regardless, researchers may want to use generalized linear models to link continuous latent space on phylogenetic trees to discrete phenotypes. Additionally, there are many continuous traits that are difficult to transform to normality, such as the SARS-CoV-2 spike protein binding affinities in Section 5.4.4 where the distribution of each affinity across multiple taxa has a large number of true zeros. Such cases may require more flexible, non-Gaussian model extensions that may disrupt the ability to analytically marginalize latent traits on the tree.

These models, where the latent variables evolve along a phylogenetic tree according to the same underlying Gaussian process but the model extension connecting the process on the tree to the data is non-Gaussian, may require returning to data augmentation of the latent traits at the tips of the tree. However, assuming the gradient of the augmented likelihood with respect to these latent traits is tractable (and I believe it is), then it may be possible to sample efficiently from the full conditional posterior of this latent space. Assuming such an approach were computationally efficient, it would open the door to even more flexible models of phenotypic evolution.

Bibliography

Adams, D. C. (2014a). A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Systematic Biology 63*(5), 685–697.

Adams, D. C. (2014b). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution 68*(9), 2675–2688.

Adams, D. C. (2014c). Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. *Systematic Biology 63*(2), 166–177.

Afshar, H. M. and J. Domke (2015). Reflection, refraction, and Hamiltonian Monte Carlo. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3007–3015.

Aguilar, O. and M. West (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics 18*(3), 338–357.

Alfaro, M. E., S. Zoller, and F. Lutzoni (2003). Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution 20*(2), 255–266.

Alizon, S., V. von Wyl, T. Stadler, R. D. Kouyos, S. Yerly, B. Hirschel, J. Böni, C. Shah, T. Klimkait, H. Furrer, A. Rauch, P. L. Vernazza, E. Bernasconi, M. Battegay, P. Bürgisser, A. Telenti, H. F. Günthard, S. Bonhoeffer, and Swiss HIV Cohort Study (2010, Sep). Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens 6*(9), e1001123.

Allen, G. and R. Tibshirani (2010). Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics 4*, 764–790.

Aßmann, C., J. Boysen-Hogrefe, and M. Pape (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics 192*(1), 190–206.

Aptekmann, A. A. and A. D. Nadra (2018). Core promoter information content correlates with optimal growth temperature. *Scientific Reports 8*(1), 1313.

Aristide, L., S. F. Dos Reis, A. C. Machado, I. Lima, R. T. Lopes, and S. I. Perez (2016). Brain shape convergence in the adaptive radiation of New World monkeys. *Proceedings of the National Academy of Sciences 113*(8), 2158–2163.

Aristide, L., A. L. Rosenberger, M. F. Tejedor, and S. I. Perez (2015). Modeling lineage and phenotypic diversification in the New World monkey (Platyrrhini, Primates) radiation. *Molecular Phylogenetics and Evolution 82*, 375–385.

Ayres, D. L., M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, and M. A. Suchard (2019). BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic Biology 68*(6), 1052–1061.

Ayres, D. L., A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, P. O. Lewis, J. P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, et al. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology 61*(1), 170–173.

Baele, G., P. Lemey, A. Rambaut, and M. A. Suchard (2017, 02). Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics 33*(12), 1798–1805.

Baele, G., M. A. Suchard, A. Rambaut, and P. Lemey (2017, 01). Emerging concepts of data integration in pathogen phylodynamics. *Systematic Biology 66*(1), e47–e65.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics 19*(6), 345–351.

Barido-Sottani, J., T. G. Vaughan, and T. Stadler (2020). A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. *Systematic Biology 69*(5), 973–986.

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

Bastide, P., L. S. T. Ho, G. Baele, P. Lemey, and M. A. Suchard (2021). Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. *The Annals of Applied Statistics 15*(2), 971 – 997.

Bedford, T. and R. M. Cooke (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031–1068.

Bernardi, G. and G. Bernardi (1986). Compositional constraints and genome evolution. *Journal of Molecular Evolution 24*(1-2), 1–11.

Berry, V. and O. Gascuel (1996). On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Molecular Biology and Evolution 13*(7), 999–1011.

Bertels, F., A. Marzel, G. Leventhal, V. Mitov, J. Fellay, H. F. Günthard, J. Böni, S. Yerly, T. Klimkait, V. Aubert, M. Battegay, A. Rauch, M. Cavassini, A. Calmy, E. Bernasconi, P. Schmid, A. U. Scherrer, V. Müller, S. Bonhoeffer, R. Kouyos, R. R. Regoes, and Swiss HIV Cohort Study (2018, Jan). Dissecting HIV virulence: Heritability of setpoint viral load, CD4+ T-cell decline, and per-parasite pathogenicity. *Molecular Biology and Evolution 35*(1), 27–37.

Bhattacharya, A. and D. B. Dunson (2011). Sparse Bayesian infinite factor models. *Biometrika 98*(2), 291–306.

Bielby, J., G. Mace, O. Bininda-Emonds, M. Cardillo, J. Gittleman, K. Jones, C. Orme, and A. Purvis (2007). The fast-slow continuum in mammalian life history: An empirical reevaluation. *The American Naturalist 169*(6), 748–757.

Blackburn, T. (1991). Evidence for a 'fast-slow' continuum of life-history traits among parasitoid Hymenoptera. *Functional Ecology 5*(1), 65–74.

Blanquart, F., C. Wymant, M. Cornelissen, A. Gall, M. Bakker, D. Bezemer, M. Hall, M. Hillebregt, S. H. Ong, J. Albert, N. Bannert, J. Fellay, K. Fransen, A. J. Gourlay, M. K. Grabowski, B. Gunsenheimer-Bartmeyer, H. F. Günthard, P. Kivelä, R. Kouyos, O. Laeyendecker, K. Liitsola, L. Meyer, K. Porter, M. Ristola, A. van Sighem, G. Vanham, B. Berkhout, P. Kellam, P. Reiss, C. Fraser, and B. collaboration (2017, 06). Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in europe. *PLoS Biology 15*(6), 1–26.

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

Bouchard-Côté, A., S. J. Vollmer, and A. Doucet (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association 113*(522), 855–867.

Bouckaert, R. (2016). Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ 4*, e2406.

Boukal, D. S., U. Dieckmann, K. Enberg, M. Heino, and C. Jørgensen (2014). Life-history implications of the allometric scaling of growth. *Journal of Theoretical Biology 359*, 199–207.

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *Journal of Anatomy 212*(4), 337–353.

Brito, A. F., E. Semenova, G. Dudas, G. W. Hassler, C. C. Kalinich, M. U. Kraemer, S. C. Hill, E. C. Sabino, O. G. Pybus, C. Dye, et al. (2021). Global disparities in SARS-CoV-2 genomic surveillance. *medRxiv*.

Bryja, J., O. Mikula, R. Šumbera, Y. Meheretu, T. Aghová, L. A. Lavrenchenko, V. Mazoch, N. Oguge, J. S. Mbau, K. Welegerima, et al. (2014). Pan-African phylogeny of Mus (subgenus Nannomys) reveals one of the most successful mammal radiations in Africa. *BMC Evolutionary Biology 14*(1), 1–20.

Byrne, S. and M. Girolami (2013). Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Cantet, R. J., A. N. Birchmeier, and J. P. Steibel (2004). Full conjugate analysis of normal multiple traits with missing records using a generalized inverted Wishart distribution. *Genetics, Selection and Evolution 36*, 49–64.

Capellini, I., J. Baker, W. Allen, S. Street, and C. Vendetti (2015). The role of life history traits in mammalian invasion success. *Ecology Letters 18*, 1099–1107.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1).

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In *Compstat*, pp. 227–232. Springer.

Chipman, H., E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes-Monograph Series*, 65–134.

Clavel, J., L. Aristide, and H. Morlon (2019). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to New-World monkeys brain evolution. *Systematic Biology 68*(1), 93–116.

Clobert, J., T. Garland, and R. Barbault (1998). The evolution of demographic tactics in lizards: A test of some hypotheses concerning life history evolution. *Journal of Evolutionary Biology 11*(3), 329–364.

Comas, I., M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, et al. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature Genetics 45*(10), 1176–1182.

Cybis, G., J. Sinsheimer, T. Bedford, A. Mather, P. Lemey, and M. Suchard (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Annals of Applied Statistics 9*, 969 – 991.

Dang, T. and H. Kishino (2019). Stochastic variational inference for Bayesian phylogenetics: a case of CAT model. *Molecular Biology and Evolution 36*(4), 825–833.

De Maio, N., P. Kalaghatgi, Y. Turakhia, R. Corbett-Detig, B. Q. Minh, and N. Goldman (2022). Maximum likelihood pandemic-scale phylogenetics. *bioRxiv*.

De Maio, N., C.-H. Wu, K. M. O'Reilly, and D. Wilson (2015). New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genetics 11*(8), e1005421.

Dearlove, B. L., A. J. Cody, B. Pascoe, G. Méric, D. J. Wilson, and S. K. Sheppard (2016). Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. *The ISME Journal 10*(3), 721–729.

Dellicour, S., G. Baele, G. Dudas, N. R. Faria, O. G. Pybus, M. A. Suchard, A. Rambaut, and P. Lemey (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications 9*(1), 1–9.

Dellicour, S., K. Durkin, S. L. Hong, B. Vanmechelen, J. Martí-Carreras, M. S. Gill, C. Meex, S. Bontems, E. André, M. Gilbert, et al. (2021). A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Molecular Biology and Evolution 38*(4), 1608–1613.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B 39*(1), 1–38.

Dill, K. A., K. Ghosh, and J. D. Schmit (2011). Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences 108*(44), 17876–17882.

Dinh, V., A. Bilge, C. Zhang, and F. A. Matsen IV (2017). Probabilistic path Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pp. 1009–1018. PMLR.

Dodd, M. S., D. Papineau, T. Grenne, J. F. Slack, M. Rittner, F. Pirajno, J. O'Neil, and C. T. Little (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature 543*(7643), 60–64.

Doebley, J. F., B. S. Gaut, and B. D. Smith (2006). The molecular genetics of crop domestication. *Cell 127*(7), 1309–1321.

Dominici, F., G. Parmigiani, and M. Clyde (2000). Conjugate analysis of multivariate normal data with incomplete observations. *Canadian Journal of Statistics 28*, 533–550.

Doolittle, W. F. and C. Sapienza (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature 284*(5757), 601.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology 4*(5), e88.

Drummond, A. J. and M. A. Suchard (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology 8*(1), 1–12.

Du Plessis, L., J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, et al. (2021). Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science 371*(6530), 708–712.

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature 544*(7650), 309–315.

Edwards, A. and L. Cavalli-Sforza (1964). Reconstruction of evolutionary trees. In V. H. Heywood and J. McNeill (Eds.), *Phenetic and Phylogenetic Classification*, pp. 67–76. The Systematics Association.

Efron, B., E. Halloran, and S. Holmes (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences 93*(14), 7085–7090.

Ehichioya, D. U., M. Hass, B. Becker-Ziaja, J. Ehimuan, D. A. Asogun, E. Fichet-Calvet, K. Kleinsteuber, M. Lelke, J. ter Meulen, G. O. Akpede, et al. (2011). Current molecular epidemiology of lassa virus in Nigeria. *Journal of Clinical Microbiology 49*(3), 1157–1161.

Ernest, S. (2003). Life history characteristics of placental nonvolant mammals. *Ecology 84*(12), 3402.

Fabreti, L. G. and S. Höhna (2021). Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *bioRxiv*.

Falconer, D. S. (1960). *Introduction to Quantitative Genetics.* Oliver And Boyd; Edinburgh; London.

Faulkner, J. R., A. F. Magee, B. Shapiro, and V. N. Minin (2020). Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. *Biometrics 76*(3), 677–690.

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

Felsenstein, J. (1973b). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics 25*(5), 471.

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

Felsenstein, J. (1985a). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution 39*(4), 783–791.

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

Felsenstein, J. (2004). *Inferring Phylogenies*, Volume 2. Sinauer Associates Sunderland, MA.

Felsenstein, J. and H. Kishino (1993). Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. *Systematic Biology 42*(2), 193–200.

Fernández-Sánchez, J., J. G. Sumner, P. D. Jarvis, and M. D. Woodhams (2015). Lie Markov models with purine/pyrimidine symmetry. *Journal of Mathematical Biology 70*(4), 855–891.

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

Flouri, T., F. Izquierdo-Carrasco, D. Darriba, A. J. Aberer, L.-T. Nguyen, B. Minh, A. Von Haeseler, and A. Stamatakis (2015). The phylogenetic likelihood library. *Systematic Biology 64*(2), 356–362.

Fourment, M., B. C. Claywell, V. Dinh, C. McCoy, F. A. Matsen IV, and A. E. Darling (2018). Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. *Systematic Biology 67*(3), 490–502.

Fourment, M. and A. E. Darling (2019). Evaluating probabilistic programming and fast variational Bayesian inference in phylogenetics. *PeerJ 7*, e8272.

Fourment, M., A. F. Magee, C. Whidden, A. Bilge, F. A. Matsen IV, and V. N. Minin (2020). 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Systematic Biology 69*(2), 209–220.

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution 3*(5), 940–947.

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

Gallone, B., J. Steensels, S. Mertens, M. C. Dzialo, J. L. Gordon, R. Wauters, F. A. Theßeling, F. Bellinazzo, V. Saels, B. Herrera-Malaver, et al. (2019). Interspecific hybridization facilitates niche adaptation in beer yeast. *Nature Ecology & Evolution 3*(11), 1562–1575.

Gallone, B., J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, et al. (2016). Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell 166*(6), 1397–1410.

Gaya, E., B. D. Redelings, P. Navarro-Rosinés, X. Llimona, M. De Cáceres, and F. Lutzoni (2011). Align or not to align? resolving species complexes within the Caloplaca saxicola group as a case study. *Mycologia 103*(2), 361–378.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association 95*(452), 1300–1304.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology 253*(4), 769–778.

Geweke, J. and G. Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies 9*(2), 557–587.

Giovannoni, S. J., J. C. Thrash, and B. Temperton (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal 8*(8), 1553.

Glanz, H. and L. Carvalho (2018). An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis 167*, 31–48.

Goberna, M. and M. Verdú (2016). Predicting microbial traits with phylogenies. *The ISME Journal 10*(4), 959.

Goolsby, E. W. (2017). Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. *Ecology and Evolution 7*(8), 2791–2797.

Goolsby, E. W., J. Bruggeman, and C. Ané (2017). Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution 8*(1), 22–27.

Greaney, A. J., A. N. Loes, K. H. Crawford, T. N. Starr, K. D. Malone, H. Y. Chu, and J. D. Bloom (2021). Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host & Microbe 29*(3), 463–476.e6.

Guindon, S. and N. De Maio (2021). Accounting for spatial sampling patterns in Bayesian phylogeography. *Proceedings of the National Academy of Sciences 118*(52).

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli*, 223–242.

Han, P., L. Li, S. Liu, Q. Wang, D. Zhang, Z. Xu, P. Han, X. Li, Q. Peng, C. Su, et al. (2022). Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. *Cell 185*(4), 630–640.

Hannaford, N. E., S. E. Heaps, T. M. Nye, T. A. Williams, and T. M. Embley (2020). Incorporating compositional heterogeneity into Lie Markov models for phylogenetic inference. *The Annals of Applied Statistics 14*(4), 1964–1983.

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

Harrington, S. M., V. Wishingrad, and R. C. Thomson (2021). Properties of Markov chain Monte Carlo performance across many empirical alignments. *Molecular Biology and Evolution 38*(4), 1627–1640.

Harvey, W. T., A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, S. J. Peacock, et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology 19*(7), 409–424.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*(2), 160–174.

Hassler, G., M. R. Tolkoff, W. L. Allen, L. S. T. Ho, P. Lemey, and M. A. Suchard (2020). Inferring phenotypic trait evolution on large trees with many incomplete measurements. *Journal of the American Statistical Association 0*(0), 1–15.

Hassler, G. W., B. Gallone, L. Aristide, W. L. Allen, M. R. Tolkoff, A. J. Holbrook, G. Baele, P. Lemey, and M. A. Suchard (2022). Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis. *Methods in Ecology and Evolution.*

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. von Krosigk (1959, June). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*(2), 192–218.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Reviews 23*(1), 53–60.

Hillis, D. M. and J. J. Bull (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology 42*(2), 182–192.

Ho, L. S. T. and C. Ané (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology 63*(3), 397–408.

Hobolth, A. and J. L. Jensen (2011). Summary statistics for endpoint-conditioned continuous-time Markov chains. *Journal of Applied Probability 48*(4), 911–924.

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

Hodcroft, E., J. D. Hadfield, E. Fearnhill, A. Phillips, D. Dunn, S. O'Shea, D. Pillay, A. J. Leigh Brown, UK HIV Drug Resistance Database, and UK CHIC Study (2014). The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathogens 10*(5), e1004112.

Hodcroft, E. B., M. Zuber, S. Nadeau, T. G. Vaughan, K. H. Crawford, C. L. Althaus, M. L. Reichmuth, J. E. Bowen, A. C. Walls, D. Corti, et al. (2021). Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature 595*(7869), 707–712.

Hoff, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics 18*(2), 438–456.

Höhna, S. and A. J. Drummond (2012). Guided tree topology proposals for Bayesian phylogenetic inference. *Systematic Biology 61*(1), 1–11.

Höhna, S., W. A. Freyman, Z. Nolen, J. P. Huelsenbeck, M. R. May, and B. R. Moore (2019). A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv*, 555805.

Holbrook, A., A. Vandenberg-Rodes, N. Fortin, and B. Shahbaba (2017). A bayesian supervised dual-dimensionality reduction model for simultaneous decoding of LFP and spike train signals. *Stat 6*(1), 53–67.

Holbrook, A., A. Vandenberg-Rodes, and B. Shahbaba (2016). Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*.

Holbrook, A. J., P. Lemey, G. Baele, S. Dellicour, D. Brockmann, A. Rambaut, and M. A. Suchard (2021). Massive parallelization boosts big Bayesian multidimensional scaling. *Journal of Computational and Graphical Statistics 30*(1), 11–24.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science 288*(5475), 2349–2350.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*(8), 754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science 294*(5550), 2310–2314.

Hurst, L. D. and A. R. Merchant (2001). High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London B: Biological Sciences 268*(1466), 493–497.

Izquierdo-Carrasco, F., N. Alachiotis, S. Berger, T. Flouri, S. P. Pissis, and A. Stamatakis (2013). A generic vectorization scheme and a GPU kernel for the phylogenetic likelihood library. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*, pp. 530–538. IEEE.

Jauch, M., P. D. Hoff, and D. B. Dunson (2021). Monte Carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics 30*(3), 622–631.

Jeschke, J. and H. Kokko (2009). The roles of body size and phylogeny in fast and slow life histories. *Evolutionary Ecology 23*(6), 867–878.

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

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis 97*(10), 2177–2189.

Joe, H. and D. Kurowicka (2011). *Dependence modeling: vine copula handbook.* World Scientific.

Jones, K. E., J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, C. Connolly, M. J. Cuttis, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. Bininda-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology 90*(9), 2648.

Junker, D., M. Becker, T. R. Wagner, P. D. Kaiser, S. Maier, T. M. Grimm, J. Griesbaum, P. Marsall, J. Gruber, B. Traenkle, C. Heinzel, Y. T. Pinilla, J. Held, R. Fendel, A. Kreidenweiss, A. Nelde, Y. Maringer, S. Schroeder, J. S. Walz, K. Althaus, G. Uzun, M. Mikus, T. Bakchoul, K. Schenke-Layland, S. Bunk, H. Haeberle, S. Göpel, M. Bitzer,

H. Renk, J. Remppis, C. Engel, A. R. Franz, M. Harries, B. Kessel, B. Lange, M. Strengert, G. Krause, A. Zeck, U. Rothbauer, A. Dulovic, and N. Schneiderhan-Marra (2022, 06). Antibody binding and angiotensin-converting enzyme 2 binding inhibition is significantly reduced for both the BA.1 and BA.2 Omicron variants. *Clinical Infectious Diseases*. ciac498.

Kalkauskas, A., U. Perron, Y. Sun, N. Goldman, G. Baele, S. Guindon, and N. De Maio (2021). Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Computational Biology 17*(1), e1008561.

Karcher, M. D., J. A. Palacios, T. Bedford, M. A. Suchard, and V. N. Minin (2016). Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Computational Biology 12*(3), e1004789.

Kay, K. M. and R. D. Sargent (2009). The role of animal pollination in plant speciation: integrating ecology, geography, and genetics. *Annual Review of Ecology, Evolution, and Systematics 40*, 637–656.

Ki, C. and J. Terhorst (2022). Variational phylodynamic inference using pandemic-scale data. *bioRxiv*.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications 13*(3), 235–248.

Kschischang, F. R., B. J. Frey, and H.-A. Loeliger (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory 47*(2), 498–519.

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Lakner, C., P. Van Der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist (2008). Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology 57*(1), 86–103.

Lanfear, R., X. Hua, and D. L. Warren (2016). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biology and Evolution 8*(8), 2319–2332.

Larget, B. and D. L. Simon (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution 16*(6), 750–759.

Larson, G. and D. Q. Fuller (2014). The evolution of animal domestication. *Annual Review of Ecology, Evolution, and Systematics 45*, 115–136.

Latinne, A., B. Hu, K. J. Olival, G. Zhu, L. Zhang, H. Li, A. A. Chmura, H. E. Field, C. Zambrana-Torrelio, J. H. Epstein, et al. (2020). Origin and cross-species transmission of bat coronaviruses in China. *Nature Communications 11*(1), 1–15.

Lee, S.-Y. and X.-Y. Song (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences.* John Wiley & Sons.

Lemey, P., S. L. Hong, V. Hill, G. Baele, C. Poletto, V. Colizza, Á. O'toole, J. T. McCrone, K. G. Andersen, M. Worobey, et al. (2020). Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nature Communications 11*(1), 1–14.

Lemey, P., A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathogens 10*(2), e1003932.

Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology 5*(9), e1000520.

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

Lemey, P., N. Ruktanonchai, S. L. Hong, V. Colizza, C. Poletto, F. Van den Broeck, M. S. Gill, X. Ji, A. Levasseur, B. B. Oude Munnink, et al. (2021). Untangling introductions and persistence in COVID-19 resurgence in europe. *Nature 595*(7869), 713–717.

Lemey, P., M. Salemi, and A.-M. Vandamme (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Cambridge University Press.

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

Li, S., D. K. Pearl, and H. Doss (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association 95*(450), 493–508.

Little, R. and D. Rubin (1987). *Statistical Analysis With Missing Data.* Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Liu, D., W. Shi, Y. Shi, D. Wang, H. Xiao, W. Li, Y. Bi, Y. Wu, X. Li, J. Yan, et al. (2013). Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. *The Lancet 381*(9881), 1926–1932.

Liu, J. S., W. H. Wong, and A. Kong (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 157–169.

Lopes, H. F. and M. West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica 14*, 41–67.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research 32*(4), 1363–1371.

Lupala, C. S., Y. Ye, H. Chen, X.-D. Su, and H. Liu (2022). Mutations on RBD of SARS-CoV-2 Omicron variant result in stronger binding to human ACE2 receptor. *Biochemical and Biophysical Research Communications 590*, 34–41.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution 45*(5), 1065–1080.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences 104*(suppl 1), 8597–8604.

MacPherson, A., S. Louca, A. McLaughlin, J. B. Joy, and M. W. Pennell (2022). Unifying phylogenetic birth–death models in epidemiology and macroevolution. *Systematic Biology 71*(1), 172–189.

Magee, A. F., M. D. Karcher, F. A. Matsen IV, and V. N. Minin (2021). How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error. *arXiv preprint arXiv:2109.07629*.

Mau, B., M. A. Newton, and B. Larget (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics 55*(1), 1–12.

May, M. R., S. Höhna, and B. R. Moore (2016). A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods in Ecology and Evolution 7*(8), 947–959.

Meyer, X. (2021). Adaptive tree proposals for Bayesian phylogenetic inference. *Systematic Biology 70*(5), 1015–1032.

Minin, V. N., E. W. Bloomquist, and M. A. Suchard (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution 25*(7), 1459–1471.

Minin, V. N. and M. A. Suchard (2008a). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology 56*(3), 391–412.

Minin, V. N. and M. A. Suchard (2008b). Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences 363*(1512), 3985–3995.

Mitov, V., K. Bartoszek, G. Asimomitis, and T. Stadler (2020). Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology 131*, 66–78.

Mitov, V., K. Bartoszek, and T. Stadler (2019). Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models. *Proceedings of the National Academy of Sciences 116*(34), 16921–16926.

Mitov, V. and T. Stadler (2018). A practical guide to estimating the heritability of pathogen traits. *Molecular Biology and Evolution 35*(3), 756–772.

Mohasel Afshar, H. and J. Domke (2015). Reflection, refraction, and Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems 28.*

Moretti, A. K., L. Zhang, C. A. Naesseth, H. Venner, D. Blei, and I. Pe'er (2021). Variational combinatorial sequential Monte Carlo methods for Bayesian phylogenetic inference. *arXiv preprint arXiv:2106.00075*.

Müller, N. F., D. A. Rasmussen, and T. Stadler (2017). The structured coalescent and its approximations. *Molecular Biology and Evolution 34*(11), 2970–2981.

Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, and G. Bernadi (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS letters 573*(1-3), 73–77.

Nakagawa, S. and R. P. Freckleton (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution 23*(11), 592–596.

Nascimento, F. F., M. d. Reis, and Z. Yang (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution 1*(10), 1446–1454.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press.

Nee, S., R. M. May, and P. H. Harvey (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344*(1309), 305–311.

Nishimura, A., D. B. Dunson, and J. Lu (2020, 03). Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. *Biometrika 107*(2), 365–380.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology 29*(1), 59–75.

Okada, K. and S.-i. Mayekawa (2018). Post-processing of Markov chain Monte Carlo output in Bayesian latent variable models with application to multidimensional scaling. *Computational Statistics 33*(3), 1457–1473.

Orgel, L. E. and F. H. Crick (1980). Selfish DNA: the ultimate parasite. *Nature 284*(5757), 604.

Pagel, M. (1999, October). Inferring the historical patterns of biological evolution. *Nature 401*, 877–884.

Pakman, A. and L. Paninski (2013). Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. *Advances in Neural Information Processing Systems 26.*

Papastamoulis, P. and I. Ntzoufras (2022). On the identifiability of Bayesian factor analytic models. *Statistics and Computing 32*(2), 1–29.

Payne, R., M. Muenchhoff, J. Mann, H. E. Roberts, P. Matthews, E. Adland, A. Hempenstall, K.-H. Huang, M. Brockman, Z. Brumme, et al. (2014). Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proceedings of the National Academy of Sciences 111*(50), E5393–E5400.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 133–136. Association for the Advancement of Artificial Intelligence.

Peck, K. M., A. S. Lauring, and C. S. Sullivan (2018). Complexities of viral mutation rates. *Journal of Virology 92*(14), e01031–17.

Peters, E. A. J. F. and G. de With (2012, Feb). Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E 85*, 026703.

Pianka, E. R. (1970). On r-and K-selection. *The American Naturalist 104*(940), 592–597.

Planas, D., N. Saunders, P. Maes, F. Guivel-Benhassine, C. Planchais, J. Buchrieser, W.-H. Bolland, F. Porrot, I. Staropoli, F. Lemoine, et al. (2022). Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature 602*(7898), 671–675.

Plantier, J.-C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, V. Lemée, F. Damond, D. L. Robertson, and F. Simon (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine 15*(8), 871–872.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News 6*(1), 7–11.

Pourzanjani, A. A., R. M. Jiang, B. Mitchell, P. J. Atzberger, and L. R. Petzold (2021). Bayesian inference over the Stiefel manifold via the Givens representation. *Bayesian Analysis 16*(2), 639–666.

Press, S. J. and K. Shigemasu (1989). Bayesian inference in factor analysis. In *Contributions to Probability and Statistics*, pp. 271–287. Springer.

Pruesse, E., J. Peplies, and F. O. Glöckner (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics 28*(14), 1823–1829.

Pybus, O. G., M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences 109*(37), 15066–15071.

Rambaut, A. and A. Drummond (2015). TreeAnnotator v1. 8.2. *MCMC Output Analysis*.

Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology 67*(5), 901.

Rambaut, A., T. T. Lam, L. Max Carvalho, and O. G. Pybus (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution 2*(1), vew007.

Rannala, B. and Z. Yang (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution 43*(3), 304–311.

Réale, D., D. Garant, M. M. Humphries, P. Bergeron, V. Careau, and P.-O. Montiglio (2010). Personality and the emergence of the pace-of-life syndrome concept at the population level. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1560), 4051–4063.

Revell, L. J. (2009). Size-correction and principal components for interspecific comparative studies. *Evolution 63*(12), 3258–3268.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution 3*(2), 217–223.

Revell, L. J. and L. J. Harmon (2008). Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research 10*(3), 311–331.

Reynolds, J. (2003). Life histories and extinction risk. In T. Blackburn and K. Gaston (Eds.), *Macroecology: Concepts and Consequences*, pp. 195–217. Oxford: Blackwell Publishing Ltd.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rodríguez, C. E. and S. G. Walker (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics 23*(1), 25–45.

Roff, D. A. (2002). *Life History Evolution.* Sunderland, Massachusetts. Sinauer Associates.

Rohlf, F. J. (2001). Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution 55*(11), 2143–2160.

Ronquist, F. (2004). Bayesian inference of character evolution. *Trends in Ecology & Evolution 19*(9), 475–481.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Sabath, N., E. Ferrada, A. Barve, and A. Wagner (2013). Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biology and Evolution 5*(5), 966–977.

Sæther, B.-E. and Ø. Bakke (2000). Avian life history variation and contribution of demographic traits to the population growth rate. *Ecology 81*(3), 642–653.

Salguro-Gómez, R. (2017). Applications of the fast-slow continuum and reproductive strategy framework of plant life histories. *New Phytologist 213*(3), 1618–1624.

Sánchez-Baracaldo, P., J. A. Raven, D. Pisani, and A. H. Knoll (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences 114*(37), E7737–E7745.

Sanmartín, I., P. Van Der Mark, and F. Ronquist (2008). Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the canary islands. *Journal of Biogeography 35*(3), 428–449.

Sansalone, G., K. Allen, J. Ledogar, S. Ledogar, D. Mitchell, A. Profico, S. Castiglione, M. Melchionna, C. Serio, A. Mondanaro, et al. (2020). Variation in the strength of allometry drives rates of evolution in primate brain shape. *Proceedings of the Royal Society B 287*(1930), 20200807.

Shang, J., Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, and F. Li (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences 117*(21), 11727–11734.

Shapiro, A. (1985). Identifiability of factor analysis: Some results and open problems. *Linear Algebra and its Applications 70*, 1–7.

Shuter, B. J., J. Thomas, W. D. Taylor, and A. M. Zimmerman (1983). Phenotypic correlates of genomic dna content in unicellular eukaryotes and other cells. *The American Naturalist 122*(1), 26–44.

Siepel, A., K. S. Pollard, and D. Haussler (2006). New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pp. 190–205. Springer.

Sinsheimer, J. S., J. A. Lake, and R. J. Little (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, 193–210.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology 267*(3), 396–404.

Stadler, T. (2011). Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences 108*(15), 6187–6192.

Stadler, T., R. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, et al. (2012). Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution 29*(1), 347–357.

Stadler, T., D. Kühnert, S. Bonhoeffer, and A. J. Drummond (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences 110*(1), 228–233.

Starr, T. N., A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, and J. D. Bloom (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell 182*(5), 1295–1310.e20.

Stearns, S. C. (1989). Trade-offs in life-history evolution. *Functional Ecology 3*(3), 259–268.

Stewart, G. (1998). On the adjugate matrix. *Linear Algebra and its Applications 283*(1), 151–164.

Strimmer, K. and O. G. Pybus (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution 18*(12), 2298–2305.

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution 4*(1), vey016.

Suchard, M. A. and A. Rambaut (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics 25*(11), 1370–1376.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution 18*(6), 1001–1013.

Sullivan, J. and P. Joyce (2005). Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics 36*(1), 445–466.

Sumner, J. G., J. Fernández-Sánchez, and P. D. Jarvis (2012). Lie Markov models. *Journal of Theoretical Biology 298*, 16–31.

Swiss HIV Cohort Study, F. Schoeni-Affolter, B. Ledergerber, M. Rickenbach, C. Rudin, H. F. Günthard, A. Telenti, H. Furrer, S. Yerly, and P. Francioli (2009). Cohort profile: the swiss hiv cohort study. *International Journal of Epidemiology 39*(5), 1179–1189.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences 17*(2), 57–86.

Teh, Y., H. Daume III, and D. M. Roy (2007). Bayesian agglomerative clustering with coalescents. In J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems*, Volume 20. Curran Associates, Inc.

Terra-Araujo, M. H., A. D. de Faria, A. Vicentini, S. Nylinder, and U. Swenson (2015). Species tree phylogeny and biogeography of the neotropical genus Pradosia (Sapotaceae, Chrysophylloideae). *Molecular Phylogenetics and Evolution 87*, 1–13.

Thompson, E. A., K. Thompson, E. Thompson, et al. (1975). *Human evolutionary trees.* CUP Archive.

Thorne, J. L., H. Kishino, and I. S. Painter (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution 15*(12), 1647–1657.

To, T.-H., M. Jung, S. Lycett, and O. Gascuel (2016, Jan). Fast dating using least-squares criteria and algorithms. *Systematic Biology 65*(1), 82–97.

Tolkoff, M. R., M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard (2018). Phylogenetic factor analysis. *Systematic Biology 67*(3), 384–399.

Uhlenbeck, G. E. and L. S. Ornstein (1930). On the theory of the Brownian motion. *Physical Review 36*(5), 823.

Uyeda, J. C. and L. J. Harmon (2014). A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology 63*(6), 902–918.

Van der Niet, T. and S. D. Johnson (2012). Phylogenetic evidence for pollinator-driven diversification of angiosperms. *Trends in ecology & evolution 27*(6), 353–361.

Vaughan, T. G., D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond (2014, 04). Efficient Bayesian inference under the structured coalescent. *Bioinformatics 30*(16), 2272–2279.

Visscher, P. M., W. G. Hill, and N. R. Wray (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics 9*(4), 255.

Volz, E. M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics 190*(1), 187–201.

Vrancken, B., P. Lemey, A. Rambaut, T. Bedford, B. Longdon, H. F. Günthard, and M. A. Suchard (2015). Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods in Ecology and Evolution 6*, 67–82.

Wang, H.-C., E. Susko, and A. J. Roger (2006). On the correlation between genomic G + C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochemical and Biophysical Research Communications 342*(3), 681–684.

Wang, L., A. Bouchard-Côté, and A. Doucet (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *Journal of the American Statistical Association 110*(512), 1362–1374.

Wang, L., S. Wang, and A. Bouchard-Côté (2020). An annealed sequential Monte Carlo method for Bayesian phylogenetics. *Systematic Biology 69*(1), 155–183.

Ward, M., C. Gibbons, P. McAdam, B. Van Bunnik, E. Girvan, G. Edwards, J. R. Fitzgerald, and M. Woolhouse (2014). Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Applied and Environmental Microbiology 80*(23), 7275–7282.

Whidden, C. and F. A. Matsen IV (2015). Quantifying MCMC exploration of phylogenetic tree space. *Systematic Biology 64*(3), 472–491.

Whittall, J. B. and S. A. Hodges (2007). Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nature 447*(7145), 706–709.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wiedmann, M., R. Primicerio, A. Dolgov, C. Ottensen, and M. Aschan (2014). Life history variation in Barents Sea fish: Implications for sensitivity to fishing in a changing environment. *Ecology and Evolution 4*(18), 3596–3611.

Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics 3*, 163–195.

Worobey, M., T. D. Watts, R. A. McKay, M. A. Suchard, T. Granade, D. E. Teuwen, B. A. Koblin, W. Heneine, P. Lemey, and H. W. Jaffe (2016). 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature 539*(7627), 98–101.

Wright, J. K., V. L. Naidoo, Z. L. Brumme, J. L. Prince, D. T. Claiborne, P. J. Goulder, M. A. Brockman, E. Hunter, and T. Ndung'u (2012). Impact of HLA-B* 81-associated mutations in HIV-1 Gag on viral replication capacity. *Journal of Virology 86*(6), 3193–3199.

Wu, H., Z. Zhang, S. Hu, and J. Yu (2012). On the molecular mechanism of GC content variation among eubacterial genomes. *Biology Direct 7*(1), 2.

Wu, L., L. Zhou, M. Mo, T. Liu, C. Wu, C. Gong, K. Lu, L. Gong, W. Zhu, and Z. Xu (2022). SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2. *Signal Transduction and Targeted Therapy 7*(1), 1–3.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution 39*(3), 306–314.

Yang, Z. and B. Rannala (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution 14*(7), 717–724.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution 8*(1), 28–36.

Zhang, C. (2020). Improved variational Bayesian phylogenetic inference with normalizing flows. *Advances in Neural Information Processing Systems 33*, 18760–18771.

Zhang, C., J. P. Huelsenbeck, and F. Ronquist (2020). Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. *Systematic Biology 69*(5), 1016–1032.

Zhang, C. and F. A. Matsen IV (2018a). Generalizing tree probability estimation via Bayesian networks. *Advances in Neural Information Processing Systems 31*.

Zhang, C. and F. A. Matsen IV (2018b). Variational Bayesian phylogenetic inference. In *International Conference on Learning Representations*.

Zhang, Z., A. Nishimura, P. Bastide, X. Ji, R. P. Payne, P. Goulder, P. Lemey, and M. A. Suchard (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.

Zhang, Z., A. Nishimura, X. Ji, P. Lemey, and M. A. Suchard (2022). Hamiltonian zigzag speeds up large-scale learning of direct effects among mixed-type biological traits. *arXiv preprint arXiv:2201.07291*.

Zhao, T., Z. Wang, A. Cumberworth, J. Gsponer, N. de Freitas, and A. Bouchard-Côté (2016). Bayesian analysis of continuous time Markov chains with application to phylogenetic modeling. *Bayesian Analysis 11*(4), 1203–1237.

Zuckerkandl, E. and L. Pauling (1962). *Molecular Disease, Evolution and Genetic Heterogeneity*. Academic Press.