UNIVERSITY OF CALIFORNIA

Los Angeles

Phylogenetic Factor Analysis and Natural Extensions

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Max Ryan Tolkoff

2018

ABSTRACT OF THE DISSERTATION

Phylogenetic Factor Analysis and Natural Extensions

by

Max Ryan Tolkoff

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2018

Professor Marc Adam Suchard, Chair

Frequently in evolutionary biology we are interested in how different quantitative traits of an organism evolve together over time. In order to properly understand these relationships, we need to adjust for the shared evolutionary history of these organisms. Previous methods rely on modeling quantitative traits as undergoing a high dimensional, correlated multivariate Brownian diffusion (MBD) down a phylogenetic tree. In order to present a more nuanced approach to understanding these trait relationships, we develop a phylogenetic factor analysis (PFA) model on these quantitative traits by assuming that the relatively low dimensional factors, rather than the traits themselves, undergo independent Brownian diffusion down a phylogenetic tree. Additionally, we develop a novel method for inferring the marginal likelihood estimates of probit models which allows for accurate model selection in the presence of discrete data. We demonstrate using Bayes factors that this PFA model is a more probable model than the MBD model. We then continue to develop this PFA method by relying on a shrinkage prior on the loadings matrix. This shrinkage prior consists of a normal prior with a global and local standard deviation component, and a half cauchy prior on these standard deviation

components. With this we can distinguish trait relationships which would otherwise remain hidden using a standard normal prior on the loadings. Lastly, when we wish to incorporate a large number of taxa in our MBD and PFA models, obtaining a complete suite of measurements is difficult. These missing measurements make these analyses relatively inefficient and difficult to use for larger problems. To rectify this, we develop a method by which we can evaluate the likelihood of an MBD model by analytically integrating out missing values, and then apply similar principles to integrate out the factors in a PFA model. These innovations allow for massive speedup in our inference.

The dissertation of Max Ryan Tolkoff is approved.

Michael Edward Alfaro

Donatello Telesca

Janet S Sinsheimer

Marc Adam Suchard, Committee Chair

University of California, Los Angeles

2018

To my family, whose support for me has never wavered throughout this process.

Table of Contents

List of Figures

List of Tables

Acknowledgments

It is impossible to go through a process like this without the help of those around you. First, I would like to thank Lauren Harrell for showing me the ropes when I first arrived. I would also like to thank Hillary Aralis, who was a like a sister to me from the moment I came to Los Angeles, and Daniel Conn who has been unfailingly eager to talk statistics with me. I would like to thank Will Harris, Bonnie Chan, and Martin Flores for helping me through the first few years of classes, and for being good friends ever since, even if distance has kept us apart. I would like to thank Sibon Li for helping me get started on my research, and to Gabriella Cybis for inspiring my future projects.

Throughout the bulk of my time doing research, Trevor Shaddox and Mandev Gill have been there as bulwarks of support. Trevor has always been there to support me when times got tough, and made sure I stayed fit in the process, helping me build habits that support me to this day. Mandev has always been there to keep me grounded, even when he is trolling me. Ho Si Tung Lam has helped me maintain my focus, and provided substantial contributions to Chapter 5 in particular. Yuxi Tian always knows what's important and has helped me keep my perspective on the world and in research.

I would like to thank Zhenyu Zhang who has been excited to take on my research after I'm gone. In particular I'd like to thank Akihiko Nishimura for his "Deus ex 'Aki'-na," after he descended from the heavens to help me resolve a seemingly impassible road block in Chapter 4, clearing the way for my graduation.

I would like to thank the Bruin Alliance of Skeptics and Secularists (BASS) for the three wonderful years of leadership they bestowed upon me, and the six and a half years of membership which taught me how to be wrong gracefully, so that I can get things right in the future. I would like to thank Sunday Assembly

Los Angeles for providing religion to the irreligious and for supporting BASS for the foreseeable future. And, of course, I would like to thank my girlfriend, Jenny Chieu, who has stuck with me through this long process, and who has graciously cut vegetables for me to stave off my dying from malnutrition.

I would like to thank Ron Brookmeyer for being my first academic advisor at UCLA. I would also like to thank my committee members. Janet Sinsheimer has always been around for support and has helped me navigate some of the ins and outs of a Ph.D program. Donatello Telesca has made considerable contributions to Chapter 4 and has been invaluable for discussions regarding model selection. Michael Alfaro has graciously allowed me to sit in on his lab meetings where I can learn the details of the science of evolutionary biology which would be hard to come by sitting in front of a computer screen just looking a statistical models. Marc Suchard has been a mentor to me all of these years and has patiently bestowed me all of the knowledge and research experience humanly possible in my time in his lab.

Lastly, I would like to thank the funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 and the National Institutes of Health (R01 AI107034, R01 AI117011 and R01 HG006139) and the National Science Foundation (DMS 1264153). Without the support of the grants from institutions like these, scientific progress would be impossible.

Chapter 5 is a version of MR Tolkoff, ME Alfaro, G Baele, P Lemey and MA Suchard. *Phylogenetic factor analysis*. Systematic Biology, doi:10.1093/sysbio/syx066.

Vita

2011–2018	PhD, Biostatistics (Expected) Fielding School of Public Health at UCLA.
2008–2011	B.S., (Math & Physics) Tufts University.
2017-2018	UCLA Dissertation Year Fellowship
May 2015	Systems & Integrative Biology Training Grant, Genomic Analysis Training Program, Burroughs Wellcome Fund Symposium Poster Competition Winner
March 2015	Carol Newton Memorial Symposium Poster Competition Winner
2013-2015	Systems & Integrative Biology Training Grant
2011	N. Hobbs Knight Prize Scholarship in Physics

Publications and Presentations

2017, MR Tolkoff, ME Alfaro, G Baele, P Lemey and MA Suchard. *Phylogenetic factor analysis*. Systematic Biology, doi:10.1093/sysbio/syx066

2012, G Widmer, Y Lee, P Hunt, A Martinelli, MR Tolkoff, and K Bodi. *Comparative genome analysis of two cryptosporidium parvum isolates with different host range.* Infection, Genetics and Evolution, 12(6):1213 — 1221

2016, MR Tolkoff, MA Suchard. Sparse Phylogenetic Factor Analysis. Presented at Joint Statistical Meetings, Chicago, Illinois

2015, MR Tolkoff, MA. Suchard. Phylogenetic Factor Analysis. Presented at Joint Statistical Meetings, Seattle, Washington

2014, MR Tolkoff, MA Suchard. Measuring Correlation of Evolution Rates Across Multiple Loci. Presented at Joint Statistical Meetings, Boston, Massachusetts

CHAPTER 1

Introduction

In comparative biology, we are frequently concerned with the way different characteristics or traits of an organism relate to each other. The principle confounder in these relationships stems from the reality that closely related species tend to have similar characteristics. Therefore any inference on trait relationships must adjust for this evolutionary history [Felsenstein, 1985]. Felsenstein [1985] resolved this challenge in the bivariate case by treating each trait as undergoing correlated Brownian diffusion down a phylogenetic tree. There are many adaptations to this method including but not limited to [Huelsenbeck and Rannala, 2003, Lemey et al., 2010, Pybus et al., 2012, Cybis et al., 2015, Vrancken et al., 2015, Adams, 2014, Revell, 2009]. Many of the methods described in those papers, as well as this thesis, rely heavily on Bayesian phylogenetic models and Bayesian phylogenetic inference. Therefore, we begin this thesis in Chapter 2 with a review of phylogenetics models, specifically how to construct likelihoods from evolutionary histories and sequence data. We then continue on to discuss Bayesian methods, and methods of Bayesian inference via Markov chain Monte Carlo, specifically the Metropolis-Hastings and Gibbs sampling algorithms.

Many existing phylogenetic trait methods, such as phylogenetic principle components analysis (pPCA) from Revell [2009] scale cubicly with the number of taxa and/or traits and are therefore difficult to use in larger examples. Others, such as phylogenetic least squares (PGLS) [Adams, 2014], allow for statistical testing of overall associations, but fall short when trying to evaluate

pairwise relationships. The Bayesian multivariate Brownian diffusion (MBD) [Lemey et al., 2010] infers a correlation matrix in the context of a Brownian diffusion, and does not have these same scaling issues, scaling linearly with the number of taxa. In spite of this, there are numerous areas where the MBD can be improved. For example, the MBD model hinges on the assumption that there is only a single process for all traits which is constant throughout the tree. Relatedly, the MBD method is unable to account for naturally occurring variance independent of the evolutionary process.

In order to tackle these issues, in Chapter 5, we develop the phylogenetic factor analysis (PFA) method. This method is a factor analysis method which decomposes the trait measurements across taxa into a loadings matrix, a factor matrix, and an error term independent of evolution. We then make the assumption that these independent factors undergo Brownian diffusion down a phylogenetic tree. This allows us to describe different independent evolutionary processes for a suite of traits through the loadings matrix. Because this method is Bayesian, we can take advantage both of simultaneous inference on the phylogeny using Suchard et al. [2001] as well as adapt the probit model described by Cybis et al. [2015] to handle discrete data. In order to evaluate the number of independent evolutionary processes, we develop a novel [Heaps et al., 2014] method for finding the exact marginal likelihood estimate (MLE) of a probit model using path sampling, by creating a threshold which softens over the course of the path. This allows us to rely on the path sampling method described by Gelman and Meng [1998] and implemented by Baele et al. [2013a,b] to learn about the MLE of these models and use these estimates to estimate the relative Bayes factors to determine the optimal number of independent processes. We apply these methods to the examples of reproduction of *Aquilegia* flowers, reproduction of fish of the family *Poeciliidae* in the presence of missing data, as well as fin morphometry of fish of the family

Balistidae including ancestral state reconstructions at any point along the tree. We find, using Bayes factors that the PFA fits the data better than the MBD method in all of these examples, with modest computational advantages as well.

In Chapter 4, we are interested in the case where the loadings matrix has a large number of cells whose posterior mass is concentrated around 0. This situation often occurs when the number of traits is comparable to or much larger than the number of taxa. In order to facilitate identifiability, we impose shrinkage priors, also known as horseshoe priors [Gelman, 2006, Carvalho et al., 2009] on our loadings matrix which force the marginal posterior of those loadings cells whose marginal posterior is centered around 0, closer to 0 with a tighter variance. The shrinkage prior replaces the independent normal prior on the loadings used in Chapter 5 with independent normal priors with a shared standard deviation component and a cell specific standard deviation component. To complete the shrinkage prior, we place a half Cauchy distribution on these standard deviation components. I apply this new prior scheme to the examples of *Aquilegia* flower reproduction and find a more parsimonious result than for the dense model. We also look at the diversification of *Anoles* lizards and the morphometry of *Plethodon* salamanders and find results using shrinkage priors, which we are unable to recover using the dense model.

Finally, in Chapter 5 we look at the situation where the number of taxa is large. Most methods such as pPCA [Revell, 2009] scale cubically with the number of taxa. PFA, by contrast, scales quadratically which is unfortunately still too slow for practical use when the number of taxa is large. A concurrent problem with increasing the number of taxa analyzed arises from the difficulty in obtaining a full suite of measurements in these circumstances. Other competitors such as PGLS [Adams, 2014] are difficult to adapt in these situations with many missing measurements. By relying on the MBD model from Lemey

et al. [2010] which scales linearly with the number of taxa, we can integrate out the missing values analytically by assuming a missing value has infinite variance, and a measured value has no variance. Additionally, this structure allows us to integrate out latent variables, such as those used to describe the factors in the PFA model, allowing us to draw inference on the PFA in time linear to the number of factors. Both of these methods were tested using measurements of the reproduction of mammals with the hope of being able to determine whether or not we can categorize mammalian reproduction into so called fast cycles with many offspring that grow quickly and slow cycles with few offspring that grow slowly. We find massive speedup for both the integrated MBD and PFA methods.

Lastly, in Chapter 6 we provide a list of methods which we would like to see implemented in the future. This includes a structural Equation model designed to allow us to separate traits into those which are affected by elements inside the model and those which are not. Additionally, to improve sampling we would like to explore methods such as Hamiltonian Monte Carlo [Neal, 2010] to improve performance, and a bouncy particle sampler [Bouchard-Côte and Vollmer, 2017] to sample the probit model in situations with a mix of discrete and continuous measurements with a relatively large percentage of discrete traits. Lastly, we would like to explore how to incorporate repeated measures at the tips of the tree into my integrated analysis laid out in Chapter 5.

CHAPTER 2

Review

2.1 Phylogenetics

2.1.1 Representation of Evolutionary History Via a Phylogenetic Tree

We begin this review by discussing how to formally define the evolutionary history of a series of organisms. We define a phylogeny $\mathscr{F}$ on N organisms or taxa in two parts.

The first part, known as a topology τ, is a bifurcating directed acyclic graph with a single point of origin. Each vertex in the graph is referred to as a node. Those nodes with degree 1, i.e. at the end of the graph, represented by $(\nu_1, \ldots, \nu_N)$ are known as external nodes or tips and generally represent taxa for which we have measurements pertinent to whatever analysis we may be interested in performing. Additional nodes $(\nu_{N+1}, \ldots, \nu_{2N-2})$ are of degree 3 and are known as internal nodes. Each internal node represents the most recent common ancestor between the two species immediately further down on the graph. A phylogenetic tree may or may not also have an additional node of degree 2, represented by ν_{2N-1}. Such a node, if it exists is known as the root and represents the most recent common ancestor of all species in the tree. Trees with a root node are known as rooted trees. For the analyses in this thesis, all trees are rooted.

The second part of $\mathscr{F}$ is a series of edge weights $\mathbf{B} = (B_1, \ldots, B_{2N-2})$, which represent branch times. These edge weights are a function of both measured

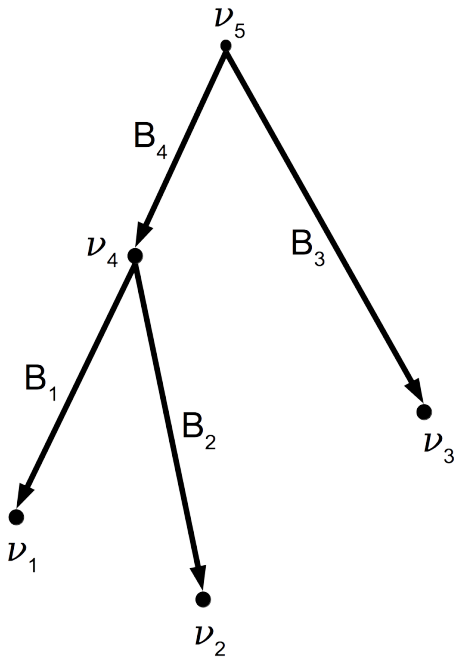

Figure 2.1: *An example of a rooted Phylogenetic tree with 3 tips at different heights.* ν_1, ν_2, *and* ν_3 *are external nodes,* ν_4 *is an internal node and* ν_5 *is the root node.* B_i *refers to the rate adjusted branch time between node i and its parent.*

time as well as a reflection of how different evolutionary processes may speed up or slow down over the course of history. Note that the distance from the root to tip of the tree need not be the same for each tip, especially in the case where the measurements at each tip were taken at significantly different time points. Figure 2.1 shows an example of a rooted phylogenetic tree where $N = 3$ and where the tips are of different heights.

2.1.2 Likelihood of a Phylogenetic Tree

Suchard et al. [2001] devises a method which uses Bayesian analysis to reconstruct not only evolutionary history, but also the parameters related to rates of evolution and state transitions. In order to do this, we need to discuss how to efficiently compute the likelihood of a phylogeny with molecular sequence data at the tips. The measured sequences, {A,G,C,T/U} in the case of DNA

or RNA respectively, from the N taxa are aligned, with a "-" representing the spaces which can occur across evolution because of an insertion or deletion event. These aligned sequences can be arranged into an $N \times S$ matrix $\mathbf{S}$, where S is the sequence length.

In order to evaluate the likelihood of a tree using sequence data, we must have a generating model for the sequence data given the tree. One model, popularized by Felsenstein [1981] relies on a continuous time Markov chain (CTMC), where we define an infinitesimal transition rate matrix Q such that the probability of a transition between two characters at a branch time B is

$$p(B) \propto \exp(BQ). \tag{2.1}$$

In general we assume that this chain has achieved a steady state. Because of the Markovian nature of this process, if we define $v_{i\ell}, v_{j\ell}$ and $v_{k\ell}$ as the state of nodes i, j and k respectively at sequence element ℓ, then the $p(v_{j\ell} \mid v_{i\ell}, B_j)$ is independent of $p(v_{k\ell} \mid v_{i\ell}, B_k)$, if node i is the parent of the nodes j and k, and where B_j and B_k are the rate adjusted branch times for nodes j and k respectively.

We can use Equation 2.1 to compute the probability of $p(v_{i\ell} \mid v_{j\ell}, v_{k\ell}, B_j, B_k)$. By using Bayes rule we see that

$$p(v_{i\ell} \mid v_{j\ell}, v_{k\ell}, B_j, B_k) \propto p(v_{j\ell} \mid v_{i\ell}, B_j)\, p(v_{k\ell} \mid v_{i\ell}, B_k). \tag{2.2}$$

Since the internal states are unknown, we average over the potential states of the internal nodes. Therefore, as first described by Felsenstein [1973], the

likelihood of a phylogeny for sequence element ℓ can be defined as

$$p\left(\mathscr{F}_\ell\right) = \sum_{\nu_{2N-1}}^{4} \ldots \sum_{\nu_{N+1}}^{4} p\left(\nu_{(N+1)\ell} \mid \nu_{1\ell}, \nu_{2\ell}, B_1, B_2\right) \ldots \\ p\left(\nu_{(N-1)\ell} \mid \nu_{(2N-2)\ell}, \nu_{(2N-3)\ell}, B_{(2N-2)}, B_{(2N-3)}\right), \tag{2.3}$$

where node $N+1$ has children 1 and 2 and node $N-1$ has children $N-2$ and $N-3$. A naive evaluation of this likelihood would require 4^{N-1} evaluations, however Felsenstein [1981] discovered that a post order traversal where we compute the probability of each sequence possibility at each parent node, and treated those probabilities as the state of the sequence at that node could allow us to compute the value of this likelihood in $\mathcal{O}\left(NS\right)$ time and $\mathcal{O}\left(N\right)$ time if these independent sequence calculations are done in parallel through a program such as BEAGLE [Browning and Browning, 2016].

It is also worth noting that in order to draw inference on Q we need to make some simplifying assumptions. One model, by Jukes and Cantor [1969] assumes that all transition probabilities are equally likely at all times. Hasegawa et al. [1985] describes a model where the transition probabilities are dependent on the resultant state, with a special penalty for transversions. Other models are also available, however in general for this thesis when tree reconstruction is necessary we will rely on Hasegawa et al. [1985]. By placing priors on the elements of this model and using this likelihood computation, we can learn about these parameters using standard Bayesian inference tools such as Markov chain Monte Carlo (MCMC).

2.2 Bayesian Analysis

In general, this thesis will rely on Bayesian analysis, and therefore we will review the fundamental aspects, definitions and evaluation methods of Bayesian

statistics. We define $\boldsymbol{\Theta}$ as the list of parameters we are interested in learning about, and $\mathbf{Y}$ as a list of observations. We wish to evaluate $p(\boldsymbol{\Theta} \mid \mathbf{Y})$, also known as the posterior distribution. In order to do this, we must apply Bayes theorem, such that

$$p(\boldsymbol{\Theta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{Y})}. \tag{2.4}$$

The expression $p(\mathbf{Y} \mid \boldsymbol{\Theta})$ is defined as the likelihood because of its analogue in frequentist statistics. The expression $p(\boldsymbol{\Theta})$ is known as the prior since it reflects our prior beliefs as to what distribution we believe the parameter takes. The expression $p(\mathbf{Y})$ reflects the probability of observing the data given the model. In order to evaluate $p(\mathbf{Y})$, we must evaluate $\int p(\mathbf{Y} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})\, d\boldsymbol{\Theta}$. This expression can often be intractable to evaluate analytically. Additionally, for a highly multidimensional posterior it may be difficult to learn about the marginal distribution of any individual parameter. Therefore, we must rely on approximations to learn about these marginal posterior distributions.

The standard method of approximation used in this context is called Markov chain Monte Carlo. In MCMC, we draw a series of samples from the posterior distribution of $\boldsymbol{\Theta}$ such that draw k, $\boldsymbol{\Theta}^{(k)}$, is dependent on the previous draw, and is independent of all other earlier draws given the previous draw. Two different forms of MCMC are described in the next sections.

2.2.1 Metropolis-Hastings

One MCMC method, first defined in Metropolis et al. [1953] is known as Metropolis-Hastings and relies on randomly selecting a proposal for $\boldsymbol{\Theta}^{(k+1)}$ based on a proposal density $q(\boldsymbol{\Theta}^{(k)})$. This proposal, $\boldsymbol{\Theta}^*$, is accepted with probability

$$min\left\{1, \frac{q(\boldsymbol{\Theta}^*)\, p(\boldsymbol{\Theta}^* \mid \mathbf{Y})}{q\left(\boldsymbol{\Theta}^{(k)}\right) p\left(\boldsymbol{\Theta}^{(k)} \mid \mathbf{Y}\right)}\right\}. \tag{2.5}$$

If we reject this proposal then $\mathbf{\Theta}^{(k)} = \mathbf{\Theta}^{k+1}$. The expression $q(\mathbf{\Theta}^*)/q(\mathbf{\Theta}^{(k)})$ is known as the Hastings ratio, since it was first proposed in Hastings [1970] and is included to compensate for asymmetry in the proposal distribution.

2.2.2 Gibbs Sampling

In some cases, with carefully chosen likelihoods and priors, the posterior distribution is known, and therefore we can sample from the posterior directly [Geman and Geman, 1984]. However, sometimes we only know the conditional posterior of $p\left(\mathbf{\Theta}_i \mid \mathbf{Y}, \mathbf{\Theta}_{(-i)}\right)$, where $\mathbf{\Theta}_i$ is parameter i, and $\mathbf{\Theta}_{(-i)}$ is the list of parameters without parameter i. For such situations, we may Gibbs sample for those parameters whose conditional posterior is known, and rely on the Metropolis-Hastings algorithm otherwise. This is known as Metropolis-within-Gibbs sampling [Liu et al., 1995], and is featured heavily throughout this thesis.

CHAPTER 3

Phylogenetic factor analysis

Phylogenetic comparative methods explore the relationships between quantitative traits adjusting for shared evolutionary history. This adjustment often occurs through a Brownian diffusion process along the branches of the phylogeny that generates model residuals or the traits themselves. For high-dimensional traits, inferring all pair-wise correlations within the multivariate diffusion is limiting. To circumvent this problem, we propose phylogenetic factor analysis (PFA) that assumes a small unknown number of independent evolutionary factors arise along the phylogeny and these factors generate clusters of dependent traits. Set in a Bayesian framework, PFA provides measures of uncertainty on the factor number and groupings, combines both continuous and discrete traits, integrates over missing measurements and incorporates phylogenetic uncertainty with the help of molecular sequences. We develop Gibbs samplers based on dynamic programming to estimate the PFA posterior distribution, over three-fold faster than for multivariate diffusion and a further order-of-magnitude more efficiently in the presence of latent traits. We further propose a novel marginal likelihood estimator for previously impractical models with discrete data and find that PFA also provides a better fit than multivariate diffusion in evolutionary questions in columbine flower development, placental reproduction transitions and triggerfish fin morphometry.

3.1 Introduction

Phylogenetic comparative methods revolve around uncovering relationships between different characteristics or traits of a set of organisms over the course of their evolution. One way to gain insight into these interactions is to analyze unadjusted correlations between traits across taxa. However, as insightfully noted by Felsenstein [1985], unadjusted analyses introduce the inherent challenge that any association uncovered may reflect the shared evolutionary history of the organisms being studied, and hence their similar traits values, rather than processes driving traits to co-vary over time. Thus, studies to identify co-varying evolutionary trait processes must simultaneously adjust for shared evolutionary history.

There have been many attempts to accomplish this goal. Felsenstein [1985] and Ives and Jr. [2010] are two such important examples, but they rely on a known evolutionary history described by a fixed phylogenetic tree and consider univariate evolutionary processes giving rise to only single traits. Felsenstein [1985] treats continuous traits as undergoing conditionally independent, Brownian diffusion down the branches of the phylogenetic tree and Ives and Jr. [2010] posit a regression model where the tree determines the error structure in the univariate outcome model. Huelsenbeck and Rannala [2003] adapt the Brownian diffusion description in a Bayesian framework with the goal of drawing simultaneous inference on both the tree from molecular sequence data as well as the correlations of interest related to a small number of traits through a multivariate Brownian diffusion process. Lemey et al. [2010] extend the multivariate process by relaxing the strict Brownian assumption along distinct branches in the tree using a scale mixture of normals representation. Cybis et al. [2015] jointly model molecular sequence data and multiple traits using a multivariate latent liability formulation to combine both continuous

and discrete observations and determine their correlation structure while adjusting for shared ancestry. This method is effective, but inference remains computationally expensive and estimates of the high-dimensional correlation matrix between traits only allows us to explain the evolution of these traits through a single process. Additional frequentist methods include, Revell [2009] who use a phylogenetically adjusted principal components analysis, Adams [2014] who use a phylogenetic least squares analysis, and Clavel et al. [2015] who also use a multivariate diffusion method. All of these methods, however require large matrix inversions which make them ill suited to adaptations to full Bayesian inference, or bootstrapping to provide measures of uncertainty.

One way to alleviate these problems lies with dimension reduction through exploratory factor analysis [Aguilar and West, 2000]. Factor analysis is the inferred decomposition of observed data into two matrices, a factor matrix representing a set of underlying unobserved characteristics of the subject which give rise to the observed characteristics and a loadings matrix which explains the relationship between the unobserved and observed characteristics. Another form of dimension reduction through matrix decomposition is an eigen decomposition known as a principal components analysis (PCA). Santos [2009] provides a method for constructing PCA adjusted for evolutionary history. This method, however, has the same problems typically associated with PCA, namely that it is not invariant to the scaling of the data and is not conducive to Bayesian analysis since it is not a likelihood based method. In a frequentist setting, the author also provides no approach for simultaneous inference on the phylogenetic tree that is rarely known without error [Huelsenbeck and Rannala, 2003]. In addition, there lacks a reasonable prescription for measuring uncertainty about which traits contribute to which principle components. Rai and Daume [2008] design a factor analysis method which uses a Kingman coalescent to construct a dendrogram across a factor analysis for genetic data.

While this is similar to the idea we will employ, this specific method uses a dendrogram between, rather than within, factors and is thus ill suited to handle the important problem we tackle in this paper. Namely, researchers often seek to identify a small number of relatively independent evolutionary processes, each represented by a factor changing over the tree, that ultimately give rise to a large number of observed, dependent traits. This paper provides such a dimension reduction tool by introducing phylogenetic factor analysis (PFA).

To formulate such a PFA model, we begin with usual Bayesian factor analysis, as posited by Lopes and West [2004] and Quinn [2004], which represents underlying latent characteristics of a group of organisms through a factor matrix and maps those latent characteristics to observed characteristics via a loadings matrix. In a standard factor analysis, the underlying factors for each species would be assumed to be independent of each other, however this does nothing to adjust for evolutionary history. Vrancken et al. [2015] describe how a high-dimensional Brownian diffusion can be used to describe the relationship between all of these observed traits, however the signal strength of the results of analyzing this model can be quite poor. By using independent Brownian diffusion priors on our factors, our PFA model groups traits into a parsimonious number of factors while successfully adjusting for phylogeny. Scientifically, these diffusions represent independent evolutionary processes. We use Markov chain Monte Carlo (MCMC) integration in order to draw inference on our model through a Metropolis-within-Gibbs approach. This facilitates both a latent data representation [Cybis et al., 2015] for integrating discrete and continuous traits and a natural method to handle missing data relevant to our problems. We further rely on path sampling methods [Gelman and Meng, 1998] to determine the appropriate number of factors [Ghosh and Dunson, 2009]. Since the latent, probit model necessitates the use of hard thresholds, we now have introduced an inherent difficulty in path sampling. In order to get

around this difficulty, we employ a novel method which relies on softening the threshold necessitated by the probit model slowly over the course of the path. We additionally develop a novel method by which to handle identifiability issues inherent to factor analysis by taking advantage of the fact that correlated elements in the loadings matrix tend to be correlated across the MCMC chain.

We show that our PFA method performs superiorly to a high-dimensional Brownian diffusion in both model fit, specifically through Bayes factors, and, when we are inferring large numbers of latent traits, speed using the examples of the evolution of the flower genus *Aquilegia*, as well as the reproduction of the fish family *Poeciliidae* that involves trait measurements missing at random. Lastly, we explore the dorsal, anal and pectoral fin shapes of the fish family *Balistidae* in order to explore this method's ability to handle situations where the number of traits are large compared to the number of species and to explore the simultaneous inference on our method along with the evolutionary history of these organisms with the aid of sequence data. The PFA model and its inference tools will be released in the popular phylogenetic inference package BEAST [Drummond et al., 2012].

3.2 Methods

3.2.1 Phenotypic Trait Evolution

Consider a collection of N biological entities (taxa). From each taxon $i = 1, \ldots, N$, we observe a P-dimensional measurement $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iP})$ of traits and, if available, a molecular sequence $\mathbf{S}_i$. We organize these phenotypic traits into an $N \times P$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)'$ and an aligned sequence matrix $\mathbf{S}$. These taxa are related to each other through an evolutionary history $\mathscr{F}$, informed through $\mathbf{S}$, and we are interested in learning about the evolutionary processes along this history that give rise to observed traits $\mathbf{Y}$.

The history $\mathscr{F}$ consists of a tree topology τ and a series of branch lengths $\mathbf{B}$. The tree topology is a bifurcating directed acyclic graph with a single generating point called the root, representing the most recent common ancestor of the given taxa, and with end points, each of which corresponds to a different taxon. The branch lengths correspond to edge weights of the graph, reflecting the evolutionary time before bifurcations. The history $\mathscr{F}$ may be known and fixed, or unknown and jointly inferred using $\mathbf{Y}$ and $\mathbf{S}$. For further details on constructing the sequence-informed prior distribution $p(\mathscr{F} \mid \mathbf{S})$ and integrating over $\mathscr{F}$ when unknown, see, e.g., Suchard et al. [2001] or Drummond et al. [2012].

In order to simultaneously model continuous, binary and ordinal traits, we adapt a latent data representation through the partially observed, standardized matrix $\mathbf{Z}$ with entries

$$Z_{ij} = \begin{cases} (Y_{ij} - \hat{Y}_j)/\hat{\sigma}_j & \text{if trait } j \text{ is continuous} \\ Z_{ij} & \text{if trait } j \text{ is binary or ordinal,} \end{cases} \tag{3.1}$$

where $\hat{Y}_j$ is the mean of trait j across taxa, $\hat{\sigma}_j$ is its standard deviation for $j = 1, \ldots, P$ and, more importantly, $Z_{ij} \in \mathbb{R}$ is an unknown random variable that satisfies the restrictions

$$\gamma_{j(c-1)} < Z_{ij} \leq \gamma_{jc} \text{ given } Y_{ij} = c \tag{3.2}$$

and $c \in \{1, \ldots, m_j\}$ for m_j-valued binary/ordinal data for trait j. For identifiability, latent trait cut-points $\boldsymbol{\gamma}_j = (\gamma_{j0}, \ldots, \gamma_{jm_j})$ take on the restrictions $\gamma_{j0} = -\infty$, $\gamma_{j1} = 0$ and $\gamma_{jm_j} = \infty$ or are otherwise random and jointly inferred. Grouping cut-points for all binary or ordinal traits into $\boldsymbol{\gamma}$, Cybis et al. [2015] suggest assuming that differences between the small number of successive, random cut-points are *a priori* exponentially distributed with mean $\frac{1}{2}$ to de-

fine their density $p(\gamma)$. Cybis et al. [2015] also discuss in detail how to treat categorical data in this sort of analysis. Since we do not use examples which contain non-ordered categorical data we elect not to describe those methods in these sections, but we will mention that they are implemented in BEAST and are easily adapted to fit the methods described in this paper.

In order to uncover the biological relationships amongst traits in $\mathbf{Z}$ while controlling for evolutionary history, previous work relies on a Gaussian process generative model induced through considering conditionally independent Brownian diffusion along each branch in $\mathscr{T}$ [Felsenstein, 1985]. In a multivariate setting, a $P \times P$ variance matrix $\mathbf{\Sigma}$ and unobserved, P-dimensional root trait value $\boldsymbol{\mu}_R$ characterize the process. Pybus et al. [2012] identify that analytic integration of $\boldsymbol{\mu}_R$ is possible by assuming that $\boldsymbol{\mu}_R$ is *a priori* multivariate normally distributed with a fixed hyperprior mean $\boldsymbol{\mu}_0$ and variance equal to $\kappa_0^{-1}\mathbf{\Sigma}$, where κ_0 is a fixed hyperprior sample-size. Consequentially, given $\mathscr{T}$ and $\mathbf{\Sigma}$, the latent traits $\mathbf{Z}$ are distributed according to a matrix-normal (MN)

$$\mathbf{Z} \sim \mathrm{MN}\Big(\boldsymbol{\mu}_0, \mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}, \mathbf{\Sigma}\Big), \tag{3.3}$$

where $\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}$ is the across-taxa (row) variance and a deterministic function of phylogeny $\mathscr{T}$, $\mathbf{\Sigma}$ is the across-trait (column) variance, and $\mathbf{J}$ is a $N \times N$ matrix of ones [Vrancken et al., 2015]. Traits $\mathbf{Z}$ have density function

$$p(\mathbf{Z} \mid \mathscr{T}, \mathbf{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\left(\mathbf{Z} - \mathbf{1}'_{\boldsymbol{\mu}_0}\right)'\left(\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}\right)^{-1}\left(\mathbf{Z} - \mathbf{1}'_{\boldsymbol{\mu}_0}\right)\right]\right\}}{(2\pi)^{NP/2}\left|\mathbf{\Sigma}\right|^{N/2}\left|\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}\right|^{P/2}}, \tag{3.4}$$

where $\mathrm{tr}[\cdot]$ is the trace operator and $\mathbf{1}$ is a N-dimensional column vector of ones. Tree variance matrix $\mathbf{\Psi}$ contains diagonal elements that are equal to the sum of the adjusted branch lengths in $\mathscr{T}$ between the root node and taxon i, and off-diagonal elements (i, i') that are equal to the sum of the adjusted

branch lengths between the root node and the most recent common ancestor of taxa i and i', where the adjusted branch lengths represent a function of wall time and a branch rate accounting for variation in evolutionary rate over the course of the tree. For our diffusion model, we scale our tree such that from the root to the most recent tip we say that the process has undergone one diffusion unit.

Placing a conjugate prior distribution on $\mathbf{\Sigma}$, such as $\mathbf{\Sigma}^{-1} \sim \text{Wishart}_{\nu}(\mathbf{\Lambda}_{R_0})$ where ν is the hyperprior degrees of freedom and $\mathbf{\Lambda}_{R_0}$ is the hyperprior belief on the structure of the inverse of the variance matrix $\mathbf{\Sigma}$, enables inference about its posterior distribution, shedding light on how the evolution of these traits relate to each other. Such inference often requires repeated evaluation of density (3.4), especially when the phylogeny $\mathscr{F}$ or variance $\mathbf{\Sigma}$ is random. This evaluation suggests a computational order $\mathcal{O}\left(N^3 + P^3\right)$, arising from the inversion of the $N \times N$ variance matrix $\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}$ and $P \times P$ variance matrix $\mathbf{\Sigma}$. One easily avoids the latter by parameterizing the model in terms of $\mathbf{\Sigma}^{-1}$ [Lemey et al., 2010]. To address the former, Pybus et al. [2012] provide an $\mathcal{O}\left(NP^2\right)$ dynamic programming algorithm to evaluate (3.4) without inversion of the across-taxa variance matrix, similar to Freckleton [2012]. This advance certainly makes for more tractable inference under these diffusion models as N grows large, but the quadratic dependence on P still hampers their use for high-dimensional traits. Inference can often be slow, taking as long as a day for problems with a dozen traits and about 30 taxa to mix properly [Cybis et al., 2015]. Finally, direct inference on $\mathbf{\Sigma}$ can often fail to produce a coherent and interpretable conclusion about the number of independent evolutionary processes generating the traits if the matrix cannot be reordered to form approximately separated blocks especially if the signal is too weak to produce many statistically significant cells.

3.2.2 Factor Analysis

To infer potentially low dimensional evolutionary structure among traits, we rely on dimension reduction via a phylogenetic factor analysis (PFA). This model builds on the premise that a small, but unknown number $K \ll \min(N, P)$ of *a priori* independent univariate Brownian diffusion processes along $\mathscr{F}$ provides a more parsimonious description of the covariation in $\mathbf{Z}$ than a P-dimensional multivariate diffusion. We parameterize the PFA in terms of an $N \times K$ factor matrix $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_K)$ whose K columns $\mathbf{F}_k = (F_{1k}, \ldots, F_{Nk})'$ for $k = 1, \ldots, K$ represent the unobserved independent realizations of univariate diffusion at each of the N tips in $\mathscr{F}$, a $K \times P$ loadings matrix $\mathbf{L} = \{L_{kj}\}$ that relates the independent factor columns to $\mathbf{Z}$, and an $N \times P$ model error matrix $\boldsymbol{\epsilon}$, such that

$$\mathbf{Z} = \mathbf{F}\mathbf{L} + \boldsymbol{\epsilon}. \tag{3.5}$$

To inject information about and control for shared evolutionary history $\mathscr{F}$, we specify that

$$\begin{aligned} \mathbf{F} &\sim \mathrm{MN}\Big(\mathbf{0}, \boldsymbol{\Psi} + \kappa_0^{-1}\mathbf{J}, \mathbf{I}_K\Big), \text{ and} \\ \boldsymbol{\epsilon} &\sim \mathrm{MN}\Big(\mathbf{0}, \boldsymbol{\Omega}, \boldsymbol{\Lambda}^{-1}\Big), \end{aligned} \tag{3.6}$$

where $\mathbf{I}_{(\cdot)}$ is the identity matrix of appropriate dimension and the residual column precision $\boldsymbol{\Lambda}$ is a diagonal matrix with entries $(\Lambda_1, \ldots, \Lambda_P)$. Lastly, since K is unknown, we place a reasonably conservative zero-truncated-Poisson prior on it, such that $p(K = 1) = 1/2$.

To better appreciate the details of the PFA model, we briefly compare it to a typical Bayesian factor analysis. Typical factor analyses assume that all entries of $\mathbf{F}$ are independent and identically distributed (iid) as $\mathrm{N}(0, 1)$, normal random variables with mean 0 and variance 1. In PFA, the shared evolutionary history $\mathscr{F}$ specifies the correlation structure within the N entries of column $\mathbf{F}_k$. Often, one refers to a given column as a "factor." Across factors, the column

variance remains $\mathbf{I}_K$ to reflect our assertion that the underlying evolutionary processes generating $\mathbf{F}_k$ are independent of each other. Note that in this model the number of parameters undergoing Brownian Diffusion is assumed to be of dimension K as opposed to of dimension P in the previous model.

To complete model specification of the loadings $\mathbf{L}$ and residual error $\boldsymbol{\epsilon}$, we assume

$$\begin{aligned} L_{kj} &\sim \mathrm{N}(0,1) \text{ for all } k \leq j, \\ \Lambda_j &\sim \Gamma(\alpha_{\boldsymbol{\Lambda}}, \beta_{\boldsymbol{\Lambda}}) \text{ for all trait } j \text{ continuous, and} \end{aligned} \tag{3.7}$$

otherwise $\Lambda_j = 1$ to preserve identifiability under the scale-free latent model for discrete traits. Here, $\Gamma(\alpha_{\boldsymbol{\Lambda}}, \beta_{\boldsymbol{\Lambda}})$ signifies a gamma distributed random variable with hyperparameter scale $\alpha_{\boldsymbol{\Lambda}}$ and rate $\beta_{\boldsymbol{\Lambda}}$.

Without further restrictions on $\mathbf{L}$, any factor analysis remains over-specified. For example, given an orthogonal $K \times K$ matrix $\mathbf{T}$, one may rotate $\mathbf{F}$ in one direction and $\mathbf{L}$ in the other and arrive at the same data likelihood, since $\mathbf{FL} = \mathbf{FTT}'\mathbf{L}$. To address this identifiability issue, we fix lower triangular entries $L_{kj} = 0$ for $k > j$ [Geweke and Zhou, 1996, Aguilar and West, 2000]. It is also standard practice to apply the restriction $L_{kk} > 0$, since otherwise $\mathbf{FL} = (\text{-}\mathbf{F})(\text{-}\mathbf{L})$. While the constraint yields an identifiable posterior distribution with respect to $\mathbf{F}$ and $\mathbf{L}$, we do not pursue it here because it introduces bias into our scientific inference on $\mathbf{L}$ and, instead, search for an alternative.

The diagonal and upper triangular entries L_{kj} for $k \leq j$ of the loadings $\mathbf{L}$ inform the magnitude and effect-direction that the evolutionary process captured in factor $\mathbf{F}_k$ contributes to trait j. As a quantitative measure of uncertainty about these relationships, we define p_{kj} to equal the absolute difference between the posterior probability that $L_{kj} > 0$ and the posterior probability that $L_{kj} < 0$; this measure ranges from 0 when L_{kj} is centered around 0 to 1 when L_{kj} is either strictly positive or strictly negative with probability 1. It is possible, and we would argue likely, that $\mathbf{F}_k$ has little or no

influence on the trait arbitrarily labeled k, such that most of the posterior mass of L_{kk} lies around and close to 0. Artificially restricting $L_{kk} > 0$ forces all of this mass above 0, signifying a positive association with prior, and hence posterior, probability 1.

To combat this bias, we recouch these identifiability conditions as a label switching problem in a mixture model and propose a *post hoc* relabeling algorithm [Stephens, 2000]. We require K sign constraints, one for each column-row outer-product in forming $\mathbf{FL}$, for posterior identification. In our prior, we modify Equation (3.7) to further assign one non-zero entry $L_{kj} > 0$ per row, but do not specify which one; this assignment mirrors the mixture model labeling. Hence, we allow the data, not an arbitrary decision, to determine which entry per row reflects a positive association with probability 1, decreasing potential bias.

Recalling that continuous traits are standardized in $\mathbf{Z}$ to have mean 0 and variance 1 affords several benefits. First, we can posit a $\mathbf{0}$-matrix mean for $\mathbf{F}$ in Equation (3.6) without loss of information. But, more importantly, when we draw inference on $\mathbf{\Lambda}$, we can interpret traits which have precision elements that demonstrate considerable posterior mass at or below 1 to be described insufficiently by the model, since the factors provide no insight beyond a random normal model. A third advantage is that standardization helps us select reasonable scales for the non-zero entries in $\mathbf{L}$, namely that these have variance 1, and hyperparameters for $\mathbf{\Lambda}$, specifically that $\frac{\alpha_{\mathbf{\Lambda}}}{\beta_{\mathbf{\Lambda}}} = 1$. In practice, $\alpha_{\mathbf{\Lambda}} = \frac{1}{3}$ and $\beta_{\mathbf{\Lambda}} = \frac{1}{3}$ for analyses in this paper. While these hyperparameter choices are by no means perfect we feel that, under the paradigm of data scaling, they are reasonable and generalizable across a variety of problems.

This model is a simplified form of the item factor analysis models that are described by Quinn [2004] in the political science literature and Beguin and Glas [2001] in the psychology literature with a tree as a prior on the factors instead

of an independent normal distribution. In fact, the methods for treating binary and ordinal data described in Quinn [2004] are the same as those described in Cybis et al. [2015], making for a convenient adaptation of this factor analysis model to phylogenetics using existing software in BEAST.

3.2.3 Inference

Given the trait measurements $\mathbf{Y}$ and aligned sequences $\mathbf{S}$, we strive to learn about the joint posterior distribution of the number of evolutionary processes K, factors $\mathbf{F}$, loadings $\mathbf{L}$, column precisions $\mathbf{\Lambda}$, latent trait cut-points $\boldsymbol{\gamma}$ and evolutionary history $\mathscr{F}$

$$\begin{aligned} p\left(K, \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\gamma}, \mathscr{F} \mid \mathbf{Y}, \mathbf{S}\right) &\propto p\left(\mathbf{Y} \mid K, \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\gamma}\right) \times p\left(\mathbf{F} \mid K, \mathscr{F}\right) \times p\left(\mathscr{F} \mid \mathbf{S}\right) \\ &\quad \times p\left(\mathbf{L} \mid K\right) \times p\left(\mathbf{\Lambda}\right) \times p\left(\boldsymbol{\gamma}\right) \times p\left(K\right) \\ &= \left(\int p\left(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma}\right) p\left(\mathbf{Z} \mid K, \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}\right) d\mathbf{Z} \right) p\left(\mathbf{F} \mid K, \mathscr{F}\right) \\ &\quad \times p\left(\mathscr{F} \mid \mathbf{S}\right) \times p\left(\mathbf{L} \mid K\right) \times p\left(\mathbf{\Lambda}\right) \times p\left(\boldsymbol{\gamma}\right) \times p\left(K\right), \end{aligned} \tag{3.8}$$

where $p\left(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma}\right) \propto \mathbf{1}\left(\mathbf{Y} | \mathbf{Z}, \boldsymbol{\gamma}\right)$ is the indicator function that the restrictions in Equation (3.2) hold. We accomplish this inference through MCMC, using a random-scan Metropolis-within-Gibbs scheme [Liu et al., 1995] for fixed K and a modification of path sampling to then estimate the marginal posterior $p\left(K \mid \mathbf{Y}, \mathbf{S}\right)$. For fixed K, our Metropolis-within-Gibbs scheme employs transition kernels described in Cybis et al. [2015] and references therein to integrate over the evolutionary history $\mathscr{F}$ and unobserved, latent traits Z_{ij} and cut-points $\boldsymbol{\gamma}_j$ where trait j is discrete.

Here, we focus on transition kernels within the scheme to integrate over the factors $\mathbf{F}$, loadings $\mathbf{L}$ and residual column precision $\mathbf{\Lambda}$. Lopes and West [2004] derive full conditional distributions for the columns of $\mathbf{L}$ and diagonals

of $\mathbf{\Lambda}$ under a traditional factor analysis. These full conditional distributions do not change under a PFA and we use them for Gibbs sampling. Specifically, for column j of $\mathbf{L}$, the first $k' = \min(j, K)$ entries are non-zero and, given all other random variables, distributed according to a multivariate normal (MVN)

$$\left(L_{1j}, \ldots, L_{k'j}\right)' \mid \mathbf{Z}, \mathbf{F}, \mathbf{\Lambda} \sim \text{MVN}\left(\mathbf{M}_j^{(\mathbf{L})}, \mathbf{V}_j^{(\mathbf{L})}\right) \text{ for } j = 1, \ldots, P, \tag{3.9}$$

parameterized in terms of its mean

$$\mathbf{M}_j^{(\mathbf{L})} = \mathbf{V}_j^{(\mathbf{L})} \Lambda_j \mathbf{F}'_{1:k'} \mathbf{Z}\, \mathbf{e}_j \tag{3.10}$$

and variance

$$\mathbf{V}_j^{(\mathbf{L})} = \left(\Lambda_j \mathbf{F}'_{1:k'} \mathbf{F}_{1:k'} + \mathbf{I}_{k'}\right)^{-1}, \tag{3.11}$$

where $\mathbf{F}_{1:k'} = (\mathbf{F}_1, \ldots, \mathbf{F}_{k'})$ is the first k' columns of $\mathbf{F}$ and $\mathbf{e}_j$ is the unit-vector in the direction of trait j. Further,

$$\Lambda_j \mid \mathbf{Z}, \mathbf{F}, \mathbf{L} \sim \Gamma\left(\alpha_{\mathbf{\Lambda}} + \frac{N}{2}, \beta_{\mathbf{\Lambda}} + \frac{1}{2}\, \mathbf{e}_j' \left(\mathbf{Z} - \mathbf{FL}\right)' \left(\mathbf{Z} - \mathbf{FL}\right) \mathbf{e}_j\right), \tag{3.12}$$

if trait j is continuous. Appendix A provides derivations of these full conditional distributions. Gibbs sampling all columns of $\mathbf{L}$ carries a computation order $\mathcal{O}\left(NK^2P\right)$, arising from the matrix multiplication of $\mathbf{F}'_{1:k'}\mathbf{F}_{1:k'}$ for each trait. The matrix inversion is not rate-limiting here since $N \gg K$. Likewise, Gibbs sampling $\mathbf{\Lambda}$ remains very light-weight at $\mathcal{O}\left(NKP\right)$, stemming from the sparse multiplication of $\mathbf{FLe}_j$ for each trait. While we write that the order of both Gibbs samplers depend on P to be clear that we must iterate over all traits, the astute reader has already recognized the conditional independence of updates between traits, such that we may execute updates for each trait in parallel.

The traditional Gibbs sampler for $\mathbf{F}$ fails in the phylogenetic setting for more than a handful of taxa, since determining the full conditional distribution of $\mathbf{F}$ requires inverting the matrix $\left(\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}\right)$. As mentioned previously, but worth repeating, this task stands as prohibitive with a computational order $\mathcal{O}\left(N^3\right)$ and presents a major challenge for PFA.

We circumvent this difficulty by exploiting the structure of the phylogenetic tree $\mathscr{F}$. Probability models on directed, acyclic graphs lend themselves well to dynamic programming for determining marginalized data likelihoods, such as Felsenstein's pruning algorithm for sequence data [Felsenstein, 1973] and related work for Brownian diffusion [Pybus et al., 2012], and conditional predictive distributions, like those obtained for (ancestral) sequence reconstruction.

In extending these conditional distributions to Brownian diffusion, first let $\mathbf{F}_{i\cdot} = (F_{i1}, \ldots, F_{iK})$ identify row i of $\mathbf{F}$, more specifically all latent factor values attributed to taxon i, and let $\mathbf{F}_{\text{-}i\cdot}$ concatenate the remaining rows. Given that $\mathbf{F}$ is matrix-normally distributed with an across-taxa (row) variance that depends on the phylogeny $\mathscr{F}$, Cybis et al. [2015] provide a tree-traversal-based algorithm to determine $p\left(\mathbf{F}_{i\cdot} \mid \mathbf{F}_{\text{-}i\cdot}, \mathscr{F}\right)$ that remains a multivariate normal distribution. The algorithm requires first a post-order tree-traversal to determine the joint distribution of all tip-values descendent to each internal node and then a pre-order tree-traversal back to taxon i to compute its prior conditional mean $\boldsymbol{\mu}_{\mathbf{F}_{\text{-}i\cdot}}$ and precision $\mathbf{\Lambda}_{\mathbf{F}_{\text{-}i\cdot}}$. Since the across-factor (column) variance on $\mathbf{F}$ is diagonal, the dynamic programming algorithm runs quickly in $\mathcal{O}\left(NK\right)$. Using this result, we determine the full conditional distribution

$$\mathbf{F}'_{i\cdot} \mid \mathbf{Z}, \mathbf{F}_{\text{-}i\cdot}, \mathbf{L}, \mathbf{\Lambda}, \mathscr{F} \sim \text{MVN}\left(\mathbf{M}_i^{(\mathbf{F})}, \mathbf{V}_i^{(\mathbf{F})}\right) \text{ for } i = 1, \ldots, N, \tag{3.13}$$

with mean

$$\mathbf{M}_i^{(\mathbf{F})} = \mathbf{V}_i^{(\mathbf{F})} \left(\mathbf{L}\boldsymbol{\Lambda}\mathbf{Z}'\mathbf{e}_i + \boldsymbol{\Lambda}_{\mathbf{F}_{\text{-}i\cdot}} \boldsymbol{\mu}_{\mathbf{F}_{\text{-}i\cdot}} \right) \tag{3.14}$$

and variance

$$\mathbf{V}_i^{(\mathbf{F})} = \left(\mathbf{L}\boldsymbol{\Lambda}\mathbf{L}' + \boldsymbol{\Lambda}_{\mathbf{F}_{\text{-}i\cdot}} \right)^{-1}, \tag{3.15}$$

where $\mathbf{e}_i$ is the unit-vector in the direction of taxon i. Appendix A delivers a derivation of this full conditional distribution. The evaluation of this full conditional distribution runs in $\mathcal{O}\left(K^2P\right)$, where the term $\mathbf{L}\boldsymbol{\Lambda}\mathbf{L}'$ is rate limiting.

Employing Equations (3.13) - (3.15), we can cycle over i to fabricate a tractable Gibbs sampler for $\mathbf{F}$ with total computational order $\mathcal{O}\left(N^2K + NK^2P\right)$. It is fruitful to compare this work with the rate-limiting step for inference under the non-sparse model. Here, sampling the precision matrix $\boldsymbol{\Sigma}^{-1}$ carries a computational cost of $\mathcal{O}\left(NP^2\right)$. From these bounds, it is clear that increasing numbers of taxa N should limit PFA, while increasing numbers of traits P should limit the non-sparse model from a computational work per MCMC iteration perspective. However, per-iterative arguments ignore the posterior correlation between model parameters and its influence on MCMC mixing times.

Finally, to maintain identifiability with respect to $\mathbf{F}$ and $\mathbf{L}$ in the posterior, we propose a simple *post hoc* relabeling algorithm [Stephens, 2000]. We sample $(\mathbf{F}t, \mathbf{L}t)$ from $p\left(K, \mathbf{F}, \mathbf{L}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}, \mathscr{F} \mid \mathbf{Y}, \mathbf{S}\right)$ for MCMC iteration $m = 1, \ldots, M$ assuming a sign-unconstrained prior. From this unconstrained sample, we select for each row k in $\mathbf{L}$ the column element with the fewest number of sign changes between iterations. Assume for row k, this is column j_k. We then constrain our sample by multiplying $\mathbf{F}_k t$ and row k of $\mathbf{L}t$ by the sign of $L_{kj_k}t$. No further sample reweighing is necessary because $p\left(\mathbf{F} \mid K, \mathscr{F}\right) = p\left(\text{-}\mathbf{F} \mid K, \mathscr{F}\right)$

is also invariant to reflection.

Model Selection To estimate the marginal posterior density $p\left(K \mid \mathbf{Y}, \mathbf{S}\right)$, we rely on a variant of path sampling that we equip to successfully integrate latent variable $\mathbf{Z}$ when traits are discrete. We employ our variant to approximate each marginal likelihood $p\left(\mathbf{Y}, \mathbf{S} \mid K = k\right)$ for $k = 1, \ldots, S$, where S is a relatively small number such as $\min\{P, 10\}$, after which we approximate $p\left(\mathbf{Y}, \mathbf{S} \mid K > S\right) = 0$. Then, invoking Bayes theorem, $p\left(K = k \mid \mathbf{Y}, \mathbf{S}\right) \propto p\left(\mathbf{Y}, \mathbf{S} \mid K = k\right) p\left(K = k\right)$. Moreover, through this approach, we can address the model selection problem of how many independent factors do the data support through Bayes factors [Jeffreys, 1935]:

$$\frac{p\left(K = k \mid \mathbf{Y}, \mathbf{S}\right)}{p\left(K = k' \mid \mathbf{Y}, \mathbf{S}\right)} = \frac{p\left(\mathbf{Y}, \mathbf{S} \mid K = k\right)}{p\left(\mathbf{Y}, \mathbf{S} \mid K = k'\right)} \frac{p\left(K = k\right)}{p\left(K = k'\right)}. \tag{3.16}$$

Lopes and West [2004] and Ghosh and Dunson [2009] have been strong proponents of Bayes factors to determine the optimal number of factors in a traditional factor analysis, where Lopes and West [2004] employ a simple harmonic mean estimator [Newton and Raftery, 1994] to estimate their marginal likelihoods. This estimator performs poorly in highly structured phylogenetic models and path sampling has largely supplanted it [Baele et al., 2012].

Path sampling is an MCMC-based integration technique to estimate marginal likelihoods, such as $p\left(\mathbf{Y}, \mathbf{S} \mid K\right)$. The technique constructs a series of power posteriors [Friel and Pettitt, 2008] at various temperatures $\beta \in [0, 1]$, where $\beta = 1$ corresponds to a joint density $l(\mathbf{Y}, \mathbf{S}, \mathbf{Z}, \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}, \boldsymbol{\gamma} \mid K)$ proportional, but with an unknown constant, to $p\left(\mathbf{Y}, \mathbf{S} \mid K\right)$ and $\beta = 0$ yields a normalized density $\hat{p}(\mathbf{Z}, \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}, \mathscr{F}, \boldsymbol{\gamma} \mid K)$ that does not depend on the data, often a combination of the prior and other working distributions, see e.g. [Baele et al., 2016]. The usual power posterior path is $q(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta}) = l(\mathbf{Y}, \mathbf{S}, \mathbf{\Theta})^{\beta} \times \hat{p}(\mathbf{\Theta})^{1-\beta}$, where $\mathbf{\Theta}$ is the set of all parameters in the model we are considering. For example, in

PFA, $\boldsymbol{\Theta} = \{\mathbf{Z}, \mathbf{F}, \mathbf{L}, \boldsymbol{\Lambda}, \mathscr{F}, \boldsymbol{\gamma}\}$.

In latent models with discrete traits, however, the support of the latent variable $\mathbf{Z}$ changes when the data are observed [Heaps et al., 2014]. In particular, our unnormalized joint density $l(\mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta})$ is zero for values of $\mathbf{Z}$ that are incompatible with $\mathbf{Y}$ because $p(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma}) = 0$, therefore a trait Z_{ij} only has support over $(\gamma_{i(c-1)}, \gamma_{ic}]$ if $Y_{ij} = c$, while $\hat{p}(\cdot)$ places non-zero density over all possible values $Z_{ij} \in (-\infty, \infty)$. Our working distribution, for example, assumes $Z_{ij} \sim N(0, 1)$ when Z_{ij} is random. If we factor $l(\mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta})$ into a support condition $\mathbf{1}(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\gamma})$ and the remaining likelihood $h(\mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta})$, then the standard path used in this scenario [Heaps et al., 2014] is

$$q(\beta, \mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta}) = \mathbf{1}(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\gamma}) \times h(\mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta})^{\beta} \times \hat{p}(\boldsymbol{\Theta})^{1-\beta}. \tag{3.17}$$

For the power posterior method to yield the marginal likelihood $p(\mathbf{Y} \mid K)$, it is necessary [Friel and Pettitt, 2008] that

$$\int \left\{ \lim_{\beta \to 0} q(\beta, \mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta}) \right\} d\boldsymbol{\Theta} = 1. \tag{3.18}$$

Plugging (3.17) into (3.18), we find

$$\int \left\{ \lim_{\beta \to 0} q(\beta, \mathbf{Y}, \mathbf{S}, \boldsymbol{\Theta}) \right\} d\boldsymbol{\Theta} = \int \mathbf{1}(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\gamma}) \times \hat{p}(\boldsymbol{\Theta}) d\boldsymbol{\Theta}. \tag{3.19}$$

If we define $\boldsymbol{\Omega}$ as the region where $\mathbf{1}(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\gamma}) = 1$, then we see that

$$\int_{\boldsymbol{\Omega}} \hat{p}(\boldsymbol{\Theta}) d\boldsymbol{\Theta} < 1, \tag{3.20}$$

since $\boldsymbol{\Omega} \subsetneq$ the support of $\boldsymbol{\Theta}$. While it is theoretically possible to construct $\hat{p}(\boldsymbol{\Theta})$ such that it is normalized to 1 over $\boldsymbol{\Omega}$, previous attempts to do so have failed. Alternatively, Heaps et al. [2014] attempt to approximate such a distribution by

fixing γ and ignoring the corresponding integral.

We posit an exact solution by proposing a new path that relies on a softening threshold. Consider the modified path

$$q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta}) = \{1 - [1 - \mathbf{1}(\mathbf{Y}|\mathbf{Z}, \gamma)]\beta\} \times h(\mathbf{Y}, \mathbf{S}, \mathbf{\Theta})^{\beta} \times \hat{p}(\mathbf{\Theta})^{1-\beta}. \quad (3.21)$$

Following from (3.18), we find that

$$\int \left\{ \lim_{\beta \to 0} q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta}) \right\} d\mathbf{\Theta} = \int \hat{p}(\mathbf{\Theta}) d\mathbf{\Theta} = 1, \quad (3.22)$$

by construction.

Lastly, in order to adapt the power posterior method, at each step in the series we need to compute the derivative of $\log q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta})$ with respect to β. From Equation (3.21), we see that

$$\begin{aligned} \frac{\partial}{\partial \beta} \log q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta}) = &- \frac{1 - \mathbf{1}(\mathbf{Y}|\mathbf{Z}, \gamma)}{1 - [1 - \mathbf{1}(\mathbf{Y}|\mathbf{Z}, \gamma)]\beta} \\ &+ \log h(\mathbf{Y}, \mathbf{S}, \mathbf{\Theta}) - \log \hat{p}(\mathbf{\Theta}), \end{aligned} \quad (3.23)$$

and observe that there is no singularity at $\beta = 1$ since, at that point in the path, latent variable $\mathbf{Z}$ only assumes values in $\mathbf{\Omega}$, such that $\mathbf{1}(\mathbf{Y}|\mathbf{Z}, \gamma) = 1$.

3.3 Empirical Examples

3.3.1 Columbine Flower Development

Columbine genus *Aquilegia* flowers have attracted at least three different pollinators across their evolutionary history: bumblebees (Bb), hawkmoths (Hm) and hummingbirds (Hb). Whittall and Hodges [2007] question the role that these pollinators play in the tempo of columbine flower evolution, tracked

through the color, length and orientation of different anatomical floral features, and are particularly interested in how transitions between pollinators relate to spur length. Cybis et al. [2015] take up this question by examining $P = 12$ different traits for $N = 30$ monophyletic populations from the genus *Aquilegia* that include 10 continuously valued traits, a binary trait that indicates presence or absence of anthocyanin pigment and a final ordinal trait indicating the primary pollinator for that population. Whittall and Hodges [2007] propose a Bb-Hm-Hb ordering and we use the fixed phylogenetic tree the authors employ in their analysis. Through fitting a latent multivariate Brownian diffusion (LMBD) model parameterized in terms of a 12×12 variance matrix $\mathbf{\Sigma}$, Cybis et al. [2015] find the data strongly support the proposed ordering over alternative orderings. We return to the relationship between pollinator and the other traits and test whether a PFA returns a better understanding of the evolutionary factors driving their interrelated change compared to an LMBD model.

Under our PFA, the most probable number of independent evolutionary processes is $K = 2$, with a log Bayes factor > 7 over the neighboring $K = 1$ or $K = 3$ factor parameterizations (Table 3.1). Further, the PFA with $K = 2$ is favored over the LMBD model with a log Bayes factor > 24 when assuming equal prior probabilities over these two models.

The PFA has high explanatory power for all continuous traits (Table 3.2) and Figure 3.1 presents our inference on the relationships between traits under the PFA with $K = 2$ and compares these findings to inference under the LMBD model. The first evolutionary process $\mathbf{F}_1$ approximately partitions the traits into two groups. One group includes: orientation, blade brightness, spur brightness, sepal length, blade length, pollinator type, spur hue, spur length, blade hue, and expected trait values increase (displayed loadings entries L_{kj} in purple) as the factor grows over the phylogeny. The other group includes: blade

	Model	Log marginal likelihood
Aquilegia	$K = 1$	-385.4
	$K = 2$	-366.9
	$K = 3$	-374.3
	LMBD	-391.1
Poeciliidae	$K = 2$	-536.0
	$K = 3$	-500.7
	$K = 4$	-501.0
	$K = 5$	-505.9
	LMBD	-592.3
Balistidae	$K = 4$	-15622.0
	$K = 5$	-15603.5
	$K = 6$	-15610.4
	MBD	-15673.2

Table 3.1: *Log marginal likelihood estimates for the number K of independent factors driving evolution under a phylogenetic factor analysis (PFA) and a latent multivariate Brownian diffusion (LMBD) model in Aquilegia, and Poeciliidae and multivariate Brownian diffusion (MBD) in Balistidae. The $K = 2$ model for Aquilegia, the $K = 3$ and $K = 4$ model for Poeciliidae and the $K = 5$ model for Balistidae achieve the highest marginal likelihoods.*

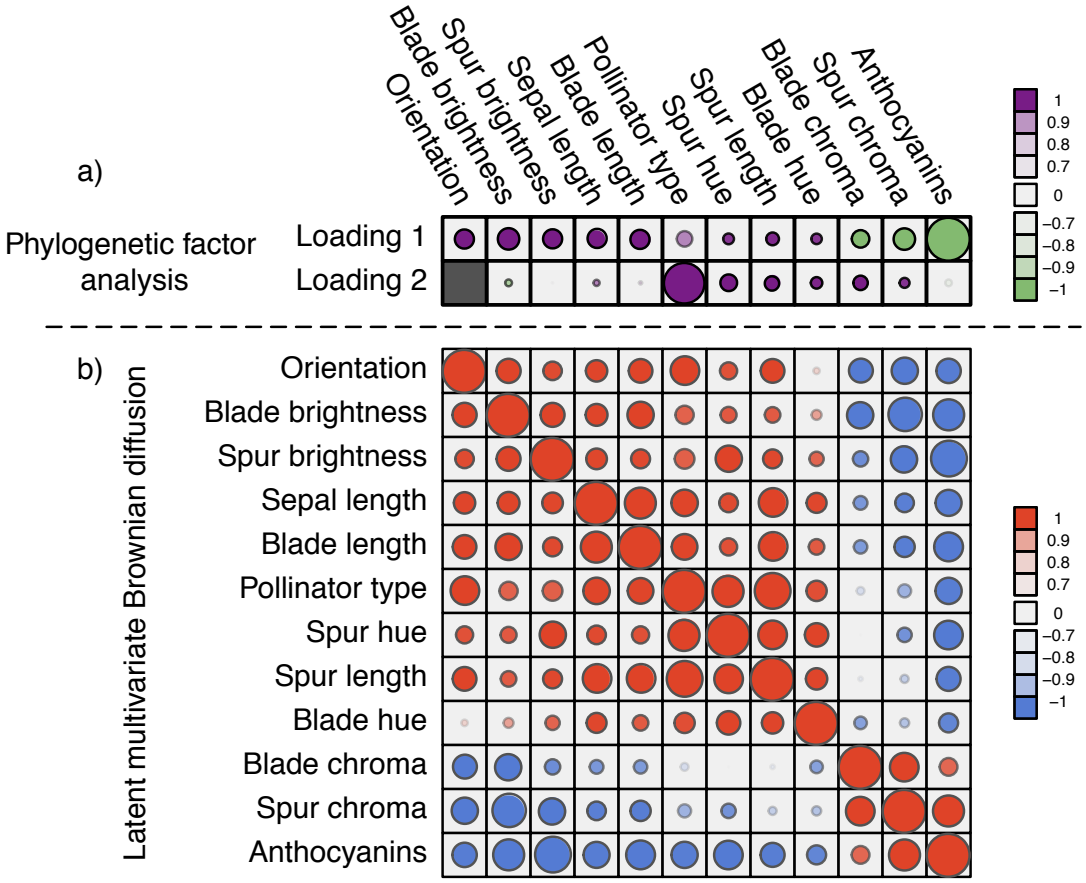

Figure 3.1: *Processes driving columbine flower evolution inferred through phylogenetic factor analysis (PFA) or latent multivariate Brownian diffusion (LMBD). a) Loadings* $\mathbf{L}$ *estimates from a* $K = 2$ *factor PFA model. Purple circles represent traits positively associated with traits represented by other purple circles within a loading, and negatively associated with traits represented by green circles within a loading. Similarly, traits represented by green circles are positively associated with traits represented by green circles within a loading. Size represents the magnitude of the value of the loadings. Opacity represents the posterior probability that the sign of the given element is equal to the sign of the posterior mean. The greyed out cell represents a structural 0 introduced for identifiability reasons. The magnitude for anthocyanins and pollinator type is less relevant since those measurements are discrete. b) Correlation matrix estimate from a LMBD model. Red represents positive correlation, blue represents anti-correlation, and opacity represents the absolute difference in posterior probability of being greater than 0 and less than 0. Size of the circle represents the magnitude of the correlation. The PFA captures well two independent processes, while the LMBD groups these processes together.*

chroma, anthocyanins pigment presence and, with less posterior probability, spur chroma, and expected trait values decrease (green) as the factor grows. A possible exception to the $\mathbf{F}_1$ partitioning is the pollinator trait, where we estimate only a 0.92 absolute difference in posterior probability of being greater than 0 versus less than 0.

Ignoring the uncertainty in pollinator trait inclusion for the moment, this partitioning recapitulates the block structure that Cybis et al. [2015] report using an LMBD model and an arbitrary thresholding on the posterior mean estimates of the individual pairwise correlation entries in $\mathbf{\Sigma}$. However, in Figure 3.1 we quantify the LMBD uncertainty by shading our inference using the same probability measure as we do for our PFA model. Taking correlation uncertainty into consideration we see that, for example the LMBD model would assert that there is no correlation between blade chroma and spur hue. The PFA model by contrast offers the more nuanced assessment that these traits are related through two independent underlying processes, one process of which has a positive association between these traits, the other of which has a negative association.

	Trait	Posterior mean	95% Bayesian credible interval
Aquilegia	Orientation	2.1	[1.0, 3.3]
	Spur length	4.4	[2.0, 7.1]
	Blade length	3.0	[1.4, 4.8]
	Sepal length	2.6	[1.3, 4.1]
	Spur chroma	4.2	[1.8, 6.9]
	Spur hue	6.2	[2.6, 10.5]
	Spur brightness	2.7	[1.2, 4.3]

	Blade chroma	2.3	[1.1, 3.7]
	Blade hue	2.1	[1.0, 3.2]
	Blade brightness	3.3	[1.4, 0.6]
Poeciliidae (K = 3)	Matrotrophy index	14.3	[5.6, 23.2]
	Gonopodium length	9.3	[4.3, 16.1]
	Male body length	3.5	[2.4, 4.6]
	Male body weight	2.8	[1.9, 3.7]
	Female body length	10.5	[5.7, 15.5]
	Female body weight	15.1	[8.0, 24.3]
Poeciliidae (K = 4)	Matrotrophy index	13.8	[5.5, 22.7]
	Gonopodium length	9.1	[4.4, 15.5]
	Male body length	3.5	[2.3, 4.8]
	Male body weight	2.8	[1.9, 3.8]
	Female body length	10.5	[5.8, 15.5]
	Female body weight	14.7	[8.2, 22.5]

Table 3.2: *Precision* $\mathbf{\Lambda}$ *posterior mean and 95% Bayesian credible interval estimates under the latent factor model for the traits in Aquilegia, in Poeciliidae and in Balistidae. The PFA model explains all of the continuous traits in these models better than a* $N(0,1)$ *distribution on the standardized traits.*

In addition to improved uncertainty quantification in the block structure of traits, our PFA returns a second independent evolutionary process $\mathbf{F}_2$ that relates pollinator with spur length and, in addition, spur and blade chroma and hue, with posterior probability approaching 1. The existence of two distinct

processes, one of which directly connects pollinator and spur length, sheds additional insight into the original hypothesis that Whittall and Hodges [2007] pose. The LMBD model fails to pick up on this, in addition to returning a worse fit to the data.

3.3.2 Transitions to Placental Reproduction

The freshwater fish *Poeciliidae* represent a family of model organisms in which one can study the transition from non-placental to placental reproduction and the evolutionary pressures associated with placental introduction. Pollux et al. [2014] define a matrotrophy index to be the log-ratio of the dry weight of newborn fish to the dry weight of eggs at fertilization as a proxy measure of how reliant a fish species is on its placenta for reproduction. Using phylogenetic generalized least squares (PGLS) [Ives and Jr., 2010], Pollux et al. [2014] find that *Poeciliidae* dichromatism, courtship behavior, superfetation, and a sexual selection index are all correlated over evolutionary history with the matrotrophy index. Unlike PFA, PGLS as used by Pollux et al. [2014] does not adjust for potential evolutionary relationships between the traits. Failure to do so can lead to false positive measures of association between individual traits and the matrotrophy index.

Pollux et al. [2014] collect from the literature or measure 14 life-history traits and compile from GenBank or sequence 28 different genes across *Poeciliidae* species. In our analysis, we only use $P = 11$ traits since three of the original traits are functions of the included ones. Of these traits, five are discrete-valued: dimorphic coloration (dichromatism), courtship behavior, superfetation, the presence or absence of ornamental display traits and a count composite of the presence or absence of three other male behaviors (sexual selection index). Six are continuous-valued: log weight and log length for males and females, gonopodium length and matrotrophy index. Considering species with at least

one trait measurement, there are $N = 98$ taxa, for which we assume the same fixed phylogenetic tree that Pollux et al. [2014] estimate and similarly condition on in their PGLS analysis. Importantly, 182 trait measurements remain missing. We treat these measurements as missing-at-random in our PFA and do not need to further prune the tree or impute values that may further introduce bias.

Pollux et al. [2014] find that dichromatism, courtship behavior, superfetation, and sexual selection index are all correlated with the matrotrophy index. Figure 3.2 shows that this concurs with the results of a $K = 2$ factor PFA. This small model fit also highlights a weakness of traditional factor analysis assumptions that fix the diagonal elements of the loadings matrix to be positive. In particular, dichromatism is unrelated to the other traits in the second factor, while the positivity constraint would have forced its inclusion. However, the most probable number of independent evolutionary processes is $K = 3$ or $K = 4$, with a log Bayes factor in favor $K = 3$ over $K = 2$ of 35.3 and a log Bayes factor in favor of $K = 4$ over $K = 5$ of 4.9 (Table 3.1). Since a log Bayes factor of only 0.3 separates the $K = 3$ and $K = 4$ models, we include both models in our results, and the data strongly support these PFA models over the LMBD model (log Bayes factor ≈ 92).

Loadings for the independent evolutionary process factors $\mathbf{F}_k^{(3)}$ and $\mathbf{F}_k^{(4)}$ under the $K = 3$ and $K = 4$ PFA models, respectively, recapitulate a negative assocation between the matrotrophy index and dichromatism, courtship behavior, and sexual selection index, and a positive association with superfetation (Figure 3.2, first loading). Uncertainty measures p_{kj} are > 0.94 for all of these trait-factor relationships. However, unlike in Pollux et al. [2014], the PFA does not recover with high posterior probability a relationship between matrotrophy index and gonopodium length nor with body weights and lengths, suggesting that these were false positive findings. For both PFA models, second independent processes $\mathbf{F}_2^{(3)}$ and $\mathbf{F}_2^{(4)}$ drive dichromatism, courtship behavior,

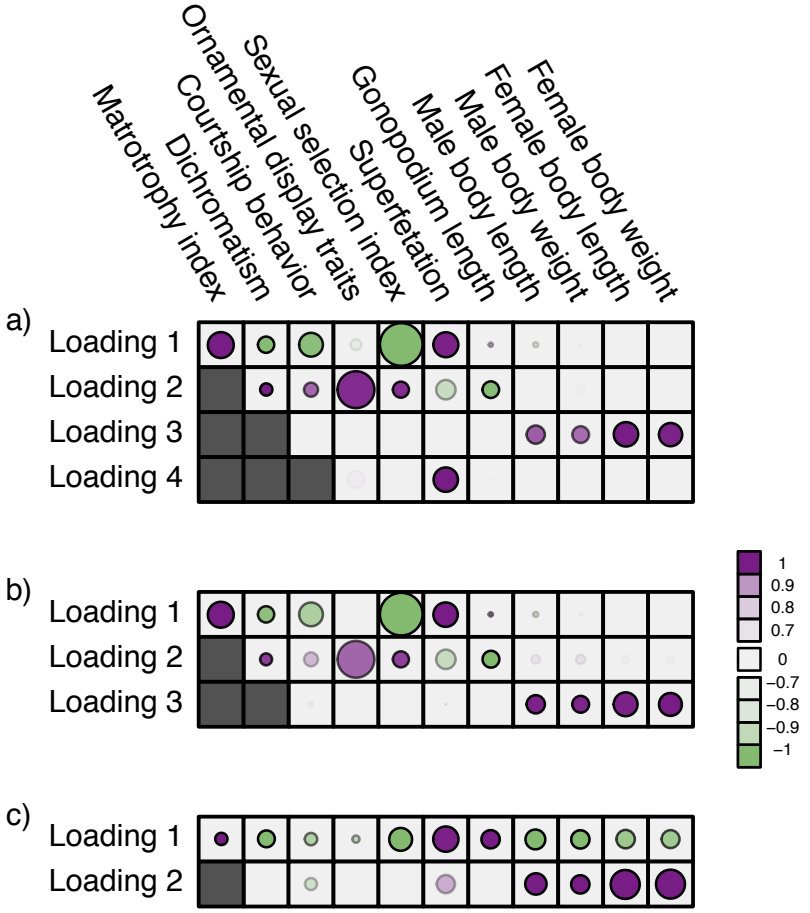

Figure 3.2: *Processes driving transitions to placental reproduction inferred through PFAs. Loading* **L** *estimates from the a)* $K = 4$*, b)* $K = 3$ *and c)* $K = 2$ *factor models. Loadings size, coloring and density follow those of Figure 3.1. Note that the magnitude for dichromatism, courtship behavior, ornamental display traits, sexual selection index and superfetation is less relevant since those data are discrete. We include the two factor model for direct comparison to the results of Pollux et al. [2014]. Loadings in the more probable* $K = 3$ *and* $K = 4$ *factor models do not support an association between matrotrophy index and gonopodium length nor body weights and lengths.*

ornamental display traits and sexual selection index positively and superfetation and gonopodium length negatively, where $p_{2j} > 0.9$ for each of these relationships except involving superfetation ($p_{2j} = 0.88$) and for courtship behavior in $\mathbf{F}_2^{(3)}$ ($p_{2j} = 0.84$). Both models also identify similar third independent processes $\mathbf{F}_3^{(3)}$ and $\mathbf{F}_3^{(4)}$ relating body lengths and weights. We do however find more posterior certainty in the $\mathbf{F}_3^{(3)}$ relationships (all $p_{3j} > 0.99$) than in the $\mathbf{F}_3^{(4)}$ relationships (all $p_{3j} > 0.94$). It is perhaps surprising that these size measurements are unrelated to any of the other reproductive characteristics. The only marked difference between the $K = 3$ and $K = 4$ factor models exists in the presence of a fourth evolutionary process $\mathbf{F}_4^{(4)}$ in the $K = 4$ factor model that controls the presence or absence of superfetation independently of all other traits.

The precision elements $\mathbf{\Lambda}$ for both the $K = 3$ and $K = 4$ factor models are all significantly greater than 1 and therefore indicate that, for both models, our PFA provides good insight into the relationship of the continuous traits (Table 3.2). Further, the precision elements are in broad agreement between the $K = 3$ and $K = 4$ factor models, as we expect due to the negligible difference in marginal likelihoods.

Frequentist-based factor analysis is only identifiable if the number of parameters inferred for a variance/covariance matrix is greater than the number of parameters that need to be inferred for the factor analysis. Interestingly, our PFA model produces interpretable results in spite of the fact that the correlation model has 66 free parameters as opposed to 333 free parameters for the $K = 3$ factor model, and 436 free parameters for the $K = 4$ factor model.

3.3.3 Triggerfish Fin Shape

The fish family *Ballistidae*, commonly know as triggerfish, live mostly in reefs; however, the particular part of the reef in which they live can vary. This variability affects not only their diet, but also their mobility needs that fin shapes well reflect [Dornburg et al., 2011]. To model shape changes through evolution, phylogenetic morphometrics often relies heavily on principle components analysis (PCA) [Revell, 2009, Polly et al., 2013]. However, deterministic data reduction via PCA can introduce bias [Uyeda et al., 2015] and, more importantly, inference of principal components while simultaneously adjusting for an uncertain evolutionary history remains a continuing challenge. PFA offers an alternative approach.

For $N = 24$ triggerfish species, Dornburg et al. [2011] sequence and align 12S (833 nucleotides, nt) and 16S (563 nt) mitochondrial genes and RAG1 (1471 nt), rhodopsin (564 nt) and Tmo4C4 (575 nt) nuclear genes, and Dornburg et al. [2008] digitally photograph and mark 13 semi-landmark Cartesian coordinates for pectoral, dorsal and anal fins, generating $P = 78$ measurements per species. Among these morphometric measurements, the species *Balistapus undulatus* is missing dorsal and anal fins landmarks, and the species *Rhinecanthus assasi* lacks pectoral fin landmarks. For these, we assume the missing data are missing at random.

To accommodate phylogenetic uncertainty within $p(\mathscr{F} \mid \mathbf{S})$, we concatenate gene alignments into $\mathbf{S}$ and model nucleotide sequence substitution along the unknown evolutionary history $\mathscr{F}$ through the Hasegawa et al. [1985] continuous-time Markov chain with unknown transition:transversion rate ratio κ and stationary distribution $\boldsymbol{\pi}$. We incorporate across-site rate variation using a discretized, one-parameter Gamma distribution [Yang, 1994] with unknown shape α and proportion p_{inv} of invariant sites. To specify prior $p(\mathscr{F}, \kappa, \boldsymbol{\pi}, \alpha, p_{\text{inv}})$,

we make relatively uninformative choices, documented in the BEAST extensible markup language (XML) file in the Supplementary Material.

These triggerfish sequences and traits favor the $K = 5$ factor model with a log Bayes factor of 18.5 over the $K = 4$ factor model and 6.9 over the $K = 6$ factor model (Table 3.1). Further, these data favor the $K = 5$ factor model over the multivariate Brownian diffusion (MBD) model with a log Bayes factor of 69.7. Even if this support were equivocal, we caution against using a MBD to model these traits. The unknown variance matrix Σ carries $P(P+1)/2 = 3081$ degrees-of-freedom that dwarfs the $N \times P = 1872$ possible measurements.

For 2 of the 5 factors in the $K = 5$ model, Figure 3.3 demonstrates how fin shape changes as a function of latent factor values. We vary $\mathbf{F}_1$ and $\mathbf{F}_3$ between -1 and 1 that approximates their highest posterior density range over their reconstructed evolutionary history. For $\mathbf{F}_1$, increasing values lead to dorsal and anal fins that become less pointed and more rounded. For $\mathbf{F}_3$, increasing values lead to a counterclockwise rotation of the dorsal fin. Our credible band decreases in size as the factor value gets closer to 0 since the standard deviation of the posterior inference on our loadings is multiplied by these factor values as well.

We also include the corresponding maximum clade credibility (MCC) tree, colored by factor value, with purple representing positive values and green representing negative values for the first factor $\mathbf{F}_1$, and the blue representing positive factor values and orange representing negative factor values for $\mathbf{F}_3$ in figure 3.4. This tree shows us that the species *Balistes polylepsis* and *Balistes vetula*, have negative factor values for $\mathbf{F}_1$, but those species as well as the rest of the clade with the genus *Balistes* and species *Pseudobalistes fuscus* have positive factor values for $\mathbf{F}_3$, whereas the clade containing the genus *Rhinecanthus* has negative factor values for $\mathbf{F}_1$, but a close to 0 factor value for $\mathbf{F}_3$. Conversely, the genus *Xanthichthys* has a negative factor value for $\mathbf{F}_3$, and a closer to 0 factor

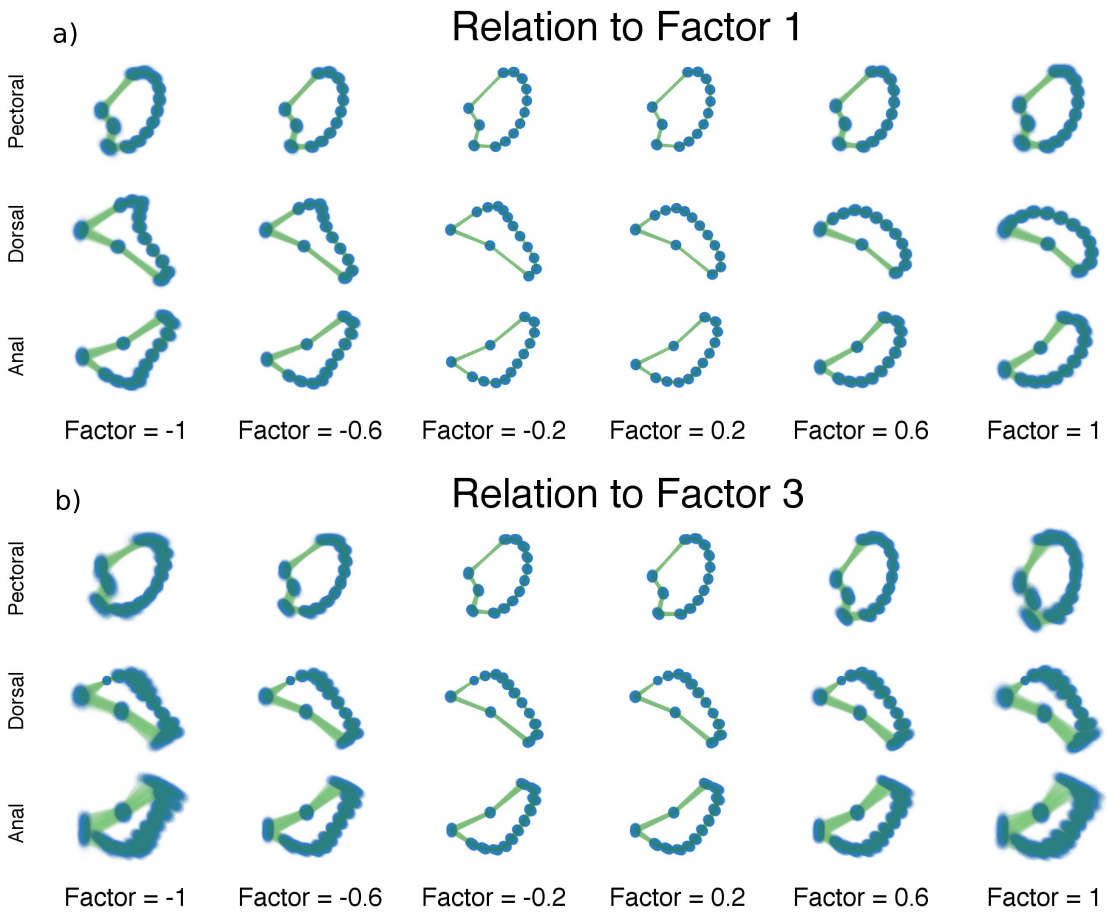

Figure 3.3: *Expected triggerfish fin shape given a range of a) first factor values* $\mathbf{F}_1$ *and b) third factor values* $\mathbf{F}_3$*, holding all others constant. Purple dots estimate semi-landmark locations. Green lines are interpolated to present a clearer outline of the fin shape. For the relation represented by* $\mathbf{F}_1$ *the dorsal and anal fins go from more pointed to less pointed. For the relation represented by* $\mathbf{F}_3$*, we see a rotation in the pectoral fin.*

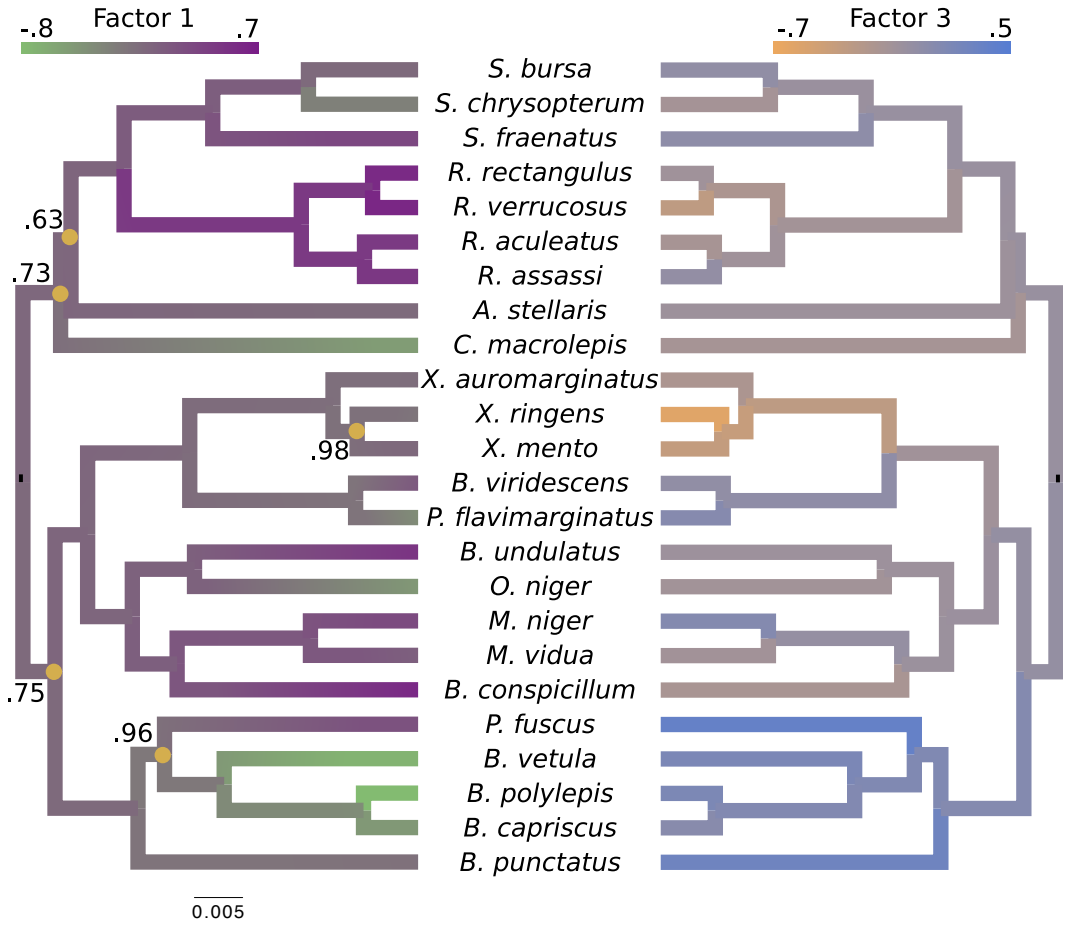

Figure 3.4: *Evolution of independent factors* **F** *driving triggerfish fin morphology along inferred phylogeny. The colorings display contemporary and ancestral first* $\mathbf{F}_1$ *and third* $\mathbf{F}_3$ *factor values under a* $K = 5$ *factor PFA model. For* $\mathbf{F}_1$*, green represents positive values and purple represents negative values. For* $\mathbf{F}_3$*, the scale is orange to blue. Appendix B contains plots for* $\mathbf{F}_2$*,* $\mathbf{F}_4$ *and* $\mathbf{F}_5$*. Balistes polylepis and Balistes vetula have negative factor values for the first factor* $\mathbf{F}_1$*, whereas the clade containing genus Rhinecanthus has positive factor values. In the third factor* $\mathbf{F}_3$*, the Balistes genus and the species Pseudobalistes fuscus have positive factor values whereas the genus Rhinecanthus has near 0 factor values. Conversely, the genus Xanthichthys has a negative factor value for* $\mathbf{F}_3$*, and has a near 0 value for* $\mathbf{F}_1$*. We display the posterior clade probabilities for probabilities* $< 99\%$*.*

value for $\mathbf{F}_1$. We also display posterior clade probabilities for those clades with probability $< 99\%$.

For brevity, we have only considered two factors in this section. We selected $\mathbf{F}_1$ and $\mathbf{F}_3$ since these factors relate distinctive information, however we include the results for the remaining factors in Appendix C. We additionally include our inference on the precision elements in Appendix D as well as our results on the inference on the other tree parameters in Appendix E.

Lastly, PFA facilitates ancestral shape reconstruction. Figure 3.5 depicts inferred pectoral, dorsal and anal fin shapes for ancestors of *Xanthichthys mento* and *Balistes capriscus* at arbitrary points into their evolutionary past. We choose reconstructions at the most recent common ancestor (MRCA) of all 24 species in our study and $1/4$, $1/2$ and $3/4$ of the expected sequence substitution distance between the MRCA and both contemporaneous species. Typically, high aspect ratio fins, or long fins with a small area, are associated with swimming quickly over large distances. The diet of *Xanthichthys mento* consists mostly of plankton and swims above reefs and has a high aspect ratio, perhaps reflecting a need to hunt down more evasive prey. We see that these low aspect ratio dorsal and anal fins arose from a moderate MRCA which flatten as the species evolved. The pectoral fin rotated clockwise as this species evolved. By contrast, *Balistes capriscus* has low aspect ratio dorsal and anal fins, reflecting the fact that it swims more towards the reef floors which may be more useful in navigating the complex habitat. This species evolved from a species with a moderate aspect ratio in its dorsal and anal fins which became broader and more pointed as it evolved. However, the aspect ratio increases again about $3/4$ of the way through its evolution. The pectoral fin rotated counterclockwise as it evolved.

This ancestral reconstruction can provide new insights into the trajectories of shape change that could be further investigated with biomechanical and fluid dynamic models.

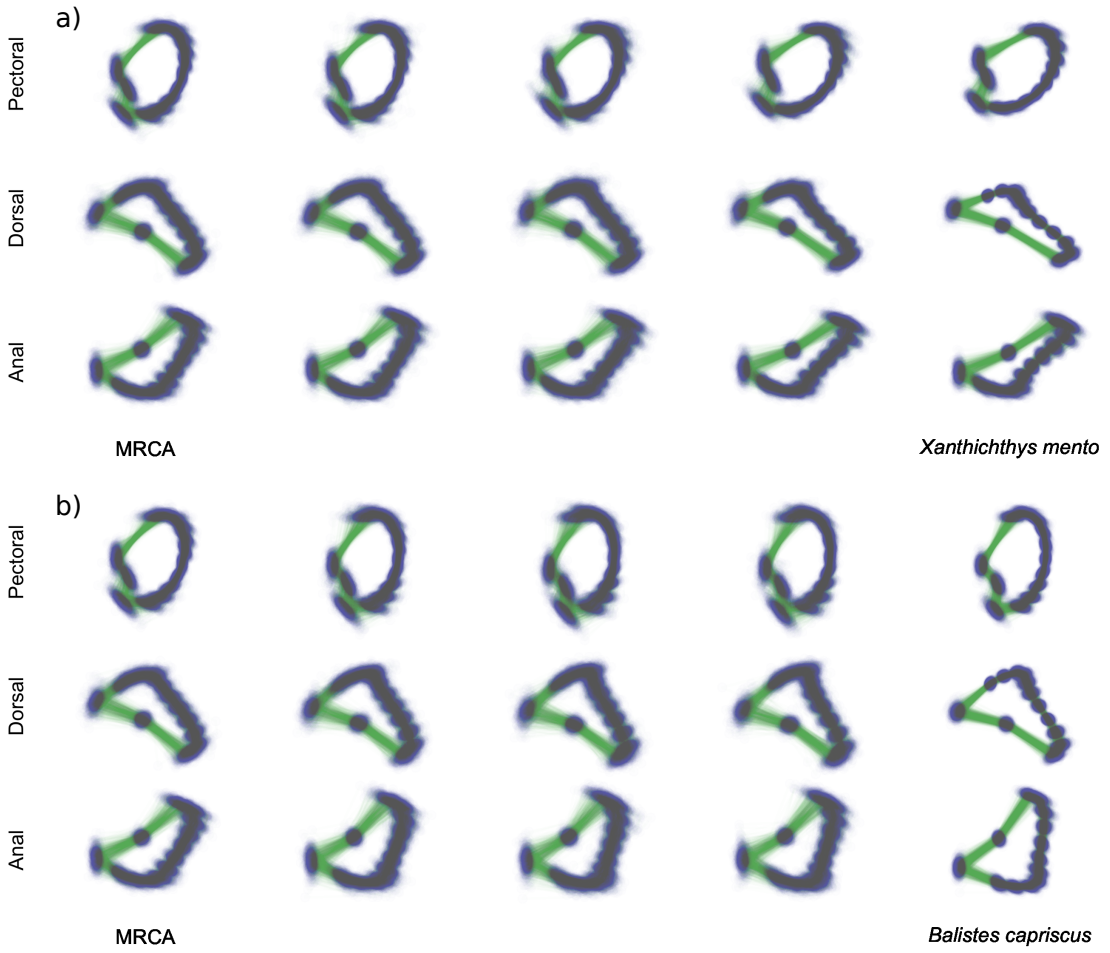

Figure 3.5: *Inferred ancestral fin shapes at the most recent common ancestor (MRCA) and* $1/4$, $1/2$ *and* $3/4$ *of the expected substitution distance between the MRCA and two contemporaneous triggerfish species. In a), Xanthichthys mento has a flat dorsal and anal fin with a point, and a clockwise rotated pectoral fin relative to its ancestors. The dorsal and anal fins become rounder and the pectoral fin rotates counterclockwise moving backwards in time. In contrast, in b), Balistes capriscus has a broad pointed dorsal and anal fin, and a counterclockwise anal fin. The dorsal and anal fins become more pointed and then round out, while the pectoral fin rotates clockwise.*

3.4 Simulation

We briefly evaluate the performance of PFA in estimating the number of factors K and loadings matrix $\mathbf{L}$ under conditions similar to the *Aquilegia* example. Assuming $K = 2$ and fixing $\mathbf{L}$ and $\mathbf{\Lambda}$ to their posterior mean estimates for the $P = 12$ traits and using the fixed tree $\mathscr{T}$ from this example, we simulate 100 dataset replicates under the PFA model. For each replicate, we then perform posterior inference and, collectively, examine our ability to recover the true generative values.

Under these simulation conditions, we recover the true number of factors with relatively high probability (0.89). With the remaining probability, we recover the more parsimonious $K = 1$ model.Averaged across all entries in $\mathbf{L}$, we achieve 79.1% coverage using the 95% highest probability density (HPD) interval estimates; while slightly below nominal, this coverage stands as reasonable in practice, especially since HPD intervals incorporate prior information and, in general, return no frequentist guarantees. We also consider the power to detect an uncertainty measure $p_{kj} > 0.95$ and find that we deem an arbitrary loading entry L_{kj} significant only with probability 0.18. However, the magnitude of the entries in $\mathbf{L}$ vary widely in our examples. Figure 3.10 in Appendix F describes how the true value of L_{kj} influences coverage, power and mean-squared bias. As expected given a prior centered around 0, coverage is lowest for the largest values of $|L_{kj}|$ (e.g. > 1.5) and power increases with increasing $|L_{kj}|$. Importantly for the interpretation of our results, many loadings entries had $p_{kj} > 0.95$, suggesting that their true values were likely larger in magnitude than our posterior estimates under the PFA model.

3.5 Computational Aspects

To draw posterior inference, we simulate MCMC chains of between 200M and 1B steps, subsampling every 10K steps to eliminate unnecessary overhead and ensure the rate-limiting computation remains the PFA and L/MDB transition kernels. For path sampling, we employ 100 path points based on the quantiles of a beta $\beta(0.3, 1)$ random variable [Xie et al., 2011], with warm-started chains of 10M steps at each point. In our examples, the PFA chains generate draws three- to five-fold faster than the L/MBD chains. Further, with the relatively large ratio of latent to non-latent traits in the *Aquilegia* example, we find an approximately 27-fold larger median effective sample size (ESS) across $\mathbf{L}$, $\mathbf{F}$ and γ than in the latent components of $\mathbf{\Sigma}$, demonstrating both faster and more efficient sampling.

3.6 Discussion

This paper merges traditional factor analysis with phylogenetics to provide a new inference tool for comparative studies. The key connection rests on modeling each factor independently as a Brownian diffusion along a phylogeny. The tool we provide not only serves as a dimension reduction technique in the face of high-dimensional traits, but directly addresses the principal scientific questions that many comparative studies raise – specifically, how many independent evolutionary processes are driving these traits? Set in a Bayesian framework, we succeed in inferring these processes for combinations of discrete and continuous traits through model selection, while simultaneously accounting for missing measurements and possible phylogenetic uncertainty.

To make inference under PFA practical, we develop two new MCMC integration techniques. While we rely on previously proposed Gibbs samplers for

integrating the loading matrix $\mathbf{L}$ and residual trait precisions $\mathbf{\Lambda}$, we require an original algorithm based on dynamic programming to integrate the factors $\mathbf{F}$ along the phylogeny efficiently. Second, we extend path sampling through a softening threshold to handle discrete traits, in which their latent support depends on the path location β. Such changing support previously has limited marginal likelihood estimation across many Bayesian models with latent random variables to combine discrete and continuous observations.

In examples involving columbine flower and fish families *Poeciliidae* and *Balistidae* evolution, inference under the PFA is notably quicker under the presence of latent traits, more interpretable and consistently favored via model selection over competing LMBD / MBD models. Interestingly, this success even holds in the *Poeciliidae* example, where one might expect an LMBD model to outperform. Here, the number of parameters inferred in the variance matrix is small relative to the number of parameters that form a PFA. The *Poeciliidae* and *Balistidae* examples also demonstrate our Bayesian approach's ability to integrate missing data if we make a simple missing-at-random assumption.

Unlike many univariate comparative methods, the PFA simultaneously adjusts for correlation between all traits. This advantage reveals that some previously identified trait relationships in *Poeciliidae* evolution may be spurious. Further, as demonstrated in the columbine flower example, the inferred factors and their associated loadings probabilistically cluster traits into independent processes that provide additional scientific insight, often hard to discern from the correlation matrix that a LMBD model provides.

An important computational limitation of PFA arises when the number of taxa N is much greater than the number of traits P. For the PFA, computational cost of our current MCMC integration scales as $\mathcal{O}\left(N^2K + NK^2P\right)$, while the cost is $\mathcal{O}\left(NP^2\right)$ for the LMBD / MBD models. Nonetheless, the *Poeciliidae* example carries $N/P \approx 9$ and, still, the PFA model integrates about $3\times$ more

efficiently due to the example's large ratio of latent traits. For larger N/P ratios, we are currently devising algorithms that remain linear in N as future work.

Arguably, PFA reaches its greatest computational potential when the number of traits stands large relative to the number of taxa – the reputed "large P, small N" setting.This setting arises commonly in the field of geometric morphometrics where very long series of Cartesian, (semi-) land-mark coordinate measurements define the shape of the organism. In our *Balistidae* example, the PFA identifies a number of independent evolutionary processes driving pectoral, dorsal and anal fin shapes. With the help of sequence data, the PFA also simultaneously infers the phylogeny and reconstructs ancestral shapes. We believe that morphometrics stands poised as a prime beneficiary of PFA.

One potential extension of this method comes from Lemey et al. [2010], where they place different diffusion rates on different branches. Additionally we can adapt the methods in Gill et al. [2017] that allow us to incorporate inference on drift in our factors whose direction changes at different points in the evolutionary process. Ornstein-Uhlenbeck processes are nested within the union of both methods that are implemented in BEAST and are therefore easily adapted for use in PFA.

3.7 Appendix

A Phylogenetic factor analysis Gibbs sampling

While the Gibbs samplers for a standard factor analysis are known and well documented [Lopes and West, 2004], there are two aspects of our phylogenetic model that differ sufficiently to require a fresh look at how to draw posterior inference. First, our prior on $\mathbf{F}$ is based on a phylogenetic tree and therefore requires particular consideration in order to produce an efficient Gibbs sampler.

Second, our inference on K uses a path sampling approach where we need to infer $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{\Lambda}$ at each point along the path $q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta})$, and deriving a Gibbs sampler that works for any point in the path β will aid this process.

Sampling factors In a standard Bayesian factor analysis, the prior on each element F_{ij} is $N(0,1)$, and so the entire matrix $\mathbf{F}$ can be Gibbs sampled efficiently in a single step [Lopes and West, 2004]. For the phylogenetic factor analysis model, the prior on the factors is defined by Brownian motion on a phylogenetic tree as defined in (3.6). Thus the conditional density of $\mathbf{F}|\mathbf{Z}, \mathbf{L}, \mathbf{\Lambda}$ in our model is proportional to

$$\begin{aligned} p(\mathbf{Z} \mid \mathbf{F}, \mathbf{L}, \mathbf{\Lambda})\, p(\mathbf{F}) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{Z}-\mathbf{FL})\,\mathbf{\Lambda}\,(\mathbf{Z}-\mathbf{FL})'\right]\right\} \\ \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{FF}'\left(\mathbf{\Psi}+\kappa_0^{-1}\mathbf{J}\right)^{-1}\right]\right\}. \end{aligned} \tag{3.24}$$

This expression does not appear to represent a distribution from which we can easily sample, principally stemming from the fact that $\mathbf{\Lambda}$ is a between-column precision and $\mathbf{\Psi}+\kappa_0^{-1}\mathbf{J}$ is a between-row precision.

Fortunately, Cybis et al. [2015] devise a pre-order tree-traversal algorithm to determine the conditional distribution $\mathbf{F}_{i.}'|\mathbf{F}_{-i.}$ of the factors at a single tip given all other tip values. This distribution is multivariate normal $\mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{F}\text{-}i.}, \mathbf{\Lambda}_{\mathbf{F}\text{-}i.})$ with conditional mean $\boldsymbol{\mu}_{\mathbf{F}\text{-}i.}$ and conditional precision $\mathbf{\Lambda}_{\mathbf{F}\text{-}i.}$. Further, in order to numerically estimate $\mathbf{F}$ at any point along the path $q^*(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta})$, we define

$$q^*(\mathbf{F}_{i.}|\beta, \mathbf{e}_i\mathbf{Z}, \mathbf{F}_{-i.}, \mathbf{L}, \mathbf{\Lambda}) \propto l(\mathbf{e}_i\mathbf{Z}|\mathbf{F}_{i.}, \mathbf{L}, \mathbf{\Lambda})^{\beta}\hat{p}(\mathbf{F}_{i.}|\mathbf{F}_{-i.}). \tag{3.25}$$

Substituting in the appropriate densities and completing the square, we find

that this path is proportional to

$$
\begin{aligned}
& q^* \left(\mathbf{F}_{i.} | \beta, \mathbf{e}_i \, \mathbf{Z}, \mathbf{F}_{-i.}, \mathbf{L}, \mathbf{\Lambda} \right) \\
& \quad \propto \exp \left\{ -\frac{1}{2} \beta \left(\mathbf{e}_i' \mathbf{Z} - \mathbf{F}_{i.} \mathbf{L} \right) \mathbf{\Lambda} \left(\mathbf{e}_i' \mathbf{Z} - \mathbf{F}_{i.} \mathbf{L} \right)' \right\} \\
& \qquad \times \exp \left\{ -\frac{1}{2} \left(\mathbf{F}_{i.}' - \boldsymbol{\mu}_{\mathbf{F}_{\text{-}i.}} \right)' \mathbf{\Lambda}_{\mathbf{F}_{\text{-}i.}} \left(\mathbf{F}_{i.}' - \boldsymbol{\mu}_{\mathbf{F}_{\text{-}i.}} \right) \right\} \\
& \quad \propto \exp \left\{ -\frac{1}{2} \mathbf{F}_{i.} \left(\beta \mathbf{L} \mathbf{\Lambda} \mathbf{L}' + \mathbf{\Lambda}_{\mathbf{F}_{\text{-}i.}} \right) \mathbf{F}_{i.}' - 2 \mathbf{F}_{i.} \left(\beta \mathbf{L} \mathbf{\Lambda} \mathbf{Z}' \mathbf{e}_i + \mathbf{\Lambda}_{\mathbf{F}_{\text{-}i.}} \boldsymbol{\mu}_{\mathbf{F}_{\text{-}i.}} \right) \right\} \\
& \quad \propto \exp \left\{ -\frac{1}{2} \left(\mathbf{F}_{i.}' - \mathbf{M}(\beta)_i^{(\mathbf{F})} \right)' \left(\mathbf{V}(\beta)_i^{(\mathbf{F})} \right)^{-1} \left(\mathbf{F}_{i.}' - \mathbf{M}(\beta)_i^{(\mathbf{F})} \right) \right\},
\end{aligned}
\tag{3.26}
$$

where

$$
\mathbf{M}(\beta)_i^{(\mathbf{F})} = \mathbf{V}(\beta)_i^{(\mathbf{F})} \left(\beta \mathbf{L} \mathbf{\Lambda} \mathbf{Z}' \mathbf{e}_i + \mathbf{\Lambda}_{\mathbf{F}_{\text{-}i.}} \boldsymbol{\mu}_{\mathbf{F}_{\text{-}i.}} \right) \tag{3.27}
$$

and

$$
\mathbf{V}(\beta)_i^{(\mathbf{F})} = \left(\beta \mathbf{L} \mathbf{\Lambda} \mathbf{L}' + \mathbf{\Lambda}_{\mathbf{F}_{\text{-}i.}} \right)^{-1}. \tag{3.28}
$$

Equation (3.26) is proportional to the density of a MVN $\left(\mathbf{M}(\beta)_i^{(\mathbf{F})}, \mathbf{V}(\beta)_i^{(\mathbf{F})} \right)$; therefore, in order to sample $\mathbf{F}$ at a particular point in the path β, we can draw a row $\mathbf{F}_{i.}$ from the distribution MVN $\left(\mathbf{M}(\beta)_i^{(\mathbf{F})}, \mathbf{V}(\beta)_i^{(\mathbf{F})} \right)$.

Sampling loadings The loadings matrix can be Gibbs sampled using the same method described by Lopes and West [2004] with an additional adaptation for use in path sampling. For the examples provided in this paper, we place a $N(0, 1)$ prior on each cell in the loadings matrix; however, in this section we prove the Gibbs Sampler for a generic $N(\mu, \lambda)$ prior. To begin, we again define for a point on the path β,

$$
q^* \left(\mathbf{L} | \beta, \mathbf{Z}, \mathbf{F}, \mathbf{\Lambda}, \mu, \lambda \right) = l(\mathbf{Z} | \mathbf{L}, \mathbf{F}, \mathbf{\Lambda}, \mu, \lambda)^{\beta} \hat{p}(\mathbf{L}). \tag{3.29}
$$

Plugging in the proper values for the sampling density and priors, rearranging and completing the square, we find that

$$
\begin{aligned}
& q^{*}\left(\mathbf{L}|\beta, \mathbf{Z}, \mathbf{F}, \boldsymbol{\Lambda}, \mu, \lambda\right) \\
& \quad \propto \exp\left\{-\frac{1}{2}\beta \mathrm{tr}\left[\left(\mathbf{Z}-\mathbf{F}\mathbf{L}\right)\boldsymbol{\Lambda}\left(\mathbf{Z}-\mathbf{F}\mathbf{L}\right)'\right]\right\} \\
& \qquad \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(\mathbf{L}-\mu\mathbf{1}\right)\lambda\boldsymbol{I}\left(\mathbf{L}-\mu\mathbf{1}\right)'\right]\right\} \\
& \quad \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\beta\mathbf{F}\mathbf{L}\boldsymbol{\Lambda}\mathbf{L}'\mathbf{F}'-2\beta\mathbf{Z}\boldsymbol{\Lambda}\mathbf{L}'\mathbf{F}'+\lambda\mathbf{L}\mathbf{L}'-2\lambda\mu\mathbf{1}\mathbf{L}'\right]\right\} \\
& \quad = \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\beta\mathbf{L}'\mathbf{F}'\mathbf{F}\mathbf{L}\boldsymbol{\Lambda}+\lambda\mathbf{L}'\mathbf{L}-2\left(\beta\boldsymbol{\Lambda}\mathbf{Z}'\mathbf{F}\mathbf{L}+\lambda\mu\mathbf{1}'\mathbf{L}\right)\right]\right\} \\
& \quad \propto \prod_{j=1}^{P}\exp\left\{-\frac{1}{2}\left(\mathbf{L}_{\cdot j}-\mathbf{M}(\beta)_{j}^{(\mathbf{L})}\right)'\left(\mathbf{V}(\beta)_{j}^{(\mathbf{L})}\right)^{-1}\left(\mathbf{L}_{\cdot j}-\mathbf{M}(\beta)_{j}^{(\mathbf{L})}\right)\right\},
\end{aligned}
\tag{3.30}
$$

where $\mathbf{L}_{\cdot j}=\left(L_{1j},\ldots,L_{k'j}\right)$, $\mathbf{1}$ is a matrix of 1's with the same dimensions as $\mathbf{L}$,

$$
\mathbf{M}(\beta)_{j}^{(\mathbf{L})}=\mathbf{V}_{j}^{(\mathbf{L})}\beta\Lambda_{j}\mathbf{F}'_{1:k'}\mathbf{Z}\,\mathbf{e}_{j} \tag{3.31}
$$

and

$$
\mathbf{V}(\beta)_{j}^{(\mathbf{L})}=\left(\beta\Lambda_{j}\mathbf{F}'_{1:k'}\mathbf{F}_{1:k'}+\mathbf{I}_{k'}\right)^{-1}. \tag{3.32}
$$

Hence we find the expression in (3.30) is proportional to a product of independent MVN $\left(\mathbf{M}(\beta)_{j}^{(\mathbf{L})},\mathbf{V}(\beta)_{j}^{(\mathbf{L})}\right)$ densities. Therefore, if we wish to numerically sample a loadings column $\mathbf{L}_{\cdot j}$ at a point on the path β then we can sample from the distribution MVN $\left(\mathbf{M}(\beta)_{j}^{(\mathbf{L})},\mathbf{V}(\beta)_{j}^{(\mathbf{L})}\right)$. Since the densities across columns are independent, we may sample from them in parallel.

Sampling residual precision We wish to sample $\boldsymbol{\Lambda}$ at any point in our path $q^{*}\left(\beta,\mathbf{Y},\mathbf{S},\boldsymbol{\Theta}\right)$. Let $\boldsymbol{\Lambda}_{c}$ be a matrix equivalent to $\boldsymbol{\Lambda}$ with rows and columns corresponding to discrete traits removed. We then say that $\boldsymbol{\Lambda}_{c}=\left(\Lambda_{(1)},\ldots,\boldsymbol{\Lambda}_{(P')}\right)'$

where $\Lambda_{(j)}$ models continuous trait j and P' is the number of continuous traits in our model. If we define $\mathbf{L}_c$ and $\mathbf{Z}_c$ as the matrices $\mathbf{L}$ and $\mathbf{Z}$ with the columns corresponding to discrete traits removed, then we can say $\mathbf{Z}_c \sim \text{MVN}(\mathbf{FL}_c, \mathbf{\Lambda}_c)$. Our prior on $\Lambda_{(j)}$ is i.i.d. for different values of j and has distribution $\Gamma(\alpha_{\mathbf{\Lambda}}, \beta_{\mathbf{\Lambda}})$. For an arbitrary point β in our path $q^*\left(\beta, \mathbf{Y}, \mathbf{S}, \mathbf{\Theta}\right)$, we then define

$$q^*\left(\mathbf{\Lambda}_c | \beta, \mathbf{Z}_c, \mathbf{F}, \mathbf{L}_c\right) \propto l(\mathbf{Z}_c | \mathbf{\Lambda}_c, \mathbf{F}, \mathbf{L}_c)^{\beta} \hat{p}(\mathbf{\Lambda}_c), \tag{3.33}$$

with density

$$\begin{aligned}
&q^*\left(\mathbf{\Lambda}_c | \beta, \mathbf{Z}_c, \mathbf{F}, \mathbf{L}_c\right) \\
&\quad \propto \prod_{j=1}^{P'} \Lambda_{(j)}^{\beta N/2} \times \exp\left\{-\frac{\beta}{2}\left[\mathbf{e}_j' \left(\mathbf{Z}_c - \mathbf{FL}_c\right)' \left(\mathbf{Z}_c - \mathbf{FL}_c\right) \mathbf{e}_j \Lambda_{(j)}\right]\right\} \\
&\qquad \times \prod_{j=1}^{P'} \Lambda_{(j)}^{\alpha_{\mathbf{\Lambda}}-1} \times \exp\left\{-\beta_{\mathbf{\Lambda}} \Lambda_{(j)}\right\} \qquad (3.34) \\
&\quad = \prod_{i}^{P'} \Lambda_{(j)}^{\alpha_{\mathbf{\Lambda}}+\beta N/2-1} \\
&\qquad \times \exp\left\{-\left(\beta_{\mathbf{\Lambda}} + \frac{\beta}{2} \mathbf{e}_j' \left(\mathbf{Z}_c - \mathbf{FL}_c\right)' \left(\mathbf{Z}_c - \mathbf{FL}_c\right) \mathbf{e}_j\right) \Lambda_{(j)}\right\}.
\end{aligned}$$

The expression in (3.34) is proportional to the density of a gamma $\Gamma\left(\alpha_{\mathbf{\Lambda}} + \frac{\beta N}{2}, \beta_{\mathbf{\Lambda}} + \frac{\beta}{2} \mathbf{e}_j' \left(\mathbf{Z} - \mathbf{FL}\right)' \left(\mathbf{Z} - \mathbf{FL}\right) \mathbf{e}_j\right)$ random variable, and therefore we can sample from this gamma distribution in order to sample $\Lambda_{(j)}$ at a given point in the path.

Supplementary figures and tables We perform a phylogenetic factor analysis (PFA) on $N = 24$ triggerfish species with 13 (x, y) coordinate measurements on the pectoral, dorsal and anal fins $(P = 78)$ obtained by Dornburg et al. [2011]. We additionally use 12S (833 nucleotides, nt), and 16S (563 nt) mitochondrial genes and RAG1 (1471 nt), rhodopsin (564 nt) and Tmo4C4 (575 nt) nuclear genes obtained by Dornburg et al. [2008] with a Kingman coalescent prior on the tree topology [Kingman, 1982], an HKY substitution model [Hasegawa et al., 1985] as well as a discretized, one-parameter Gamma distribution with unknown shape and proportion of invariant sites [Yang, 1994]. We settle on a $K = 5$ factor model.

Additionally we ran 100 replicates simulating $\mathbf{F}$ given the fixed phylogenetic tree used in our *Aquilegia* example, and subsequently simulate the data matrix $\mathbf{Y}$ given $\mathbf{F}, \mathbf{L}$ and $\mathbf{\Lambda}$, fixing $\mathbf{L}$ and $\mathbf{\Lambda}$ to the posterior mean estimates of the corresponding matrices from our *Aquilegia* example.

B Remaining loadings plots for triggerfish example

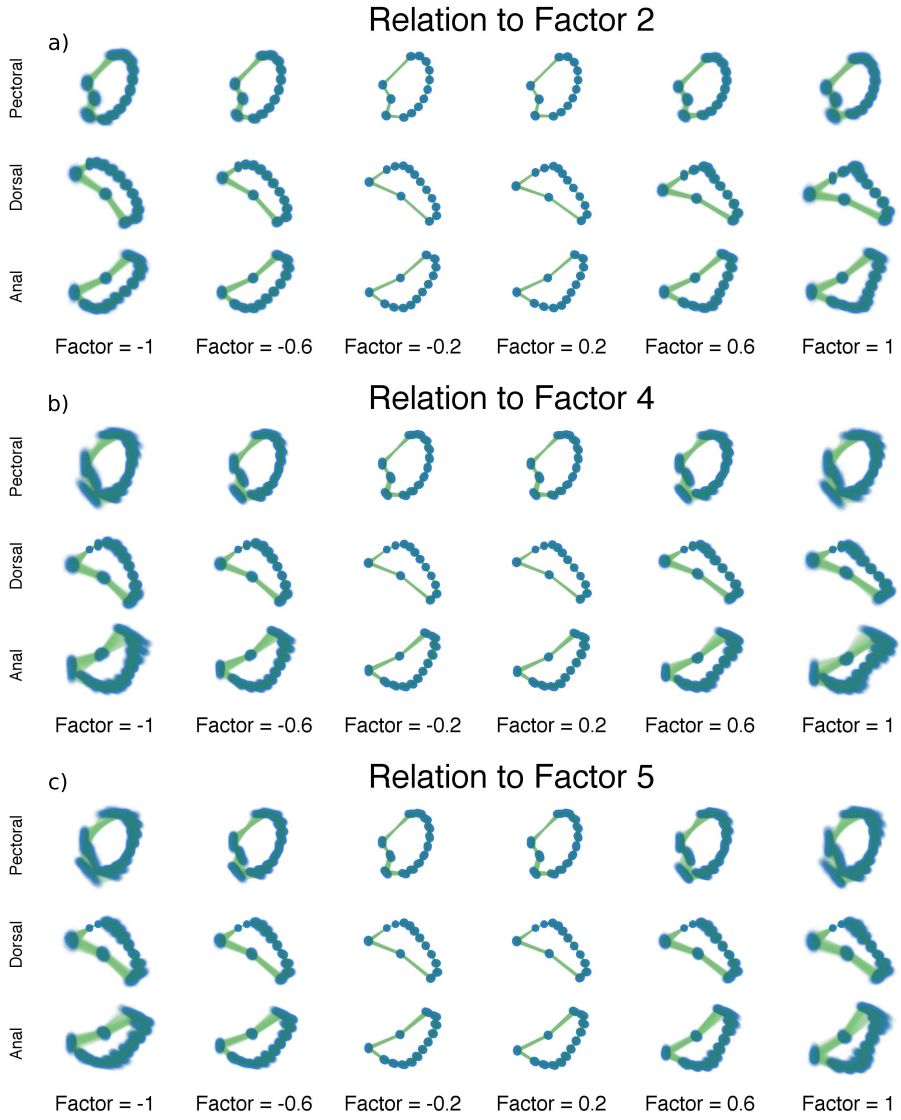

Figure 3.6: *Expected triggerfish fin shape given a range of a)* $\mathbf{F}_2$*, b)* $\mathbf{F}_4$ *and c)* $\mathbf{F}_5$ *values, holding other factor values constant. Purple dots estimate semi-landmark locations. Green lines are interpolated to present a clearer outline of the fin shape.*

C Remaining factor tree plots for triggerfish example

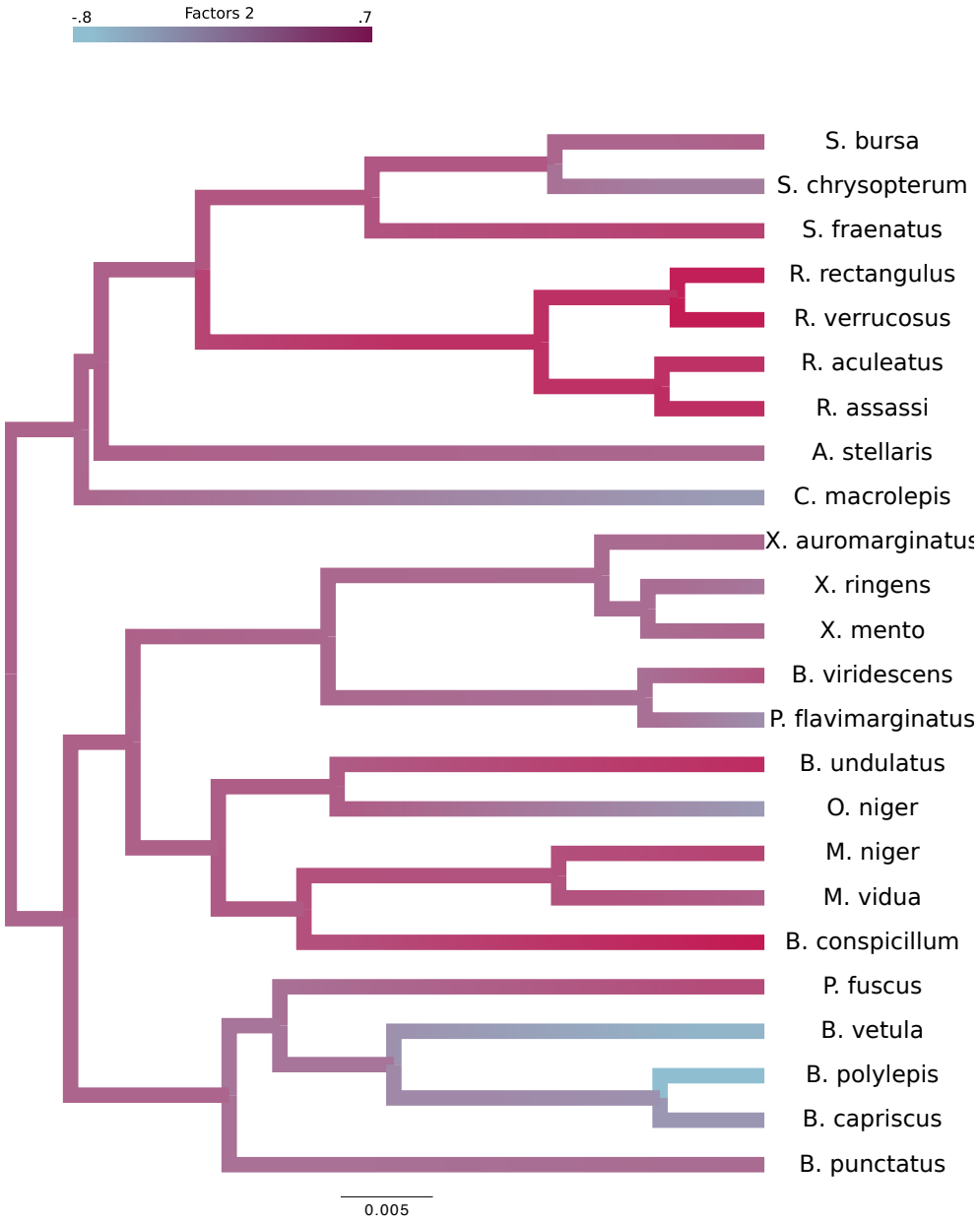

Figure 3.7: *Maximum clade credibility tree for triggerfish species with teal representing a negative factor value and red-purple representing larger factor values.*

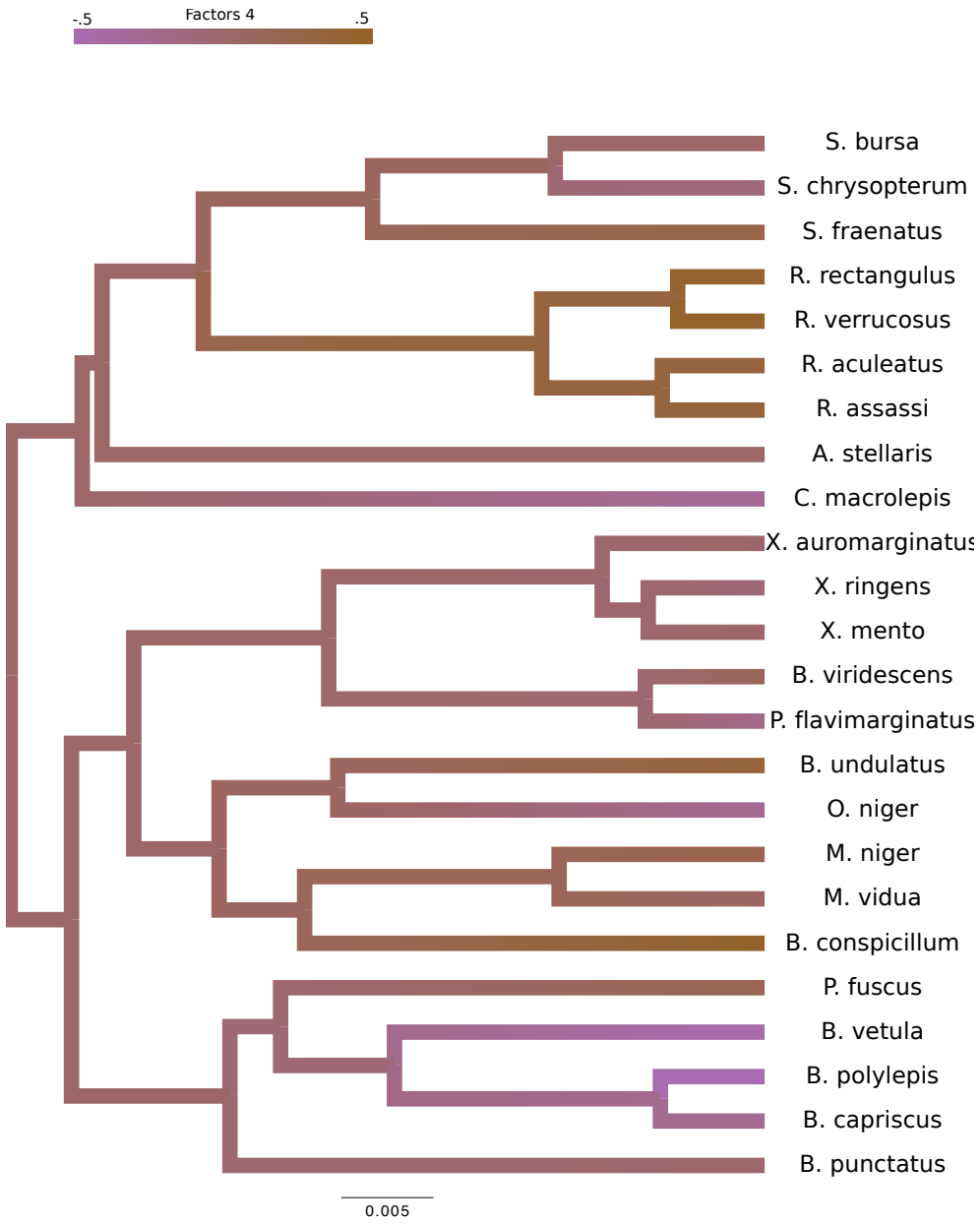

Figure 3.8: *Maximum clade credibility tree for triggerfish species with light purple representing a negative factor value and brown representing larger factor values.*

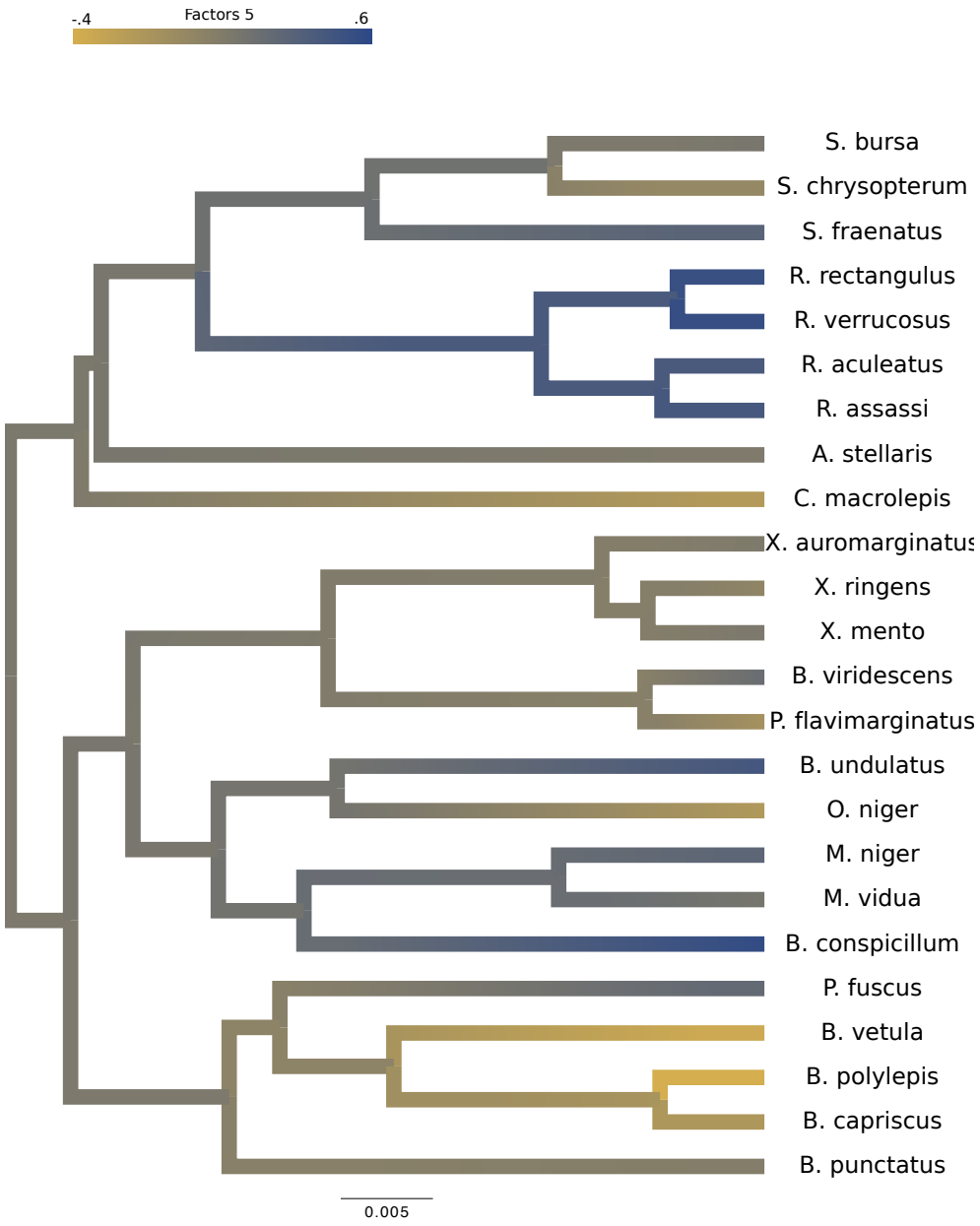

Figure 3.9: *Maximum clade credibility tree for triggerfish species with yellow representing a negative factor value and navy blue representing larger factor values.*

D Triggerfish Fin Precision Elements

Label	X-mean	X-95% BCI	Y-mean	Y-95% BCI
Pectoral Pt. 1	12.90	[5.27, 21.95]	1.13	[0.46, 1.83]
Pectoral Pt. 2	14.29	[5.72, 24.25]	2.29	[0.99, 3.80]
Pectoral Pt. 3	19.34	[7.53, 32.08]	7.14	[2.80, 11.97]
Pectoral Pt. 4	12.27	[4.97, 20.30]	10.84	[3.90, 18.14]
Pectoral Pt. 5	2.84	[1.14, 4.64]	14.86	[5.93, 25.05]
Pectoral Pt. 6	0.95	[0.43, 1.55]	13.43	[5.56, 22.71]
Pectoral Pt. 7	2.43	[1.01, 4.04]	8.86	[3.39, 15.41]
Pectoral Pt. 8	9.52	[3.80, 16.06]	3.93	[1.37, 7.00]
Pectoral Pt. 9	15.20	[6.36, 26.11]	1.80	[0.63, 3.13]
Pectoral Pt. 10	12.07	[4.53, 20.23]	4.92	[1.96, 8.47]
Pectoral Pt. 11	6.07	[2.62, 10.24]	11.00	[4.87, 18.79]
Pectoral Pt. 12	2.75	[1.11, 4.51]	6.22	[2.55, 10.50]
Pectoral Pt. 13	1.09	[0.49, 1.85]	10.86	[4.33, 18.27]
Dorsal Pt. 1	12.56	[4.82, 21.75]	6.55	[1.89, 12.08]
Dorsal Pt. 2	11.26	[3.83, 18.93]	7.30	[2.48, 12.71]
Dorsal Pt. 3	10.89	[3.88, 18.53]	3.69	[1.43, 6.05]
Dorsal Pt. 4	3.64	[1.40, 6.20]	2.83	[1.22, 4.70]
Dorsal Pt. 5	2.18	[0.80, 3.82]	2.46	[1.02, 4.11]
Dorsal Pt. 6	6.38	[2.01, 11.38]	3.24	[1.20, 5.75]
Dorsal Pt. 7	14.76	[5.16, 25.55]	7.55	[2.19, 13.70]
Dorsal Pt. 8	13.62	[5.09, 22.87]	5.33	[1.53, 10.30]
Dorsal Pt. 9	12.12	[4.19, 21.13]	2.89	[1.04, 5.02]
Dorsal Pt. 10	8.62	[2.19, 16.12]	3.15	[1.24, 5.26]
Dorsal Pt. 11	5.21	[1.50, 9.91]	3.70	[1.44, 6.16]
Dorsal Pt. 12	2.43	[0.95, 4.03]	3.86	[1.55, 6.38]

Dorsal Pt. 13	1.99	[0.81, 3.33]	3.37	[1.33, 5.80]
Anal Pt. 1	6.32	[2.44, 10.71]	8.15	[2.86, 14.14]
Anal Pt. 2	8.77	[3.46, 15.15]	7.40	[2.87, 13.12]
Anal Pt. 3	10.49	[3.91, 17.73]	2.28	[0.90, 3.74]
Anal Pt. 4	11.81	[4.37, 20.06]	1.70	[0.74, 2.85]
Anal Pt. 5	4.79	[1.67, 8.26]	3.24	[1.22, 5.57]
Anal Pt. 6	3.01	[1.04, 5.11]	4.04	[1.36, 7.08]
Anal Pt. 7	4.34	[1.77, 7.54]	6.13	[2.00, 11.02]
Anal Pt. 8	6.69	[2.56, 11.28]	9.65	[3.27, 16.94]
Anal Pt. 9	14.89	[5.50, 25.09]	9.95	[3.71, 17.29]
Anal Pt. 10	15.39	[6.22, 26.70]	7.76	[2.88, 13.26]
Anal Pt. 11	1.40	[0.58, 2.35]	4.45	[1.68, 7.52]
Anal Pt. 12	4.20	[1.71, 7.04]	3.24	[1.22, 5.46]
Anal Pt. 13	8.29	[3.12, 14.11]	5.50	[2.15, 9.30]

Table 3.3: *Triggerfish pectoral, dorsal and anal fin precision element posterior mean and 95% Bayesian credible interval (BCI) estimates.*

E Phylogenetic character substitution estimates

Trait	Posterior mean	95% Bayesian credible interval
π_A	0.275	$[0.262, 0.288]$
π_C	0.259	$[0.247, 0.271]$
π_G	0.221	$[0.231, 0.231]$
π_T	0.245	$[0.232, 0.256]$
κ	4.304	$[3.852, 4.816]$
α	0.552	$[0.382, 0.753]$
P_{inv}	0.673	$[0.627, 0.725]$

Table 3.4: *Posterior estimates of HKY substitution model [Hasegawa et al., 1985], discretized Gamma shape α, and proportion of invariant sites P_{inv} [Yang, 1994]. For the HKY model, $(\pi_A, \pi_C, \pi_G, \pi_T)$ represent the nucleotide stationary distribution, and κ represents the rate ratio of transitions to transversions.*

F Simulation study

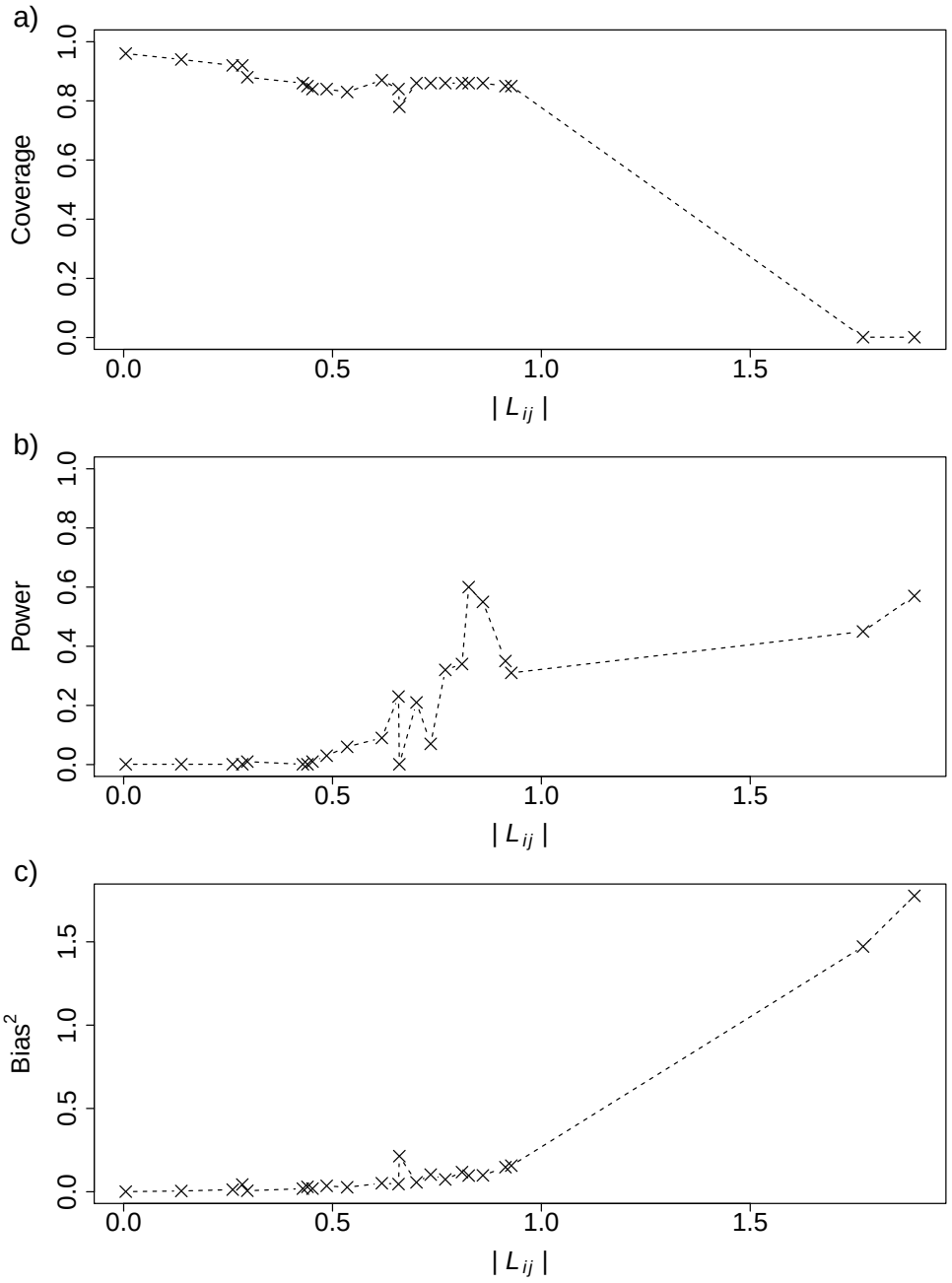

Figure 3.10: *Relationship between loadings matrix element $|L_{kj}|$ and a) coverage of 95% high probability density interval b) power and c) bias^2 from our simulation study with known loadings matrix* **L**.

CHAPTER 4

Bayesian model selection for phylogenetic factor analysis

Phylogenetic factor analysis (PFA) provides a Bayesian approach to characterize the relationships between high-dimensional biological traits measured across multiple organisms, while simultaneously controlling for their uncertain evolutionary history. Inference under PFA can suffer in cases where the number of taxa is not overwhelmingly large, as the posterior distribution across the loadings matrix entries tends to be overly diffuse. To address this shortcoming, we introduce sparsity into the loadings matrix through a shrinkage prior, whose variance collapses as the conditional value of the loadings entries approaches zero. This shrinkage prior consists of a normal prior with a half-Cauchy hyperprior on the variance. This grants a sparse and identifiable loadings matrix that we can efficiently and reliably infer using Markov chain Monte Carlo. We evaluate inference under this PFA shrinkage prior through three examples. In examining *Aquilegia* flower diversity, the sparse model returns a more parsimonious description of the evolutionary processes guiding flower evolution as compared to the original dense model. The sparse model delineates different morphometric evolutionary processes in *Anolis* lizards that associate with different habitats; whereas, the dense model fails to make these determinations. Lastly, we are now able to produce morphometric inference on *Plethodon* salamanders allowing us to more confidently assert the lack of relationship between size and shape on these organisms.

4.1 Introduction

Comparative biology aims to understand the relationships between measured characteristics or traits across a set of organisms or taxa. These traits may display similar patterns across taxa, spuriously owing to their shared evolutionary history, or may co-vary through biologically relevant processes along their history. Therefore, to properly estimate these processes, one must control for the shared evolutionary history through a phylogeny [Felsenstein, 1985]. Several methods that attempt phylogenetic adjustment while estimating trait covariation stand out. These include phylogenetic principal components analysis (PPCA) [Revell, 2009], phylogenetic least squares (PGLS) [Adams, 2014], multivariate Brownian diffusion (MBD) along a phylogeny [Felsenstein, 1985, Huelsenbeck and Rannala, 2003, Lemey et al., 2010] and, recently, phylogenetic factor analysis (PFA) [Tolkoff et al., 2017]. Unfortunately, PPCA suffers from the same limitations as principal components analysis generally, namely that results are not scale-invariant and there is no underlying data generative model from which to easily infer measures of uncertainty on the associations. Further, the estimation methods for PPCA currently necessitate large matrix inversions [Revell, 2009] and, as such, do not scale for large numbers of taxa, specially when entertaining bootstrapped measures of sampling variability. PGLS provides analysis of variance style statistical tests to identify associations, but falls short on measuring pairwise uncertainty between all traits. Finally, like many frequentist approaches in phylogenetics, these two methods experience difficulty in simultaneously incorporating uncertainty about the underlying phylogeny [Suchard et al., 2001].

Readily formulated in a Bayesian framework, MBD and PFA both posit the observed traits as realizations of diffusion processes running along the branches of a possible unknown phylogeny to furnish a full probabilistic data

generate process and simultaneously adjust for shared evolutionary history. MBD asserts a single, multivariate correlated process and attempts to uncover pairwise correlation between traits through the marginal posterior distribution of the off-diagonal entries of the correlation matrix. PFA assumes a small unknown number of independent, univariate evolutionary processes arise along the phylogeny and these factors generate clusters of dependent traits. For large numbers of traits, such as those arising from high-throughput biological experiments, a smaller number of biologically interpretable processes often represent a more parsimonious model than the full multivariate diffusion variance matrix. Model selection using Bayes factors [Jeffreys, 1935] generally favors PFA over MBD, and the multiple independent groupings of traits under PFA can offer more nuance and therefore more explanatory power [Tolkoff et al., 2017]. Despite this improvement, theoretical and practical statistical issues with PFA remain outstanding, particularly as the number of observed traits continues to grow.

Standard factor analysis presumes that each factor influences, via non-zero loadings matrix entries, most traits in a phylogenetic setting. In many biological applications, our inference on each independent evolutionary process is likely to be reflected in only a relatively small number of highly correlated traits when the total number of taxa is small relative to the number of traits. As a consequence, many loadings entries have small point-estimates or diffuse marginal posterior distributions in a Bayesian setting, as Tolkoff et al. [2017] observe. This clouds inference when researchers wish to identify biologically relevant, covarying clusters of traits. We intend to solve this by inducing sparsity in the loadings matrix. However, we wish to avoid the typical pitfalls of sparse factor analyses. Namely, Indian buffet process (IBP) priors [Griffiths and Ghahramani, 2005] on the loading matrix [Knowles and Ghahramani, 2007] tend to produce estimates with a small number of factors which explain

relationships between most of the traits along with a large number of factors which explain only a small number of characteristics [Xu et al., 2016]. An alternative is the determinental point process prior that penalizes sparsity patterns across loading matrix rows that are too similar [Xu et al., 2016]. The resulting sparse posterior distribution is often multi-modal [George and McCulloch, 1993, Ročková and George, 2014, Li and Pati, 2017] and, for large numbers of traits, Markov chain Monte Carlo (MCMC) estimation of the posterior may mix poorly.

Instead, we rely on a shrinkage priors, also known as horseshoe priors [Gelman, 2006, Carvalho et al., 2009]. These distributions, rather than setting firm sparseness measures, instead shrink the estimate of the posterior variance towards 0 when the parameter is near 0. This allows both the benefits of having a sparse distribution, without the drawback of getting stuck in local modes during MCMC-based posterior estimation.

With more biologically realistic, sparsity-inducing priors , we explore the benefits of shrinkage PFA (sPFA) over its dense PFA (dPFA) counterpart in terms of inferring the number of independent processes driving evolution and how these processes relate to trait inference. Our examples cover associations between physical characteristics of columbine flowers and their pollinators, diversification of lizards of the genus *Anolis*, and allometry of *Plethodon* salamanders. We make our open-source software implementing our methods freely available through the BEAST package [Drummond et al., 2012].

4.2 Phylogenetic Factor Analysis

Across a set of N biological entities (taxa), we observe P continuous, binary or ordinal traits $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iP})$ and, if available, a molecular sequence $\mathbf{S}_i$ for each taxon $i = 1, \ldots, N$. We arrange this information into an $N \times P$ trait matrix

$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)'$ and an aligned sequence matrix $\mathbf{S}$. Scientific interest lies in understanding the underlying associations between traits after controlling for the evolutionary history $\mathscr{F}$, informed through $\mathbf{S}$, that relates these taxa.

To model continuous, binary and ordinal traits simultaneously, we first define a partially observed, standardized matrix $\mathbf{Z} = \{Z_{ij}\}$. If trait j is continuous for $j = 1, \ldots, P$, we set $Z_{ij} = \left(Y_{ij} - \hat{\mu}_j\right) / \hat{\sigma}_j$, where $\hat{\mu}_j$ is the sample mean, and $\hat{\sigma}_j$ is the sample standard deviation of trait j. If trait j is discrete, we take Z_{ij} to be an unknown, continuous random variable. Given unknown cutoffs $\boldsymbol{\gamma}_j = \{\gamma_{j0}, \ldots, \gamma_{jm_j}\}$, where m_j is the number of different discrete values trait j realizes, we restrict random Z_{ij} such that $\gamma_{jc-1} < Z_{ij} < \gamma_{jc}$ for observed $Y_{ij} = c$. For identifiability, we fix $\gamma_{j0} = -\infty$, $\gamma_{j1} = 0$, and $\gamma_{jm_j} = \infty$, group all remaining cutoffs into $\boldsymbol{\gamma}$, and assume random cutoffs are *a priori* i.i.d. exponentially distributed with mean $\frac{1}{2}$ to define their density $p(\boldsymbol{\gamma})$.

4.2.1 Phylogenetic Adjustment

Our factor analysis model posits that a small, but unknown number $K \ll \min(N, P)$ of *a priori* independent, univariate Brownian diffusion processes along $\mathscr{F}$ provides a parsimonious description of the covariation in $\mathbf{Z}$. To accomplish this, we first construct an $N \times K$ factor matrix $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_K)$ whose K columns $\mathbf{F}_k = (F_{1k}, \ldots, F_{Nk})'$ for $k = 1, \ldots, K$ identify these unobserved, independent realizations at each of the N tips in $\mathscr{F}$.

The phylogeny $\mathscr{F}$ consists of a topology τ and a series of branch lengths $\mathbf{B}$. The topology τ is a bifurcating directed acyclic graph for which the initiating point, called the root, represents the most recent common ancestor of the observed taxa and the end points, called the tree tips, represent the observed taxa. The branch lengths $\mathbf{B}$ are edge weights that at their simplest represent physical time between bifurcations, but often modulate physical time by relative

rates of evolutionary change. The phylogeny may be known and fixed, or unknown and jointly inferred using $\mathbf{Y}$ and $\mathbf{S}$. The latter requires constructing the sequence-informed prior distribution $p(\mathscr{F}|\mathbf{S})$ and integrating over $\mathscr{F}$; see, e.g., Suchard et al. [2001] or Drummond et al. [2012].

In its typical characterization in evolutionary biology [Felsenstein, 1985], a univariate Brownian diffusion process along $\mathscr{F}$ generates normally distributed increments across each branch with variance equaling its branch length. If we assume that the unobserved root value of the process is also normally distributed with hyperprior mean 0 and variance κ_0^{-1} and match $\mathbf{F}_k$ to the tip realizations, then jointly $\mathbf{F}$ are distributed according to a matrix-normal (MN)

$$\mathbf{F} \sim \mathrm{MN}\Big(\mathbf{0}, \rho^{-1}\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}_N, \mathbf{I}_K\Big)\,, \tag{4.1}$$

where $\rho^{-1}\mathbf{\Psi} + \kappa_0^{-1}\mathbf{J}_N$ is the across-taxa (row) variance, $\mathbf{J}_N$ is the $N \times N$ matrix of ones, and $\mathbf{I}_K$ is the across-traits (column) variance equaling the K-dimensional identity matrix [Vrancken et al., 2015]. The phylogenetic variance matrix $\mathbf{\Psi} = \{\Psi_{ii'}\}$ is a deterministic function of $\mathscr{F}$. Diagonal elements Ψ_{ii} measure the sum of the branch lengths between the root and taxon i and off-diagonal elements $\Psi ii'$ sum the branch lengths between the root and the most recent common ancestor of taxa i and i'. We fix $\rho = \max\{\Psi_{11}, \ldots, \Psi_{NN}\}$ to say that the process undergoes one diffusion unit, in line with our standardization of the continuous traits in $\mathbf{Z}$.

4.2.2 Trait Factorization

We decompose $\mathbf{Z}$ such that

$$\mathbf{Z} = \mathbf{FL} + \boldsymbol{\epsilon} \tag{4.2}$$

where $\mathbf{L}$ is a $K \times P$ loadings matrix and $\boldsymbol{\epsilon}$ is an $N \times P$ residual matrix. We model the residuals via

$$\boldsymbol{\epsilon} \sim \text{MN}\left(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\Lambda}^{-1}\right), \tag{4.3}$$

where $\boldsymbol{\Lambda} = \text{diag}\left(\Lambda_1, \ldots, \Lambda_P\right)$ and, for identifiability, we restrict $\Lambda_j = 1$ if trait j is discrete. Otherwise we assume *a priori* Λ_j are gamma distributed with shape $\alpha_{\boldsymbol{\Lambda}}$ and rate $\beta_{\boldsymbol{\Lambda}}$. In practice, we set $\alpha_{\boldsymbol{\Lambda}} = \beta_{\boldsymbol{\Lambda}} = \frac{1}{3}$.

4.3 Shrinkage priors

We now introduce priors on $\mathbf{L}$ that favor parsimonious models. One of the more popular sparsity-inducing priors for factor analysis arises from the Indian buffet process [Griffiths and Ghahramani, 2005, Knowles and Ghahramani, 2007]. This prior conveniently provides a density on both K and a sparse loadings matrix along with Gibbs sampling approaches to integrate over K as well as sparse loadings components of $\mathbf{L}$ [Griffiths and Ghahramani, 2011]. The IBP prior, however, carries the undesirable feature that the mass of the density is overwhelmingly concentrated towards realizations where either all or none of the random elements within a row of $\mathbf{L}^{(0,1)}$ are 1, signifying model inclusion. The IBP framework also tends to reward structures with many factors and few traits within each element [Xu et al., 2016]. Another option is to use determinental point processes [Xu et al., 2016] mixed with non-local priors [Johnson and Rossell, 2012], however in our experience with this framework, we could not produce consistent inference on this model using MCMC chains with different random seed values.

Instead, we mimic the advantages of a sparse $\mathbf{L}$ by using shrinkage priors [Gelman, 2006, Carvalho et al., 2009]. Shrinkage priors place an i.i.d.

$$L_{kj} \sim N\left(0, \phi_{kj}^2 \tau^2\right), \tag{4.4}$$

distribution on $\mathbf{L}_{kj}$, where ϕ_{kj}^2 is the local component of the variance and τ^2 is the global component of the variance. We place i.i.d. hyper priors on ϕ_{kj} and τ such that

$$\phi_{kj}, \tau \sim \mathcal{C}^+ (0, 1) , \tag{4.5}$$

where $\mathcal{C}^+ (\cdot, \cdot)$ is a half-Cauchy distribution with a centrality parameter and a scale parameter.

Since sampling the posterior of $\boldsymbol{\phi} = (\phi_{11}, \ldots, \phi_{KP})$ and τ is difficult in this form, we rely on the equivalent augmented model [Makalic and Schmidt, 2016] such that

$$\begin{aligned} \phi_{kj}^2 &\sim \text{IG} \left(\frac{1}{2}, \nu_{kj} \right) \\ \tau^2 &\sim \text{IG} \left(\frac{1}{2}, \xi \right) , \end{aligned} \tag{4.6}$$

where $\boldsymbol{\nu} = (\nu_{11}, \ldots, \nu_{KP})$ are augmented local parameters and ξ is an augmented global parameter. Finally, we place i.i.d. priors on ν_{kj} and ξ, such that

$$\nu_{kj}, \xi \sim \Gamma \left(\frac{1}{2}, 1 \right) . \tag{4.7}$$

4.4 Inference

We aim to draw inference on the joint posterior distribution of the factors $\mathbf{F}$, loadings $\mathbf{L}$, column precisions $\boldsymbol{\Lambda}$, and, in theory, evolutionary history $\mathscr{F}$, given

trait measurements $\mathbf{Y}$ and aligned sequences $\mathbf{S}$,

$$\begin{aligned} & p\left(\mathbf{F}, \mathbf{L}, \boldsymbol{\Lambda}, \mathscr{F} \mid K, \mathbf{Y}, \mathbf{S}\right) \\ & \quad \propto p\left(\mathbf{Y} \mid \mathbf{F}, \mathbf{L}, \boldsymbol{\Lambda}\right) \times p\left(\mathbf{F} \mid K, \mathscr{F}\right) \\ & \qquad \times p\left(\mathscr{F} \mid \mathbf{S}\right) \times p\left(\mathbf{L} \mid K\right) \times p\left(\boldsymbol{\Lambda}\right) \\ & \quad = \left(\iint p\left(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma}\right) p\left(\mathbf{Z} \mid \mathbf{F}, \mathbf{L}, \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\gamma}\right) d\mathbf{Z} d\boldsymbol{\gamma} \right) p\left(\mathbf{F} \mid K, \mathscr{F}\right) \\ & \qquad \times p\left(\mathscr{F} \mid \mathbf{S}\right) \times p\left(\mathbf{L} \mid K\right) \times p\left(\boldsymbol{\Lambda}\right), \end{aligned} \tag{4.8}$$

where $p\left(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma}\right) \propto \mathbf{1}(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\gamma})$ is the indicator function that the conditions on $\mathbf{Z}$ from section 4.2 hold, and where

$$p\left(\mathbf{L} \mid K\right) = \iint p\left(\mathbf{L} \mid K, \boldsymbol{\phi}, \tau\right) \iint p\left(\boldsymbol{\phi}, \tau \mid \boldsymbol{\nu}, \xi\right) p\left(\boldsymbol{\nu}, \xi\right) d\boldsymbol{\nu} d\xi \times d\boldsymbol{\phi} d\tau. \tag{4.9}$$

To draw inference on this distribution, we rely on a random-scan Metropolis-within-Gibbs Markov chain Monte Carlo (MCMC) framework [Liu et al., 1995]. The Metropolis-Hastings [Metropolis et al., 1953, Hastings, 1970] portion of this framework is invoked through transition kernels on the evolutionary history $\mathscr{F}$ as well as the cut points $\boldsymbol{\gamma}$ described in Cybis et al. [2015]. To draw inference on Z_{ij} for trait j discrete, we can express $Z_{ij}|\mathbf{L}, \mathbf{F}$ as proportional to $\mathrm{N}\left(0, \mathbf{F}_{i\cdot}\mathbf{L}_{\cdot j}, 1\right) \mathbf{1}\left(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\gamma}\right)$, where $\mathbf{F}_{i\cdot}$ is row i of $\mathbf{F}$ and $\mathbf{L}_{\cdot j}$ is column j of $\mathbf{L}$, and easily simulate from this distribution.

For inference on $\mathbf{F}$, $\mathbf{L}$ and $\boldsymbol{\Lambda}$ we rely on the Gibbs samplers described in Tolkoff et al. [2017]. For inference on ϕ_{kj} we can use a conditional Gibbs sampler [Makalic and Schmidt, 2016] such that

$$\phi_{kj}^2|\tau, \nu_{kj}, L_{kj} \sim \mathrm{IG}\left(1, \nu_{kj} + \frac{L_{kj}^2}{2\tau^2}\right), \tag{4.10}$$

and similarly for τ,

$$\tau^2|\phi, \xi, \mathbf{L} \sim \text{IG}\left(\frac{1+KP}{2}, \xi + \frac{1}{2}\sum_{kj}\frac{L_{kj}^2}{\phi_{kj}^2}\right). \tag{4.11}$$

Additionally, we can Gibbs sample $\boldsymbol{\nu}$ and ξ through the samplers

$$\nu_{kj}|\phi_{kj}^2 \sim \Gamma\left(1, 1 + \frac{1}{\phi_{kj}^2}\right) \tag{4.12}$$

and

$$\xi|\tau^2 \sim \Gamma\left(1, 1 + \frac{1}{\tau^2}\right) \tag{4.13}$$

respectively.

Lastly, we need to draw inference on $K|\mathbf{Y}$. Since these shrinkage priors rely on a Cauchy distribution on the variance, ϕ and τ have no expectation, making MLE inference on models using these priors difficult. Instead, we infer K under the dPFA model described by Tolkoff et al. [2017].

4.5 Examples

4.5.1 Columbine Flowers

The perennial plants of the *Aquilegia* genus bloom across the Northern Hemisphere and their petals display an enormous range of diversity. Whittall and Hodges [2007] explore how different anatomical floral features adapt to changing pollinators. The authors examine $P = 12$ traits consisting of 10 continuous measures (orientation, blade brightness, spur brightness, sepal length, blade length, spur hue, spur length, blade hue, blade chroma, and spur chroma) and 2 discrete measures (the presence/absence of anthocyanins pigment and ordinal pollinator type) across $N = 30$ monophyletic populations from the

genus on a pre-determined (and fixed) phylogeny $\mathscr{F}$. The authors posit that over evolutionary time-scales the mode of pollination tends to progress from bumble bee- to hawkmoth- to hummingbird-mediated and that nectar spur length changes as these plants transition through these pollinators, possibly owing to a co-evolutionary 'race' between spurs and pollinator tongues. Given the data, the MBD model strongly supports the proposed pollinator ordering over alternative orderings [Cybis et al., 2015]. Further, dPFA uncovers $K = 2$ independent processes giving rise the traits [Tolkoff et al., 2017]. Still, some ambiguity lingers around the relative importance of pollinator type in the first process and of blade brightness, sepal length, blade length and anythocyanins presence in the second process.

Mentioned earlier, but worth restating is that we use the same marginal likelihood estimates from Tolkoff et al. [2017] in order to circumvent issues relating to the infinite variance of the half Cauchy prior. Therefore, once again, we favor the $K = 2$ factor model.

We present the results from the sPFA in figure 4.1. The results for $\mathbf{\Lambda}$ are shown in Appendix A. We only show results whose Bonferroni corrected 95% HPD does not contain 0. We find that the first loading, $\mathbf{L}_1$ shows a positive relationship between spur length, spur hue and pollinator type, as displayed by the green circles. This affirms the relationship of interest from Whittall and Hodges [2007]. The second loading, $\mathbf{L}_2$ shows a positive relationship between spur chroma, blade chroma and the presence of anthocyanins pigment, as well as a negative association, as shown in purple, with orientation, blade length, spur brightness and blade brightness.

This contrasts significantly with the results from Tolkoff et al. [2017], which found important relationships between all traits, with the possible exception of pollinator type for one factor, and a positive relationship between pollinator type, spur hue, spur length, blade hue, and blade chroma. We, by contrast,

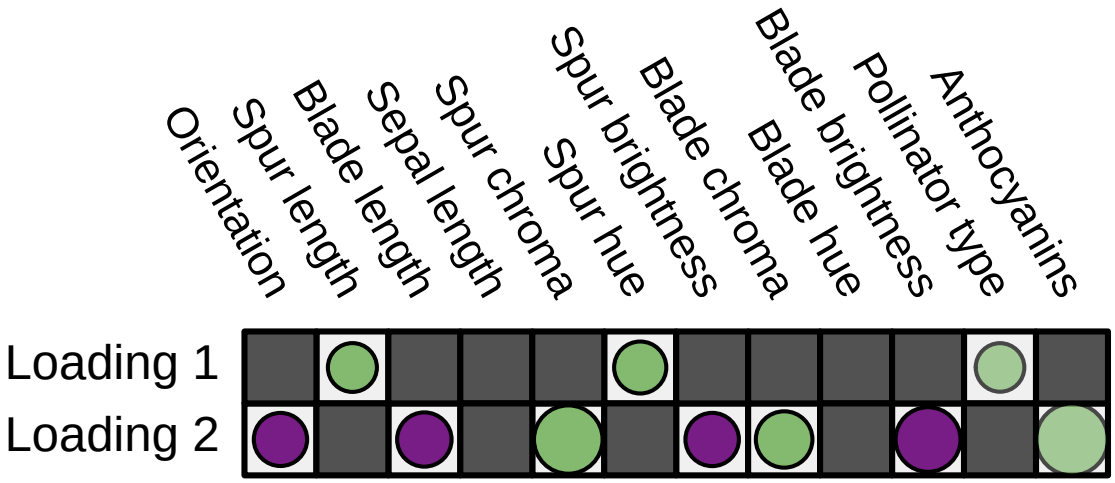

Figure 4.1: *Posterior loadings estimate,* **L***, of the process driving the evolution of Aquilegia flowers under shrinkage phylogenetic factor analysis (sPFA) assumptions. Within a loading, purple circles are positively associated with other purple circles, and negatively associated with green circles, and green circles are positively associated with other green circles. Grayed out cells represent Bonferroni corrected 95% HPD intervals which contain 0. Size represents the magnitude of the inferred loadings values for included elements. Size for discrete traits pollinator type and presence or absence of anthocyanins pigment is less relevant, and therefore we resize these circles to fit the figure, and displayed these traits in a lighter color. We find significantly fewer relationships than found in Tolkoff et al. [2017].* $\mathbf{L}_1$ *preserves the relationship between spur length and pollinator type, as well as the relationship with spur hue.* $\mathbf{L}_2$ *finds a positive relationships between orientation, blade length, spur brightness, and blade brightness, with a negative relationship with spur chroma, blade chroma, and presence and absence of anthocyanins pigment. Notably, sepal length and blade hue are unexplained by evolution in this model.*

find that sepal length and blade hue are poorly explained through evolutionary processes, and spur length, spur hue and pollinator type occupy their own independent evolutionary process, rather than exist to countervail a factor which uniformly effects all measurements. Therefore, it appears that adding sparseness to our modeling assumptions increases the parsimony of our results.

4.5.2 Anoles Lizards

Lizards of the genus *Anolis*, found across South and Central America as well as in the islands of the Greater Antilles region, are often used as a model organism

	Model	MLE
Anolis	$K = 3$	165.38
	$\boxed{K = 4}$	259.03
	$K = 5$	219.75
Plethodon	$K = 4$	-1136.7
	$\boxed{K = 5}$	-1126.7
	$K = 6$	-1131.7

Table 4.1: *Marginal likelihood estimate (MLE) for the Anolis and Plethodon examples. In order to circumvent issues related to the Cauchy distribution used in the shrinkage distribution, we use the phylogenetic factor analysis PFA model described by Tolkoff et al. [2017]. Similarly, the MLE for the Aquilegia example is found in Tolkoff et al. [2017] and favors the $K = 2$ model. The Anolis example favors the $K = 4$ model and the Plethodon example favors the $K = 5$ model.*

to study diversification [Losos and Schneider, 2009]. Even though mainland species have as much species richness as their island cousins [Pinto et al., 2008], the island species are more commonly studied. Mahler et al. [2010] use these island species to study rates and patterns of phenotypic diversification. Since certain characteristics, such as climate [Velasco et al., 2015], affect the species richness of a given niche, we feel that by measuring phenotype relatedness we can gain a clearer understanding of how those niches manifest themselves in terms of the morphometry of *Anolis* lizards.

We rely on the $N = 100$ and $P = 22$ continuous log-transformed morphometric measurements collected by Mahler et al. [2010], as well as the phylogenetic tree which they inferred using mitochondrial sequence data.

We favor the $K = 4$ model (table 4.1) , shown in figure 4.2 over the $K = 3$ and $K = 5$ models with log Bayes factors >93 and >39 respectively. In this example, all relevant traits within a loadings row are positively associated with each other. Once again, we gray out results whose Bonferroni corrected 95%

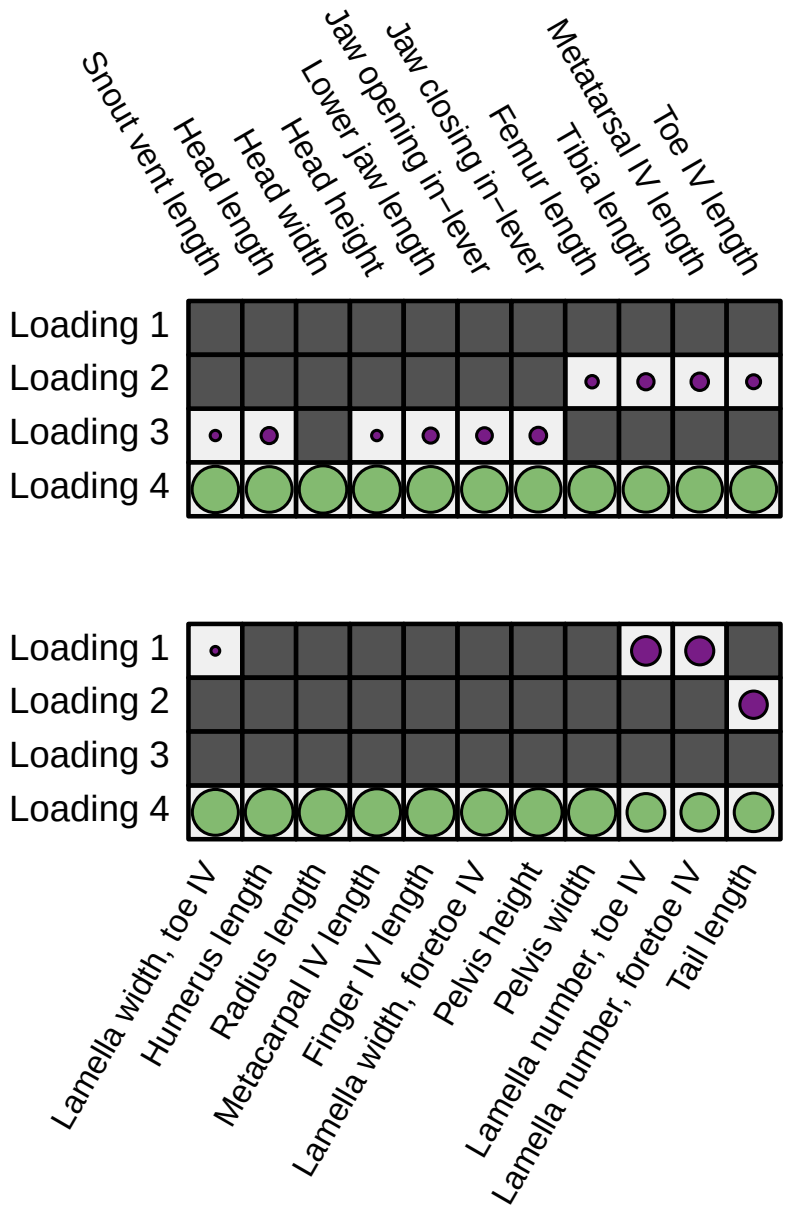

Figure 4.2: *The process driving the evolution of Anolis lizards inferred through the shrinkage phylogenetic factors analysis (sPFA). Colors and sizes follow from figure 4.1. We settle on the $K = 4$ model. The first loading, $\mathbf{L}_1$, represents a positive association between lamella number from toe and foretoe IV, as well as lamella width from toe IV. While lamella width of foretoe IV is excluded from $\mathbf{L}_1$, this may be due to the conservative nature of the Bonferroni correction. $\mathbf{L}_2$ represents leg, toe and tail length. The third loading, $\mathbf{L}_3$, represents head length and height, but not width. Lastly, $\mathbf{L}_4$ represents a general factor of size.*

HPD contains 0. Loading $\mathbf{L}_1$ represents the evolutionary process of the lamella, specifically the lamella number on toe IV and foretoe IV, as well as the lamella width on toe IV. We exclude the lamella width on foretoe IV, although this may be due to the conservative nature of the Bonferroni correction. Loading $\mathbf{L}_2$ represents length for the legs, toes, and the tail. Loading $\mathbf{L}_3$ represents head length and height, as well as associated characteristics such as snout vent length, lower jaw length, jaw opening in-lever length and jaw closing in-lever length. Loading $\mathbf{L}_4$ represents a generic factor of size. The results for $\mathbf{\Lambda}$ are shown in Appendix A. These results leave us to expect that lamella characteristics, leg and tail length, head size (but not width), and overall size are each independent evolutionary factors which may be reflective of how *Anolis* species filled different niches when they migrated to the islands of the Greater Antilles region.

For the dPFA, we find that if we use the identifiability post processing described by Tolkoff et al. [2017], and use the inclusion criteria described in this paper, then each loading only contains a single trait. The sPFA produces more scientifically interesting results than the dPFA, which, in this example produces results which are difficult to draw scientific meaning from.

4.5.3 Plethodon Salamanders

One issue comparative biology concerns itself with is one of allometry, or the relation between size and shape. Adams et al. [2009] look at this question in an attempt to study the diversification rates of *Plethodon* salamanders. Adams [2014] look at the question of allometry more specifically in order to analyze their new PGLS method, as well as to gain a greater understanding of the diversification rates and competitive effectiveness of *Plethodon* salamanders. Adams and Collyer [2015] develop a distance based PGLS (D-PGLS) and compare it to the permutation test (PIC_{rand}) developed by Kingenberg and

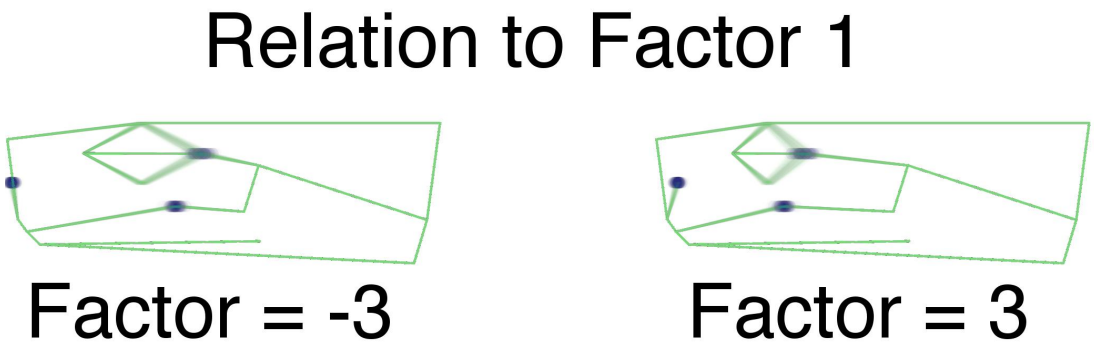

Figure 4.3: *The process driving the evolution of Plethodon salamanders inferred through the shrinkage phylogenetic factors analysis (sPFA) for the first loading* $\mathbf{L}_1$*. Purple dots represent points whose Bonferroni corrected 95% high probability density (HPD) intervals do not contain 0 in either the x or y coordinate. Green lines are interpolated for clarity. We can see that* $\mathbf{L}_1$ *affects eye size, nostril location as well as mouth shape.*

Marugán-Lobón [2013] and find that D-PGLS affirms a significant relationship between shape and size, versus PIC_{rand}, which does not find a significant relationship, and then further go on to show that D-PGLS produces a false positive rate which is closer to the nominal rate.

We use the measurements compiled by Adams and Collyer [2015], which consist of $N = 42$ taxa, an associated phylogeny, 11 location measurements and a snout to vent length which functions as a proxy for head size, for a total of $P = 23$ traits. We favor the $K = 5$ model over the $K = 4$ and $K = 6$ models with log Bayes factors ~ 10 and ~ 5 respectively (table 4.1). We show our results for the first loading $\mathbf{L}_1$ the sPFA in figure 4.3. Our results for $\mathbf{\Lambda}$ are shown in Appendix A, and our results for $\mathbf{L}_2 - \mathbf{L}_5$ are shown in the Appendix B.

We find that there is no relationship between size and shape for any of the five loadings. For $\mathbf{L}_1$ we see a difference in nostril location, mouth shape and eye size. By contrast, the dense model only gives us a single point and a single direction in each loading, specifically the point/direction selected by the algorithm described by Tolkoff et al. [2017].

4.6 Discussion

Our goal is to create a sparse version of the dPFA methodology for comparative biology devised by Tolkoff et al. [2017]. We maintain the general framework from Tolkoff et al. [2017], however we introduce a shrinkage prior in order to induce sparseness on the loadings matrix. This sparseness allows us to cluster the non-zero posterior mass for a given trait in fewer loadings cells. This allows us to produce either more parsimonious results, or allows us to tease apart scientific meaning from results which would otherwise fail to elucidate any trait relationships. We use the examples of *Aquilegia* flowers, *Anolis* lizards and *Plethodon* salamanders to study the efficacy of this method. While none of these examples examine simultaneous inference on the phylogeny, this is easily adapted in the Bayesian framework using the BEAST software. We include the XML code for our examples in the supplementary material.

Up to this point, we have not mentioned the typical identifiability issues associated with factor analysis. These issues arise from the fact that

$$\mathbf{FL} = \mathbf{FT}^t\mathbf{TL}, \tag{4.14}$$

where $\mathbf{T}$ is an arbitrary orthonormal matrix. When we use a sparse $\mathbf{L}$, we constrain our model somewhat, however we are still vulnerable to the situation where

$$\mathbf{FL} = \mathbf{FR}^t\mathbf{RL}, \tag{4.15}$$

where $\mathbf{R}$ is a reflection and rotation matrix. In other words, we can freely rearrange the order of our factors as well as reverse the sign of a loadings row and corresponding factor column without affecting our evaluation of $\mathbf{FL}$. In practice we switch between modes infrequently enough that we can excise a part of the chain, or do post processing to readjust the chain to be in a

single mode when the loadings MCMC chains switch labels. Tolkoff et al. [2017] develops a method to handle the situation where the factor column and loadings row changes sign by negating a draw from a given loadings row based on criterion from how our posterior was sampled. We are able to adapt this with shrinkage priors as well. In the future, we hope to adapt the methods for inducing identifiability in a factor analysis that are described by Holbrook et al. [2016] and Holbrook et al. [2017]

Future model extensions involve missing data and the situation where a large percentage of the traits are discrete. In the former case, we can easily integrate out these values analytically and similarly adapt our sampler. In the latter case we can similarly integrate out these latent values, however Gibbs sampling in this framework becomes difficult. Instead, we can rely on a bouncy particle sampler for sampling latent values [Bouchard-Côte and Vollmer, 2017]. Additionally, future work involves integrating out the latent factors so we no longer have to infer all of the information at the tips, particularly when the number of taxa is large. Lastly, we think it is important to be able to infer the number of factors K under the shrinkage prior model, and also consider this future work.

4.7 Appendix

We perform a sparse phylogenetic factor analysis on *Aquilegia* flowers, *Anolis* lizards and *Plethodon* salamanders with $N = 30, 100,$ and 42 respectively, and $P = 12, 22,$ and 42 respectively.

A Column Precisions

	X-Point Estimate	X-95% Credible Interval	Y-Point Estimate	Y-95% Credible Interval
Pt. 1	9.32	[3.97, 15.59]	5.3868	[2.71, 9.22]
Pt. 2	4.63	[2.01, 7.78]	2.78	[1.29, 4.35]
Pt. 3	6.72	[3.20, 10.96]	5.38	[2.78, 8.41]
Pt. 4	1.33	[0.79, 2.02]	3.50	[1.74, 5.37]
Pt. 5	3.08	[1.67, 4.77]	23.80	[11.35, 37.96]
Pt. 6	8.97	[3.84, 14.97]	22.78	[11.03, 36.34]
Pt. 7	2.97	[1.50, 4.65]	2.75	[1.36, 4.24]
Pt. 8	2.83	[1.41, 4.43]	4.46	[2.24, 7.26]
Pt. 9	4.03	[1.80, 6.81]	2.49	[1.32, 3.76]
Pt. 10	1.90	[1.06, 2.97]	2.70	[1.46, 3.98]
Pt. 11	10.68	[4.28, 18.44]	13.37	[5.27, 22.79]

Table 4.3: *Inference on $\mathbf{\Lambda}$ for the x and y coordinates for the Plethodon salamander morphometrics, along with the associated 95% credible intervals.*

		Point Estimate	95% HPD Interval
Aquilegia	Orientation	2.31	[1.13, 3.72]
	Spur length	5.07	[1.90, 8.84]
	Blade length	3.14	[1.44, 5.12]
	Sepal length	2.69	[1.27, 4.29]
	Spur chroma	4.50	[1.71, 7.92]
	Spur hue	5.66	[1.95, 10.09]
	Spur brightness	2.49	[1.14, 3.94]
	Blade chroma	2.80	[1.13, 4.68]
	Blade hue	1.98	[0.95, 3.10]
	Blade brightness	3.03	[1.46, 5.52]
Anolis	Snout vent length	57.30	[41.75, 74.66]
	Head length	93.81	[66.91, 122.6]
	Head width	53.51	[32.23, 70.02]
	Head height	48.26	[33.64, 62.71]
	Lower jaw length	101.12	[71.95, 130.66]
	Jaw opening in-lever	105.58	[76.71, 138.73]
	Jaw closing in-lever	95.09	[67.99, 123.74]
	Femur length	57.12	[40.15, 74.54]
	Tibia length	63.65	[43.23, 83.80]
	Metatarsal IV length	78.47	[54.40, 103.63]
	Toe IV length	46.46	[32.89, 61.30]
	Lamella width, toe IV	45.48	[31.52, 59.52]
	Humerus length	74.58	[54.13, 98.79]
	Radius length	62.70	[44.67, 83.52]
	Metacarpal length	40.58	[29.05, 53.22]
	Finger IV length	58.96	[41.82, 76.93]
	Lamella width, foretoe IV	43.20	[30.25, 57.23]

	Pelvis height	37.97	[26.97, 49.25]
	Pelvis width	54.57	[39.43, 72.88]
	Lamella number, toe IV	24.46	[15.07, 34.73]
	Lamella number, foretoe IV	25.47	[15.70, 35.76]
	Tail length	10.82	[7.34, 14.56]
Plethodon	Snout-vent length	1.96	[1.10, 2.90]

Table 4.4: *Inference on $\mathbf{\Lambda}$ and associated 95% high probability density (HPD) intervals for the examples of Aquilegia flowers, Anolis lizards, and the snout-vent length for Plethodon salamanders.*

B Plethodon Loadings

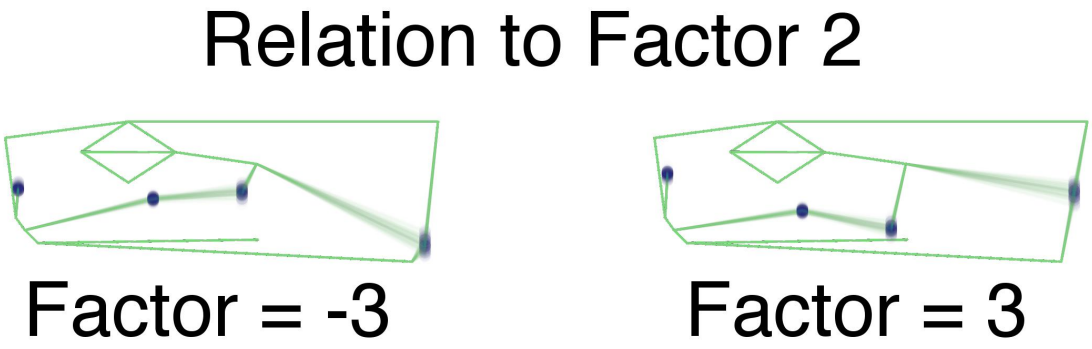

Figure 4.4: *Process driving Plethodon salamanders for the second loading. Purple dots represent points whose Bonferroni corrected 95% high probability density (HPD) intervals do not contain 0 in either the x or y coordinate. Green lines are interpolated for clarity.*

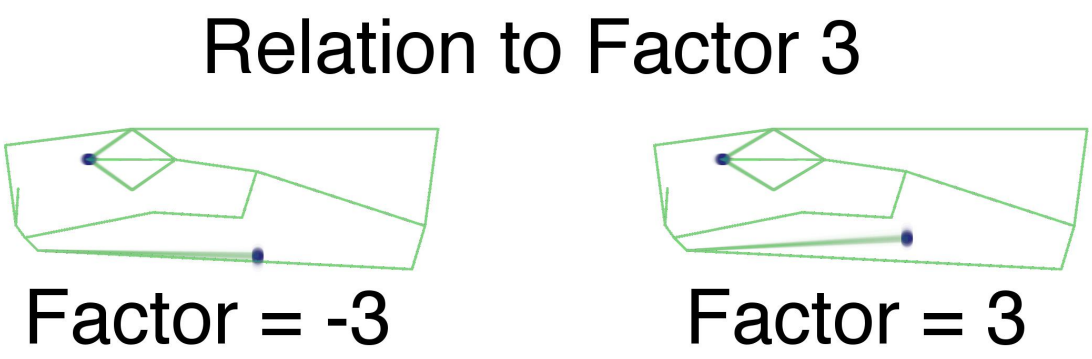

Figure 4.5: *Process driving Plethodon salamanders for the third loading. The colors in the figure follow those from figure 4.4*

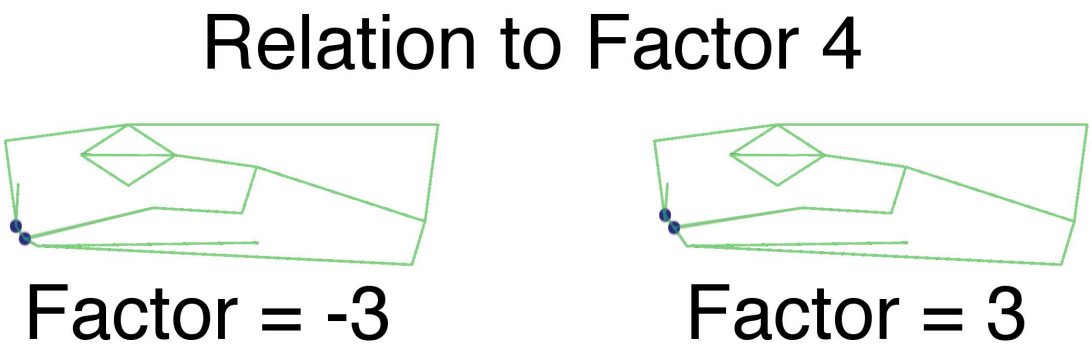

Figure 4.6: *Process driving Plethodon salamanders for the fourth loading. The colors in the figure follow those from figure 4.4*

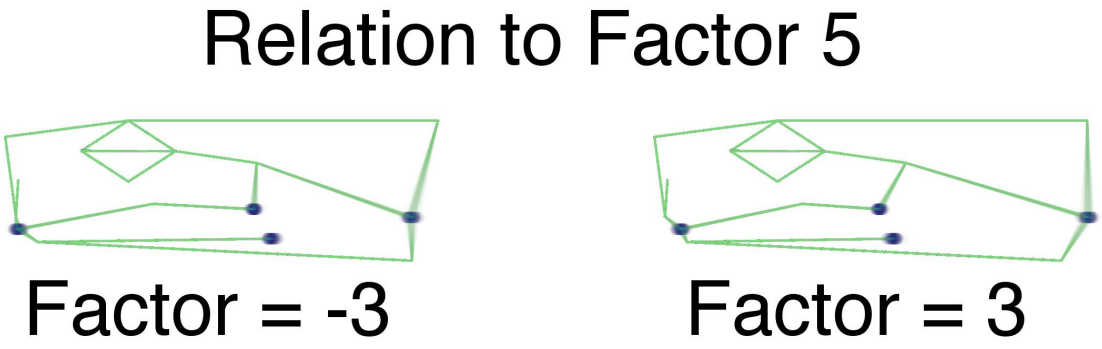

Figure 4.7: *Process driving Plethodon salamanders for the fifth loading. The colors in the figure follow those from figure 4.4*

CHAPTER 5

Phenotypic evolution on large trees with many missing measurements

In comparative biology, we are often interested in learning about the relationships between different characteristics or traits. In order to properly understand these relationships we need to study many different taxa while controlling for the evolutionary history of the organisms.

While there are many methods which can accomplish this, they almost universally scale poorly as the number of taxa increases. The other challenge arises from the fact that obtaining a full suite of measurements becomes increasingly difficult as the number of taxa increases.

Therefore, we propose a model in which we can integrate out missing and latent values analytically and which scales in linear time by using a post-order traversal algorithm based on a multivariate Brownian diffusion (MBD) model. Additionally, we adapt this method to integrate out the factors and missing values in a phylogenetic factor analysis (PFA) model to obtain results which take into account multiple independent processes, account for traits whose variance is not explained through evolution, and scales linearly with the number of traits.

We test these methods using the example of life history covariation in the mammalian class. We find support for the hypothesis that there are species with short reproductive cycles which produce large litters, and vice versa. We

obtain these results $\sim 10000\times$ faster for the MBD model and $\sim 14.8\times$ faster for the PFA model.

5.1 Introduction

In comparative biology, we are often interested in learning about the relationships between different characteristics, or traits of an organism. Felsenstein [1985] showed, however, that in order to properly understand these phenotypic relationships, we must also adjust for the evolutionary history of these organisms. New genetic sequencing methods have allowed for a greater view of the tree of life than ever before, allowing us to incorporate phenotypic measurements from more species than previously possible while properly adjusting for evolutionary history. Analyses incorporating dense taxonomic sampling at large scales creates potential for new research examining very general patterns in phenotypic evolution, key differences between subgroups, and the relationship between macro and micro-evolutionary dynamics.

Unfortunately, many popular techniques are poorly equipped to handle situations where the number of taxa are large. The methods described by Revell [2009] and Adams [2014] scale cubicly with the number of taxa, making them infeasible for large problems. Similarly, the phylogenetic factor analysis (PFA) model described by Tolkoff et al. [2017] scales quadratically with the number of taxa, which is still infeasible.

One of the most difficult challenges that arises as the number of taxa grows large is that obtaining a complete suite of phenotypic data for each species becomes increasingly challenging. Without a method for handling missing data in comparative analyses the consequence is a reduction in the number of species and/or traits that can be analysed such that only a fraction of the data relevant to a problem is utilised. One of the ways to handle missing data is to integrate out the missing values numerically via Markov chain Monte Carlo (MCMC) [Tolkoff et al., 2017]. While this method produces accurate results if the values are missing at random, this method is slow, particularly when the

number of missing traits grows large.

We rely on the multivariate Brownian diffusion (MBD) model described by Lemey et al. [2010] as the basis for our method, since it scales linearly with the number of taxa. This method assesses the between trait variance/covariance matrix of a Brownian diffusion down a phylogenetic tree. In this paper, we develop a novel technique which allows us to integrate out the missing observations analytically, greatly speeding up our analyses. We accomplish this by treating our observed measurements as infinitely precise at the tips, and unobserved measurements with no precision, and propagating these assumptions up the tree.

In addition to observed and unobserved measurements, we sometimes have latent measurements we can integrate out. This is the case with the PFA model from Tolkoff et al. [2017], where we construct latent factors to help us explain trait measurements at the tips of the tree. Using this method of integration, we can integrate out the factors analytically in a way which allows us to infer our model in time linear to the number of taxa, even with a large quantity of missing measurements.

We draw inference on the MBD and PFA using the example of life history covariation across the mammalian class, which is too large to be inferred efficiently with other methods. We find support for the hypothesis that we can group the mammalian class into those which have fast reproductive cycles and produce many offspring, and those with slow reproductive cycles which produce few offspring. The PFA model reproduces these results but also finds an additional independent process which shows that an increase in litter size reduces neonatal body mass and number of litters per year. We were able to obtain these results with a $\sim 10000\times$ speedup in the MBD model, and a less spectacular but still impressive $\sim 14.8\times$ speedup in the PFA model. We make this software freely available in the program BEAST [Drummond et al., 2012].

5.2 Phylogenetic Trait Analysis

5.2.1 Multivariate Brownian Diffusion

When biologists wish to study the phenotype relationships between organisms, it is important to adjust for the shared evolutionary history of these organisms [Felsenstein, 1985]. To do this, we rely on the method described by Lemey et al. [2010], which treats each trait as undergoing Brownian diffusion down a phylogenetic tree and draws inference on the covariance matrix over these traits. In order to adjust for evolutionary history, we need to first define how we represent this history. We define a phylogeny $\mathscr{F}$ in two components, a tree topology τ and a series of branch rates $\mathbf{B}$. τ is a bifurcating directed acyclic graph with ends called tips which represent the observed species, and an originating point called the root which represents the most recent common ancestor of these species. $\mathbf{B}$ are a series of lengths on this graph which represent a function of wall time and rates of evolution. In theory we can follow Suchard et al. [2001] using genetic sequence data to construct a tree simultaneously along with our inference, however since we are considering a large number of taxa in our example, we rely on a fixed tree for our analysis.

We arrange our phenotypic measurements into an $N \times P$ matrix, $\mathbf{Y}$, where N is the number of taxa and P is the number of traits, and standardize $\mathbf{Y}$ by the mean and variance of the measured traits. Following from Felsenstein [1985], we assume that our matrix $\mathbf{Y}$ is generated through a Brownian diffusion process where the traits at the root $\mathbf{Y}_{2N-1}$ are *a priori* assumed to be

$$\mathbf{Y}_{2N-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}\right), \tag{5.1}$$

where $\mathcal{N}(.,.)$ is a multivariate normal distribution with a mean and a variance, $\boldsymbol{\mu}_0$ is the prior mean at the root, which in our example we assume to be 0, $\boldsymbol{\Sigma}$ is

the $P \times P$ variance/covariance matrix detailing the relationships between traits and κ_0 is the prior sample size. We describe this Brownian process across our tree by defining the $N \times N$ tree variance/covariance matrix $\mathbf{\Psi}$ where element ii' is the distance from the root to the most recent common ancestor when $i \neq i'$ and is the distance from the root to tip i when $i = i'$. Following from Vrancken et al. [2015] and using this setup, we can describe the generative process on $\mathbf{Y}$ as

$$\mathbf{Y} \sim \text{MN}\left(\mathbf{J}_N \boldsymbol{\mu}_0', \mathbf{\Psi} + \frac{1}{\kappa_0}\mathbf{J}, \mathbf{\Sigma}\right), \tag{5.2}$$

where MN is a matrix normal with a mean, a between row variance and a between column variance, $\mathbf{J}$ is a conformable matrix of 1's and $\mathbf{J}_N$ is an $N \times 1$ vector of 1's. We place a Wishart$_P(\nu, \mathbf{\Lambda}_0)$ prior on $\mathbf{\Sigma}^{-1}$, with degrees of freedom $\nu = P$ and where $\mathbf{\Lambda}_0$ is a prior matrix which in practice we say is the identity. The algorithm developed by Lemey et al. [2010] allows us to evaluate the likelihood of this distribution in $\mathcal{O}\left(NP^2\right)$ time when we have complete data.

5.2.2 Partially Missing Traits

One of the greatest challenges for inference on large scale problems is the prevalence of missing observations. Despite the ever-increasing availability of trait data, as $\mathbf{Y}$ increases in size, the difficulty in obtaining measurements for each cell increases commensurately. Therefore, in order to analyze problems with large N, it is necessary to have a way to efficiently manage missing observations in our analysis.

Typically, in Bayesian problems, these values are integrated out numerically. The naive way of integrating out these values is to place a random walk transition kernel on the missing values in $\mathbf{Y}$. This naive method can be slow, with run time of $\mathcal{O}\left(NP^2M\right)$, where M is the number of missing values. Cybis et al. [2015] developed an algorithm to compute the conditional distributions

at the tips of a tree given the values at the other tips. By using this method we can draw inference at a pace of $\mathcal{O}\left(N^2P^2\right)$, however when the number of taxa is large, this method is still prohibitively slow.

Our goal is to integrate out these missing values analytically using a dynamic programming algorithm in order to bring our run time down to a much more manageable $\mathcal{O}\left(NP^2\right)$.

5.2.3 Algorithms

Let $\mathbf{Y}^{\text{obs}}$ be the trait values that we can observe and $\mathbf{Y}^{\text{mis}}$ be the missing trait values. Here, we are interested in evaluating the likelihood $p\left(\mathbf{Y}^{\text{obs}} \mid \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0\right)$. We compute the likelihood $p\left(\mathbf{Y}^{\text{obs}} \mid \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0\right)$, using a pre-order traversal and use many of the resulting intermediate calculations to sample $\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0$. Lastly, we use a post-order traversal to sample
$\boldsymbol{\Sigma}^{-1} \mid \mathbf{Y}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0$.

To develop an algorithm to efficiently compute the data likelihood, we first introduce some useful abstractions and notation. At each tip in $\mathscr{T}$, information about each of the P traits comes in one of three forms: a trait value that 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 zero variance (or 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 infinite variance (or zero precision). To formalize this, for tip $i = 1, \ldots, N$, we construct a permutation matrix $\mathbf{Q}_i$ that groups traits in directly observed, latent and

completely missing order and populate a pseudo-precision matrix

$$\mathbf{P}_i = \mathbf{Q}_i \operatorname{diag}\left[\infty\mathbf{I}, \mathbf{R}_i, 0\mathbf{I}\right] \mathbf{Q}_i', \tag{5.3}$$

where $\operatorname{diag}[\cdot]$ is a function that arranges its constituent elements into block-diagonal form, any block may be 0-dimensional depending on the data and $\mathbf{R}_i$ is the latent block precision. 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. After permutation via $\mathbf{Q}_i$, we partition $\mathbf{\Sigma}$ into sub-matrix $\mathbf{\Sigma}_{++}$ that describes the variance between traits with partial information at node i, sub-matrix $\mathbf{\Sigma}_{00}$ for exactly observed traits, sub-matrix $\mathbf{\Sigma}_{\infty\infty}$ for completely missing traits, and their corresponding cross-terms. We additionally define a series of operations which we will find useful for defining this algorithm. We define the pseudo-inverse

$$\mathbf{P}_i^{-} = \mathbf{Q}_i \operatorname{diag}\left[0\mathbf{I}, \mathbf{R}_i^{-1}, \infty\mathbf{I}\right] \mathbf{Q}_i'. \tag{5.4}$$

We define the rank as the number of non-zero singular values and the determinant $\hat{\det}()$ as the product of the non-zero singular values. Lastly, we define $\boldsymbol{\delta}_i$ as a diagonal matrix of indicators whose diagonal elements take the value 1 for traits which are either observed or latent and 0 for traits which are completely missing for node i.

Post-order full precision likelihood: Following from Pybus et al. [2012], we select a node k, condition on its children, i, j and its parent ℓ, and integrate the trait values at this node. This allows us to build a pre-order traversal algorithm to compute the likelihood $p\left(\mathbf{Y}^{\text{obs}} \mid \mathbf{\Sigma}, \mathbf{\Psi}, \boldsymbol{\mu}_0, \kappa_0\right)$. Recursively, for node $k = N+1, \ldots, 2N-1$ and its two daughter nodes i and j, we first compute

branch-deflated pseudo-precisions

$$\begin{aligned}\mathbf{P}_i^\star &= \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \boldsymbol{\Sigma} \boldsymbol{\delta}_i\right)^- \text{ and} \\ \mathbf{P}_j^\star &= \left(\mathbf{P}_j^- + t_j \boldsymbol{\delta}_j \boldsymbol{\Sigma} \boldsymbol{\delta}_j\right)^-\end{aligned} \tag{5.5}$$

along the descent branches. We discuss in detail how to evaluate the pseudo-inverse in Equation 5.5 in Appendix A. Then, we combine information about daughters at the current node via

$$\begin{aligned}\mathbf{P}_k &= \mathbf{P}_i^\star + \mathbf{P}_j^\star \text{ and} \\ \boldsymbol{\delta}_k &= \boldsymbol{\delta}_i \vee \boldsymbol{\delta}_j,\end{aligned} \tag{5.6}$$

where $\vee$ is the element-wise "logical or" operation, and solve for partial mean $\mathbf{m}_k$ as a (not necessarily unique) solution to

$$\mathbf{P}_k \mathbf{m}_k = \mathbf{P}_i^\star \mathbf{m}_i + \mathbf{P}_j^\star \mathbf{m}_j \tag{5.7}$$

using an LU decomposition or singular value decomposition depending on the rank of $\mathbf{P}_k$. Finally, we update integration remainder

$$\begin{aligned}\log r_k = \log r_i + \log r_j + \Delta_{ijk} \log\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{1}{2}\log\left(\frac{\hat{\det}\left(\mathbf{P}_i^\star\right)\hat{\det}\left(\mathbf{P}_j^\star\right)}{\hat{\det}\left(\mathbf{P}_k\right)}\right) \\ - \frac{1}{2}\left[\mathbf{m}_i' \mathbf{P}_i^\star \mathbf{m}_i + \mathbf{m}_j' \mathbf{P}_j^\star \mathbf{m}_j - \mathbf{m}_k' \mathbf{P}_k \mathbf{m}_k\right],\end{aligned} \tag{5.8}$$

where the change of informative dimensions $\Delta_{ijk} = \text{rank}\left(\mathbf{P}_i^\star\right) + \text{rank}\left(\mathbf{P}_j^\star\right) - \text{rank}\left(\mathbf{P}_k\right)$ in the remainder. At the tips, r_i is 1 if all traits are either measured or missing. In cases where some or all of the elements are latent, the remainder at the tips is dependent on the problem.

Let $\mathbf{Y}_k$ be the trait values at node k and $\mathbf{Y}_k^{\text{obs}}$ be the observations restricted to all descendant species of k, then we can compute the likelihood of the tree

from node k downwards using

$$\log p\left(\mathbf{Y}_k^{\text{obs}} \mid \mathbf{Y}_k, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\right) = \log r_k - \frac{1}{2}(\mathbf{Y}_k - \mathbf{m}_k)'\mathbf{P}_k(\mathbf{Y}_k - \mathbf{m}_k) - \frac{1}{2}\text{rank}\left(\mathbf{P}_k\right)\log(2\pi) + \frac{1}{2}\hat{\text{det}}\left(\mathbf{P}_k\right). \quad (5.9)$$

At the root, we can use this distribution, combined with the prior on $\mathbf{Y}_{2N-1}$ from Equation 5.1 to find that

$$\begin{aligned}\log p\left(\mathbf{Y}_{2N-1}^{\text{obs}} \mid \mathbf{Y}_{2N-1}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\right) + \log p\left(\mathbf{Y}_{2N-1}^{\text{obs}} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0\right) = \\ \log r_{\text{full}} + \log\left\{\mathcal{N}\left(\mathbf{Y}_{2N-1}; \left[\kappa_0\boldsymbol{\Sigma}^{-1} + \mathbf{P}_{2N-1}\right]^{-1} \times \right.\right. \\ \left.\left.\left[\kappa_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \mathbf{P}_{2N-1}\mathbf{m}_{2N-1}\right], \left[\kappa_0\boldsymbol{\Sigma}^{-1} + \mathbf{P}_{2N-1}\right]^{-1}\right)\right\} \end{aligned} \quad (5.10)$$

where $\mathcal{N}(x; y, z)$ signifies the multivariate normal density function with argument x, mean y and variance z, and

$$\begin{aligned}\log r_{\text{full}} = \log r_{2N-1} - \frac{1}{2}\left[\boldsymbol{\mu}_0'\kappa_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \mathbf{m}_{2N-1}'\mathbf{P}_{2N-1}\mathbf{m}_{2N-1}\right. \\ - \left(\mathbf{P}_{2N-1}\mathbf{m}_{2N-1} + \kappa_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right)'\left(\mathbf{P}_{2N-1} + \kappa_0\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{P}_{2N-1}\mathbf{m}_{2N-1} + \kappa_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right) \\ \left. + \text{rank}\left(\mathbf{P}_{2N-1}\right)\log(2\pi) + \log\left(\frac{\hat{\text{det}}\left(\mathbf{P}_{2N-1}\right)\hat{\text{det}}\left(\kappa_0\boldsymbol{\Sigma}^{-1}\right)}{\hat{\text{det}}\left(\mathbf{P}_{2N-1} + \kappa_0\boldsymbol{\Sigma}^{-1}\right)}\right)\right]. \end{aligned} \quad (5.11)$$

By integrating out the values $\mathbf{Y}_{2N-1}$ at the root, we find that the log likelihood of the tree is $\log r_{\text{full}}$.

Pre-order missing data augmentation: In the case where all of our data is either measured or missing, and not latent, in order to Gibbs sample $\boldsymbol{\Sigma}$, we wish to compute

$$p\left(\boldsymbol{\Sigma} \mid \mathbf{Y}^{\text{obs}}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0\right) = \int p\left(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0\right) \times p\left(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Lambda}_0\right) \times d\mathbf{Y}^{\text{mis}}. \quad (5.12)$$

In order to integrate out $\mathbf{Y}^{\text{mis}}$ in our analysis, we sample $\mathbf{Y}^{\text{mis}}$ at the tips of the tree before computing the conditional distribution of $\boldsymbol{\Sigma}$. To do this, from Equation 5.10 we see that at the root, we can sample $\mathbf{Y}_{2N-1}$ from the normal distribution with mean $(\kappa_0\boldsymbol{\Sigma}^{-1} + \mathbf{P}_{2N-1})^{-1}(\kappa_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \mathbf{P}_{2N-1}\mathbf{m}_{2N-1})$ and covariance matrix $\kappa_0\boldsymbol{\Sigma}^{-1} + \mathbf{P}_{2N-1}$. For internal node $k = 2N-2, \ldots, N+1$ we can condition on their parent nodes ℓ, and sample $\mathbf{Y}_k$ from the normal distribution with mean $[(t_k\boldsymbol{\Sigma})^{-1} + \mathbf{P}_k]^{-1}[(t_k\boldsymbol{\Sigma})^{-1}\mathbf{Y}_\ell + \mathbf{P}_k\mathbf{m}_k]$ and covariance matrix $(t_k\boldsymbol{\Sigma})^{-1} + \mathbf{P}_k$. At the tips, we sample $\mathbf{Y}_k^{\text{mis}}$ from the normal distribution with mean $\mathbf{Y}_\ell^{\text{mis}} + \boldsymbol{\Sigma}'_{0\infty}\boldsymbol{\Sigma}_{00}^{-1}[\mathbf{Y}_k^{\text{obs}} - \mathbf{Y}_\ell^{\text{obs}}]$ and covariance matrix $\boldsymbol{\Sigma}_{\infty\infty} - \boldsymbol{\Sigma}'_{0\infty}\boldsymbol{\Sigma}_{00}^{-1}\boldsymbol{\Sigma}_{0\infty}$, where $\mathbf{Y}_\ell^{\text{mis}}$ and $\mathbf{Y}_\ell^{\text{obs}}$ are the elements of $\mathbf{Y}_\ell$ which correspond to the missing and observed elements of its child node respectively.

From Equation 5.12 we see that the posterior distribution of $\boldsymbol{\Sigma}$ can be expressed as

$$\boldsymbol{\Sigma}^{-1} \mid \mathbf{Y}, \boldsymbol{\Psi}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0 \sim \text{Wishart}_P\left[\boldsymbol{\Lambda}_0 + \left(\mathbf{Y} - \mathbf{J}_N\boldsymbol{\mu}_0'\right)'\left(\boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}\right)^{-1}\left(\mathbf{Y} - \mathbf{J}_N\boldsymbol{\mu}_0'\right), \nu + N\right]. \quad (5.13)$$

We apply the post-order computation method proposed by Ho and Ané [2014] to compute $(\mathbf{Y} - \mathbf{J}_N\boldsymbol{\mu}_0')'\left(\boldsymbol{\Psi} + \frac{1}{\kappa_0}\mathbf{J}\right)^{-1}(\mathbf{Y} - \mathbf{J}_N\boldsymbol{\mu}_0')$. The computational complexity of this method is $\mathcal{O}(NP^2)$.

5.3 Phylogenetic Factors Analysis

The algorithms described in the previous section can easily be adapted for other purposes in models which rely on Brownian diffusion. As an example, we will look at the phylogenetic factor analysis method described by Tolkoff

et al. [2017]. We assume that

$$\mathbf{Y} = \mathbf{FL} + \epsilon, \tag{5.14}$$

where $\mathbf{F}$ is a $N \times K$ factor matrix, $\mathbf{L}$ is a $K \times P$ upper triangular loadings matrix, and

$$\epsilon \sim \text{MN}\left(0, \mathbf{I}, \mathbf{\Lambda}^{-1}\right), \tag{5.15}$$

where $\mathbf{\Lambda}$ is a diagonal precision matrix. We place an i.i.d. $\text{N}(0, 1)$ prior on the non-zero cells of $\mathbf{L}$ and i.i.d $\Gamma(1, 1)$ prior on the diagonal elements of $\mathbf{\Lambda}$.

We incorporate Brownian diffusion in this model through $\mathbf{F}$ by making the assumption that $\mathbf{F}$ follows the form of Equation 5.2, with $\mathbf{\Sigma} = \mathbf{I}$ and is therefore a composed entirely of latent values. Using the algorithm described in the previous section, we can modify our starting assumptions by treating the factors as latent unobserved elements, as described in Appendix B. We can then use this form to integrate out the latent factors analytically.

5.4 Mammalian Life History

A major task for life history theory is to understand the ecological and evolutionary significance of patterns of covariation between life history traits such as age at sexual maturity, the number of offspring produced at each reproductive event, and reproductive lifespan [Roff, 2002]. This requires establishing exactly what these patterns are. Currently, the dominant hypothesis is that most trait variation occursb on a fast-slow axis [Reynolds, 2003]. Archetypical fast organisms follow a strategy that invests in current reproduction by possessing a suite of traits such as early maturity, large broods of small offspring, and frequent reproduction over a short lifespan, whereas slow species invest in future reproduction by having the opposite pattern of traits. Numerous

investigations have used comparative life history data to investigate whether a fast-slow axis is the primary dimension of trait covariation at the species level (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]), with mixed support for the hypothesis. This may reflect important taxonomic differences in life history evolution, but there is concern that differences are a consequence of different methodological approaches and decisions taken by researchers about how to interpret patterns [Jeschke and Kokko, 2009].

One key issue is that to date the methods for analyzing patterns of covariation between multiple life history traits have required complete data for each species. As complete estimates 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. Reducing the number of traits risks missing important complex structure in covariance patterns, while reducing the number of species reduces ability to detect structure. The MBD and PFA models, by integrating out missing values, provide a solution to this issue. Here we present a reanalysis of Bielby et al. [2007], who analyzed covariation of mammalian life history data taken from an early version of the PanTHERIA database [Jones et al., 2009] using standard factor analysis on (a) body size corrected residuals and (b) phylogenetic independent contrasts [Felsenstein, 1985]. Following data cleaning, this resulted in analysis of 267 species with complete data on six traits. Results suggested that mammalian life history variation lay on two separate axes; the first described the timing of reproduction, running from early maturing, frequently reproducing and early weaning species at one end, to those with the opposite pattern at the other. The second axis described reproductive output per reproductive event, ranging from species with small

litters of large offspring and long gestation periods, to those with large litters of small offspring with short gestation times.

Here we 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 data from Ernest [2003] and additional sources Capellini et al. [2015]. The 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 also include female body mass in the PFA analysis rather than analyze size corrected residuals, as doing so is known to introduce biases in analysis of covarying data [Freckleton, 2009]. All traits are log10 transformed prior to analysis to normalize trait distributions. The analysis uses the phylogeny of Fritz et al. [2009], which remains the most complete phylogeny for mammals. In total, 3690 species in the phylogeny have data on at least one trait and are included in analyses. The number of species with data on each trait is: body mass = 3508 (4.93% missing); gestation length = 1427 (61.33% missing); weaning age = 1253 (66.04% missing); neonatal body mass = 1108 (69.97% missing); litter size = 2538 (31.22% missing); litter frequency = 1231 (66.64% missing); age at first birth = 945 (74.39% missing) and reproductive lifespan = 748 (79.73% missing), for a total of 12758 data points. 518 species have data on all 8 traits, thus the ability to include species with partially missing traits enables inclusion of 208% more data points.

For the MBD model our results are shown in figure 5.1. We color this such that red is positive correlation, blue is negative correlation, and circle size reflects magnitude of the correlation. We shade our results based on the posterior probability of the parameter being of the same sign as the posterior mean, adjusted to be on a 0 to 1 scale.

We find that we can break these traits into two different categories. The first

is a positively correlated group of body mass, gestation length, weaning age, neonatal body mass, age at first birth and reproductive lifespan. The second group is a positive correlation between litter size and litters per year, and these two groups are anti-correlated with each other. This supports the hypothesis of the fast species which produce many offspring quickly, and slow species which produce few offspring more slowly. While larger species tend to be on the fast track, this relationship is generally rather weak.

For the PFA model, we favor the full $K = 8$ model with log Bayes factors of 51.4 over the $K = 7$ model. This is in contrast with the > 5000 log Bayes factors in favor of the factor model over the MBD model. Our results for PFA are also shown in figure 5.1. Purple circles are positively associated with other purple circles within a loading, and are negatively associated with green circles. Similarly, green circles are positively associated with other green circles. Size represents magnitude. We once again shade based on the posterior probability of the parameter having the same sign as the posterior mean, adjusted to be on a 0 to 1 scale. We find that loadings 1 through 3 reaffirm the MBD model with the exception of the structural 0's. Loading 4 finds a positive relationship between neonatal body mass and litters per year, and an anti-association with litter size. Loading 5 finds a weak anti association between litters per year and litter size, age at first birth and reproductive lifespan.

Inference on $\mathbf{\Lambda}$ is shown in table 5.1. We say that those traits whose 95% high probability density intervals do not contain 1 are traits which are explained by evolution. This is true of all of the traits in this analysis. We will emphasize that $\mathbf{\Lambda}$ is a model variance, and is therefore in theory not affected by the large sample size.

The key advantage to this method is the ability to obtain these results faster than before. We compare this integrated method with a simple random walk transition kernel on the missing values, with the likelihood evaluated using

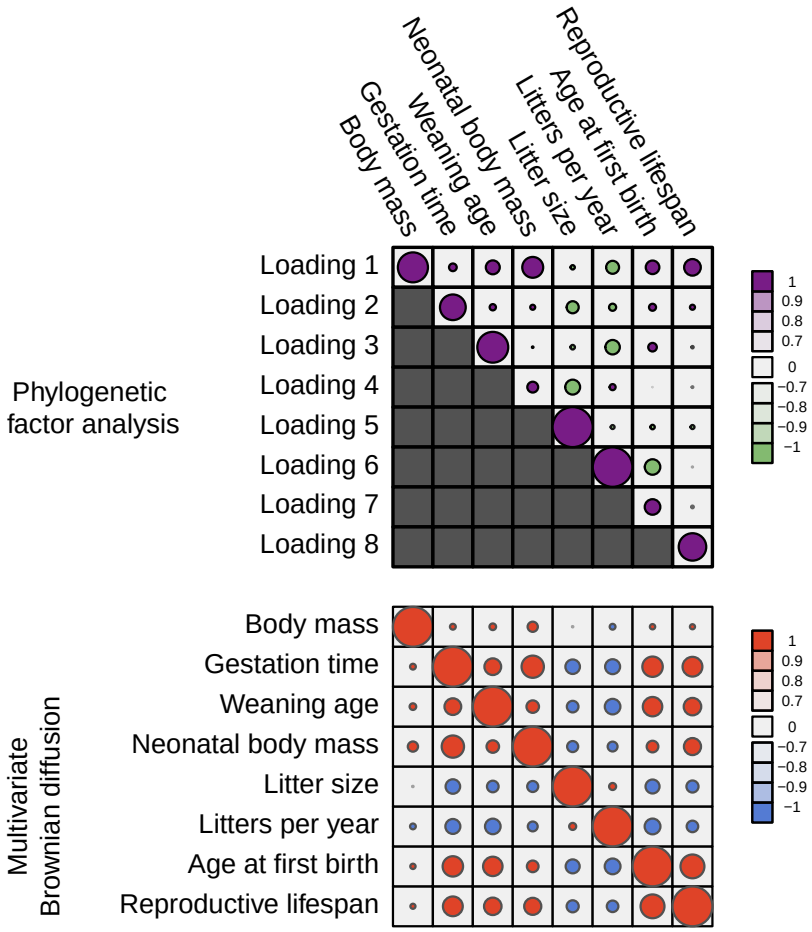

Figure 5.1: *Processes driving evolution under assumptions for the multivariate Brownian diffusion (MBD), and the $K = 8$ model for the phylogenetic factor analysis (PFA). For the MBD model, red represents positive correlation, and blue represents negative correlation. Size represents the estimate of the posterior mean of the correlation between these traits. For PFA, purple traits are positively associated with other purple traits within a loading, and negatively associated with green traits. Similarly, green traits are positively associated with green traits. Size represents magnitude. We shade these cells by the posterior probability that the parameter value of the cell is of the same sign as the posterior mean, adjusted to be on a 0, 1 scale. In the MBD model we find that we can separate the traits into two groups. Body mass, gestation length, weaning age, neonatal body mass, age at first birth and reproductive lifespan are all positively correlated with each other. Additionally, litter size and litters per year are positively correlated with each other, however these two groups are anti-correlated with each other. In the PFA model, we recover these relationships in the first three loadings with the exception of structural 0's included. Loading 4 shows a relationship between neonatal body mass and litters per year, but a negative relationship with litter size. Loading 5 also shows a weak negative relationship with litter size and the remaining traits.*

	Point Estimate	95% HPD Interval
Body mass	153.9	[132.5, 180.6]
Gestation time	91.6	[73.1, 110.5]
Weaning age	10.4	[9.2, 12.0]
Neonatal body mass	105.3	[87.4, 120.8]
Litter size	17.7	[15.0, 20.3]
Litters per year	6.9	[5.6, 7.9]
Age at first birth	8.9	[7.9, 10.2]
Reproductive lifespan	8.1	[6.6, 9.6]

Table 5.1: *Inferred value of $\mathbf{\Lambda}$ for the mammals example. We present both point estimate and 95% high probability density interval (HPD). We say that an HPD interval that does not contain 1, which is true of each trait, indicates a significant influence of evolution on these traits.*

the method described by Pybus et al. [2012]. We find that the median effective sample size per hour for our analytically integrated method is 16938.8 versus 1.8 for the numerically integrated method resulting in a speedup of almost 5 orders of magnitude. Similarly, for the PFA method, we use the Gibbs samplers described by Tolkoff et al. [2017] with a simple random walk transition kernel on each of the missing values. While the speedup for the integrated PFA method versus the method described by Tolkoff et al. [2017] is not as dramatic, we still find an improvement to 26.9 effective samples per hour versus 1.8 samples per hour, a speedup of about 14.8 times.

5.5 Discussion

Oftentimes we 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 trait relationships given a tree while integrating out missing values analytically. We compare the speed of these integrated methods with the traditional method of integrating out these values numerically using the example of mammals.

This along with an improved Gibbs sampler allows us to draw inference both a MBD model $\sim 10000\times$ more quickly. Additionally, we are able to integrate out the missing values and factors in a PFA model allowing us to draw inference on the PFA model $\sim 14.8\times$ faster.

In Tolkoff et al. [2017] and Cybis et al. [2015], discrete values were included in the analysis using a latent probit model. While traditional methods of inferring these latent values were too slow to be included in this analysis, we still believe that it is important to be able to consider discrete data in these methods. One candidate method for exploring these latent values is Hamiltonian Monte Carlo (HMC) [Neal, 2010]. Additionally, we may be able to use HMC to infer the values of $\mathbf{L}$ in the PFA. We have begun exploring this inference method and consider this future work.

Additionally, we may encounter situations where we have varying numbers of repeated measures at the tips. In practice, at this point we average together these measurements, however as future work we consider how to elegantly and efficiently handle these situations.

5.6 Appendix

A Matrix inversion computations used in the peeling algorithm.

We need to compute:

$$
\begin{aligned}
\mathbf{P}_i^{\star} &= \left(\mathbf{P}_i^{-} + t_i \boldsymbol{\delta}_i \boldsymbol{\Sigma} \boldsymbol{\delta}_i\right)^{-} \\
&= \mathbf{Q}_i \left(\left(\operatorname{diag}\left[\infty\mathbf{I}, \tilde{\mathbf{P}}_i, 0\mathbf{I}\right]\right)^{-} + \operatorname{diag}\left[t_i \begin{pmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{+0} \\ \boldsymbol{\Sigma}'_{+0} & \boldsymbol{\Sigma}_{++} \end{pmatrix}, 0\mathbf{I} \right] \right)^{-} \mathbf{Q}_i' \\
&= \mathbf{Q}_i \left(\operatorname{diag}\left[0\mathbf{I}, \tilde{\mathbf{P}}_i{}^{-1}, \infty\mathbf{I}\right] + \operatorname{diag}\left[t_i \begin{pmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{+0} \\ \boldsymbol{\Sigma}'_{+0} & \boldsymbol{\Sigma}_{++} \end{pmatrix}, 0\mathbf{I} \right] \right)^{-} \mathbf{Q}_i' \\
&= \mathbf{Q}_i \left(\operatorname{diag}\left[\begin{pmatrix} t_i\boldsymbol{\Sigma}_{00} & t_i\boldsymbol{\Sigma}_{+0} \\ t_i\boldsymbol{\Sigma}'_{+0} & \tilde{\mathbf{P}}_i{}^{-1} + t_i\boldsymbol{\Sigma}_{++} \end{pmatrix}, \infty\mathbf{I} \right] \right)^{-} \mathbf{Q}_i' \\
&= \mathbf{Q}_i \operatorname{diag}\left[\begin{pmatrix} t_i\boldsymbol{\Sigma}_{00} & t_i\boldsymbol{\Sigma}_{+0} \\ t_i\boldsymbol{\Sigma}'_{+0} & \tilde{\mathbf{P}}_i{}^{-1} + t_i\boldsymbol{\Sigma}_{++} \end{pmatrix}^{-1}, 0\mathbf{I} \right] \mathbf{Q}_i',
\end{aligned}
\tag{5.16}
$$

An attentive reader should remark that $\tilde{\mathbf{P}}_i{}^{-1}$ may not exist since $\tilde{\mathbf{P}}_i$ can be singular. This situation occurs, for example, in a factor analysis when the number of observed traits for a given tip is less than the number of factors, such that the latent tip precision $\mathbf{R}_i$ becomes rank-deficient. However, this is not problematic; we can determine the final matrix inverse in Equation (5.16) as a function of $\tilde{\mathbf{P}}_i$ directly through two algebraic slights of hand.

The first trick invokes commonly used expressions for the inverse of a partitioned matrix. Letting $\mathbf{D}_i = \tilde{\mathbf{P}}_i{}^{-1} + t_i\boldsymbol{\Sigma}_{++}$ and $\mathbf{A}_i = t_i\boldsymbol{\Sigma}_{00} - t_i\boldsymbol{\Sigma}_{+0}\mathbf{D}_i^{-1}t_i\boldsymbol{\Sigma}_{+0}$,

we can re-write

$$\begin{aligned}
&\begin{pmatrix} t_i\boldsymbol{\Sigma}_{00} & t_i\boldsymbol{\Sigma}_{+0} \\ t_i\boldsymbol{\Sigma}_{+0} & \tilde{\mathbf{P}}_i^{-1} + t_i\boldsymbol{\Sigma}_{++} \end{pmatrix}^{-1} \\
&\quad = \begin{pmatrix} \mathbf{A}_i^{-1} & -\mathbf{A}_i^{-1}t_i\boldsymbol{\Sigma}_{+0}\mathbf{D}_i^{-1} \\ -\mathbf{A}_i^{-1}t_i\boldsymbol{\Sigma}_{+0}\mathbf{D}_i^{-1} & \mathbf{D}_i^{-1} + \mathbf{D}_i^{-1}t_i\boldsymbol{\Sigma}_{+0}\mathbf{A}_i - 1 t_i\boldsymbol{\Sigma}_{+0}\mathbf{D}_i^{-1} \end{pmatrix}
\end{aligned} \tag{5.17}$$

if $\mathbf{A}_i$ and $\mathbf{D}_i$ are non-singular.

The second trick revolves around the inverse of the sum of two matrices and has been discovered and forgotten several times, see e.g. [Henderson et al., 1959, Henderson and Searle, 1981]. Namely,

$$\left(\tilde{\mathbf{P}}_i^{-1} + t_i\boldsymbol{\Sigma}_{++}\right)^{-1} = \tilde{\mathbf{P}}_i - \tilde{\mathbf{P}}_i\left(\tilde{\mathbf{P}}_i + \frac{1}{t_i}\boldsymbol{\Sigma}_{++}^{-1}\right)^{-1}\tilde{\mathbf{P}}_i. \tag{5.18}$$

Notice that in Equation (5.18), $\tilde{\mathbf{P}}_i + \frac{1}{t_i}\boldsymbol{\Sigma}_{++}^{-1}$ represents the sum of a positive-semi-definite and a positive-definite matrix for $t_i > 0$ and the sum is almost surely a positive-definite matrix and non-singular. Hence, $\left(\tilde{\mathbf{P}}_i^{-1} + \frac{1}{t_i}\boldsymbol{\Sigma}_{++}\right)^{-1}$ exists and $\mathbf{A}_i$ is also non-singular, demonstrating that Equation (5.17) also holds. In the degenerate case where $t_i = 0$, we note $\mathbf{P}_i^\star = \mathbf{P}_i$. Most commonly, we encounter cases where either all of our tips are missing or observed or the case where all of the tips are latent. In the first example, this inverse becomes $\frac{1}{t_i}\boldsymbol{\Sigma}_{00}^{-1}$, and in the latter case it follows to the form of Equation (5.18).

B Integrating factors using tree recursion

In this section we aim to reformulate the PFA model [Tolkoff et al., 2017] in such a way that we can integrate out the factors using the methods described in this paper. Following from the fact that the across-row variance of $\boldsymbol{\epsilon}$ equals $\mathbf{I}$, then if we define $\mathbf{F}_i$ as the vector corresponding to the factors at tip i and $\boldsymbol{\Lambda}_{mis_i}$

as $\mathbf{\Lambda}$ where those diagonal elements corresponding to the missing elements of $\mathbf{Y}_i$ are set to 0, then we can factorize

$$p\left(\mathbf{Y} \mid \mathbf{F}, \mathbf{L}, \mathbf{\Lambda}_{mis_i}\right) = \prod_{i=1}^{N} p\left(\mathbf{Y}_i \mid \mathbf{F}_i, \mathbf{L}, \mathbf{\Lambda}_{mis_i}\right). \tag{5.19}$$

Breaking this down by column,

$$\begin{aligned} p\left(\mathbf{Y}_i \mid \mathbf{F}_i, \mathbf{L}, \mathbf{\Lambda}_{mis_i}\right) =& \mathcal{N}\Big(\mathbf{Y}_i; \left(\mathbf{F}_i'\mathbf{L}\right)', \mathbf{\Lambda}_{mis_i}^{-}\Big) \\ =& \sqrt{\frac{\hat{\det}\left(\mathbf{\Lambda}_{mis_i}\right)}{\hat{\det}\left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\right)} \frac{1}{2\pi^{P-K}}} \\ & \times \exp\left\{-\frac{1}{2}\left(\mathbf{Y}_i'\mathbf{\Lambda}_{mis_i}\mathbf{Y}_i - \boldsymbol{\mu}_i'\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\boldsymbol{\mu}_i\right)\right\} \\ & \times \mathcal{N}\Big(\mathbf{F}_i; \boldsymbol{\mu}_i, \left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\right)^{-}\Big), \end{aligned} \tag{5.20}$$

where the conditional mean $\boldsymbol{\mu}_i = \left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}^t\right)^{-}\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{Y}_i$. Note conveniently that we have already solved the integral

$$\int \prod_{i=1}^{N} \mathcal{N}\Big(\mathbf{F}_i; \boldsymbol{\mu}_i, \left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\right)^{-}\Big)\, p\left(\mathbf{F} \mid \cdot\right)\, \mathrm{d}\mathbf{F} \tag{5.21}$$

in Section (5.2.3) by setting $\mathbf{m}_i = \boldsymbol{\mu}_i$ $\mathbf{V}_i = \left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\right)^{-}$ for tips $i = 1, \ldots, N$ and recursively invoking the post-order full precision likelihood traversal. Under this condition the remainder at tip i

$$\begin{aligned} r_i =& \sqrt{\frac{\hat{\det}\left(\mathbf{\Lambda}_{mis_i}\right)}{\hat{\det}\left(\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\right)} \frac{1}{2\pi^{P-K}}} \\ & \times \exp\left\{-\frac{1}{2}\left(\mathbf{Y}_i'\mathbf{\Lambda}_{mis_i}\mathbf{Y}_i - \boldsymbol{\mu}_i'\mathbf{L}\mathbf{\Lambda}_{mis_i}\mathbf{L}'\boldsymbol{\mu}_i\right)\right\}. \end{aligned} \tag{5.22}$$

CHAPTER 6

Future Directions

6.1 Structural Equation Models

6.1.1 Introduction

One of the advantages of a phylogenetic factor analysis model is its ability to scale down large problems. Scientific questions in the field of morphometrics will regularly produce large numbers of covariates and require phylogenetic adjustments in order to formulate a proper analysis. One could argue, however, that covariates related to the specific coordinates of landmarks of an organism should not be grouped together in a factor analysis with covariates related to other aspects of an organism such as its size, environment, or any measure of blood chemistry. However, we may nevertheless wish to infer some relationship between these two types of variables. Similarly, one may have measurements on the traits of a series of organisms and may want to make causal statements about how the genetic makeup of these organisms affect these traits. The phylogenetic factor analysis model as it is currently constructed does not handle different types of covariates in a sufficiently sophisticated manner to distinguish between variable types or infer causal relationships. To deal with this problem, we wish to implement a phylogenetically structural Equation model (PSEM).

6.1.2 Defining a Structural Equation Model

There are many different forms of structural Equation models, however we choose to proceed with the one described in the review by Dunson et al. [2005] since it has a similar form to the factor analysis model described in Chapter 5. We define endogenous variables as those variables whose state can be described by elements within our system. There may be some underlying factor or factors which give rise to our collection of endogenous variables, but we assume that these factors can be influenced by other elements within our system. Exogenous variables are those variables whose state cannot be explained by elements within our system. Exogenous variables may also be grouped together through underlying factors.

We intend for our endogenous variables to consist of morphometric information, whereas our exogenous variables contain information regarding genetics. Trait information is typically exogenous, however it can be endogenous if there is genetic information, since we anticipate genetic information to be causal to the traits of an organism. The ultimate goal of this model is to provide covariate adjustments for phylogenetic factor analysis.

We arrange the endogenous measurements into a $N \times P$ matrix of outcomes $\mathbf{Y}$ and the exogenous variables into a $N \times \ell$ matrix of covariates $\mathbf{X}$ both of whose continuous traits are scaled to have mean 0 and variance 1. We again break down these matrices such that

$$\mathbf{Y} = \mathbf{F_Y}\mathbf{L_Y} + \boldsymbol{\epsilon_Y} \tag{6.1}$$

and

$$\mathbf{X} = \mathbf{F_X}\mathbf{L_X} + \boldsymbol{\epsilon_X} \tag{6.2}$$

where $\mathbf{L_Y}$ and $\mathbf{L_X}$ are $K \times P$ and $M \times \ell$ loadings matrices respectively, $\mathbf{F_Y}$

and $\mathbf{F_X}$ are $N \times K$ and $N \times M$ factor matrices respectively and $\boldsymbol{\epsilon_Y}$ and $\boldsymbol{\epsilon_X}$ are residual matrices. We say that K and M are the number of factors for the endogenous and exogenous variables respectively. The matrix $\boldsymbol{\epsilon_Y}$ is an error matrix distributed $\text{MN}(0, \mathbf{I}, \boldsymbol{\Lambda_Y})$ and similarly $\boldsymbol{\epsilon_X}$ is an error matrix distributed $\text{MN}(0, \mathbf{I}, \boldsymbol{\Lambda_X})$, where MN is a matrix normal distribution with a mean, a row precision and a column precision, and where $\boldsymbol{\Lambda_Y}$ and $\boldsymbol{\Lambda_X}$ are $P \times P$ and $\ell \times \ell$ diagonal matrices respectively.

This model then relates these two separate factor analyses via a regression, specifically

$$\mathbf{F_Y} = \mathbf{F_Y}\boldsymbol{\Gamma} + \mathbf{F_X}\boldsymbol{\eta} + \boldsymbol{\xi}, \tag{6.3}$$

where $\boldsymbol{\Gamma}$ is a $K \times K$ graph relating the outcomes to themselves, $\boldsymbol{\eta}$ is an $M \times K$ graph relating the covariates to the various outcomes of interest, and $\boldsymbol{\xi}$ is a residual term distributed $\text{MN}(0, \mathbf{I}, \boldsymbol{\Lambda_\xi})$ and $\boldsymbol{\Lambda_\xi}$ is a $K \times K$ diagonal matrix. The same priors and restrictions as mentioned previously in Chapter 5 apply to both loadings matrices, both factor matrices, as well as all three residual terms. The diagonal elements on the graph $\boldsymbol{\Gamma}$ are 0 whereas the diagonal elements on the graph $\boldsymbol{\eta}$ are positive. Both graphs have i.i.d. $\text{N}(0, 1)$ priors for most elements, and i.d.d. $\text{N}^+(0, 1)$ priors on the diagonal elements of $\boldsymbol{\eta}$, where N^+ is positive half-normal distribution.

6.2 Hamiltonian Monte Carlo

In the future, we would like to implement Hamiltonian Monte Carlo (HMC) methods for more efficient sampling [Neal, 2010]. HMC relies on treating the posterior distribution as the negative of a potential map. It then treats the current location of the chain as the location of a particle, gives the particle some momentum and samples the posterior based on this particle's ultimate location. The advantage of this method is that it is useful as long as the potential gradient

(derivative of the negative log likelihood) is simple to compute.

If we look at the integrated model described in Chapter 5, to compute an efficient HMC sampler, we need the potential of the posterior. If we can obtain this, then we can hopefully adapt our HMC to work with Holbrook et al. [2016, 2017] in order to handle the identifiability issues inherent to factor analysis.

The generic HMC [Neal, 2010] relies on treating the values of $\mathbf{L}$ and as the location of a particle in a Hamiltonian system. We draw a momentum, $\boldsymbol{\rho}$, associated with this particle from a series of i.i.d. $N(0,1)$ distributions and move it across a potential U. This momentum $\boldsymbol{\rho}$ is conformable to the matrix $\mathbf{L}$. The potential used to infer $\mathbf{L}$, based on the model in Chapter 5 is

$$U = -\log\left(p\left(\mathbf{L} \mid \mathbf{Z}, \boldsymbol{\Lambda}\right) p\left(\mathbf{L}\right)\right). \tag{6.4}$$

Generally, we use the "leapfrog method" where

$$\boldsymbol{\rho}^{(i+1/2)} = \boldsymbol{\rho}^{(i)} - \left[\frac{\epsilon}{2}\left(\nabla_{(\mathbf{L},\mathbf{F})} U\right)\Big|_{\mathbf{L}^{(i)}}\right] \tag{6.5}$$

$$\left(\mathbf{L}\right)^{(i+1)} = \boldsymbol{\rho}^{(i+1/2)} + \epsilon\left(\mathbf{L}\right)^{(i)} \tag{6.6}$$

$$\boldsymbol{\rho}^{(i+1)} = \boldsymbol{\rho}^{(i+1/2)} - \left[\frac{\epsilon}{2}\left(\nabla_{(\mathbf{L},\mathbf{F})} U\right)\Big|_{\mathbf{L}^{(i+1)}}\right], \tag{6.7}$$

where $\boldsymbol{\rho}^{(i)}$ and $\mathbf{L}^{(i)}$ are iteration i for $\boldsymbol{\rho}$ and $\mathbf{L}$ respectively, and $\nabla_{(\mathbf{L},\mathbf{F})}$ is the derivative with respect to $\mathbf{L}$. We run this algorithm for s steps, and accept this state with probability

$$\min\left[1, \exp\left(-U\left(\mathbf{L}'\right) + U\left(\mathbf{L}\right) - K\left(\boldsymbol{\rho}'\right) + K\left(\boldsymbol{\rho}\right)\right)\right], \tag{6.8}$$

where $\mathbf{L}'$ and $\boldsymbol{\rho}'$ are the draw from the HMC of $\mathbf{L}$ and $\boldsymbol{\rho}$ respectively, and

$$K\left(\boldsymbol{\rho}\right) = \frac{1}{2}\text{vec}\boldsymbol{\rho}'\text{vec}\boldsymbol{\rho}. \tag{6.9}$$

For my future work, we would like to be able to derive the appropriate potential for the integrated model described in Chapter 5.

6.3 Bouncy Particle Sampler

In Cybis et al. [2015], these authors lay out a latent liability model for handling discrete data in the context of a multivariate Brownian diffusion. In Chapter 5 we adapt this probit model to work with the phylogenetic factor analysis model. While Chapter 5 deals with integrating out latent values, this method is difficult to adapt to the latent liability model because of the discrete cutpoints entailed. Instead, we would like to sample these latent values from the tips of a tree using a bouncy particle sampler [Bouchard-Côte and Vollmer, 2017], which moves particles in a way similar to Hamiltonain Monte Carlo [Neal, 2010] but allows for bouncing off of cutpoints. I hope this innovation leads to more efficient computation in these models.

6.4 Repeated Measures

Much of this thesis is devoted to measuring the relationships between traits. Oftentimes these traits for a given taxa are obtained by measuring many samples for that taxa and averaging these samples. While this works well as an approximation, doing this loses the within taxa variance inherent in these species.

In Chapter 5 we laid out a method for integrating out latent measurements. By treating each species as having an unknown, latent mean and variance from which these measurements are drawn we hope to be able to integrate out these underlying normal distributions in order to gain insight into the relationships between these traits across species.

Bibliography

DC Adams. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution*, pages 2675–2688, 2014.

DC Adams and ML Collyer. Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: What you shuffle matters. *Evolution*, 69(3):823–829, March 2015.

DC Adams, CM Berns, KH Kozak, and JJ Wiens. Are rates of species diversification correlated with rates of morphological evolution? *Proceedings of the Royal Society of London B: Biological Sciences*, 2009.

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

G Baele, P Lemey, T Bedford, A Rambaut, MA Suchard, and AV Alekseyenko. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Evolutionary Biology*, 29(9):2157–2167, 2012.

G Baele, P Lemey, and S Vansteelandt. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics*, 14(85), March 2013a.

G Baele, WLS Li, AJ Drummond, MA Suchard, and P Lemey. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Evolutionary Biology*, 30(2):239–243, 2013b.

G Baele, P Lemey, and MA Suchard. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. *Systematic Biology*, 65 (2):250–264, 2016.

AA Beguin and CAW Glas. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4):541–562, December 2001.

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

TM Blackburn. Evidence for a 'fast-slow' continuum of life-history traits among parasitoid hymenoptera. *Functional Ecology*, 5(1):65–74, 1991.

A Bouchard-Côte and SJ Vollmer. The bouncy particle sampler: A non-reversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 2017.

BL Browning and SR Browning. Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, 98(1):116–126, January 2016.

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

CM Carvalho, NG Polson, and JG Scott. Handling Sparsity via the Horseshoe. In David V. Dyk and Max Welling, editor, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09)*, volume 5, pages 73–80. Journal of Machine Learning Research - Proceedings Track, 2009.

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

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

GB Cybis, JS Sinsheimer, T Bedford, AE Mather, P Lemey, and MA Suchard. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Annals of Applied Statistics*, 9:969 – 991, 2015.

A Dornburg, F Santini, and ME Alfaro. The influence of model averaging on clade posteriors: An example using the triggerfishes (family *Balistidae*). *Systematic Biology*, 57(6):905–919, 2008.

A Dornburg, B Sidlauskas, F Santini, L Sorenson, TJ Near, and ME Alfaro. The influence of an innovative locomotor strategy on the phenotypic diversification of triggerfish (family: *Balistidae*). *Evolution*, 65(7):1912–1926, July 2011.

AJ Drummond, MA Suchard, D Xie, and A Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973, 2012.

DB Dunson, J Palomo, and K Bollen. Bayesian structural equation modeling. Technical Report 5, Statistical and Applied Mathematical Sciences Institute, July 2005.

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

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

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

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

RP Freckleton. The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, 22(7):1367–1375, 2009.

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

N Friel and AN Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3): 589–607, 2008.

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

A Gelman. Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

A Gelman and XL Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2): 163–185, 1998.

S Geman and D Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Annals of Machine Intelligence*, 6(6):721–741, June 1984.

EI George and RE McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, September 1993.

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

J Ghosh and DB Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, pages 306–320, 2009.

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

TL Griffiths and Z Ghahramani. Infinite latent feature models and the indian buffet process. *MIT Press*, pages 475–482, 2005.

TL Griffiths and Z Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, March 2011.

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

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

SE Heaps, RJ Boys, and M Farrow. Computation of marginal likelihoods with data-dependent support for latent variables. *Computational Statistics & Data Analysis*, 71:392–401, 2014.

CR Henderson, O Kempthorne, SR Searle, and CM von Krosigk. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218, June 1959.

HV Henderson and SR Searle. On deriving the inverse of a sum of matrices. *SIAM Reviews*, 23(1):53–60, 1981.

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

A Holbrook, A Vandenberg-Rodes, and B Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv:1606.04478v1*, 2016.

A Holbrook, A Vandenberg-Rodes, N Fortin, and B Shahbaba. A Bayesian supervised dual-dimensionality reduction model for simultaneous decoding of LFP and spike train signals. *Stat*, 2017.

JP Huelsenbeck and B Rannala. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, 57(6):1237–1247, 2003.

AR Ives and T Garland Jr. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, 2010.

H Jeffreys. Some tests of signficance, treated by the theory of probability. *Procedings of the Cambridge Philosophical Society*, 29:83–87, 1935.

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

VE Johnson and D Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498), 2012.

KE Jones, J Bielby, M Cardillo, SA Fritz, J O'Dell, CDL Orme, K Safi, W Sechrest, EH Boakes, C Carbone, C Connolly, MJ Cuttis, JK Foster, R Grenyer, M Habib, CA Plaster, SA Price, EA Rigby, J Rist, A Teacher, ORP Bininda-Emonds, JL Gittleman, GM Mace, and A Purvis. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648, 2009.

TH Jukes and CR Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969.

CP Kingenberg and J Marugán-Lobón. Evolutionary covariation in geometric morphometric data: Analyzing integration, modularity, and allometry in a phylogenetic context. *Systematic Biology*, 62(4):591–610, 2013.

JFC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

D Knowles and Z Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In Mike E. Davies, Christopher J. James, Samer A. Abdallah, and Mark D Plumby, editors, *Independent Component Analysis and Signal Separation*, volume 4666, pages 381–388, 2007.

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

H Li and D Pati. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119, March 2017.

JS Liu, WH Wong, and A Kong. Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–169, 1995.

HF Lopes and M West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

JB Losos and CJ Schneider. *Anolis* lizards. *Current Biology*, 19(19):R316–R318, April 2009.

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

E Makalic and DF Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, January 2016.

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

RM Neal. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press, 2010.

MA Newton and AE Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 56(1):3–48, 1994.

G Pinto, DL Mahler, LJ Harmon, and JB Losos. Testing the island effect in adaptive radiation: rates and patterns of morphological diversification in Caribbean and mainland *Anolis* lizards. *Procedings of the Biological Sciences B*, 275(1652):2749–2757, December 2008.

BJA Pollux, RW Meredith, MS Springer, and DN Reznick. The evolution of the placenta drives a shift in sexual selection in livebearing fish. *Nature*, 13451, 2014.

PD Polly, AM Lawing, AC Fabre, and A Goswami. Phylogenetic principal components analysis and geometric morphometrics. *Hystrix, the Italian Journal of Mammalogy*, 24(1):33–41, 2013.

OG Pybus, MA Suchard, P Lemey, FJ Bernardin, A Rambaut, FW Crawford, RR Gray, N Arinaminpathy, SL Stramer, MP Busch, and EL Delwart. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Procedings of the National Academy of Sciences*, 109(37):15066–15071, 2012.

KM Quinn. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, 12(4):338–353, 2004.

P Rai and H Daume. The infinite hierarchical factor regression model. *Advances in Neural Information Processing Systems*, 2008.

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

JD Reynolds. *Macroecology: concepts and consequences*, chapter Life histories and extinction risk, pages 195–217. Oxford: Blackwell Publishing Ltd., 2003.

V Ročková and EI George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, June 2014.

DA Roff. *Life history evolution*. Sunderland, Massachusetts. Sinauer Associates, January 2002.

BE Sæther and Ø Bakke. Avian life history variation and contribution of demographic traits to the population growth rate. *Ecology*, 81(3):642–653, 2000.

R Salguro-Gómez. Applications of the fast-slow continuum and reproductive strategy framework of plant life histories. *New Phytologist*, 213(3):1618–1624, 2017.

JC Santos. The implementation of phylogenetic structural equation modeling for biological data from variance-covariance matrices, phylogenies, and comparative analyses. Master's thesis, University of Texas at Austin, 2009.

M Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809, May 2000.

MA Suchard, RE Weiss, and JS Sinsheimer. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18 (6):1001–13, June 2001.

MR Tolkoff, ME Alfaro, G Baele, P Lemey, and MA Suchard. Phylogenetic factor analysis. *Systematic Biology*, 2017.

JC Uyeda, DS Caetano, and MW Pennell. Comparative analysis of principal components can be misleading. *Systematic Biology*, 64:677–689, 2015.

J Velasco, E Martínez-Meyer, O Flores-Villela, A García, AC Algar, G Köhler, and JM Daza. Climatic niche attributes and diversification in *Anolis* lizards. *Journal of Biogeography*, 43(1):134–144, January 2015.

B Vrancken, P Lemey, A Rambaut, T Bedford, B Longdon, HF Günthard, and MA Suchard. Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods in Ecology and Evolution*, 6:67–82, 2015.

JB Whittall and SA Hodges. Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nature*, 447(7145):706–709, 2007.

MA Wiedmann, R Primicerio, A Dolgov, CAM Ottensen, and M Aschan. Life history variation in Barents Sea fish: Implications for sensitivity to fishing in a changing environment. *Ecology and Evolution*, 4(18):3596–3611, 2014.

W Xie, PO Lewis, and MH Chen. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.

Y Xu, P Müller, and D Telesca. Bayesian inference for latent biologic structure with determinental point processes (DPP). *Biometrics*, 72(3):955–964, September 2016.

Z Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.