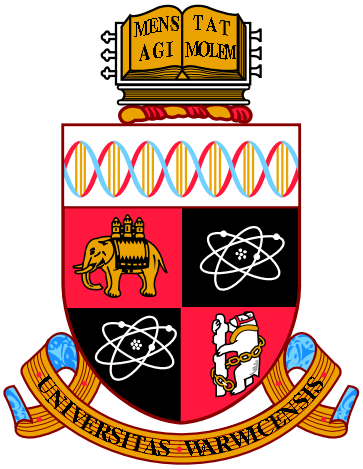


Bayesian Ancestral Reconstruction for Bat Echolocation

by

Joseph Patrick Meagher

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Statistics

June 2020

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Thank you to my supervisors on this project. To Mark Girolami, for offering me the opportunity to undertake this programme of doctoral research; to Kate Jones, for providing such a motivating problem, and to Theo Damoulas, whose mentorship made this work possible.

Thank you to both the EPSRC for their generous funding of this project and the Department of Statistics at the University of Warwick for hosting me throughout.

Thank you to my friends, old and new, who have been with me on this journey. In particular, I want to thank Arthur, five years is a long time to spend under the same roof, but you made it easy.

Thank you to my family; to my grandparents, for whom my admiration only increases with the passage of time; to Ailshe and Liam, may you learn from the mistakes of your brother, whatever you may judge them to be; to my mother, whose constant self-development is as much motivation as it is inspiration; and to my father, for demonstrating the value of consistent enthusiasm. I can never truly thank you for all I've been given in this life, but only do my best to make the most of it.

And to Suzy, on to the next chapter.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy.

I declare that it has been composed by myself and that the work contained herein is my own except where explicitly stated otherwise.

This work has been completed wholly while in candidature for a research degree at the University of Warwick and has not been submitted for any other degree or professional qualification.

Aspects of this research have been published in:

- JP Meagher, T Damoulas, KE Jones, and M Girolami. Discussion of "The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages", by Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. Journal of the Royal Statistical Society: Series C (Applied Statistics), 67(5):1103-1145, 2018.

- JP Meagher, T Damoulas, KE Jones, and M Girolami. Phylogenetic Gaussian processes for bat echolocation. Statistical Data Science, pages 111-122, 2018. doi:10.1142/q0159.

Abstract

Ancestral reconstruction can be understood as an interpolation between measured characteristics of existing populations to those of their common ancestors. Doing so provides an insight into the characteristics of organisms that lived millions of years ago. Such reconstructions are inherently uncertain, making this an ideal application area for Bayesian statistics. As such, Gaussian processes serve as a basis for many probabilistic models for trait evolution, which assume that measured characteristics, or some transformation of those characteristics, are jointly Gaussian distributed. While these models do provide a theoretical basis for uncertainty quantification in ancestral reconstruction, practical approaches to their implementation have proven challenging. In this thesis, novel Bayesian methods for ancestral reconstruction are developed and applied to bat echolocation calls. This work proposes the first fully Bayesian approach to inference within the Phylogenetic Gaussian Process Regression framework for Function-Valued Traits, producing an ancestral reconstruction for which any uncertainty in this model may be quantified. The framework is then generalised to collections of discrete and continuous traits, and an efficient approximate Bayesian inference scheme proposed, representing the first application of Variational inference techniques to the problem of ancestral reconstruction. This efficient approach is then applied to the reconstruction of bat echolocation calls, providing new insights into the developmental pathways of this remarkable characteristic. It is the complexity of bat echolocation that motivates the proposed approach to evolutionary inference, however, the resulting statistical methods are broadly applicable within the field of Evolutionary Biology.

Abbreviations

AM	Adaptive Metropolis
ARD	Automatic Relevance Determination
ASIS	Ancillarity-Sufficiency Interweaving Strategy
BM	Brownian Motion
ESS	Elliptical Slice Sampler
FA	Factor Analysis
FDA	Functional Data Analysis
FVT	Function-Valued Trait
GP	Gaussian Process
ICA	Independent Components Analysis
MCMC	Markov Chain Monte Carlo
MRCA	Most Recent Common Ancestor
OU	Ornstein-Uhlenbeck
PCA	Principal Components Analysis
PCM	Phylogenetic Comparative Method
PFA	Phylogenetic Factor Analysis
PGLS	Phylogenetic Generalised Least Squares
PGPR	Phylogenetic Gaussian Process Regression
PLVM	Phylogenetic Latent Variable Model
PMM	Phylogenetic Mixed Model
SDE	Stochastic Differential Equation
SRVF	Square Root Velocity Function
STFT	Short-time Fourier Transform

Chapter 1

Introduction

What is it like to be a bat? This question, posed by Nagel [1974] to illustrate the limitations of objectivity in the study of consciousness, is indicative of our longstanding fascination with these "*fundamentally alien*" creatures. Bats are ubiquitous in myths and folklore, from the Mayan "*death bat*" Camazotz [Miller and Taube, 1997] and Chinese "*five good fortunes*" [Sung, 2002], to more modern characterisations such as Dracula [Stoker, 1897] and Batman [Miller et al., 2002]. The first scientific studies of these creatures date back to the 1790s when Spallanzani established that blinded bats successfully avoided obstacles while deafened ones did not [Galambos, 1942]. It was Griffin and Galambos [1941] who demonstrated that bats interact with their environment by echolocation, and since then many researchers have sought to deepen our understanding of these astonishing creatures [Simmons and Stein, 1980; Simmons, 1994; Schnitzler et al., 2004; Maltby et al., 2010; Meagher et al., 2018a,b].

Advances in the sequencing and modelling of molecular data [Suchard et al., 2018] have allowed a consensus on bat's evolutionary history to emerge, with the structure and timing of ancestral relationships between many species being well-resolved [Teeling et al., 2000, 2005; Eick et al., 2005; Tsagkogeorga et al., 2013; Amador et al., 2018]. Despite this progress, describing the development of echolocation throughout this history remains a challenge. One approach has been to argue for particular developmental paths based on bats physiology [Simmons and Stein, 1980; Schnitzler et al., 2004]. Alternatively, quantitative analyses have considered various call representations and summary statistics [Eick et al., 2005; Collen, 2012; Meagher et al., 2018b]. Despite these efforts, bat echolocation represents a complex characteristic which does not easily conform to existing mathematical models for trait evolution. Thus, this thesis' contribution is the development of statistical models for the evolution of such complex phenotypes.

Since Darwin [1859] described the process of natural selection in his seminal text, "*On the Origin of Species*", characterising those origins has been central to the development of evolutionary biology. As the field has progressed, describing creatures from the ancient past and elucidating their influence on those living today has been framed as a statistical problem [Felsenstein, 1985; Martins and Hansen, 1997; Suchard et al., 2018]. For instance, it is useful to think of ancestral reconstruction as the interpolation between characteristics of extant taxa[1] given their evolutionary history [Joy et al., 2016]. Irrespective of the characteristic in question, be it a phenotype, genetic sequence, or even an entire genome, insights obtained through such analysis are only as good as the statistical model for evolution that underpins them [Joy et al., 2016]. Thus, generations of researchers have devoted themselves to the development of such models, with many theoretical and practical issues having been resolved [Cavalli-Sforza and Edwards, 1967; Felsenstein, 1973; Grafen, 1989; Hansen, 1997; Pagel, 1999b; Blomberg et al., 2003; Housworth et al., 2004; Ives and Garland Jr, 2009; Hadjipantelis et al., 2013; Cybis et al., 2015; Goolsby, 2015; Tolkoff et al., 2017; Mariñas-Collado et al., 2019]. The Phylogenetic Gaussian Process Regression (PGPR) framework provides a foundation for this contribution to statistical models for trait evolution [Jones and Moriarty, 2013]. This framework explicitly links evolutionary inference to Gaussian processes, an important research area in Statistics and Machine Learning [Rasmussen and Williams, 2006; Stein, 2012]. Extending PGPR beyond the Function-Valued Traits (FVTs) considered by Jones and Moriarty [2013] and developing state-of-the-art methods for Bayesian inference allows the development of novel approaches to ancestral reconstruction.

For any statistical method, the adage, "*garbage in, garbage out*", will hold. Thus, the representation of echolocation calls to be reconstructed requires careful consideration. These acoustic signals, precisely structured in both time and frequency, are subject to myriad constraints, due not only to the anatomy of bats call production systems [Fenton et al., 2016], but also the principles of radar and sonar [Denny, 2007]. A characterisation which not only captures the signal transmitted by these echolocation calls but also allows their comparative analysis, has proven challenging [Collen, 2012; DiCecco et al., 2013; Fu and Kloepper, 2018; Meagher et al., 2018b]. Despite this, echolocation calls remain nothing more than another acoustic signal. Thus, informed by decades of research in Bioacoustics [Hopp et al.,

[1] In taxonomy and systematics, the branches of biology that deal with the classification and nomenclature of organisms, the term taxon, and its plural taxa, refers to a taxonomic group of any rank [Campbell et al., 1997]. The methods developed in this thesis are primarily concerned with characteristics at the level of species; however, it is more convenient to use this general term throughout.

2012], signal processing [Oppenheim and Schafer, 2014], and time-frequency analysis [Cohen, 1995; Hlawatsch and Auger, 2008], such a representation is not beyond reach.

The ancestral reconstruction of bat echolocation calls is the objective of this thesis and work towards this goal begins with a review of the relevant literature, presented in Chapter 2. It begins by providing some background on the scientific study of bats, covering not only the structure and diversity of echolocation calls across the order [Fenton et al., 2016] but also the consensus which has now emerged regarding their evolutionary history [Amador et al., 2018]. As is the case for taxa in general, a phylogenetic tree represents this history, referred to as the phylogeny [Felsenstein, 2004]. It is knowledge of this object, and the implied dependence between taxa, that allows the development of statistical methods for phylogenetic comparative analysis and ancestral reconstruction [Felsenstein, 1985]. Thus, a review of Phylogenetic Comparative Methods (PCMs), charting their development from the method of independent contrasts for scalar-valued continuous characteristics [Felsenstein, 1985], to the PGPR framework for FVTs [Jones and Moriarty, 2013; Hadjipantelis et al., 2013], provides more of the context within which this work can be placed. This discussion leads to a presentation of the statistical principles and techniques underpinning the contributions made in this thesis. A general introduction to Gaussian processes is provided, demonstrating the flexibility of Gaussian process regression and highlighting the Matérn class of covariance functions [Rasmussen and Williams, 2006; Stein, 2012]. This allows Jones and Moriarty's [2013] PGPR framework, which models FVT evolution over a phylogeny in terms of a separable phylogeny-trait covariance function, to be presented in some detail. Factor Analysis, [Lopes, 2014] Functional Data Analysis [Ramsay, 2004; Srivastava and Klassen, 2016], and the Time-Frequency Analysis of acoustic signals [Cohen, 1995; Hlawatsch and Auger, 2008] are all relevant to the statistical methods developed here, and so each topic is briefly discussed. The chapter concludes with a presentation of Markov Chain Monte Carlo (MCMC) methods for Bayesian inference [Robert and Casella, 2013; Gelman et al., 2013]. In particular, an Adaptive Metropolis algorithm [Haario et al., 2001; Roberts and Rosenthal, 2009], sampling schemes for Gaussian process models [Murray et al., 2010; Murray and Adams, 2010; Yu and Meng, 2011; Filippone et al., 2013], and model comparison via Bridge Sampling [Meng and Wong, 1996; Gronau et al., 2017a], are all discussed in some detail.

Chapter 3, representing the first research contribution in this thesis, presents an MCMC sampling scheme for Bayesian inference within the PGPR framework. Ancestral reconstruction of a FVT by PGPR is based on separable phylogeny-trait

covariance functions for FVTs [Jones and Moriarty, 2013]. Hadjipantelis et al. [2013] and Meagher et al. [2018a,b], attempted to do this by first obtaining a low rank approximation to the trait covariance function under the assumption of independent trait observations. Once fixed, this allowed the phylogenetic covariance to be estimated. Here, introducing the Phylogenetic Latent Variable Model (PLVM), a model closely related to Factor Analysis [Bartholomew et al., 2011; Lopes, 2014], underpins the implementation of a Bayesian approach to learning which relaxes the assumption of separability and allows joint inference of phylogeny-trait covariance function. The development of this MCMC inference scheme, based around state-of-the-art methods for Gaussian process models [Murray et al., 2010; Murray and Adams, 2010; Yu and Meng, 2011; Filippone et al., 2013], presents many challenges. Chief amongst these is the management of the algorithm's computational expense. To this end, efficient algorithms for computing both the likelihood and conditional distribution of Brownian Motion over a phylogeny are extended to general Gauss-Markov processes [Pybus et al., 2012; Cybis et al., 2015], representing an important contribution in the development of PGPR for evolutionary inference. This generalisation, along with a novel definition of the phylogenetic covariance function, allows intra-taxon variation to be incorporated in the PLVM, an effect which is typically ignored by PCMs [Hadjipantelis et al., 2013; Cybis et al., 2015; Tolkoff et al., 2017]. The application of this inference scheme to a synthetic dataset simulated from the model allows an assessment of its performance. It offers excellent reconstruction and uncertainty quantification for ancestral FVTs while offering significant conceptual advantages over and above alternative PCMs. Despite this, its computational expense makes it wholly unsuitable for the analysis of a large dataset of bat echolocation calls. Thus, those insights gleaned from this study instead provide the basis for a more practical approach to evolutionary inference.

In Chapter 4, focus shifts from the development of a fully Bayesian model for the evolution of a FVT, to one which can fit flexibly and efficiently to any collection of traits, addressing a significant shortcoming of the PGPR framework. Typically, it is large collections of both discrete and continuous traits that are of interest in phylogenetic comparative analyses [Collen, 2012; Cybis et al., 2015; Tolkoff et al., 2017; Adams and Collyer, 2017]. FVTs are infinite dimensional objects [Kirkpatrick and Heckman, 1989], and as such, the implementation of models for their evolution is a multivariate method, however, current perspectives on the PGPR consider a single FVT only [Jones and Moriarty, 2013; Hadjipantelis et al., 2013; Goolsby, 2015; Meagher et al., 2018a,b]. This narrow focus represents a severe limitation of PGPR. Based on the threshold model for discrete trait evolution [Wright, 1934;

Felsenstein, 2011], PGPR is extended to incorporate ordinal and categorical discrete traits alongside both scalar- and function-valued continuous traits within a single model. To this end, observed manifest traits are augmented by real-valued auxiliary variables, allowing the definition of a probit likelihood, as described by Albert and Chib [1993]. Relaxing some assumptions from the formulation in Chapter 3, these auxiliary variables are then modelled as a PLVM, which results in the definition of a multi-modal posterior distribution over the parameters and hyper-parameters of the model. This multi-modal posterior, coupled with the computational expense of MCMC inference for the PLVM, precludes the implementation of a sampling scheme for this generalised PLVM. The development of a Co-ordinate Ascent Variational Inference algorithm for approximate Bayesian inference [Blei et al., 2017] addresses each of these issues. Although Variational Inference can underestimate uncertainty in the posterior distribution over parameters in the model, it fits to data far more efficiently than a simulation-based approach, making the method especially popular in Machine Learning [Jordan et al., 1999; Bishop, 2006]. The application of this model and inference scheme to another simulated dataset demonstrates its efficacy. In this instance, much of the accurate ancestral reconstruction and uncertainty quantification seen in Chapter 3, along with the inclusion of intra-taxon variation, is preserved. Furthermore, the model fits to the dataset in a fraction of the time required by the MCMC scheme proposed in the previous chapter. Thus, the method is eminently applicable for the ancestral reconstruction of bat echolocation calls, as indeed it is for the phylogenetic comparative analysis of any collection of traits.

Given this general model for trait evolution, Chapter 5 considers its application to the multi-harmonic signals that are bat echolocation calls [Fenton et al., 2016]. Such signals consist of multiple components with a precise structure in both time and frequency, where each component lies at an integer multiple of the fundamental frequency, which is itself a smooth function of time [Gerhard, 2003]. The analysis of multi-component signals is a challenging problem, with Time-Frequency Analysis representing an active area of research [Hlawatsch and Auger, 2008; Huang et al., 2009; DiCecco et al., 2013; Fu and Kloepper, 2018]. While the Spectrogram underpins some recent advances in the comparative analysis of acoustic signals [Stathopoulos et al., 2018; Pigoli et al., 2018], this time-frequency representation is not suitable for ancestral reconstruction of the bat echolocation call, as will be discussed in section 2.3.5. Thus, an alternative representation is required. To this end, a harmonic model for bat echolocation calls is developed, along with a maximum-a-posteriori inference scheme [Quinn and Thomson, 1991; Gerhard, 2003; Shi et al., 2019]. Fitting this model to a publicly available set of bat echolocation call

recordings (see Stathopoulos et al. [2018]) and post-processing the output defines a feature representation for each call. Given the phylogeny describing the structure and timing of familial relationships for recorded bat species [Collen, 2012], fitting a generalised PLVM to this call representation allows ancestral reconstruction of the bat echolocation call.

Based on this analysis, the Most Recent Common Ancestor of bats included in this sample, which lived approximately 52.5 million years ago [Collen, 2012], employed a multi-harmonic call with at least two frequency components. The call consisted of a broadband sweep from approximately 40 to 30 kHz, lasting 3 to 8 ms, with the fundamental frequency most probably dominating other frequency components. A hypothetical echolocation call for the most recent common ancestor of extant bats is illustrated below.

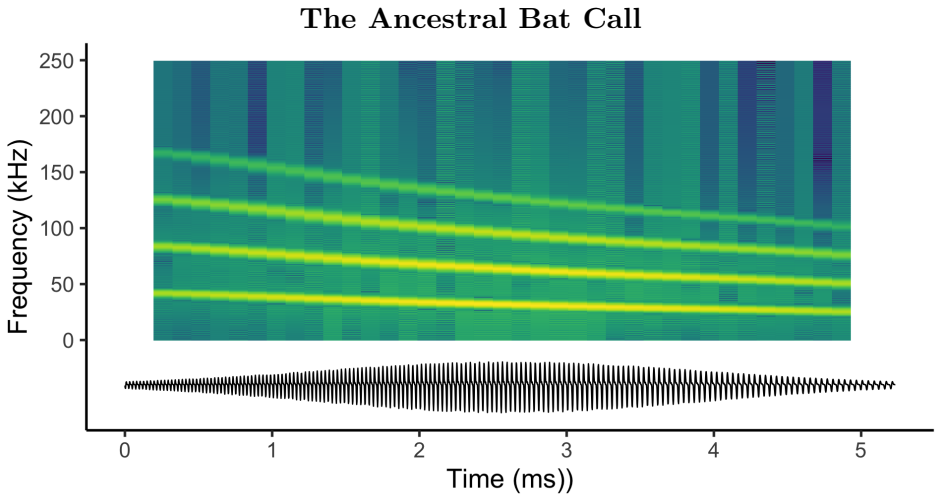


Figure 1.1

The final chapter (Chapter 6) presents a brief outline of the thesis' research findings and limitations. In particular, while conditioning trait evolution on an evolutionary history does allow ancestral reconstruction, the reality is that this history is unknown. This link with the broader field of phylogenetics, along with some other limitations, present many opportunities for future research.

In summary, Bayesian solutions to the problem of ancestral reconstruction are developed and applied to bat echolocation. Thus, while a description of bats consciousness may remain beyond our grasp, by reconstructing the echolocation calls of ancient bats, this thesis goes some way towards answering an equally fundamental question: how did these fantastic creatures come to be?

Chapter 2

Literature Review

2.1 Some Background on Bats

Over 1200 species and 21 families of extant bat (order *Chiroptera*) are currently recognised, making bats the second most speciose order of mammals, after rodents [Simmons, 2005; Amador et al., 2018]. The only mammals capable of powered flight, bats are usually crepuscular or nocturnal creatures. They are found on every continent, except Antarctica [Nowak and Walker, 1994], and are considered a keystone species in many habitats, given their roles in pollination, seed dispersal, and pest control [Jones et al., 2009].

Traditionally, bats have been split into two sub-orders. The Old World fruit bats (Pteropodidae) make up the sub-order Megachiroptera, while all other bats are considered to be Microchiroptera [Dobson, 1875]. This division is based not only on size, as the name alludes to, but also the ability to echolocate. While all Microchiropera can do so, all but a few species of Megachiroptera lack this distinguishing ability [Fenton et al., 2016].

Echolocation, the "*process of locating obstacles by means of echoes*" [Griffin, 1944] is usually, though not exclusively, associated with bats. The phenomenon has been observed in toothed whales [Surlykke et al., 2014], and, remarkably, oilbirds and cave swiftlets [Brinkløv et al., 2013], demonstrating that it is not exclusive to mammals. That bats echolocate while in flight was confirmed by Griffin and Galambos in 1941, with most species using signals produced in the larynx and emitted through the mouth or nose [Pedersen, 1998]. Again, pteropodids are an exception. Those members of the *Rousettus* genus that are capable of echolocation do so using tongue-clicks, which are broadband signals with a duration of only 50-100 μs [Holland et al., 2004].

Laryngeal echolocation calls are *tonal signals*, composed of some combination of constant frequency (CF) and frequency modulating (FM) components. The dominant component of an echolocation call, that is the one carrying most energy, ranges from 9 kHz in *Euderma maculatum* [Fullard and Dawson, 1997], to 212 kHz for *Cloeotis percivali* [Fenton and Bell, 1981], while the calls duration is typically between 3 and 50 ms [Surlykke et al., 2014]. Similarly to voiced human speech, the lowest frequency component is defined as the *fundamental frequency* [Deller Jr and Hansen, 2004]. All subsequent components occur at integer multiples of this frequency, although the dominant component may be distinct from the fundamental [Fenton et al., 2016]. This structure implies that laryngeal echolocation calls are multi-harmonic signals, where the fundamental frequency is the first harmonic [Hopp et al., 2012].

Bats can adjust aspects of their echolocation call in response to environmental conditions. For some species, calls occur through three distinct phases as they hunt and capture prey. These are the search and approach phases, followed by the terminal buzz [Moss et al., 2011]. Through each of these phases, bats will increase the rate, shorten the duration, and even lower the frequency of their calls [Griffin et al., 1960]. Despite this, the distribution of time-frequency components within each species remains broadly similar across both calls and individuals [Jones and Holderied, 2007; Jones et al., 2009]. Diversity in the call structure is manifest as between-species variation, although closely related species do have similar calls [Collen, 2012]. This diversity has driven the development of algorithms for echolocation call classification, which may be applied for biodiversity monitoring [Redgwell et al., 2009; Stathopoulos et al., 2018; Mac Aodha et al., 2018]. In fact, Collen [2012] described 11 categories of tonal echolocation call, based on their time-frequency structure. When species are assigned to a guild, that is a functional group foraging under similar ecological conditions, members of each guild tend to possess structurally similar calls, irrespective of how closely related those species are [Denzinger and Schnitzler, 2013]. Thus, echolocation calls represent an example of convergent evolution and adaptive radiation [Jones and Holderied, 2007]

The time-frequency structures observed in echolocation calls reflect the theoretical basis for radar and sonar [Denny, 2007]. The most straightforward approach is to emit short, broadband signals and wait for echoes. In engineering terms, such a signal has a low-duty cycle and allows classification of a target given the arrival time of, and frequencies reflected in, the echo. A more sophisticated method is to use a long, narrow-band signal, with Doppler shifts in the echoes due to the relative motion of emitter and target allowing detection. This approach, employing a

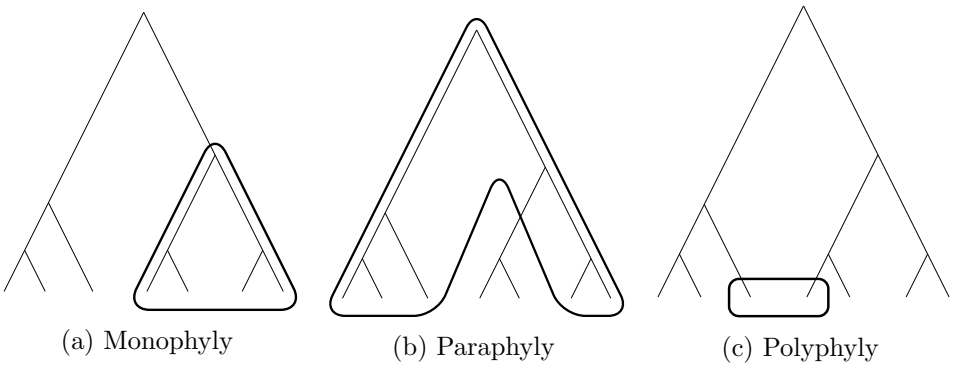
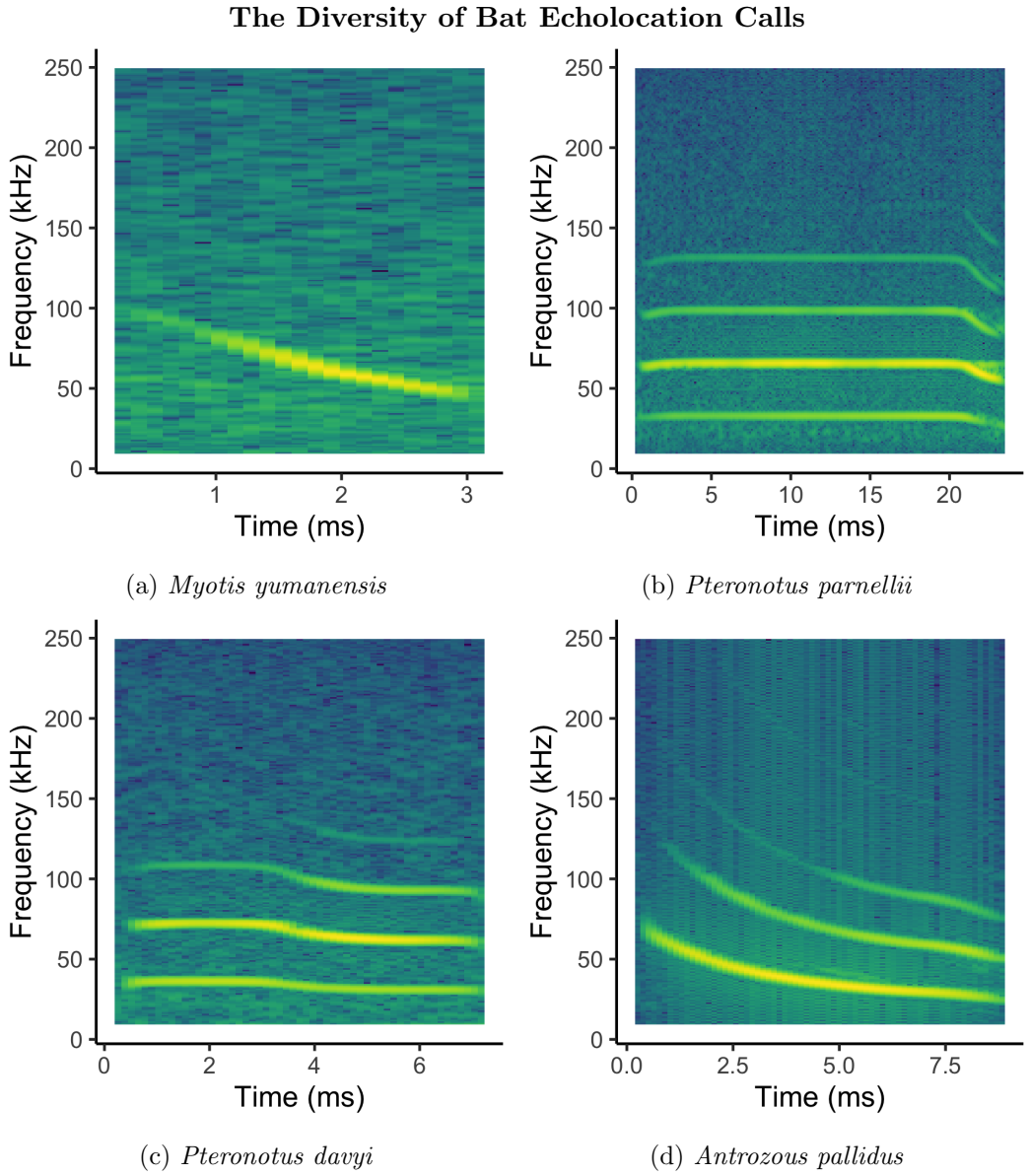


Figure 2.1: Definitions for monophyletic, paraphyletic, and polyphyletic groups. in each sub-plot the heavy black outline illustrates the taxa belonging to each definition. A monophyletic group, or monophyly, includes taxa that are all descendants of a unique common ancestor. Paraphyletic groups (paraphyly) are those where one or more monophyletic sub-groups have been kept apart from all other descendants of a unique common ancestor. Finally, the polyphyletic group (polyphyly) refers to taxa that do not share an immediate common ancestor [Felsenstein, 2004].

signal with a high-duty cycle, can provide a more extensive detection range. The implementations of each strategy found in bat echolocation calls [Jones and Teeling, 2006; Fenton et al., 2012; Collen, 2012] has been presented by Dawkins [1996] as an example of "*Good Design*" by nature. A selection of bat echolocation call spectrograms, illustrating various call structures, is presented in Figure 2.2.

Historically, the evolutionary history of bats has been a contentious issue. Debate on the topic arose when neurological data suggested that Megachiroptera were more closely related to primates and colugos (arboreal gliding mammals found in Southeast Asia) than to Microchiroptera [Pettigrew, 1986], implying that Chiroptera was, in fact, a polyphyletic group (see Figure 2.1c). This hypothesis has since been rejected as being unsupported by either morphological [Simmons, 1994] or molecular data [Ammerman and Hillis, 1992]. Another point of debate has been the position of Megachiroptera within the bat phylogeny. Phylogenetic trees based on the classification system of Miller [1907] split bats into the Megachiroptera and Microchiroptera sub-orders, based on laryngeal echolocation [Smith, 1976; Van Valen, 1979], however, modern techniques based on molecular data have consistently concluded that Megachiroptera are in fact nested within Microchiroptera [Teeling et al., 2000, 2005; Eick et al., 2005; Tsagkogeorga et al., 2013; Amador et al., 2018]. This has resulted in the consensus view that Chiroptera is a monophyletic group (Figure 2.1a) with the Most Recent Common Ancestor (MRCA) dated to 52-66 million years



(a) *Myotis yumanensis*

(b) *Pteronotus parnellii*

(c) *Pteronotus davyi*

(d) *Antrozous pallidus*

Figure 2.2: Selected bat echolocation call spectrograms, obtained via a short-time Fourier transform of call recordings, illustrating the diversity in call structures. The *Myotis yumanensis* (a) call is an example of a short duration broadband sweep, with a single frequency component. This is an example of a call having a low duty cycle. *Pteronotus parnellii* (b) has a high duty cycle, multi-harmonic call in which the second component dominates, consisting of a long constant frequency section followed by a short broadband sweep. *Pteronotus davyi* (c) and *Antrozous pallidus* (d) calls can then be described as narrowband and broadband frequency modulating multi-harmonic signals respectively.

ago, and the Microchiroptera sub-order being paraphyletic (Figure 2.1b) [Jones and Teeling, 2006; Amador et al., 2018].

Given this current understanding of bats evolutionary history, the most parsimonious explanation for the emergence of echolocation is that it evolved on a single occasion, at the root of the phylogeny. Pteropodids then lost the ability, only for echolocating species of the *Rosettus* genus to regain it [Jones and Teeling, 2006]. Patterns of fetal cochlear (the spiral cavity of the inner ear) development in pteropodids support this hypothesis [Wang et al., 2017]. As does an *Icaronycteris index* fossil, a species basal to all bats, displaying morphological characteristics similar to extant Microchiroptera [Jones and Teeling, 2006], although Eick et al. [2005] argued that the morphology of Rhinolophoidea family supports multiple origins. Furthermore, there remains debate on whether flight or echolocation evolved first, or indeed if both occurred in tandem, with no clear evidence to support any of these three hypotheses over the others [Simmons et al., 2008; Veselka et al., 2010].

To date, any attempts at the ancestral reconstruction of bat echolocation have been based either on supposition or high-level characteristics only. Simmons and Stein [1980] simply assumed that the ancestral bat used short, narrow-band, multi-harmonic signals with a low-duty cycle, based on the structure of bats larynx. On the other hand, Schnitzler et al. [2004] argued that broadband signals were ancestral. Collen [2012] performed a quantitative analysis of echolocation call characteristics for 410 species of extant bat which supported the conclusion of Schnitzler et al.; however, this analysis failed to account for the correlation structure within, and physical constraints on, the echolocation call, and reconstructions required significant post-processing before resembling those of extant species.

While the debate on bat's evolutionary history now seems to have been resolved, ancestral reconstruction of their echolocation call conditional on this history remains a challenging problem. Convergent evolution and adaptive radiation mean that distantly related taxa have developed similar call structures, which are subject to physical and design constraints, making the identification of intermediate developmental stages difficult. Furthermore, the complexity of the correlation structure within calls makes standard models for mapping traits to a phylogeny wholly unsuitable for the task. Tackling this problem requires careful consideration of both the features chosen to characterise calls and the model for traits evolving over the phylogeny.

2.2 The Phylogenetic Comparative Method

Phylogenetics is the study of evolutionary relationships between genetically related taxa, with the phylogenetic tree describing the evolution of each taxon in terms of branches radiating from a series of common ancestors [Felsenstein, 2004]. This tree is referred to as a *phylogeny*, which is derived from the Greek words *phylon*, meaning tribe or race, and *genetikós*, meaning origin or source [Ride et al., 1999; Liddell and Scott, 1897]. Early efforts at the algorithmic inference of phylogenies were based on parsimony criteria [Fitch, 1971], i.e. Occam's razor, which is to say that the phylogeny minimising character changes between observed taxa would be deemed most likely. Cavalli-Sforza and Edwards [1967] and Felsenstein [1973] were the first to develop formal statistical methods for phylogenetics, modelling the evolution of continuous characteristics as Brownian Motion (BM), which allowed maximum likelihood estimation of the phylogeny. This field has seen considerable progress in the intervening years. Modern methods take a Bayesian approach to inferring phylogenies from molecular sequences and analysis can be performed using open-source software [Drummond et al., 2002, 2012; Suchard et al., 2018; Bouckaert et al., 2019].

An important application of phylogenetics is the phylogenetic comparative analysis and ancestral reconstruction of phenotypes [Paradis, 2014; Joy et al., 2016]. A phenotype, referred to as a *trait* throughout this thesis, is some observable, measurable characteristic of an organism and is the result of interaction between that organism's genotype and environment [Campbell et al., 1997]. Therefore, as noted by Felsenstein [1985], traits sampled from genetically related taxa are not independent, due to their shared ancestry. This dependence, allowing ancestral reconstruction [Joy et al., 2016], must be accounted for when attempting to correlate traits with another variable. Any method for doing so is referred to as a Phylogenetic Comparative Method (PCM). Thus, PCMs are distinct from phylogenetics, though they are heavily dependant on the field, in that a PCM examines the distribution of traits among taxa once the phylogeny has been inferred [Paradis, 2014].

Typically, a PCM relies on some model for trait evolution. A popular choice is to model the trait as a Gauss-Markov process over the phylogeny [Rasmussen and Williams, 2006; Jones and Moriarty, 2013], that is, either as BM or an Ornstein-Uhlenbeck (OU) process [Felsenstein, 1973; Lande, 1976]. Alternatively, a heavy-tailed stable distribution could be employed [Elliot and Mooers, 2014]. Modelling trait evolution as BM is straightforward to justify and interpret for a continuous scalar-valued trait. Let $Y_t \in \mathbb{R}$ be the scalar-valued trait for the t^{th} generation,

where $\mathbb{R}$ denotes the set of real numbers. It is first assumed Y_t is independent of all earlier generations conditional on Y_{t-1} only. This is to say that the first-order Markov property holds [Billingsley, 2008], such that

$$p\left(Y_t = y_t | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots, Y_0 = y_0\right) = p\left(Y_t = y_t | Y_{t-1} = y_{t-1}\right).$$

Given that traits depend on the genotype, which is passed directly from one generation to the next, this would seem reasonable. Secondly, traits are assumed to change for each generation according to an independent and identically distributed process, with mean zero and finite variance, such that

$$\begin{aligned} \Delta y_t &\equiv y_t - y_{t-1}, \\ &= \epsilon_t, \end{aligned}$$

with $\mathbb{E}\left[\epsilon_t\right] = 0$ and $\mathbb{E}\left[\epsilon_t^2\right] = \sigma^2$. In this case, the Central Limit Theorem states that $\sqrt{t}Y_t \xrightarrow{d} \mathcal{N}\left(y_0, \sigma^2\right)$ [Casella and Berger, 2002], and so the dynamics of scalar-valued continuous traits over many generations can be modelled as BM. One problem with this model for trait evolution is that it fails to account for the fitness of an organism within its environment. It is possible that natural selection results in the trait tending towards some optimal value. To this end, Lande [1976] and Hansen [1997] proposed an OU model which incorporates this effect into the traits evolutionary dynamics. This is referred to as "*stabilising selection*" [Hansen, 1997]. For this model

$$\Delta y_t = \alpha(\mu - y_t) + \epsilon_t,$$

with changes in the trait value from one generation to the next tending towards an optimum $\mu \in \mathbb{R}$ according to the strength of selection $\alpha \in \mathbb{R}^+$, where the notation $\mathbb{R}^+ \equiv (0, \infty)$ will be employed throughout this thesis. The model can also be extended to accommodate a dynamic trait optimum by modelling μ itself as a function, either of evolutionary time or some other set of covariates.

A particularly important concept for phylogenetic comparative analysis is the notion of *phylogenetic signal*, that is, the tendency of traits from related taxa to resemble each other [Münkemüller et al., 2012]. One approach to quantifying this is to employ a Phylogenetic Mixed Model (PMM). The PMM, as defined by Housworth et al. [2004], assumes that trait evolution can be modelled as BM with variance $\sigma_h^2 \in \mathbb{R}^+$, referred to as the heritable variation. It then includes an additional parameter, $\sigma_e^2 \in \mathbb{R}^+$, which is referred to as the environmental, or non-phylogenetic, variation. This environmental variation is the variance of an independent Gaussian

noise process associated with the observed trait at each taxon, that is, variation independent of the phylogeny. Thus,

$$\kappa \equiv \frac{\sigma_h^2}{\sigma_h^2 + \sigma_e^2}, \tag{2.1}$$

defines the heritability of the process, which is the proportion of trait variation attributable to the stochastic process over the phylogeny. If this is close to 1, it implies strong heritability, and therefore a strong phylogenetic signal for the trait in question. Pagel's λ [1999a] and Blomberg's K [2003] offer two alternative approaches to assessing phylogenetic signal, each of which compares the actual variation amongst traits to that expected under a BM model for trait evolution.

Often, the objective of a phylogenetic comparative analysis is to establish the relationship between a trait and some set of covariates while controlling for dependence between taxa due to the phylogeny. Indeed, it was in this context that [Felsenstein, 1985] proposed his method of independent contrasts for real-valued traits. This approach was later generalised to phylogenetic regression by Grafen [1989], which has underpinned the development of Phylogenetic Generalised Least Squares (PGLS) [Hansen, 1997; Symonds and Blomberg, 2014]. In its simplest form, PGLS relates observations of a real valued trait for N taxa with a set of D covariates, given the phylogeny and model for trait evolution, according to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} \in \mathbb{R}^N$ are observed traits, $\mathbf{X}$ is the $N \times D$ matrix of covariates, $\boldsymbol{\beta} \in \mathbb{R}^D$ is the vector of regression coefficients, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ is the N-dimensional error vector modelling the traits random variation over the phylogeny as either BM or an OU process [Martins and Hansen, 1997]. This PCM can also be extended beyond real valued traits via a link function [Nelder and Wedderburn, 1972], with Ives and Garland Jr [2009] employing the logit link function to model a binary trait within this framework.

More recently, however, efforts have been focussed on developing PCMs which model the joint distribution of multivariate traits over a phylogeny [Adams and Collyer, 2017]. Doing so allows a covariance structure over multiple traits to be defined within a single model for trait evolution, rather than attempting to fit and interpret many instances of PGLS. Revell [2009] extended Principal Components Analysis (PCA) [Tipping and Bishop, 1999] to real-valued multivariate traits, assuming that the phylogeny and model for trait evolution is known. Furthermore,

multivariate PCMs have been generalised to collections of continuous and discrete traits by Felsenstein [2011] using the threshold model proposed by Wright [1934]. This model, which is analogous to probit regression [Albert and Chib, 1993], assumes that discrete traits are associated with some unobserved auxiliary variables, which Felsenstein [2011] refers to as liabilities. Discrete traits change state as auxiliary variables cross particular thresholds, where auxiliary variables are modelled as a Gauss-Markov process over the phylogeny. This allows both ordinal and categorical traits to be modelled alongside those that are real-valued. Markov Chain Monte Carlo (MCMC) algorithms for Bayesian inference on the threshold model have been developed by Cybis et al. [2015] and Tolkoff et al. [2017]. Each of these implementations allows integration over a distribution of phylogenies, which can be inferred from molecular sequences associated with the taxa of interest [Bouckaert et al., 2019]. Thus, uncertainty on the phylogeny can be accounted for within a PCM. Of these models, Phylogenetic Factor Analysis (PFA) is of particular interest [Tolkoff et al., 2017]. In this case, a latent variable model is assumed for auxiliary variables, similar to Factor Analysis [Bartholomew et al., 2011; Lopes, 2014], such that

$$\mathbf{X} = \mathbf{Z}\mathbf{W}^\top + \boldsymbol{\epsilon},$$

where $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the matrix of auxiliary variables, $\mathbf{Z} \in \mathbb{R}^{N \times Q}$ are factors, such that each column is assumed to be an independent BM over the phylogeny, $\mathbf{W} \in \mathbb{R}^{D \times Q}$ is the loading, and $\boldsymbol{\epsilon} \in \mathbb{R}^{N \times D}$ is independent Gaussian observation noise. A similar approach to modelling trait evolution will be employed by the models developed in this thesis.

Each of the PCMs outlined thus far is concerned with (collections of) scalar-valued continuous and discrete traits, however, some traits are best described as continuous functions of time (or some other reference variable). Such a trait is an infinite-dimensional object, in that it could be recorded an arbitrary set of points over an interval, and is referred to as a function-valued trait (FVT) [Kirkpatrick and Heckman, 1989; Kirkpatrick et al., 1990; Meyer and Kirkpatrick, 2005; Gomulkiewicz et al., 2018]. FVTs pose a particular set of challenges for evolutionary inference. They are functional data objects and as such, are subject to Functional Data Analysis (FDA) techniques such as smoothing and registration [Ramsay, 2004; Srivastava and Klassen, 2016], discussed in more detail in sub-section 2.3.4. Furthermore, FVTs are generally assumed to vary slowly and continuously with respect to time [Meyer and Kirkpatrick, 2005]. As such, there exists a covariance structure within the trait which is not explicitly modelled by methods such as phylogenetic PCA of PFA [Revell, 2009; Tolkoff et al., 2017]. To address these issues Jones and Moriarty

[2013] proposed the phylogenetic Gaussian process regression (PGPR) framework, linking Gaussian processes to the evolution of FVTs [Rasmussen and Williams, 2006]. The development of this framework, which will be discussed in greater detail in sub-section 2.3.2, is ongoing. It has been linked to PGLS [Goolsby, 2015], and approximations to PGPR applied to synthetic data [Hadjipantelis et al., 2013] and bat echolocation calls [Meagher et al., 2018a,b]. The framework has also been applied to the evolution of multi-dimensional facial curves [Mariñas-Collado et al., 2019].

As a final note, typical methods for phylogenetic comparative analysis imply some distribution over trait values for ancestral taxa [Martins and Hansen, 1997; Jones and Moriarty, 2013; Tolkoff et al., 2017]. Thus, ancestral reconstruction and the PCM can be thought of as two sides of the same coin with each offering its own perspective on the evolutionary relationships between taxa [Joy et al., 2016].

2.3 Statistical Models for Data

2.3.1 Gaussian Processes

Gaussian processes, ubiquitous in the disciplines of Statistics and Machine Learning [Rasmussen and Williams, 2006; Stein, 2012], offer an approach to non-parametric regression that is both flexible and analytically tractable. A brief discussion on Gaussian process regression and the importance of covariance functions, referred to as kernels, is presented in the following. For a full treatment of Gaussian processes, the interested reader can refer to Rasmussen and Williams [2006].

In order to understand the appeal of a Gaussian process (GP), consider $\mathbf{y} \equiv (y_1, \ldots, y_N)^\top$, the instance of a multivariate Gaussian distributed random variable, such that

$$p(\mathbf{y}) \equiv \mathcal{N}(\mathbf{y}|\mathbf{m}, \mathbf{K}) \equiv |2\pi\mathbf{K}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{m})^\top \mathbf{K}^{-1}(\mathbf{y}-\mathbf{m})\right), \tag{2.2}$$

defines the Gaussian probability density function (pdf) for mean $\mathbf{m}$ and covariance $\mathbf{K}$. Two particularly useful properties of the Gaussian distribution are that it is closed under both marginalisation and conditioning. That is to say, when

$$p(\mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}_A \\ \mathbf{y}_B\end{bmatrix} \mid \begin{bmatrix}\mathbf{m}_A \\ \mathbf{m}_B\end{bmatrix}, \begin{bmatrix}\mathbf{K}_{AA} & \mathbf{K}_{AB} \\ \mathbf{K}_{BA} & \mathbf{K}_{BB}\end{bmatrix}\right), \tag{2.3}$$

it can be shown that

$$p(\mathbf{y}_A) = \int_{\mathbb{R}} p(\mathbf{y}_A, \mathbf{y}_B)\, d\mathbf{y}_B = \mathcal{N}(\mathbf{y}_A | \mathbf{m}_A, \mathbf{K}_{AA}), \tag{2.4}$$

and

$$p(\mathbf{y}_A \mid \mathbf{y}_B) = \mathcal{N}\left(\mathbf{y}_A | \mathbf{m}_A + \mathbf{K}_{AB}\mathbf{K}_{BB}^{-1}(\mathbf{y}_B - \mathbf{m}_A), \mathbf{K}_{AA} - \mathbf{K}_{AB}\mathbf{K}_{BB}^{-1}\mathbf{K}_{BA}\right). \tag{2.5}$$

Thus, for any set of Gaussian distributed random variables, there exist analytically tractable definitions of the marginal and conditional distributions for each element. GPs extend these notions to infinite dimensions.

A Gaussian process is *a collection of random variables, any finite number of which have a joint Gaussian distribution* [Rasmussen and Williams, 2006]. Letting $\mathcal{X}$ denote the space over which a GP is observed (typically $\mathcal{X} \equiv \mathbb{R}^d$), the GP $f(\mathbf{x}) \in \mathbb{R}$, defined as

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right), \tag{2.6}$$

is fully specified by its *mean function* $m(\mathbf{x})$ and *covariance function* $k(\mathbf{x}, \mathbf{x}')$, where

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k\left(\mathbf{x}, \mathbf{x}'\right) &= \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f\left(\mathbf{x}'\right) - m\left(\mathbf{x}'\right)\right)\right]. \end{aligned}$$

Without any loss of generality, it can be assumed that $m(\mathbf{x}) = 0$, and so the process is described by its second-order statistics only.

The convenience of a GP prior can be illustrated given observation y_n indexed by $\mathbf{x}_n$ for $n = 1, \ldots, N$, which is modelled as an instantiation of a GP such that

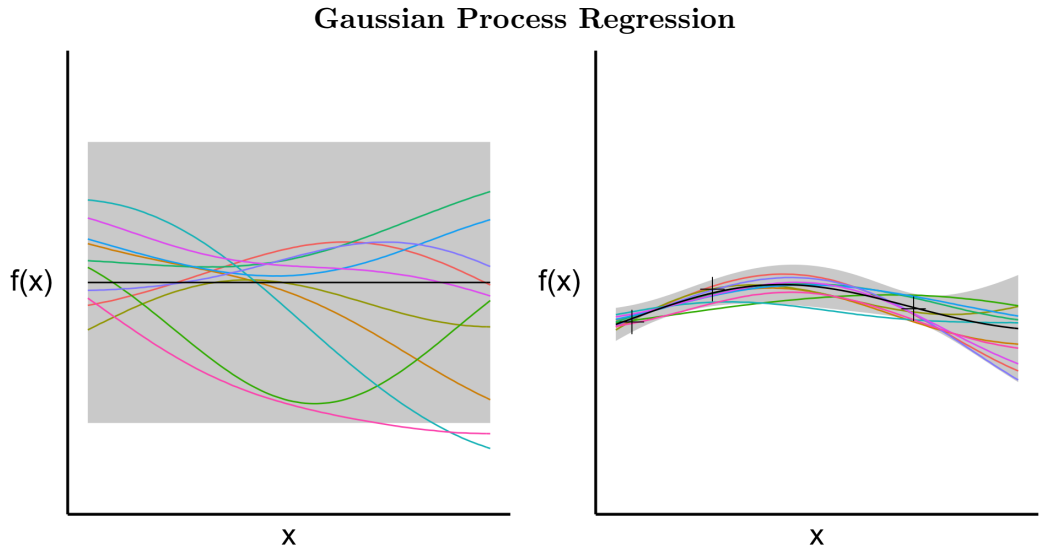
$$y_n = f(\mathbf{x}_n) + \epsilon_n$$

for $\epsilon_n \sim \mathcal{N}\left(0, \lambda^{-1}\right)$. Letting $\mathbf{f}_* \equiv f(\mathbf{x}_*)$ for the unobserved index $\mathbf{x}_* \in \mathcal{X}$, it can be shown that

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{f}\,\mathbf{f}} + \lambda^{-1}\mathbf{I}_N & \mathbf{k}_{\mathbf{f}*} \\ \mathbf{k}_{\mathbf{f}*}^\top & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

where $\mathbb{E}[f(\mathbf{x})] = 0$, $(\mathbf{K}_{\mathbf{f}\,\mathbf{f}})_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$ such that $\mathbf{K}_{\mathbf{f}\,\mathbf{f}}$ is the Gram matrix of $k(\cdot, \cdot)$ for $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ [Rasmussen and Williams, 2006], and $(\mathbf{k}_{\mathbf{f}*})_n = k(\mathbf{x}_n, \mathbf{x}_*)$. This is a joint Gaussian distribution, the pdf of which is given in (2.3), and so the distribution of $\mathbf{f}_*$ conditional on $\mathbf{y}$ is given by (2.5). Thus, Gaussian process regression allows the definition of a posterior distribution for all $\mathbf{x}_* \in \mathcal{X}$.

When modelling data as a GP, careful consideration must be given to the co-



(a) Gaussian process prior samples (b) Gaussian process posterior samples

Figure 2.3: Prior and posterior distributions for a Gaussian process with an exponentiated quadratic covariance function defined for $\mathbf{x} \in \mathbb{R}$. Grey shaded regions represent two standard deviations of about the mean, which is illustrated as a black line. Samples are then represented by coloured lines. The posterior distribution is obtained by performing Gaussian process regression given three noisy observations of the underlying Gaussian process, represented by crosses.

variance function chosen. While simple linear trends and effects from covariates can be included in the mean function, the covariance function encodes any assumptions on the underlying stochastic process. For $k(\cdot,\cdot)$ to be a valid covariance function, its Gram matrix, denoted $\mathbf{K}$, must be positive semi-definite, which is to say that $\mathbf{z}^\top \mathbf{K}\mathbf{z} \geq 0$ for all $\mathbf{z} \in \mathbb{R}^N$. When this is the case $k(\cdot,\cdot)$ is a *Mercer kernel*, where the term *kernel* refers to any function mapping two inputs to the real numbers [Scholkopf and Smola, 2001].

There are a number of properties to be considered when choosing a kernel to model any given phenomenon. Assuming that $\mathcal{X} \equiv \mathbb{R}^d$, it is often desirable for $k(\cdot,\cdot)$ to be *weakly stationary*, which is to say that it is a function of $\boldsymbol{\tau} \equiv \mathbf{x} - \mathbf{x}'$ such that $k(\boldsymbol{\tau}) \equiv k(\mathbf{x}, \mathbf{x}')$ [Rasmussen and Williams, 2006]. A more restrictive assumption is to assume that the kernel is *weakly isotropic*, in which case it is a function of $r \equiv |\boldsymbol{\tau}|$, where $|\cdot|$ denotes Euclidean distance [Rasmussen and Williams, 2006]. It is also important to consider mean square continuity and differentiability, which describe the smoothness of a stochastic process. The stochastic process $f(\cdot)$

is *mean square continuous* at $\mathbf{x} \in \mathbb{R}^d$ if

$$\lim_{\mathbf{x}' \to \mathbf{x}} \mathbb{E}\left[f\left(\mathbf{x}'\right) - f\left(\mathbf{x}\right)\right]^2 \to 0,$$

while it is *mean square differentiable* if the limit

$$\lim_{h \to 0} \mathbb{E}\left[\left(\frac{f\left(\mathbf{x} + h\mathbf{e}_i\right) - f\left(\mathbf{x}\right)}{h}\right)^2\right] = \frac{\partial f\left(\mathbf{x}\right)}{\partial x_i},$$

exists, where $\mathbf{e}_i$ is the unit vector along the i^{th} dimension [Banerjee and Gelfand, 2003; Stein, 2012].

The *Matérn* class of isotropic covariance functions is given by

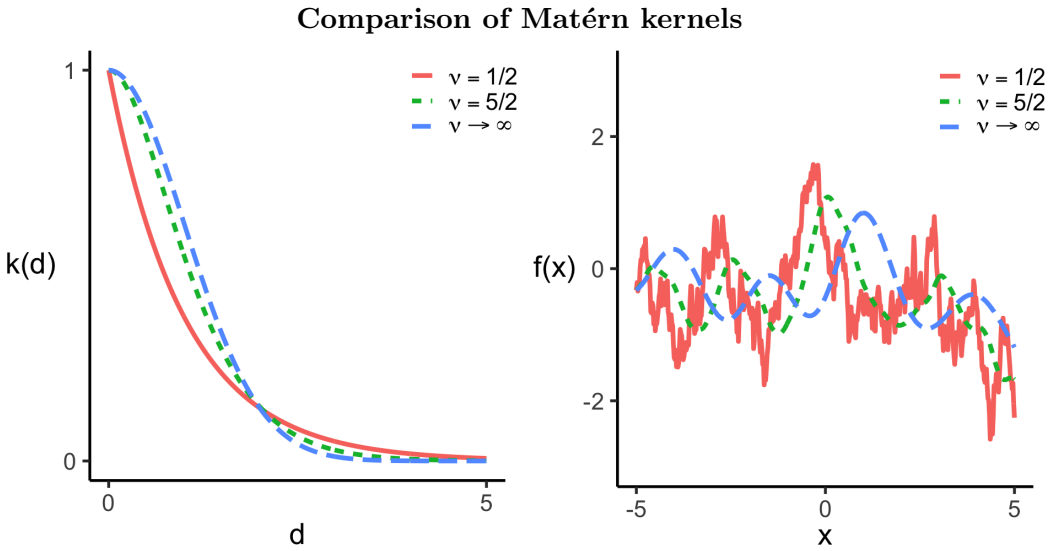
$$k_\nu\left(r \mid \sigma^2, \ell\right) \equiv \sigma^2 \frac{2^{1-\nu}}{\Gamma\left(\nu\right)} \left(\frac{\sqrt{2\nu} r}{\ell}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu} r}{\ell}\right), \tag{2.7}$$

where the variance $\sigma^2 \in \mathbb{R}^+$, smoothing parameter $\nu \in \mathbb{R}^+$, and characteristic length-scale $\ell \in \mathbb{R}^+$. $K_\nu\left(\cdot\right)$ is then a modified Bessel function [Stein, 2012]. Important properties of the Matérn class are defined with respect to the smoothing parameter ν. Firstly, the process $f\left(\mathbf{x}\right)$ is k-times mean square differentiable if and only if $k > \nu$. Furthermore, when $\nu = p + \frac{1}{2}$ for a non-negative integer p, a simplified expression of (2.7) is obtained and, when $d = 1$, the resulting model is a form of autoregressive process of order $p + 1$ [Rasmussen and Williams, 2006]. The cases $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\right\}$ and $\nu \to \infty$ are of particular interest in Machine Learning. In fact, the limiting case, when $\nu \to \infty$, is the popular *exponentiated quadratic* covariance function

$$k_{EQ}\left(r \mid \sigma^2, \ell\right) \equiv \sigma^2 \exp\left(-\frac{r^2}{2\ell^2}\right), \tag{2.8}$$

for which σ^2 defines the process amplitude and ℓ the rate at which correlation decays with increasing r, as is the case for all kernels of the Matérn class.

There exist many other kernels suitable for use as covariance functions in Gaussian processes including the polynomial, periodic, and neural network kernels [Rasmussen and Williams, 2006], however, it is important to note that a single kernel does not have to be chosen. Mercer kernels are closed under both multiplication and addition allowing multiple kernels can be combined in a single analysis [Rasmussen and Williams, 2006]. Further detail on Gaussian process regression, covariance functions, and GPs in general can be found in both [Rasmussen and Williams, 2006] and [Stein, 2012].



(a) Process covariance function

(b) Process samples

Figure 2.4: Sub-plot (a) presents a comparison of isotropic Matérn kernels for different values of smoothing parameter ν where $\sigma^2 = 1$ and $\ell = 1$. It can be seen that as ν increases the kernel decays more slowly close to 0, implying smoother function realisations. Samples from each process, instantiated with the same seed, illustrate this clearly in (b).

2.3.2 The Phylogenetic Gaussian Process Framework

Consider a FVT, defined over the *phylogeny-trait space* $\mathcal{T} \times \mathcal{X}$, where $\mathcal{T}$ denotes a phylogeny, for which branch lengths are proportional to evolutionary time between taxa, and $\mathcal{X}$ the space over which the FVT is observed. Modelling this as a GP implies that

$$f(\mathbf{x}, \mathbf{t}) \sim \mathcal{GP}\left(0, k\left((\mathbf{t}, \mathbf{x}), (\mathbf{t}, \mathbf{x})'\right)\right), \tag{2.9}$$

for $(\mathbf{t}, \mathbf{x}) \in \mathcal{T} \times \mathcal{X}$, where the Mercer kernel $k(\cdot, \cdot)$ will be referred to as the *phylogeny-trait covariance function*.[1] Thus, a model for the evolution of a FVT is fully specified by $k(\cdot, \cdot)$.

Jones and Moriarty [2013] define the PGPR framework in terms of a separable phylogeny-trait covariance function, such that

$$k\left((\mathbf{t}, \mathbf{x}), (\mathbf{t}, \mathbf{x})'\right) = k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right) k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right),$$

[1] Jones and Moriarty [2013] refer to this as the phylogenetic covariance function. The terminology has been changed in order to distinguish between covariance structures over $\mathcal{T}$ and $\mathcal{X}$

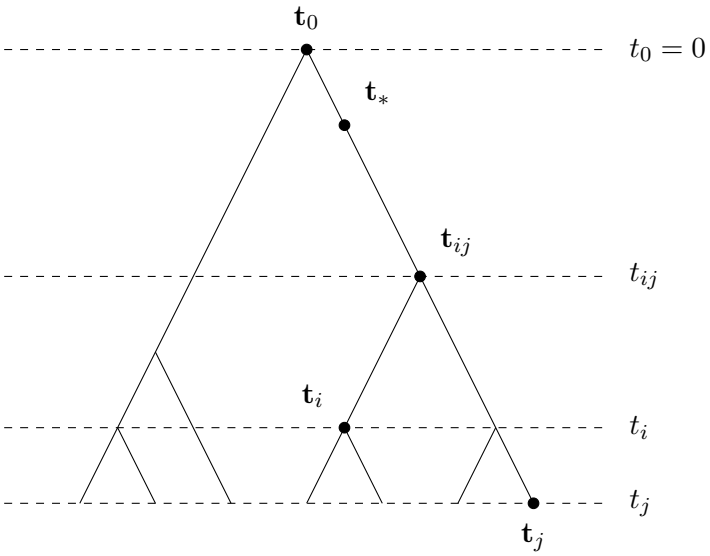


Figure 2.5: An example of a bifurcating phylogeny. Here, $\mathbf{t}_i$, $\mathbf{t}_j$, $\mathbf{t}_{ij}$, $\mathbf{t}_*$, and $\mathbf{t}_0$ each denote a position on $\mathcal{T}$. $\mathbf{t}_0$ is the taxon at the root of the phylogeny, while $\mathbf{t}_{ij}$ is the MRCA for the taxa at $\mathbf{t}_i$ and $\mathbf{t}_j$, and $\mathbf{t}_*$ is an ancestor of $\mathbf{t}_{ij}$. Furthermore, each position $\mathbf{t} \in \mathcal{T}$ is associated with a depth, denoted t, which distance of $\mathbf{t}$ from the root of $\mathcal{T}$. A more rigorous definition of a phylogeny will be provided in section 3.2.1.

where $k_{\mathcal{T}}(\cdot,\cdot)$ is the *phylogenetic covariance function* and $k_{\mathcal{X}}(\cdot,\cdot)$ the *trait covariance function*, each of which are Mercer kernels. Consider first the phylogenetic covariance function, specification of which relies on two standard assumptions in the context of evolution [Felsenstein, 1973].

Assumption 1. *Conditional on their most recent common ancestor on the phylogeny $\mathcal{T}$, traits at $\mathbf{t}$ and $\mathbf{t}'$ are statistically independent.*[2]

Assumption 2. *The statistical relationship between the trait at $\mathbf{t} \in \mathcal{T}$ and its descendants is independent of the topology of $\mathcal{T}$.*

In order to understand the implications of these assumptions, consider a

[2] Jones and Moriarty [2013] assume traits at $\mathbf{t}$ and $\mathbf{t}'$ are statistically independent given common ancestors, rather than the stronger assumption made here, suggesting that the process may also be dependant on ancestors of the MRCA. Despite this, the phylogenetic covariance functions for which the PGPR framework is developed imply that, given the MRCA, traits at $\mathbf{t}$ and $\mathbf{t}'$ are independent not only of each other but also any ancestors of the MRCA. Thus, this assumption has been made explicit here.

univariate GP over $\mathcal{T}$ such that

$$z\left(\mathbf{t}\right) \sim \mathcal{GP}\left(0, k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right)\right).$$

Assumption 1 simply states that the Markov property holds for this process over $\mathcal{T}$ such that

$$p\left(z\left(\mathbf{t}_i\right), z\left(\mathbf{t}_j\right) | z\left(\mathbf{t}_{ij}\right), z\left(\mathbf{t}_*\right)\right) = p\left(z\left(\mathbf{t}_j\right) | z\left(\mathbf{t}_{ij}\right)\right) p\left(z\left(\mathbf{t}_i\right) | z\left(\mathbf{t}_{ij}\right)\right),$$

where, throughout this sub-section, $\mathbf{t}_*$ is an ancestor $\mathbf{t}_{ij}$ on $\mathcal{T}$ and the taxon at $\mathbf{t}_{ij}$ is the MRCA of taxa at $\mathbf{t}_i$ and $\mathbf{t}_j$, as presented in Figure 2.5.

Assumption 2, on the other hand, describes the Gaussian process modelling trait evolution along the individual paths through $\mathcal{T}$ from its root to each tip. This is referred to as the *marginal process* [Jones and Moriarty, 2013] and it is assumed to be identically distributed along each path. Furthermore, in order to satisfy Assumption 1, the marginal process must have the Markov property.

These assumptions allow the phylogenetic covariance function to be defined as follows. Let the distance of position $\mathbf{t} \in \mathcal{T}$ from the root of $\mathcal{T}$ be the *depth* of $\mathbf{t}$, denoted t. The covariance function of the marginal process can then be defined as $\tilde{k}\left(t, t'\right)$ for positions $\mathbf{t}$ and $\mathbf{t}'$ lying on a single path through $\mathcal{T}$. Then, for arbitrary positions, $\mathbf{t}_i$, $\mathbf{t}_j$, and their MRCA $\mathbf{t}_{ij}$ it can be seen that

$$\begin{aligned}
k_{\mathcal{T}}(\mathbf{t}_i, \mathbf{t}_j) &= \mathbb{E}\left[z(\mathbf{t}_i) z(\mathbf{t}_j)\right] && (2.10)\\
&= \mathbb{E}\left[\mathbb{E}\left[z(\mathbf{t}_i) z(\mathbf{t}_j) | z(\mathbf{t}_{ij})\right]\right], && (2.11)\\
&= \mathbb{E}\left[\mathbb{E}\left[z(\mathbf{t}_i) | z(\mathbf{t}_{ij})\right] \mathbb{E}\left[z(\mathbf{t}_j) | z(\mathbf{t}_{ij})\right]\right], && (2.12)\\
&= \mathbb{E}\left[\tilde{k}(t_i, t_{ij}) \tilde{k}(t_{ij}, t_{ij})^{-1} z(\mathbf{t}_{ij}) \tilde{k}(t_j, t_{ij}) \tilde{k}(t_{ij}, t_{ij})^{-1} z(\mathbf{t}_{ij})\right], && (2.13)\\
&= \tilde{k}(t_i, t_{ij}) \tilde{k}(t_{ij}, t_{ij})^{-1} \tilde{k}(t_{ij}, t_j). && (2.14)
\end{aligned}$$

where (2.10) is the definition of covariance for a process with zero mean, (2.11) holds by the law of iterated expectations [Casella and Berger, 2002], (2.12) is given by Assumption 1, (2.13) is a result of the conditional mean of Gaussian random variables, and (2.14) is simply the expected value.

The covariance function for the marginal process must be defined in order to complete the specification of a phylogenetic covariance function. Two classes of continuous-time Gauss-Markov processes are considered, Brownian Motion (BM) and the Ornstein-Uhlenbeck (OU) process. The covariance function for a BM

marginal process can be expressed as

$$\tilde{k}^{bm}(t, t') = \sigma_h^2 \min(t, t'),$$

for variance $\sigma_h^2 \in \mathbb{R}^+$. This implies that

$$k_{\mathcal{T}}^{bm}(\mathbf{t}_i, \mathbf{t}_j) = \sigma_h^2 t_{ij},$$

which defines a kernel for the BM model of trait evolution [Felsenstein, 1973, 1985; Cybis et al., 2015; Tolkoff et al., 2017].

Alternatively, an OU process, the class of stationary Gauss-Markov processes [Doob, 1942], can be assumed such that

$$\tilde{k}^{ou}\left(t, t'\right) = \sigma_h^2 \exp\left(-\frac{|t - t'|}{\ell}\right)$$

for variance $\sigma_h^2 \in \mathbb{R}^+$ and characteristic length-scale $\ell \in \mathbb{R}^+$. It is worth noting that this covariance function belongs to the Matérn class for which it is equivalent to (2.7) when $\nu = 1/2$ [Rasmussen and Williams, 2006]. This allows the definition of a phylogenetic covariance function

$$\begin{aligned} k_{\mathcal{T}}^{ou}\left(\mathbf{t}_i, \mathbf{t}_j\right) &= \sigma^2 \exp\left(-\frac{|t_i - t_{ij}| + |t_j - t_{ij}|}{\ell}\right), \\ &= \sigma^2 \exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right)}{\ell}\right), \end{aligned}$$

where $\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right)$ is the *patristic distance* between $\mathbf{t}_i$ and $\mathbf{t}_j$ on $\mathcal{T}$ [Rédei, 2008; Jones and Moriarty, 2013], that is, the sum of differences in depth between each position and their MRCA. Thus, a phylogenetic covariance function for the OU model of trait evolution can also be defined [Hansen, 1997].

As a final note on the phylogenetic covariance function, introducing an independent Gaussian noise process for traits at observed taxa does not violate any model assumptions. Thus, it is straightforward to incorporate the PMM presented by Housworth et al. [2004] into these phylogenetic covariance functions.

In order to extend this univariate phylogenetic GP to a FVT, note that by Mercer's theorem [Rasmussen and Williams, 2006]

$$k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right) = \sum_{i=1}^{\infty} \xi_i^{\mathcal{X}} u_i^{\mathcal{X}}(\mathbf{x}) u_i^{\mathcal{X}}(\mathbf{x}'),$$

for eigenvalues $\xi_i^{\mathcal{X}}$ and eigenfunctions $u_i^{\mathcal{X}}(\mathbf{x})$. Then, when

$$f(\mathbf{x}, \mathbf{t}) = \sum_{i=1}^{\infty} \sqrt{\xi_i^{\mathcal{X}}} u_i^{\mathcal{X}}(\mathbf{x}) z_i(\mathbf{t})$$

for $z_i(\mathbf{t}) \sim \mathcal{GP}(0, k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}'))$, it can be shown that the FVT is being modelled as a phylogeny-trait separable GP such that

$$f(\mathbf{x}, \mathbf{t}) \sim \mathcal{GP}\left(0, k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right) k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right)\right), \tag{2.15}$$

as desired. Thus, the PGPR framework has been fully specified, providing a coherent approach to evolutionary inference for FVTs.

As a final remark on the PGPR framework, it is important to note that separability of the phylogeny-trait covariance function is a very restrictive assumption. Not only does it imply that the trait covariance function is constant with respect to the phylogeny, but it does not accommodate more standard modelling assumptions. For example, $k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}') k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') + \sigma^2 \delta(\mathbf{x} = \mathbf{x}')$, where $\delta(\cdot)$ is the indicator function, is not separable, which implies that a separable phylogeny-trait covariance function cannot include independent observation noise on traits. Furthermore, some variation in the trait covariance function over the phylogeny may be desirable. Such a model could be applied to bat echolocation calls to allow different families of bat their own family-level trait covariance functions, offering a far more flexible model for their evolution. Despite the appeal of such phylogeny-trait covariance functions however, some structure must be imposed. Ancestral reconstruction and evolutionary inference become impossible when there is no defined relationship between extant taxa and their common ancestors, separable phylogeny-trait covariance functions provide a useful tool for defining these relationships. Thus, relaxing the separability assumption, while preserving key elements of the structure and intuition it provides, allows for the development of novel methods for evolutionary inference presented later in this thesis.

2.3.3 Latent Variable Models and Factor Analysis

Solutions to a range of statistical problems, including probit regression for discrete variables [Albert and Chib, 1993] and hidden Markov models for sequential data [Rabiner, 1989], can be cast as latent variable models. Such a model relates observed *manifest variables* $\mathbf{y}_n \equiv (y_{1n}, \ldots, y_{Dn})^\top$ to unobserved *latent variables* $\mathbf{z}_n = (z_{1n}, \ldots, z_{Qn})^\top$, for $n = 1, \ldots, N$. A particularly important class of latent variable model, for which manifest variables are assumed to be independent and

identically distributed, is Factor Analysis (FA) [Bartholomew et al., 2011], where

$$\mathbf{y}_n = \boldsymbol{\mu} + \mathbf{W}\mathbf{z}_n + \boldsymbol{\epsilon}_n, \tag{2.16}$$

with mean $\boldsymbol{\mu} \in \mathbb{R}^D$, loading $\mathbf{W} \in \mathbb{R}^{D \times Q}$, factors $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_Q)$,[3] and observation noise $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, for the diagonal covariance matrix $\boldsymbol{\Psi}$.

The motivation for FA is that, when $Q << D$, factors provide a parsimonious description of the variation between manifest variables, while the loading defines variation within those manifest variables [Lopes and West, 2004]. Such a model can provide a useful interpretation for observed data. Indeed, Spearman [1904] originally formulated FA to produce an objective measure of intelligence from multiple test scores. Integrating over latent variables provides another important perspective on FA. The marginal distribution for manifest variables is

$$\mathbf{y}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{2.17}$$

where $\boldsymbol{\Omega} = \mathbf{W}\mathbf{W}^\top + \boldsymbol{\Psi}$. This demonstrates that FA is in fact modelling the covariance matrix of manifest variables, however $\boldsymbol{\Omega}$ depends on $D(Q+1)$ parameters, rather than $D(D+1)/2$ as it does in the unconstrained case. Thus, when $Q << D$, FA provides a low rank approximation to the covariance matrix of manifest variables [Lopes, 2014].

FA provides a flexible model for data, however, as defined in (2.16) the loading is non-identifiable. The marginal distribution in (2.17) is invariant to reflection and rotation of $\mathbf{W}$. Reflection invariance is a result of $(-\mathbf{W})(-\mathbf{W})^\top = \mathbf{W}\mathbf{W}^\top$, while, for the orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_Q$, rotation invariance is shown by noting that $(\mathbf{W}\mathbf{Q})(\mathbf{W}\mathbf{Q})^\top = \mathbf{W}\mathbf{W}^\top$. Correcting for reflection invariance is straightforward, simply fixing diagonal elements of $\mathbf{W}$ to be strictly positive typically does so [Geweke and Zhou, 1996; Lopes and West, 2004]. Alternatively, in the context of posterior inference using MCMC samples, post-hoc relabelling algorithms based on that developed by Stephens [2000] have also been proposed [Erosheva and Curtis, 2017; Tolkoff et al., 2017]. For the correction of rotation invariance, one approach is to specify $\mathbf{W}$ such that $\mathrm{Var}(\mathbf{z}_n|\mathbf{y}_n) = \left(\mathbf{W}^\top\boldsymbol{\Psi}^{-1}\mathbf{W} + \mathbf{I}_Q\right)^{-1}$ is diagonal [Seber, 2009]. More popular in Bayesian FA however [Lopes and West, 2004; Lopes, 2014], is to fix upper-triangular entries of $\mathbf{W}$ to 0, as introduced by Geweke and Zhou [1996]. That this constraint fixes rotation invariance is a result of the QR

[3] This assumption is not in any way restrictive of FA, if the model were parametrised by $\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ with an arbitrary covariance matrix $\mathbf{V} = \mathbf{L}\mathbf{L}^\top$, an equivalent model could be parametrised by $\mathbf{W}' = \mathbf{W}\mathbf{L}$ and $\mathbf{z}'_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_Q)$ [Lopes, 2014].

decomposition [Golub and Van Loan, 2013], which states that any square matrix $\mathbf{A}$ may be decomposed as

$$\mathbf{A} = \mathbf{QR},$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is upper triangular. Furthermore, $\mathbf{Q}$ is unique when the diagonal elements of $\mathbf{R}$ are strictly positive. The QR decomposition extends to the $D \times Q$ matrix for which upper triangular entries are 0, and as such, $\mathbf{W}$ is no longer invariant to rotation.

FA is widely applied, and serves as a basis for many useful extensions. Probabilistic principal components analysis is formulated by assuming that $\mathbf{\Psi} \equiv \sigma^2 \mathbf{I}_D$ [Tipping and Bishop, 1999], which in turn motivates Gaussian Process Latent Variable Models [Lawrence, 2005; Titsias and Lawrence, 2010] and structured principal components analysis [Skinner, 2019], while Tolkoff et al. [2017] extended FA to phylogenetic comparative analysis.

2.3.4 Functional Data Analysis

Functional Data Analysis (FDA) is the branch of statistics concerned with the study of data generated by continuous processes [Ramsay, 2004; Srivastava and Klassen, 2016]. Such data occur across many scientific disciplines and pose challenges that are not considered by standard multivariate methods. In particular, functional data typically requires smoothing and registration as part of its analysis, techniques for which are outlined in the following.

In general, the analysis of functional data starts with a set of discrete observations and associated time points $(y_d, t_d) \in \mathbb{R} \times [0,1]$ for $d = 1, \ldots, D$, from which the underlying function $f(\cdot)$ must be estimated. This estimation of the underlying function is referred to as smoothing [Ramsay, 2004].

It is assumed that $f(t) \in \mathbb{R}$ for all $t \in [0,1]$, and $\int_0^1 f^2(t)\, dt < \infty$, which is to say that $f(\cdot)$ belongs to the set of real valued, square integrable functions on the unit interval, denoted $L^2([0,1], \mathbb{R})$, or more simply L^2. In addition, equipping L^2 with the inner product

$$\langle f, g \rangle_2 = \int_0^1 f(t)\, g(t)\, dt, \qquad \text{for } f(\cdot), g(\cdot) \in L^2.$$

defines a Hilbert space with norm $||f||_2 = \sqrt{\int_0^1 f^2(t)\, dt}$ [Srivastava and Klassen, 2016]. Observations can then be modelled as

$$y_d = f(t_d) + \epsilon_d, \tag{2.18}$$

for $\mathbb{E}[\epsilon_d] = 0$ and $\mathbb{E}\left[\epsilon_d^2\right] < \infty$, which is to say that observations of the underlying function are subject to a noise process with zero mean and finite variance.

Without placing any further constraints on the underlying process, a potential solution to this problem would be to simply model $f(\cdot)$ as a piecewise linear interpolation between data points. This would define $f(\cdot)$ over the entire interval provided there exists $t_d = 0$ and $t_{d'} = 1$, however such an approach generalises very poorly in the presence of noise and does not allow continuous derivatives of $f(\cdot)$ to be estimated, objects which are often of great interest in functional data analyses [Ramsay, 2004]. A popular alternative is to instead assume $f(\cdot)$ to be the smooth, twice-differentiable function which minimises the penalised residual sum of squares

$$\mathcal{L}_{rss}(f, \lambda) \equiv \sum_{d=1}^{D} \left(y_d - f\left(t_d\right)\right)^2 + \lambda \langle f'', f'' \rangle_2, \tag{2.19}$$

where λ is the *smoothing parameter* penalising the function's second derivative [Friedman et al., 2001; Ramsay, 2004; Srivastava and Klassen, 2016]. The appeal of this approach is that it allows an estimate for $f(\cdot)$ that can model observed data well without overfitting. Special cases of (2.19) occur when $\lambda = 0$, where $f(\cdot)$ can be any function interpolating the data, and $\lambda \to \infty$, where $f(\cdot)$ must be the Ordinary Least Squares line of best fit.

A natural approach to this problem is to assume that $f(\cdot)$ is a spline function, the nomenclature for which is derived from the devices used by draughtsmen to draw smooth shapes. Introduced by Schoenberg [1946a,b], the spline function of order p

$$f(t) = \sum_{m=1}^{M} \alpha_m B_{m,p}(t), \tag{2.20}$$

is a piecewise-polynomial curve of degree $p-1$ with $p-2$ continuous derivatives, defined with respect to knots τ_m, for $\tau_m \in [0,1]$ and $\tau_m \leq \tau_{m+1}$, basis functions $B_{m,p}(\cdot)$ spanning $[0,1]$, and coefficients α_m, for $m = 1, \ldots, M$. Setting $p = 4$ ensures that $f(\cdot)$ is twice differentiable, yielding a *cubic spline* [Friedman et al., 2001]. The basis function $B_{m,4}(\cdot)$ is defined recursively by De Boor's algorithm [1972] and so, given $\mathbf{y} = (y_1, \ldots, y_D)^\top$, the $D \times M$ basis matrix $\mathbf{B}$ where $\mathbf{B}_{dm} = B_{m,4}(t_d)$, coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^\top$, and $M \times M$ penalty matrix $\mathbf{D}$ where $\mathbf{D}_{mm'} = \int_0^1 B_{m,4}(s)\, B_{m',4}(s)\, ds$, (2.20) can be rewritten as

$$\mathcal{L}(\boldsymbol{\alpha}, \lambda) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top \mathbf{D} \boldsymbol{\alpha},$$

which is minimised with respect to $\boldsymbol{\alpha}$ at

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{D}\right)^{-1} \mathbf{B}^\top \mathbf{y}.$$

Thus, the fitted spline is given by

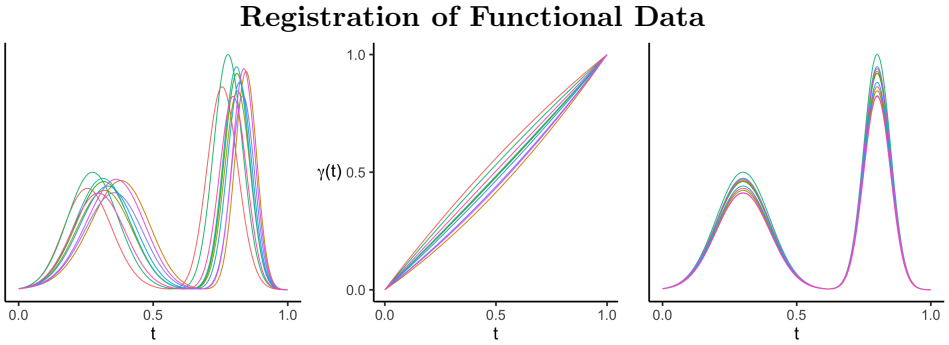
$$\hat{f}(t) = \sum_{m=1}^{M} B_{m,4}(t)\, \hat{\alpha}_m,$$

allowing $\hat{f}(t)$ approximate $f(t)$ for all $t \in [0,1]$, where λ can either be set a priori or inferred by cross-validation [Friedman et al., 2001].

This provides one approach to the smoothing of functional data, though many more have been developed. Kernel methods [Friedman et al., 2001], including GPs [Rasmussen and Williams, 2006], offer a non-parametric approach to smoothing, while wavelets offer an alternative basis to the splines outlined above [Percival and Walden, 2006]. Each of these methods defines some $\hat{f}(t)$ for $t \in [0,1]$, allowing consideration of the second analysis technique most associated with FDA, that is, function registration.

Function registration, also referred to a curve registration [Ramsay and Li, 1998], curve synchronisation [Tang and Müller, 2008], or dynamic time warping [Myers and Rabiner, 1981; Berndt and Clifford, 1994], is required when important features of some set of functions are not aligned along their time axis. This occurs when the chronological time for a particular function does not map directly to the real time scale on which it was recorded, a phenomenon which obfuscates statistical inference on a sample of functions and is referred to as phase variation [Srivastava and Klassen, 2016]. The problem can be understood by considering the height of an individual and how it changes from birth to adulthood, a seminal example in FDA [Ramsay et al., 1995]. Individuals tend to go through two separate growth spurts, one early in life and another at the onset of puberty; however, these spurts start and end at slightly different ages for each individual. If functional registration is not performed prior to the comparative analysis of growth curves, these spurts may not be reflected in any inferred mean and covariance functions for individual growth curves.

A formal description of this problem requires the definition of a warping function. For the remainder of this sub-section, allow function $f(\cdot)$ be denoted as f, providing a less cluttered notation. As such, following the approach of Srivastava and Klassen [2016], let $\gamma(t) \in [0,1]$ for $t \in [0,1]$ be the monotone increasing warping function such that $\gamma(0) = 0$, $\gamma(1) = 1$, γ is invertible, and both γ and γ^{-1} are



(a) Warped Functions (b) Warping Functions (c) Registered Functions

Figure 2.6: An illustration of the variation observed in functional data. Warped functions in (a) are subject to phase (x-axis) and amplitude (y-axis) variation. Warping functions (b) allow curves to be registered (c) such that a comparative analysis can be performed.

smooth. Then, consider a pair of twice-differentiable functions $f_n \in L^2$ for $n = 1, \ldots, N$ such that

$$f_n(t) = a_n g(\gamma_n(t)) \tag{2.21}$$

where a_n is a random variable driving amplitude variation across functions and γ_n are warping functions causing phase variation of the signal g. Functional registration is the definition of γ_n^* such that functions are being compared on a common time scale. A toy example describing the problem is presented in Figure 2.6.

Several approaches have been developed for functional registration. Early efforts were applied for speech recognition [Sakoe and Chiba, 1978] and facial biometrics [Bookstein et al., 1986]. Landmark registration, also known as Procrustes Analysis, has proven a popular approach [Kendall, 1984; Kneip and Gasser, 1992; Dryden and Mardia, 2016]; however, it is impractical for large datasets as landmarks must be identified manually. An automatic approach, similar in spirit to landmark registration is the self-modelling warping functions [Gervini and Gasser, 2004], where observations are mapped to a mean curve, while the application of pairwise curve synchronisation is relatively straightforward for larger datasets [Tang and Müller, 2008]. More recently, registration techniques based on analysis of the Square Root Velocity Function (SRVF), a transformation of observed functions, have been developed [Srivastava et al., 2011; Cheng et al., 2016; Srivastava and Klassen, 2016; Tucker, 2019]. These methods possess some particularly appealing properties.

In order to motivate registration of the SRVF, consider first the estimation of a warping function γ^* which registers f_2 to f_1 by minimisation of the L^2 norm.

In this case

$$\gamma^* = \arg\min_{\gamma} \left\{ ||f_1 - f_2 \circ \gamma||_2 \right\},$$

where $(f \circ \gamma)(t) \equiv f(\gamma(t))$. This approach is intuitively appealing, it registers functions such that the difference in amplitudes is minimised, however, it does have some undesirable characteristics. As discussed by Srivastava and Klassen [2016], the L^2 norm lacks isometry under warping, which is to say that $||f_1 - f_2||_2 \neq |f_1 \circ \gamma - f_2 \circ \gamma||_2$ for some random warping function γ. Thus, an identical warping of f_1 and f_2 does not necessarily preserve their registration, i.e. the warping function registering f_2 to f_1 may differ from that registering $f_2 \circ \gamma$ to $f_1 \circ \gamma$. Furthermore, registration by the L^2 norm may result in a pinching effect and inverse inconsistency. The pinching effect occurs when the L^2 norm can be minimised by squeezing a large part of f_2 onto a short interval, while inverse inconsistency occurs when registering f_1 to f_2 does not produce the inverse warping function of registering f_2 to f_1.

These problems are addressed by introducing the SRVF, defined as

$$q_f(t) \equiv \text{sign}\left(f'(t)\right) \sqrt{|f'(t)|},$$

for which $||q_{f_1} - q_{f_2}||_2 = |q_{f_1 \circ \gamma} - q_{f_2 \circ \gamma}||_2$. Given the SRVF, registering f_2 to f_1 with the warping function defined by

$$\gamma^*(\cdot) = \arg\min_{\gamma(\cdot)} \left\{ ||q_{f_1} - q_{f_2 \circ \gamma}||_2 \right\}, \tag{2.22}$$

$$= \arg\min_{\gamma(\cdot)} \left\{ ||q_{f_1} - (q_{f_2} \circ \gamma)(\cdot) \sqrt{\gamma'(\cdot)}||_2 \right\}, \tag{2.23}$$

provides a theoretically appealing approach to functional registration in that the loss function being minimised is now isometric under random warping functions [Srivastava et al., 2011; Srivastava and Klassen, 2016]. This approach can be further extended to allow the registration of $f_1, \ldots, f_N$ to a common time scale, for which open-source software is available [Tucker, 2019].

2.3.5 Acoustic Signal Processing

Techniques for modelling acoustic signals, be they biotic sounds [Hopp et al., 2012], human speech [Deller Jr and Hansen, 2004], or musical notes [Davy and Godsill, 2003], are built on a foundation of sine waves. Consider the real valued acoustic signal $z(\cdot)$ with duration $T \in \mathbb{R}^+$, which is to say that $z(t) \in \mathbb{R}$ for $t \in [0, T]$. In its simplest form

$$z(t) = a \cos(2\pi f t + \varphi), \tag{2.24}$$

follows a simple harmonic motion, such that it is characterised by amplitude $a \in \mathbb{R}^+$, frequency $f \in \mathbb{R}^+$, and phase shift $\varphi \in [0, 2\pi]$, where these parameter constraints have been enforced only for ease of interpretation [Radi and Rasmussen, 2012]. This deterministic process can be seen to undergo f oscillations around 0, per unit time, with the extrema of each oscillation being at $\pm a$. Thus, it is a periodic signal with constant amplitude a and period $\frac{1}{f}$.

Taking an intuitive approach to extending (2.24), consider

$$z(t) = A(t) \cos(\phi(t)), \tag{2.25}$$

where $A(t) \in \mathbb{R}^+$ for defines the *amplitude envelope*, and

$$\phi(t) = 2\pi \int_0^t f(\tau)\, d\tau + \varphi, \tag{2.26}$$

describes the *instantaneous phase*, where $f(t) \in \mathbb{R}^+$ is the slowly-varying *instantaneous frequency*. Thus, the instantaneous frequency of a signal can be thought of as the derivative of its instantaneous phase [Boashash, 1992; Cohen, 1995; Hlawatsch and Auger, 2008; Huang et al., 2009].

It is immediately obvious that the model defined by (2.25) and (2.26) is problematic. Although $z(t)$ is completely specified given $\{A(t), \phi(t)\}$, there are infinite $\{A(t), \phi(t)\}$ pairs that will yield $z(t)$. While Gabor's [1946] method does allow the estimation of unique amplitude envelope and instantaneous frequency pairs via the Hilbert transform of a signal, the results are not necessarily subject to physical interpretation [Cohen, 1995; Loughlin and Tacer, 1996], and extending the concept to multi-component signals requires much of the information regarding components of the signal be provided by the user [Olhede and Walden, 2005; DiCecco et al., 2013]. Furthermore, instantaneous frequency is a somewhat paradoxical concept [Cohen, 1995], given that frequency must be defined with respect to some interval of time. Nonetheless, instantaneous frequency is a phenomenon that is experienced on a daily basis in both colour gradients and smooth changes in the pitch of sounds [Huang et al., 2009], and the intuition described by (2.25) and (2.26) underpin practical approaches to acoustic signal processing.

A much less contentious approach to modelling acoustic signals is to assume nothing more than that the signal is locally periodic. This allows the definition of the Short-Time Fourier transform (STFT) [Allen and Rabiner, 1977; Portnoff, 1980], a representation of the signal in time and frequency. Given a square integrable window function $w(t) \in \mathbb{R}^+$ such that $\int w^2(d)\, dt < \infty$ [Kokoszka and Reimherr,

2017] , concentrated at 0, where it is at a maximum, and an acoustic signal $z(t) \in \mathbb{R}$ for $t \in [0, T]$, then

$$\mathrm{STFT}_z^w(t, f) \equiv \int_{-\infty}^{\infty} z(\tau)\, w(\tau - t) \exp(-j2\pi f\tau)\, d\tau, \tag{2.27}$$

where $j \equiv \sqrt{-1}$. This in turn allows the spectrogram, a ubiquitous tool in signal processing [Cohen, 1995; Hopp et al., 2012; Hlawatsch and Auger, 2008; Damoulas et al., 2010; Stathopoulos et al., 2018; Mac Aodha et al., 2018; Pigoli et al., 2018], be defined as

$$S_z^w(t, f) \equiv |\,\mathrm{STFT}_z^w(t, f)\,|^2, \tag{2.28}$$

which can be thought of as an energy density of $z(\cdot)$ at time t and frequency f [Hlawatsch and Auger, 2008]. Thus, the time-frequency distribution of $z(\cdot)$ can be examined by considering $S_z^w(t, f)$ for $(t, f) \in [0, T] \times [0, F]$, where F is the Nyquist frequency [Oppenheim and Schafer, 2014].

The choice of $w(\cdot)$ in the $\mathrm{STFT}_z^w(\cdot, \cdot)$ is of enormous importance, as Heisenberg's uncertainty principle imposes a limit on the time-frequency resolution that can be obtained [Hlawatsch and Auger, 2008]. Thus, the time interval defined by $w(\cdot)$ imposes a compromise between preserving resolution in time and resolution in frequency. This implies that the particular requirements for a given application of the $\mathrm{STFT}_z^w(\cdot, \cdot)$ must be carefully considered when choosing $w(\cdot)$ [Stathopoulos et al., 2018; Pigoli et al., 2018].

The spectrogram, or more typically its logarithm, which is also referred to as the spectrogram, is an essential visual tool for signal processing (see Figure 2.2), and recent advances in echolocation call classification have been driven by its analysis [Stathopoulos et al., 2018; Mac Aodha et al., 2018]. Furthermore, techniques for modelling spectrograms as functional data objects have been developed in the context of computational linguistics [Pigoli et al., 2018].

Despite this progress, there remain difficulties in the comparative analysis of spectrograms. Firstly, note that the spectrogram is defined for a grid of points over $[0, T] \times [0, F]$. For most acoustic signals, there is a limited set of frequencies with a high energy density at any particular point in time. This implies that the spectrogram includes information that could be considered redundant. Of more concern in comparative analysis, however, are the vastly different implications of characterising a signal by its spectrogram as opposed to its instantaneous frequency, in particular when an interpolation between two signals is considered. As illustrated for a toy example in Figure 2.7, a linear interpolation between spectrograms implies a signal that is impossible for the larynx to produce [Deller Jr and Hansen, 2004], given that

it consists of two intersecting frequency components. If applied to bat echolocation calls for ancestral reconstruction, such an interpolation may infer calls that are impossibilities, given the physiology of a bats larynx [Fenton et al., 2016]. Interestingly, this does not appear to be the case for instantaneous frequency, with interpolation between two signals consisting of a single frequency component resulting in a signal which itself has a single frequency component.

2.4 Bayesian Inference

Suppose there exists a collection of probabilistic models $\mathcal{M}_1, \ldots, \mathcal{M}_K$, each providing an explanation for some phenomenon. Initial beliefs on the plausibility of $\mathcal{M}_k$ are encoded in a prior probability distribution, denoted $p(\mathcal{M}_k)$, and each model is parametrised by the D_k dimensional vector θ_k. When initial beliefs on θ_k are given by the prior distribution $p(\theta_k|\mathcal{M}_k)$, Bayes' theorem describes how these beliefs should be updated when the data $\mathbf{y}$ is observed [MacKay, 1992; Gelman et al., 2013]. Typically, this updating of prior beliefs occurs on two levels. The first is model fitting, also referred to as parameter inference, for which Bayes' theorem states that

$$p(\theta_k|\mathbf{y}, \mathcal{M}_k) = \frac{p(\mathbf{y}|\theta_k, \mathcal{M}_k)\, p(\theta_k|\mathcal{M}_k)}{\int p(\mathbf{y}|\theta_k, \mathcal{M}_k)\, p(\theta_k|\mathcal{M}_k)\, d\theta_k},$$

where $p(\theta_k|\mathbf{y}, \mathcal{M}_k)$ is the posterior distribution over the parameters for $\mathcal{M}_k$ given $\mathbf{y}$, $p(\mathbf{y}|\theta_k, \mathcal{M}_k)$ is the likelihood of θ_k, and $p(\mathbf{y}|\mathcal{M}_k) = \int p(\mathbf{y}|\theta_k, \mathcal{M}_k)\, p(\theta_k|\mathcal{M}_k)\, d\theta_k$ is the evidence for $\mathcal{M}_k$. The second level of inference, referred to as model comparison, involves finding the posterior probability of each model where

$$p(\mathcal{M}_k|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_k)\, p(\mathcal{M}_k).$$

This general approach to inference for probabilistic models was described by MacKay [1992] as the evidence framework.

In general, closed-form solutions for the posterior distribution over parameters and the model evidence do not exist. This has motivated the development of simulation-based approaches to inference. Markov Chain Monte Carlo (MCMC) methods implement this strategy by sampling a Markov chain, for which the stationary distribution is equivalent to the desired target distribution [Robert and Casella, 2013]. While algorithms such as Reversible Jump MCMC do allow joint parameter inference and model comparison to be performed [Green, 1995], the standard approach to inference within the evidence framework samples from the posterior, that is $p(\theta_k|\mathbf{y}, \mathcal{M}_k)$, before estimating the evidence $p(\mathbf{y}|\mathcal{M}_k)$ from this sample. A

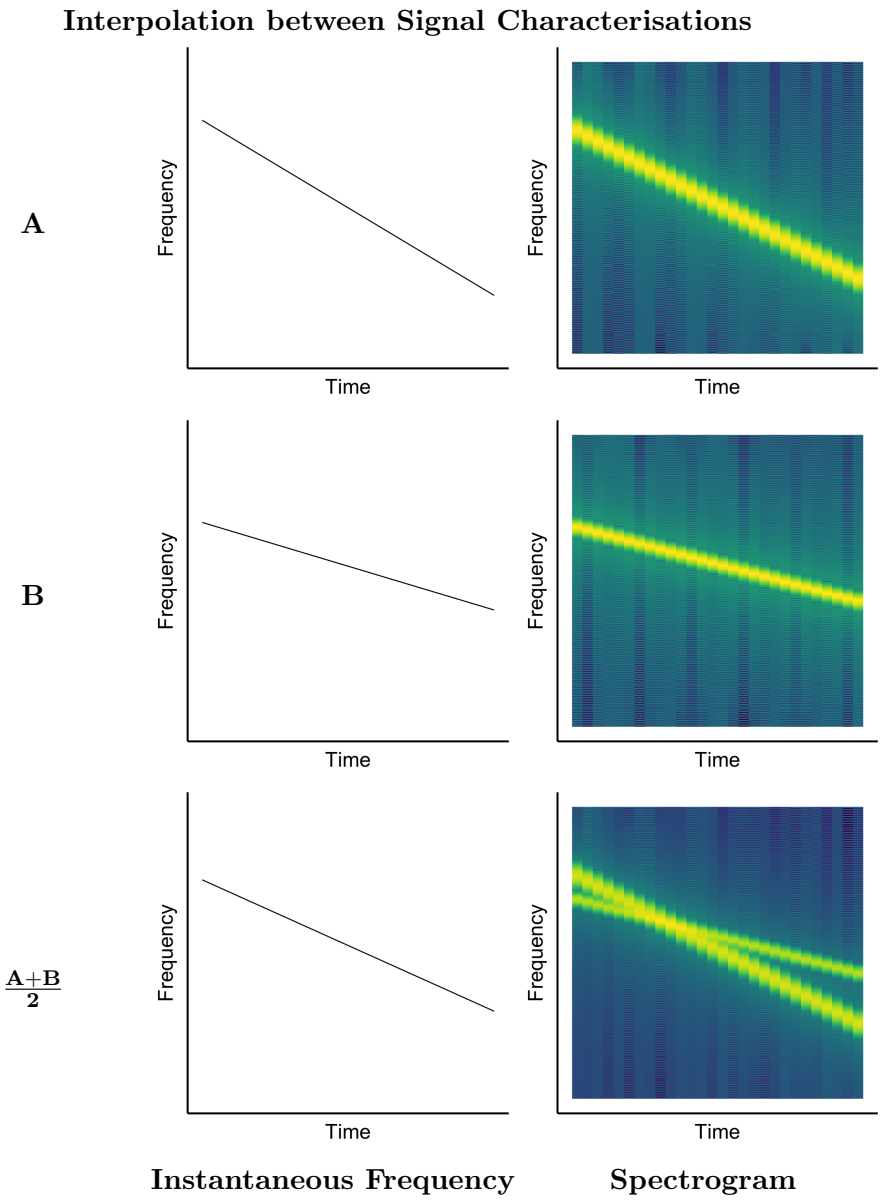


Figure 2.7: Characterising signals with the instantaneous frequency result in a signal with a single frequency component, while the spectrogram implies two. If the interpolation is meant to imply an intermediate state between existing signals, then the spectrogram interpolant would appear unlikely, if not physically impossible, in the context of bat echolocation calls.

plethora of MCMC methods have been developed for these tasks, with algorithms such as Metropolis-Hastings [Metropolis et al., 1953; Hastings, 1970], Slice Sampling [Neal, 2003], and Hamiltonian Monte Carlo [Betancourt and Girolami, 2015; Betancourt, 2017; Carpenter et al., 2017] providing a flexible set of techniques for parameter inference, while the Harmonic Mean estimator [Newton and Raftery, 1994], Candidate's estimator [Chib, 1995], Bridge Sampling [Meng and Wong, 1996; Gronau et al., 2017b], path sampling [Gelman and Meng, 1998], and the method of power posteriors [Friel and Pettitt, 2008] all allow estimation of the model evidence.

The following presents a brief review of those MCMC schemes for Bayesian inference implemented in this thesis.

2.4.1 MCMC Methods for Parameter Inference

The Metropolis-Hastings algorithm (Algorithm 1), introduced by Metropolis et al. [1953] and subsequently generalised by Hastings [1970], provides a general MCMC method for parameter inference [Chib and Greenberg, 1995; Robert and Casella, 2013]. Let $\pi(\theta)$ be the *target distribution*, which is known up to a normalising constant, and define a *proposal distribution* $q(\theta_a|\theta_b)$ which is also known up to a normalising constant. Then, given an initial value $\theta^{(0)}$, the Metropolis-Hastings algorithm samples the Markov chain $\theta^{(1)}, \ldots, \theta^{(T)}$, for which the stationary distribution is equivalent to $\pi(\theta)$, by accepting the proposal $\theta^* \sim q\left(\theta^*|\theta^{(t)}\right)$ for $\theta^{(t+1)}$ with probability

$$\alpha\left(\theta^*|\theta^{(t)}\right) = \frac{\pi(\theta^*)/q(\theta^*|\theta^{(t)})}{\pi(\theta^{(t)})/q(\theta^{(t)}|\theta^*)},$$

and setting $\theta^{(t+1)}$ to $\theta^{(t)}$ otherwise, as described by Algorithm 1.

The Metropolis-Hastings algorithm provides a conceptually straightforward and flexible approach to inferring intractable posterior distributions, however, its implementation is not without difficulties. In particular, the choice of proposal is crucial to ensuring that the target distribution is sampled efficiently. In order to address this problem, Adaptive MCMC methods such as Adaptive Metropolis (AM) algorithms have been developed [Haario et al., 2001; Roberts and Rosenthal, 2009]. Such methods are based on the Metropolis algorithm [Metropolis et al., 1953], a special case of the Metropolis-Hastings algorithm which occurs when the proposal distribution is symmetric. This results in a simplified expression for the acceptance probability, where

$$\alpha\left(\theta^*|\theta^{(t)}\right) = \frac{\pi(\theta^*)}{\pi(\theta^{(t)})}.$$

AM then relies on a relaxation of the detailed balance condition necessary for sam-

Algorithm 1: The Metropolis-Hastings Algorithm

Data: $\mathbf{y}, \theta^{(0)}, T$

Result: $\theta^{(1)}, \ldots, \theta^{(T)}$

for $t = 0, \ldots, T$ **do**

Sample θ^* from $q(\theta^*|\theta^{(t)})$;

Set

$$\theta^{(t+1)} \leftarrow \begin{cases} \theta^*, & \text{with probability } \min\left\{\alpha\left(\theta^*|\theta^{(t)}\right), 1\right\} \\ \theta^{(t)}, & \text{otherwise} \end{cases}$$

where

$$\alpha\left(\theta^*|\theta^{(t)}\right) \leftarrow \frac{\pi(\theta^*)/q(\theta^*|\theta^{(t)})}{\pi(\theta^{(t)})/q(\theta^{(t)}|\theta^*)}.$$

end

pling ergodic Markov chains with the correct stationary distribution. Under detailed balance

$$\pi\left(\theta_a\right) K\left(\theta_a, \theta_b\right) = \pi\left(\theta_b\right) K\left(\theta_b, \theta_a\right),$$

where $K\left(\theta_a, \theta_b\right) = p(\theta^{(t+1)} = \theta_b | \theta^{(t)} = \theta_a)$ is the *transition kernel* for the Markov chain. Relaxing this condition such that it holds only in the limit as $t \to \infty$ provides the flexibility which underpins AM sampling schemes. The approach taken here, adapted from that proposed by Haario et al. [2001] and suitable for any $\theta \in \mathbb{R}^D$, assumes a proposal distribution of the form

$$q\left(\theta^*|\theta^{(t)}\right) = \begin{cases} \mathcal{N}\left(\theta^*|\theta^{(t)}, \left(\frac{0.1^2}{D}\right)\mathbf{I}_D\right), & \text{for } s \leq 2D, \\ \mathcal{N}\left(\theta^*|\theta^{(t)}, \beta\left(\frac{0.1^2}{D}\right)\mathbf{I}_D + (1-\beta)\left(\frac{2.38^2}{D}\right)\hat{\Sigma}_\theta^{(s)}\right) & \text{otherwise,} \end{cases}$$

for some small $\beta \in (0, 1)$ ($\beta = 0.05$ for all implementations in this thesis). In this case $\hat{\Sigma}_\theta^{(t)}$ is the sample variance of the Markov chain, computed after discarding initial warm up samples $\theta^{(0)}, \ldots, \theta^{(\lfloor S*t \rfloor)}$ for some $S \in (0, 1)$. This is motivated by the fact that $\mathcal{N}\left(\theta^*|\theta^{(t)}, \left(\frac{2.38^2}{D}\right)\Sigma_\theta\right)$ has been shown to be the optimal proposal distribution for the Metropolis-Hastings algorithm in some settings, where $\text{Var}\left(\theta\right) = \Sigma_\theta$ [Roberts et al., 2001]. A description of this sampling scheme, which provides a flexible approach to parameter inference by automatically choosing a proposal distribution, is presented in Algorithm 2.

Algorithm 2: Adaptive Metropolis

Data: $\mathbf{y}, \theta^{(0)}, T$
Result: $\theta^{(1)}, \ldots, \theta^{(T)}$

for $t = 0, \ldots, T$ **do**
  if $t \leq 2D$ **then**
    Sample $\theta^* \sim \mathcal{N}\left(\theta^{(t)}, \left(\frac{0.1^2}{D}\right)\mathbf{I}_D\right)$;
  else
    Compute $\bar{\theta}^{(t)} \leftarrow \frac{1}{t-\lfloor S*t \rfloor}\sum_{i=\lfloor S*t \rfloor+1}^{t}\theta^{(i)}$;
    Compute $\hat{\Sigma}_\theta^{(t)} \leftarrow \frac{1}{t-\lfloor S*t \rfloor-1}\sum_{i=\lfloor S*t \rfloor+1}^{t}\left(\theta^{(i)}-\bar{\theta}^{(t)}\right)^2$;
    Sample $\theta^* \sim \mathcal{N}\left(\theta^{(t)}, \beta\left(\frac{0.1^2}{D}\right)\mathbf{I}_D + (1-\beta)\left(\frac{2.38^2}{D}\right)\hat{\Sigma}_\theta^{(s)}\right)$;
  end
  Set

$$\theta^{(t+1)} \leftarrow \begin{cases} \theta^*, & \text{with probability } \min\left\{\alpha\left(\theta^*|\theta^{(t)}\right), 1\right\} \\ \theta^{(t)}, & \text{otherwise} \end{cases}$$

  where

$$\alpha\left(\theta^*|\theta^{(t)}\right) \leftarrow \frac{\pi(\theta^*)}{\pi(\theta^{(t)})}.$$

end

2.4.2 MCMC Methods for Gaussian Processes

While the Metropolis-Hastings algorithm and AM provide general methods for parameter inference, domain specific MCMC algorithms have also been developed. Of particular interest is the Elliptical Slice Sampler (ESS) for GPs [Murray et al., 2010]. Modelling the observations $\mathbf{y} = (y_1, \ldots, y_N)^\top$ associated with index variables $\mathbf{x} = (x_1, \ldots, x_N)^\top$ as a GP typically implies that

$$y_n = f(x_n) + \epsilon_n,$$

where

$$f(x) \sim \mathcal{GP}\left(0, k\left(x, x'|\theta_f\right)\right),$$

for covariance function $k(x, x'|\theta_f)$ governed by hyper-parameters θ_f, and $\epsilon_n \sim \mathcal{N}\left(0, \sigma^2\right)$.

This model implies that $p\left(\mathbf{y}|\mathbf{f}, \sigma^2\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}_N\right)$ and $p\left(\mathbf{f}|\theta_f\right) = \mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}\right)$ for $\mathbf{f} = \left(f(x_1), \ldots, f(x_N)\right)$, where $\mathbf{K}$ is the Gram matrix of $k(\cdot, \cdot|\theta_f)$ such that $(\mathbf{K})_{ij} = k(x_i, x_j|\theta_f)$ [Rasmussen and Williams, 2006]. The ESS allows $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(T)}$,

Algorithm 3: The Elliptical Slice Sampler

Data: $\mathbf{y}, \mathbf{f}^{(0)}, T$
Result: $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(T)}$
for $t = 0, \ldots, T$ **do**
  Sample ellipse $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$;
  Sample $u \sim \mathcal{U}[0, 1]$;
  Set slice threshold $\gamma \leftarrow u\, \pi\left(\mathbf{f}^{(t)}\right)$;
  Sample ellipse angle $\phi \sim \mathcal{U}[0, 2\pi]$;
  Define sampling bracket $[\phi_{min}, \phi_{max}] \leftarrow [\phi - 2\pi, \phi]$;
  Compute proposal $\mathbf{f}^{*} \leftarrow \mathbf{f}^{(t)} \cos\phi + \boldsymbol{\nu} \sin\phi$;
  if $\pi(\mathbf{f}^{*}) > \gamma$ **then**
    $\mathbf{f}^{(t+1)} \leftarrow \mathbf{f}^{*}$
  else
    if $\phi < 0$ **then**
      $\phi_{min} \leftarrow \phi$;
    else
      $\phi_{max} \leftarrow \phi$;
    end
    Sample ellipse angle $\phi \sim \mathcal{U}[\phi_{min}, \phi_{max}]$;
    Return to 7;
  end
end

a Markov chain for which $p\left(\mathbf{f}|\mathbf{y}, \theta_f, \sigma^2\right)$ is the stationary distribution, to be drawn by setting the target distribution to $\pi(\mathbf{f}) = p\left(\mathbf{y}|\mathbf{f}, \sigma^2\right) p(\mathbf{f}|\theta_f)$ and implementing a slice sampling algorithm over an ellipse defined by $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ [Neal, 2003]. This allows for "*rejection-free*" sampling, in that $\mathbf{f}^{(t)}$ is updated with a new value at each iteration. The ESS proposed by Murray et al. [2010] is described fully in Algorithm 3.

Typically, θ_f in this model known only up to a prior distribution $p(\theta_f)$ and as such must be inferred given $\mathbf{y}$. A strongly recommended approach to this problem is to implement the Ancilliarity-Sufficiency Interweaving Strategy (ASIS) developed by Yu and Meng [2011] and referred to as "whitening" by Murray and Adams [2010] [Filippone et al., 2013; Monterrubio-Gómez et al., 2018]. This approach is based on the insight that, for $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ where $\mathbf{L}\mathbf{L}^{\top} = \mathbf{K}$, there exists $\boldsymbol{\zeta}$ for which $p(\boldsymbol{\zeta}) = \mathcal{N}(\boldsymbol{\zeta}|\mathbf{0}, \mathbf{I}_N)$, such that $\mathbf{f} = \mathbf{L}\boldsymbol{\zeta}$ [Petersen and Pedersen, 2012]. This allows the Sufficient Augmentation of the model [Yu and Meng, 2011] (referred to as "unwhitened" by Murray and Adams [2010]), such that the joint distribution can be expressed as

$$p(\mathbf{y}, \mathbf{f}, \theta_f) = p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|\theta_f)\, p(\theta_f),$$

and an Ancilliary Augmentation [Yu and Meng, 2011] ("whitened" [Murray and Adams, 2010]), such that

$$p\left(\mathbf{y}, \boldsymbol{\zeta}, \theta_f\right) = p\left(\mathbf{y}|\boldsymbol{\zeta}, \theta_f\right) p\left(\boldsymbol{\zeta}\right) p\left(\theta_f\right).$$

This allows an ASIS sampling scheme where at each iteration $p\left(\mathbf{f}|\mathbf{y}, \theta_f, \sigma^2\right)$ is first sampled by a ESS step, followed by an AM step for $p\left(\theta_f|\mathbf{f}\right)$. This in turn allows the definition of $\boldsymbol{\zeta} \equiv \mathbf{L}^{-1}\mathbf{f}$ and so an AM step can be implemented for $p\left(\theta_f|\mathbf{y}, \boldsymbol{\zeta}, \sigma^2\right)$ [Filippone et al., 2013].

The benefit of this approach is that the convergence rate for Ancilliary Augmented samples differ from those of Sufficiently Augmented samples. Moreover, it has been observed that, if one leads to the fast convergence of sampled chains, then the other is usually slow, depending on the observed data [Yu and Meng, 2011]. Thus, implementing an ASIS takes advantage of these differing convergence rates and can result in dramatic improvements to sampling efficiency in the context of GPs [Filippone et al., 2013].

2.4.3 Estimating Model Evidence

Consider once more the probabilistic models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ describing the observed data $\mathbf{y}$. Given those methods for approximating $p\left(\theta_k|\mathbf{y}, \mathcal{M}_k\right)$ described above, the second level of the evidence framework involves inference of the model evidence, that is $p\left(\mathbf{y}|\mathcal{M}_k\right)$. While a number of approaches to this problem have been developed, it is Bridge Sampling that is considered here [Meng and Wong, 1996; Gronau et al., 2017a], for which Gronau et al. [2017b] have developed the `bridgesampling` package in R.

In order to motivate Bridge Sampling, consider the identity

$$1 = \frac{\int p\left(\mathbf{y}|\theta_k, \mathcal{M}_k\right) p\left(\theta_k|\mathcal{M}_k\right) h\left(\theta_k\right) g\left(\theta_k\right) d\theta_k}{\int p\left(\mathbf{y}|\theta_k, \mathcal{M}_k\right) p\left(\theta_k|\mathcal{M}_k\right) h\left(\theta_k\right) g\left(\theta_k\right) d\theta_k},$$

where $h\left(\cdot\right)$ is the *bridge function* and $g\left(\cdot\right)$ is referred to as the *proposal distribution.* Multiplying both sides of this identity by the model evidence yields

$$p(\mathbf{y}|\mathcal{M}_k) = \frac{\mathbb{E}_{g(\theta_k)}\left[p\left(\mathbf{y}|\theta_k, \mathcal{M}_k\right) p\left(\theta_k|\mathcal{M}_k\right) h\left(\theta_k\right)\right]}{\mathbb{E}_{p(\theta_k|\mathbf{y}, \mathcal{M}_k)}\left[h\left(\theta_k\right) g\left(\theta_k\right)\right]}.$$

Thus, given samples from $p\left(\theta_k|\mathbf{y}, \mathcal{M}_k\right)$, denoted $\theta^*_{k,i}$ for $i = 1, \ldots, T_1$, and samples from $g\left(\theta_k\right)$, $\tilde{\theta}_{k,j}$ for $j = 1, \ldots, T_2$, the model evidence for $\mathcal{M}_k$ for $k = 1, \ldots, K$ can be estimated.

The optimal bridge function [Meng and Wong, 1996], with respect to the relative mean square error of the estimator, is

$$h\left(\theta_k\right) = C_k \cdot \frac{1}{s_1 p\left(\mathbf{y}|\theta_k, \mathcal{M}_k\right) p\left(\theta|\mathcal{M}_k\right) + s_2 p\left(\mathbf{y}|\mathcal{M}_k\right) g\left(\theta_k\right)}$$

where $s_1 = \frac{T_1}{T_1+T_2}$, $s_2 = \frac{T_2}{T_1+T_2}$, and C_k is a normalising constant. Given that this bridge function depends on $p\left(\mathbf{y}|\mathcal{M}_k\right)$, an iterative approach can be taken to estimating the marginal likelihood, where the estimator

$$\widehat{p(\mathbf{y}|\mathcal{M}_k)}^{(t+1)} = \frac{\frac{1}{T_2}\sum_{j=1}^{T_2} \frac{p\left(\mathbf{y}|\tilde{\theta}_j,\mathcal{M}_k\right)p\left(\tilde{\theta}_j|\mathcal{M}_k\right)}{s_1 p\left(\mathbf{y}|\tilde{\theta}_j,\mathcal{M}_k\right)p\left(\tilde{\theta}_j|\mathcal{M}_k\right)+s_2 \widehat{p(\mathbf{y}|\mathcal{M}_k)}^{(t)} g\left(\tilde{\theta}_j\right)}}{\frac{1}{T_1}\sum_{i=1}^{T_1} \frac{g\left(\theta_i^*\right)}{s_1 p\left(\mathbf{y}|\theta_i^*,\mathcal{M}_k\right)p\left(\theta_i^*|\mathcal{M}_k\right)+s_2 \widehat{p(\mathbf{y}|\mathcal{M}_k)}^{(t)} g\left(\theta_i^*\right)}} \tag{2.29}$$

is robust with respect to the tail behaviour of the proposal distribution. When tails of the proposal distribution are heavier than those of the posterior, samples from the proposal tail contribute 0 to the numerator sum in (2.29). Given that this ratio is bounded, and such samples occur only occasionally, their occurrence will not dominate the estimated evidence. Similarly, when tails of the proposal are lighter than those of the posterior distribution, samples from the posterior tail contribute 0 to the denominator sum. Again, this bounded ratio will not dominate the estimator. Thus, provided the proposal and posterior distribution share some region of overlap, (2.29) provides a robust estimate of the evidence for each model considered.

Chapter 3

A Phylogenetic Latent Variable Model for Function-valued Traits

3.1 Introduction

A bats echolocation call is a process that is continuous in time, and as such, it can be thought of as a Function-Valued Trait (FVT). Thus, in order to develop a method for the ancestral reconstruction of bat echolocation calls, the problem of evolutionary inference for FVTs must first be considered in general terms. The Phylogenetic Gaussian Process Regression (PGPR) framework proposed by Jones and Moriarty [2013] allows the definition of a prior distribution for a FVT over a phylogeny, such that a probabilistic model for trait evolution is defined by the phylogeny-trait covariance function of a phylogenetic Gaussian process. Although they have not necessarily been referred to as such, special cases of the PGPR framework have been employed for both phylogenetics and phylogenetic comparative analysis for decades [Felsenstein, 1973; Lande, 1976; Felsenstein, 1985; Grafen, 1989; Hansen, 1997; Hadjipantelis et al., 2013]. Despite this being the case, a Bayesian approach to inference of the full phylogeny-trait covariance function and its hyper-parameters had yet to be developed. Such a method allows ancestral trait reconstruction, conditional on a phylogeny describing the evolutionary relationships between taxa, while also accommodating uncertainty in the model for trait evolution. This is the contribution made in this chapter.

Consider first the phylogeny-trait covariance function within the PGPR framework, discussed in sub-section 2.3.2, and its role in ancestral trait reconstruction. In

general, the form of this object is unknown. Thus, it must either be assumed a priori or inferred from observed data. Fixing the phylogeny-trait covariance function such that each trait is modelled as a independent Brownian Motion (BM) over the phylogeny [Felsenstein, 1973, 1985] is unlikely to yield accurate trait reconstruction and uncertainty quantification for ancestral taxa, particularly when there exists a rich covariance structure within multivariate traits. This is especially true for FVTs. In an attempt to model this structure, Hadjipantelis et al. [2013] proposed a method for the ancestral reconstruction of FVTs within the PGPR framework. An Independent Principal Components Analysis (IPCA) [Yao et al., 2012] of FVT observations provide a set of basis functions which define the trait covariance structure. Given these independent components, modelling latent variables as independent phylogenetic Gaussian processes allows ancestral trait reconstruction. Despite the appeal of this method, which describes the evolution of FVTs as a linear combination of basis functions and latent variables, it is not without drawbacks. The IPCA implies that observed FVT are independent, violating the assumption of dependence between taxa due to the phylogeny that is central to phylogenetic comparative analysis [Felsenstein, 1985; Revell, 2009]. Furthermore, there is no quantification of uncertainty for the basis functions defining the independent components, nor does the model include observation noise on trait measurements. Finally, selecting the number of basis functions to include is guided by heuristics rather than any principled method for model selection. Thus, a more sophisticated approach to inference is required.

The PGPR framework offers a probabilistic model for trait evolution that is closely related those underpinning Phylogenetic Comparative Methods (PCMs) proposed by Cybis et al. [2015] and Tolkoff et al. [2017], which model collections of discrete and continuous traits over a phylogeny. These methods take the opposite perspective on inference to Hadjipantelis et al. [2013], effectively assuming the phylogenetic covariance function to be a known Brownian Motion (BM) kernel and then inferring the trait covariance structure. Markov Chain Monte Carlo (MCMC) sampling schemes implement Bayesian inference, which accommodate uncertainty on the evolutionary history between taxa by sampling from a posterior distribution of phylogenies inferred from molecular sequences. Although this Bayesian approach to inference is appealing, such methods are unsuitable for ancestral reconstruction. Assuming a BM model for trait evolution fixes the covariance structure between taxa and precludes joint inference of the phylogeny-trait covariance function. For ancestral reconstruction, the Ornstein-Uhlenbeck (OU) process offers a more appealing alternative, preserving the Markov property over the phylogeny while allow-

ing the observed data inform the hyper-parameters governing its behaviour [Jones and Moriarty, 2013]. Furthermore, the methods of Cybis et al. [2015] and Tolkoff et al. [2017] do not accommodate intra-taxon variation, preventing the inclusion of repeated measurements for any taxon. Extending key aspects of their MCMC inference schemes to the PGPR framework will address each of these shortcomings.

The first difficulty encountered when developing a Bayesian inference scheme for PGPR is the computational cost of evaluating the model's likelihood. This likelihood is a Gaussian pdf, and in general, its computation scales cubically with the number of observations [Rasmussen and Williams, 2006; Jones and Moriarty, 2013]. Implementing an MCMC inference scheme relying on such a computation is impractical for all but the smallest of datasets. This computation of a Gaussian likelihood for traits over a phylogeny constitutes a long-standing problem in phylogenetics [Felsenstein, 1973]. The BM model for trait evolution is especially well studied and many algorithms scaling linearly with the number of observed taxa have been proposed for the computation of this likelihood [Felsenstein, 1973; Pybus et al., 2012; Freckleton, 2012; Mitov and Stadler, 2017]. The insight underpinning these algorithms is that, when taxa are conditionally independent given their Most Recent Common Ancestor (MRCA), the marginal process over each branch of the phylogeny is a first-order Gauss-Markov process [Jones and Moriarty, 2013]. Thus, a post-order tree traversal, that is, a traversal of the phylogeny from tips to root, computes the model likelihood efficiently. Furthermore, based on the likelihood computation of Pybus et al. [2012], Cybis et al. [2015] employs a further post-order tree traversal to compute the conditional distribution of a trait for any extant taxon given all other extant taxa, under the BM model for trait evolution. The post-order tree traversal is a traversal from the root of the phylogeny to its tips, and as such, this computation scales quadratically with the number of observed taxa. These methods allowed Cybis et al. [2015] and Tolkoff et al. [2017] to implement efficient MCMC inference schemes for their PCMs, however, the algorithms are not without limitations. Except for Mitov and Stadler [2017], who designed an algorithm that computes the likelihood of a univariate OU Phylogenetic Mixed Model (PMM) [Housworth et al., 2004], the algorithms identified above only consider the BM model of trait evolution. Furthermore, none of the algorithms developed to date allow the inclusion of intra-taxon variation within the model. Such limitations are problematic in the application of PGPR for ancestral reconstruction. Not only is an OU process the preferred model for trait evolution, but repeated measurements for each taxon are typical. A Bayesian inference scheme should include this information explicitly. Thus, an algorithm for the efficient computation of the model likelihood

for a general Gauss-Markov model of trait evolution is required.

The objective of this chapter is to develop a Bayesian approach to the ancestral reconstruction of a FVT within the PGPR framework, given the evolutionary history linking taxa. To this end, the Phylogenetic Latent Variable Model (PLVM) is introduced, for which an MCMC sampling scheme allows inference on a phylogeny-trait separable phylogenetic Gaussian process which is subject to independent observation noise. Not only does this represent the first fully Bayesian approach to inference within the PGPR framework, but also extends PGPR beyond separable phylogeny-trait covariance functions with the inclusion of observation noise. This inference scheme also includes intra-taxon variation within the phylogenetic comparative analysis. In order to achieve this, an efficient algorithm computing the likelihood for extant taxa of a general Gauss-Markov processes over a phylogeny is developed, an important contribution in its own right. In addition, a novel algorithm computing the distribution of a general Gauss-Markov processes at each position on a phylogeny, conditional on extant taxa, allows computationally efficient ancestral trait reconstruction.

This chapter is structured as follows. After describing the phylogeny in terms of a graphical model and illustrating the inclusion of repeated measurements for extant taxa, the PLVM is defined. An outline of the efficient computation of the model likelihood follows this, although a detailed derivation of the algorithm is presented only in Appendix A.1. Specifying prior distributions for the parameters and hyperparameters in the PLVM allows the derivation of a Bayesian inference scheme. This posterior inference is based on state-of-the-art MCMC methods for Gaussian process (GP) models presented in section 2.4 [Murray et al., 2010; Murray and Adams, 2010; Yu and Meng, 2011; Filippone et al., 2013], while Bridge Sampling is proposed for model comparison [Meng and Wong, 1996; Gronau et al., 2017a]. The final method presented in this chapter outlines the efficient computation of the conditional distributions for traits over the phylogeny, allowing ancestral reconstruction of the FVT. A detailed derivation of this algorithm is relegated to Appendix A.2. A synthetic dataset drawn from a PLVM allows the assessment of this approach to ancestral reconstruction, with experiments demonstrating the methods efficacy. Discussion of the method's strengths and weaknesses, along with its implications for further research, concludes the chapter.

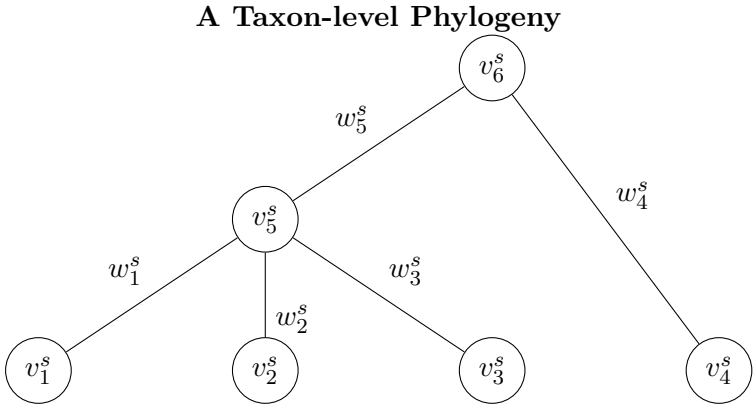


Figure 3.1: A taxon-level phylogeny with $S = 4$ and $M = 2$. This is an example of a phylogenetic tree with a polytomy at v_5^s, that is $\{v_1^s, v_2^s, v_3^s\}$ are modelled as having a single common ancestor.

3.2 Methods

3.2.1 The Phylogeny: A Graphical Model for Shared Ancestry

Consider a set of S related extant taxa, for which a rooted phylogenetic tree, denoted $\mathcal{T}_S$, represents their evolutionary history. Formally, the taxon-level phylogeny $\mathcal{T}_S = \{\mathcal{V}_S, \mathcal{W}_S\}$ is a graph with vertices $\mathcal{V}_S$ and edge weights $\mathcal{W}_S$ [Højsgaard et al., 2012], referred to as nodes and branches respectively. Assuming that the S taxa are extant and represented by terminal nodes of $\mathcal{T}_S$, there exist M internal nodes, representing ancestral taxa, such that $\mathcal{V}_S = \left\{v_1^s, \ldots, v_{S+M}^s\right\}$ and $\mathcal{W}_S = \left\{w_1^s, \ldots, w_{S+M-1}^s\right\}$, with $M = S-1$ when $\mathcal{T}_S$ is a bifurcating tree. Letting v_{N+M}^s be the root node of $\mathcal{T}_S$, that is the Most Recent Common Ancestor (MRCA) of the extant taxa in question, if nodes v_i^s and v_j^s share an edge and v_j^s is on the path from v_i^s to v_{S+M}^s then $j = \mathrm{pa}\,(i)$ and $i \in \mathrm{ch}\,(j)$, which is to say that v_j^s is the parent of v_i^s and v_i^s a child of v_j^s. The branch connecting v_i^s to v_j^s is of length $w_i^s \in \mathbb{R}^+$, for $i = 1, \ldots, S+M-1$, where w_i^s is proportional to the evolutionary time between v_i^s and v_j^s. Each terminal node $v_i^s \in \mathcal{V}_S$ for $i = 1, \ldots, S$ is of degree 1, with one internal parent node $v_{\mathrm{pa}(i)}^s$. Internal nodes $v_i^s \in \mathcal{V}_S$ for $i = S+1, \ldots, M-1$ are of degree $d_i \geq 3$, with $d_i - 1$ child nodes, and the root node v_{S+M}^s is of degree $d_{S+M} \geq 2$ with d_{S+M} children. A toy example of such a phylogeny is presented in Figure 3.1.

Suppose now that there are N_i individuals associated with each extant taxon, such that $N = \sum_{i=1}^{S} N_i$. Appending the root of a star phylogeny (a multifurcating tree with all branches connected at a single internal node) with N_i branches of length 0 to the terminal node of $\mathcal{T}_S$ corresponding to the i^{th} extant taxon, yields

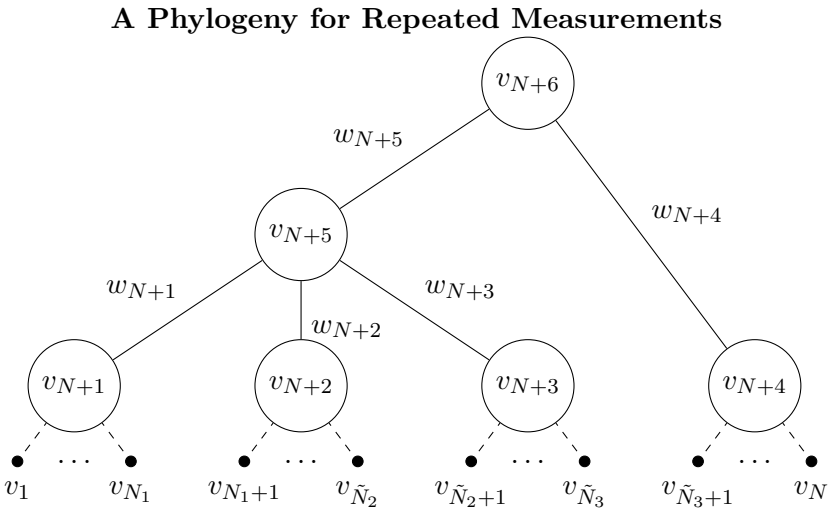


Figure 3.2: A phylogeny accommodating multiple observations in each extant taxon, based on the phylogeny in Figure 3.1, where $\tilde{N}_j = \sum_{i=1}^{j} N_i$. Solid edges are proportional to evolutionary time, dashed lines are edges of length 0, large circles represent unobserved nodes on the phylogeny, and small filled circles correspond to observations.

an individual-level phylogeny $\mathcal{T} = \{\mathcal{V}, \mathcal{W}\}$ where $\mathcal{V} = \{v_1, \ldots, v_{N+S+M}\}$ and $\mathcal{W} = \{w_1, \ldots, w_{N+S+M-1}\}$. This assumes that an evolutionary time of zero separates individuals within taxa. In this case $\mathcal{T}$ has terminal nodes $v_i \in \mathcal{V}$ for $i = 1, \ldots, N$, one for each individual, taxon-level internal nodes $v_i \in \mathcal{V}$ for $i = N+1, \ldots, N+S$, and ancestral nodes $v_i \in \mathcal{V}$ for $i = N+S+1, \ldots, N+S+M$, such that the MRCA $v_{N+S+M} \in \mathcal{V}$ is at the root of $\mathcal{T}$. An example of one such phylogeny is presented in Figure 3.2.

As a final remark, it is useful to consider $\mathbf{t}_i \in \mathcal{T}$, where $\mathbf{t}_i$ denotes the position of v_i on $\mathcal{T}$ with respect to $\mathcal{V}$ and $\mathcal{W}$. This allows the *patristic distance* operator be defined such that $\mathrm{d}_{\mathcal{T}}(\mathbf{t}_i, \mathbf{t}_j)$ is the sum of branch lengths along the shortest path from v_i to v_j, as discussed in sub-section 2.3.2 [Rédei, 2008; Jones and Moriarty, 2013]. A detailed discussion of the patristic distance is also presented by Mariñas-Collado et al. [2019]. Note that the notion of position on a phylogeny can be extended to include any point along a branch of $\mathcal{T}$, meaning that it can take a continuum of values, however, only those positions corresponding with $\mathcal{V}$ are considered within this thesis.

3.2.2 A Phylogenetic Latent Variable Model for Function-valued Traits

Consider the matrix $\mathbf{Y} \in \mathbb{R}^{N \times D}$, such that $\mathbf{Y}_{n\cdot} \equiv (y_{n1}, \dots y_{nD})^\top$ is the D-dimensional vector of manifest variables relating to a FVT for $n = 1, \dots, N$. Let

$$y_{ni} = f(\mathbf{t}_n, \mathbf{x}_i) + \epsilon_{ni}, \tag{3.1}$$

where $f(\mathbf{t}_n, \mathbf{x}_i)$, the underlying function-valued trait, and $\epsilon_{ni} \sim \mathcal{N}\left(0, \lambda_i^{-1}\right)$ are associated with position $(\mathbf{t}_n, \mathbf{x}_i) \in \mathcal{T} \times \mathcal{X}$. It is further assumed that traits are aligned on $\mathcal{X}$ such that differences between $f(\mathbf{t}_n, \mathbf{x}_i)$ and $f(\mathbf{t}_{n'}, \mathbf{x}_i)$ are due to amplitude variation only. Then, the Phylogenetic Latent Variable Model (PLVM) is defined by

$$f(\mathbf{t}_n, \mathbf{x}_i) = \sum_{j=1}^{Q} w_j(\mathbf{x}_i) z_j(\mathbf{t}_n), \tag{3.2}$$

for basis functions $w_j(\cdot)$ and latent variables

$$z_j(\mathbf{t}_n) \sim \mathcal{GP}(0, k_{\mathcal{T}}(\mathbf{t}_n, \mathbf{t}_{n'} | \theta_{\mathcal{T}})), \tag{3.3}$$

given $k_{\mathcal{T}}(\cdot, \cdot | \theta_{\mathcal{T}})$, the covariance function for a univariate Gauss-Markov process over $\mathcal{T}$ which is dependant on hyper-parameters $\theta_{\mathcal{T}}$. Because the model does not assume the trait covariance function to be from a class of stationary covariance functions, it describes a spatially inhomogeneous phylogenetic Gaussian process [Jones and Moriarty, 2013].

This model can be rewritten in matrix notation, such that

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^\top + \boldsymbol{\epsilon}, \tag{3.4}$$

where the $N \times Q$ matrix of latent variables, also referred to as *factors*, is given by $\mathbf{Z}_{nj} = z_j(\mathbf{t}_n)$, the $D \times Q$ matrix of basis functions, referred to as *loadings*, is $\mathbf{W}_{ij} = w_j(\mathbf{x}_i)$, and for the $N \times D$ observation noise matrix, $\boldsymbol{\epsilon}_{ni} = \epsilon_{ni}$. Thus,

$$\begin{aligned} p(\mathbf{Y} | \mathbf{W}, \mathbf{Z}, \boldsymbol{\Lambda}) &= \mathcal{MN}\left(\mathbf{Y} | \mathbf{Z}\mathbf{W}^\top, \mathbf{I}_N, \boldsymbol{\Lambda}^{-1}\right), \\ &= \frac{\exp\left(-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Lambda}\left(\mathbf{Y} - \mathbf{Z}\mathbf{W}^\top\right)^\top \mathbf{I}_N^{-1}\left(\mathbf{Y} - \mathbf{Z}\mathbf{W}^\top\right)\right)\right)}{(2\pi)^{ND/2} \left|\boldsymbol{\Lambda}^{-1}\right|^{N/2} |\mathbf{I}_N|^{D/2}}, \end{aligned} \tag{3.5}$$

is a Matrix-Normal pdf [Dawid, 1981], where $\boldsymbol{\Lambda}$ is a diagonal matrix such that

$\mathbf{\Lambda}_i = \lambda_i$. The prior distribution over factors is given by

$$p\left(\mathbf{Z}|\theta_{\mathcal{T}}\right) = \mathcal{MN}\left(\mathbf{Z}|\mathbf{0}, \mathbf{K}_{\mathcal{T}}, \mathbf{I}_Q\right), \tag{3.6}$$

where $\mathbf{K}_{\mathcal{T}}$ is the Gram matrix of $k_{\mathcal{T}}\left(\cdot,\cdot|\theta_{\mathcal{T}}\right)$. This allows the marginalised likelihood for the model, obtained after integrating out factors, to be defined as

$$\mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}, Q\right) = \mathcal{N}\left(\mathrm{vec}\left(\mathbf{Y}\right)|\mathbf{0}, \left(\mathbf{W}\mathbf{W}^{\top} \otimes \mathbf{K}_{\mathcal{T}}\right) + \left(\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_N\right)\right), \tag{3.7}$$

where $\mathrm{vec}\left(\cdot\right)$ is the vec operator [Petersen and Pedersen, 2012]. This makes clear that the inclusion of observation noise relaxes the assumption of separability of the phylogeny trait covariance function for observations of a FVT.

3.2.3 Efficient Computation of the Model Likelihood

Naive computation of the Gaussian likelihood in (3.7) requires the inversion of a $DN \times DN$ covariance matrix, an operation that scales with $\mathcal{O}\left(\left(DN\right)^3\right)$ operations, a prohibitively expensive cost for anything other than the smallest of datasets. Jones and Moriarty [2013], Hadjipantelis et al. [2013], and Mariñas-Collado et al. [2019] all address this problem by assuming noise-free observations of the FVT are available, which is to say that $\mathbf{\Lambda}_i = 0$ for all $i = 1, \ldots D$. This approach does reduce the computational burden to $\mathcal{O}\left(D^3 + N^3\right)$, however, it remains problematic. Firstly, assuming noise free observations may result in a rigid model, prone to overfitting, which fails to identify any phylogenetic signal. A second problem is that, while it can be argued that the $\mathcal{O}\left(D^3\right)$ expense is worth paying to model correlation within a FVT, the $\mathcal{O}\left(N^3\right)$ computational expense means Bayesian inference for $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$ remains impractical.

An alternative approach is to introduce factors and FVTs for internal nodes of the phylogeny such that $\mathbf{z}_i^* \equiv \left(z_1\left(\mathbf{t}_i\right), \ldots, z_Q\left(\mathbf{t}_i\right)\right)^{\top}$ and $\mathbf{f}_i \equiv \mathbf{W}\mathbf{z}_i^*$ for $i = N+1, \ldots, N+S+M$. This implies a joint distribution over observed and internal traits

$$\begin{aligned} &p\left(\mathbf{Y}, \mathbf{f}_{N+1}, \ldots \mathbf{f}_{N+S+M}\right) = \\ &\qquad \left(\prod_{n=1}^{N} p\left(\mathbf{Y}_{n\cdot}|\mathbf{f}_{\mathrm{pa}(n)}\right)\right)\left(\prod_{i=N+1}^{N+S+M-1} p\left(\mathbf{f}_i|\mathbf{f}_{\mathrm{pa}(i)}\right)\right) p\left(\mathbf{f}_{N+S+M}\right), \end{aligned} \tag{3.8}$$

which in turn implies that

$$\mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}\right) = \int \cdots \int p\left(\mathbf{Y}, \mathbf{f}_{N+1}, \ldots \mathbf{f}_{N+S+M}\right) d\mathbf{f}_{N+1} \ldots d\mathbf{f}_{N+S+M}, \tag{3.9}$$

Then, defining $\{\mathbf{Y}\}_h^{post}$ as the set of all observed traits descended from and including that at position $\mathbf{t}_h$, a close examination of (3.9) reveals that by iteratively solving the integral

$$p\left(\{\mathbf{Y}\}_h^{post}|\mathbf{f}_{\mathrm{pa}(h)}\right) = \int \left(\prod_{i\in\mathrm{ch}(h)} p\left(\{\mathbf{Y}\}_i^{post}|\mathbf{f}_h\right)\right) p\left(\mathbf{f}_h|\mathbf{f}_{\mathrm{pa}(h)}\right) d\mathbf{f}_h, \tag{3.10}$$

the model likelihood can be evaluated in a post-order traversal of $\mathcal{T}$.

Not only does (3.10) generalise the derivations of Felsenstein [1973], Pybus et al. [2012], Freckleton [2012], and Mitov and Stadler [2017] to a general Gauss-Markov model for trait evolution, it also includes independent Gaussian noise on observed traits. Thus, it provides a far more flexible approach to modelling trait evolution, while scaling linearly with N. Because it starts at leaf nodes and works back to the root of a tree, the quantity computed is referred to here as the *pruned likelihood*. Though the derivation of quantities required for the pruned likelihood is straightforward for a noise-free BM model for trait evolution [Pybus et al., 2012], extending this derivation to a general Gauss-Markov case is a notationally involved task, as such, it is included in Appendix A.1. Given this algorithm however, the implementation of Bayesian inference schemes for $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$ becomes more practical.

3.2.4 Prior Specification

In order to perform Bayesian inference on the PLVM for FVTs, a prior distribution for $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$ along with the form of phylogenetic covariance function $k_{\mathcal{T}}(\cdot, \cdot|\theta_{\mathcal{T}})$ must be defined. Furthermore, the model in (3.2) is in fact a generalisation of Factor Analysis, and as such, particular consideration must be given to invariance in the likelihood due to scaling, reflection, and rotation [Lopes, 2014].

Consider first $k_{\mathcal{T}}(\cdot, \cdot|\theta_{\mathcal{T}})$, which is assumed to define a univariate first-order Gauss-Markov process over $\mathcal{T}$ governed by hyper-parameters $\theta_{\mathcal{T}}$. The family of stationary Gauss-Markov processes, that is, (OU) processes [Uhlenbeck and Ornstein, 1930; Doob, 1942], are typically defined over the interval $t \in \mathbb{R}^+$ by the stochastic differential equation (SDE)

$$dz(t) = \alpha(\mu - z(t))\, dt + \beta\, dW(t), \tag{3.11}$$

where $\mu \in \mathbb{R}$ is the process mean, $\alpha \in \mathbb{R}^+$ is the central tendency, and $\beta \in \mathbb{R}^+$ scales the Weiner process $W(t)$ [Billingsley, 2008], with the process being a BM in

its limit as $\alpha \to 0$. Thus, assuming that the Gauss-Markov process over $\mathcal{T}$ belongs to this family provides a flexible approach to modelling trait evolution which can also approximate the BM model for trait evolution [Felsenstein, 1973].

As presented by Rasmussen and Williams [2006], the covariance function for the process defined in (3.11) is given by

$$k\left(t, t'\right) = \frac{\beta^2}{2\alpha} \exp\left(-\alpha|t - t'|\right), \tag{3.12}$$

which Jones and Moriarty [2013] extended to phylogenies by replacing the Euclidean distance $|t - t'|$ with the patristic distance $\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right)$ for $\mathbf{t}, \mathbf{t}' \in \mathcal{T}$. The covariance function for the process over $\mathcal{T}$ can then be completed by including a parameter for non-phylogenetic inter-taxon variation, as per the Phylogenetic Mixed Model [Housworth et al., 2004], and another for intra-taxon variation. And so, for $\mathbf{t}_i, \mathbf{t}_j \in \mathcal{T}$ with $i, j = 1, \ldots, N + S + M$, consider

$$\begin{aligned} k_{\mathcal{T}}^{tmp}\left(\mathbf{t}_i, \mathbf{t}_j\right) = \sigma_h^2 \, \exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right)}{\ell}\right) + \\ \sigma_e^2 \, \delta\left(\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right) = 0\right) \delta\left(i \leq N + S\right) + \\ \sigma_\tau^2 \, \delta\left(i = j\right) \delta\left(i \leq N\right), \end{aligned}$$

where $\sigma_h^2 \in \mathbb{R}^+$ is the heritable variance, $\ell \in \mathbb{R}^+$ is the phylogenetic length-scale, $\sigma_e^2 \in \mathbb{R}^+$ is the non-phylogenetic inter-taxon variance, $\sigma_\tau^2 \in \mathbb{R}^+$ is the intra-taxon variance, and $\delta\left(\cdot\right)$ is an indicator function. This defines the *heritability* of a process over $\mathcal{T}$ as

$$\kappa \equiv \frac{\sigma_h^2}{\sigma_h^2 + \sigma_e^2} \tag{3.13}$$

While $k_{\mathcal{T}}^{tmp}\left(\cdot, \cdot\right)$ defines the Gauss-Markov process over $\mathcal{T}$, scale invariance in (3.7) must be considered before setting $k_{\mathcal{T}}\left(\cdot, \cdot | \theta_{\mathcal{T}}\right)$. Scale invariance in Factor Analysis is typically fixed by assuming the marginal variance of each factor to equal some constant, usually one [Lopes, 2014]. Enforcing this constraint on $k_{\mathcal{T}}^{tmp}\left(\cdot, \cdot\right)$ implies that $\sigma_h^2 + \sigma_e^2 + \sigma_\tau^2 = 1$ and $\sigma_h^2, \sigma_e^2, \sigma_\tau^2 \in (0, 1)$. Then, defining $\theta_{\mathcal{T}} = \{\kappa, \tau, \ell\}$ with $\kappa \in (0, 1)$, $\tau \in (0, 1)$, and $\ell \in \mathbb{R}^+$, such that the heritable variance $\sigma_h^2 = (1 - \tau)\,\kappa$, environmental variance $\sigma_e^2 = (1 - \tau)\,(1 - \kappa)$, and intra-taxon variance $\sigma_\tau^2 = \tau$, the phylogenetic covariance function is

$$\begin{aligned} k_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j | \theta_{\mathcal{T}}\right) = (1 - \tau) \left(\kappa \, \exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right)}{\ell}\right) + \right. \\ \left. (1 - \kappa) \; \delta\left(\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right) = 0\right) \delta\left(i \leq N + S\right)\right) + \end{aligned}$$

$$\tau\,\delta\left(i=j\right)\delta\left(i\leq N\right). \tag{3.14}$$

Prior distributions for the model parameters and hyper-parameters can now be considered. Given that κ and τ are each defined over the unit interval, it is natural to assume a Beta distributed hyper-prior. That is to say,

$$\begin{aligned} p\left(\kappa\right) &= Beta\left(\kappa|a_\kappa, b_\kappa\right), \\ &= \frac{\Gamma\left(a_\kappa+b_\kappa\right)}{\Gamma\left(a_\kappa\right)\Gamma\left(b_\kappa\right)}\kappa^{a_\kappa-1}\left(1-\kappa\right)^{b_\kappa-1}, \end{aligned} \tag{3.15}$$

for shape parameters $a_\kappa \in \mathbb{R}^+$ and $b_\kappa \in \mathbb{R}^+$, while $p\left(\tau\right) = Beta\left(\tau|a_\tau, b_\tau\right)$ is defined analogously.

The hyper-prior distribution for ℓ requires somewhat more careful consideration. Firstly, note that the OU process is equivalent to the Matérn process with smoothing parameter $\nu = 1/2$ [Rasmussen and Williams, 2006]. For such models, it is impossible to estimate both the variance and length-scale consistently [Zhang, 2004], a problem typically addressed by fixing the variance to be constant [Monterrubio-Gómez et al., 2018]. Within the PLVM however, $\mathbf{W}$, κ, and τ all contribute to the variance of the stochastic process and are to be inferred from data. An alternative approach would be to fix ℓ a priori; however, this would result in a less flexible model for ancestral reconstruction. Thus, a suitably informative hyper-prior distribution must be chosen for ℓ.

Given that BM is the standard model for trait evolution, the hyper-prior distribution for ℓ is chosen such that

$$\sigma_h^2\,\exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_i,\mathbf{t}_j\right)}{\ell}\right),$$

defines a stochastic process over $\mathcal{T}$ that is similar to BM with unit variance when $\mathcal{T}$ has been scaled such that $\max\left\{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_n,\mathbf{t}_{N+S+M}\right)\right\}_{n=1}^{N}=1$. This implies that $\alpha\to 0$ and $\beta=1$ in (3.11). As such, given that (3.12) implies $\sigma_h^2=\frac{\beta^2}{2\alpha}$ and $\ell=\alpha^{-1}$, assume that $\mathbb{E}\left[\ell|\sigma_h^2\right]=2\sigma_h^2$ and let ℓ be a Gamma distributed random variable, which is to say that

$$\begin{aligned} p\left(\ell|\sigma_h^2\right) &= Gamma\left(\ell|a_\ell, b_\ell\right), \\ &= \frac{b_\ell^{a_\ell}}{\Gamma\left(a_\ell\right)}\ell^{a_\ell-1}\exp\left(-b_\ell\ell\right), \end{aligned} \tag{3.16}$$

for shape $a_\ell > 0$ and rate $b_\ell > 0$. Note that $\mathbb{E}\left[\ell|\sigma_h^2\right]=\frac{a_\ell}{b_\ell}$ and $\mathrm{Var}\left(\ell|\sigma_h^2\right)=\frac{a_\ell}{b_\ell^2}$, while $p\left(\ell|\sigma_h^2\right)$ is maximised at $\frac{a_\ell-1}{b_\ell}$ for $a_\ell \geq 1$. The shape and rate parameters

are then chosen by minimising the squared difference between the hyper-prior mean and mode, while maximising the variance. That is to say, a_ℓ and b_ℓ are obtained by minimising the objective

$$\left(\frac{a_\ell}{b_\ell} - \frac{a_\ell - 1}{b_\ell}\right)^2 - \frac{a_\ell}{b_\ell^2} = \frac{1 - a_\ell}{b_\ell^2},$$

when $2\sigma_h^2 b_\ell = a_\ell$. And so, the hyper-prior distribution for ℓ is

$$p\left(\ell|\sigma_h^2\right) = Gamma\left(\ell|2, \frac{1}{\sigma_h^2}\right). \tag{3.17}$$

With that, distributions over factors and phylogenetic Gaussian process hyper-parameters in the PLVM have been specified fully, and scale invariance in the PLVM has been addressed by the parametrisation of (3.14) and an informative hyper-prior distribution for ℓ.

A prior distribution for loadings $\mathbf{W}$ must now be defined. To do so, consider first the basis functions $w_j(\cdot)$ for $j = 1, \ldots, Q$, which are assumed to be twice mean square differentiable processes over the domain $\mathcal{X} \equiv \mathbb{R}^d$. As such, an isotropic Matérn GP prior with $\nu = 5/2$ is deemed appropriate [Stein, 2012]. That is to say

$$w_j(\mathbf{x}) \sim \mathcal{GP}\left(0, k_\mathcal{X}\left(\mathbf{x}, \mathbf{x}'|\theta_\mathcal{X}\right)\right)$$

where

$$k_\mathcal{X}\left(\mathbf{x}, \mathbf{x}'|\theta_\mathcal{X}\right) = \sigma_w^2\left(1 + \frac{\sqrt{5}r}{\ell_w} + \frac{5r^2}{3\ell_w^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell_w}\right)$$

for the hyper-parameters $\theta_\mathcal{X} = \left\{\sigma_w^2, \ell_w\right\}$, the variance and characteristic length-scale respectively, where $r = |\mathbf{x} - \mathbf{x}'|$ denotes the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$. As discussed, σ_w^2 and ℓ_w cannot be estimated consistently [Zhang, 2004], and so the prior distribution for $w_j(\cdot)$ is completed by choosing some σ_w^2 a priori and assuming that $p(\ell_w) = Gamma(\ell_w|a_w, b_w)$. This implies a prior distribution for the PLVM loadings

$$p(\mathbf{W}|\theta_\mathcal{X}, Q) = \mathcal{MN}(\mathbf{W}|\mathbf{0}, \mathbf{K}_\mathcal{X}, \mathbf{I}_Q), \tag{3.18}$$

where $\mathbf{K}_\mathcal{X}$ is the Gram matrix of $k_\mathcal{X}(\cdot, \cdot|\theta_\mathcal{X})$.

Rather than attempting to encode a solution to rotation and reflection invariance within the prior specification for $\mathbf{W}$, note that the prior distribution defined in (3.18) is itself invariant to rotation of $\mathbf{W}$. Therefore, by the LQ variant of the

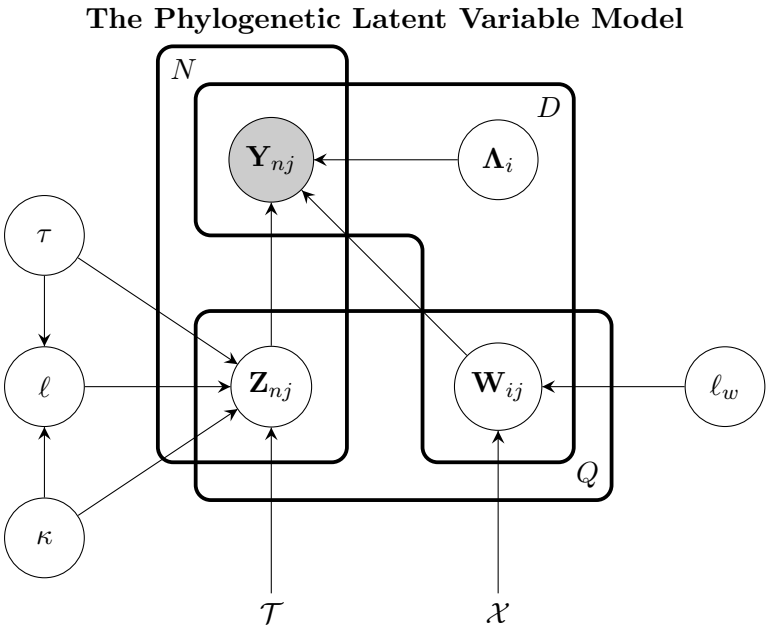


Figure 3.3: A graphical representation of the Phylogenetic Latent Variable Model described in sub-sections 3.2.2 and 3.2.4. Each circle represents a random variable, where those that are shaded grey have been observed. Boxes around circles are plates, denoting the number of existing random variables of that type. The phylogeny-trait space is included for completeness.

QR decomposition [Golub and Van Loan, 2013], for any $\mathbf{W}$, there exists a matrix $\mathbf{W}'$ for which the upper triangular elements are zero and $\mathcal{MN}\left(\mathbf{W}|\mathbf{0}, \mathbf{K}_{\mathcal{X}}, \mathbf{I}_Q\right) = \mathcal{MN}\left(\mathbf{W}'|\mathbf{0}, \mathbf{K}_{\mathcal{X}}, \mathbf{I}_Q\right)$. This implies that correcting for rotation and reflection invariance can be treated as a post-processing step in any inference scheme, while preserving the identity $\left(\mathbb{E}_{p(\mathbf{W}|\theta_{\mathcal{X}})}\left[\mathbf{W}\mathbf{W}^\top\right]\right)_{ii} = \left(\mathbb{E}_{p(\mathbf{W}|\theta_{\mathcal{X}})}\left[\mathbf{W}\mathbf{W}^\top\right]\right)_{jj}$ for all $i = 1, \ldots, D$ and $j = 1, \ldots, D$.

The prior specification for the model is then completed by assuming that $p\left(\mathbf{\Lambda}_i\right) = Gamma\left(\mathbf{\Lambda}_i|a_{\mathbf{\Lambda}}, b_{\mathbf{\Lambda}}\right)$. A graphical representation for the model is presented in Figure 3.3.

3.2.5 Posterior Inference and Model Selection

Applying the PLVM to ancestral reconstruction of a FVTs requires inference on

$$p\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}, Q\right) \propto \mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}, Q\right) p\left(\mathbf{W}\right) p\left(\theta_{\mathcal{T}}\right) p\left(\mathbf{\Lambda}\right),$$

that is, the posterior distribution over the loadings, phylogenetic hyper-parameters, and observation noise, given observed data $\mathbf{Y}$ and the number of factors Q. Inference for Q can then be treated as a model comparison problem.

In Bayesian Factor Analysis, inference on the loadings and observation noise is typically conditional on the factors and performed using a Gibbs sampler [Lopes and West, 2004]. This represents an efficient approach when factors are independent and Gaussian. Phylogenetic Factor Analysis (PFA) also relies on a Gibbs sampler for inference, the computational cost of which scales with $\mathcal{O}\left(N^2\right)$, given that factors are modelled as BM over the phylogeny [Tolkoff et al., 2017]. While this may suggest a Gibbs sampling approach to inference for the PLVM, in fact, it has been deemed inappropriate in this case. Closed form conditional distributions do exist for $\mathbf{W}$, $\mathbf{Z}$, and $\mathbf{\Lambda}$, but no such distribution is available for $\theta_{\mathcal{T}}$. Thus, some variant on the Metropolis-Hastings algorithm would be required for posterior inference. Furthermore, the computational expense of sampling the factors scales with $\mathcal{O}\left(N^2\right)$. Thus, the approach taken here is to perform inference after integrating over factors. This block-at-a-time MCMC algorithm scales linearly with N, given the efficient computation of the pruned likelihood, and samples from

$$\begin{aligned} p\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}, \theta_{\mathcal{X}} | \mathbf{Y}, Q\right) &\propto \mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda} | \mathbf{Y}, Q\right) p\left(\mathbf{W} | \theta_{\mathcal{X}}\right) p\left(\theta_{\mathcal{X}}\right) p\left(\theta_{\mathcal{T}}\right) p\left(\mathbf{\Lambda}\right), \\ &= \mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda} | \mathbf{Y}, Q\right) p\left(\mathbf{W} | \ell_w\right) \\ &\quad p\left(\ell_w\right) p\left(\ell | \kappa, \tau\right) p\left(\kappa\right) p\left(\tau\right) \prod_{i=1}^{D} p\left(\mathbf{\Lambda}_i\right), \end{aligned} \tag{3.19}$$

which is the posterior distribution over all model parameters and hyper-parameters after integrating out factors $\mathbf{Z}$.

The first block considered corresponds to $\mathbf{W}$ conditional on $\{\theta_{\mathcal{T}}, \mathbf{\Lambda}, \theta_{\mathcal{X}}\}$, which is sampled by an Elliptical Slice Sampler (ESS) [Murray et al., 2010], similar to that presented in Algorithm 3. Noting that for $\mathbf{L}\mathbf{L}^{\top} = \mathbf{K}_{\mathcal{X}}$ there exists $\boldsymbol{\zeta}$, the "whitened" representation of $\mathbf{W}$ [Petersen and Pedersen, 2012], such that $\mathbf{W} = \mathbf{L}\boldsymbol{\zeta}$ and $\mathcal{MN}\left(\mathbf{W} | \mathbf{0}, \mathbf{K}_{\mathcal{X}}, \mathbf{I}_Q\right) = \mathcal{MN}\left(\boldsymbol{\zeta} | \mathbf{0}, \mathbf{I}_D, \mathbf{I}_Q\right)$, the target distribution for the ESS is

$$\pi\left(\boldsymbol{\zeta}\right) \propto \mathcal{L}\left(\mathbf{W} = \mathbf{L}\boldsymbol{\zeta}, \theta_{\mathcal{T}}, \mathbf{\Lambda} | \mathbf{Y}, Q\right) \mathcal{MN}\left(\boldsymbol{\zeta} | \mathbf{0}, \mathbf{I}_D, \mathbf{I}_Q\right).$$

Rotation invariance is corrected in this block by simply rotating each update according to the LQ decomposition [Golub and Van Loan, 2013], yielding samples for which all upper-triangular elements are zero. Reflection invariance is also corrected for in each block, using an approach similar to that proposed by [Stephens, 2000]

and implemented in Phylogenetic Factor Analysis [Tolkoff et al., 2017]. Assuming that the desired posterior density for the j^{th} column of $\mathbf{W}$ is $\mathcal{N}(\mathbf{W}_{\cdot j}|\mathbf{m}_j, \mathbf{\Sigma}_j)$, after rotation the j^{th} column of the sample is multiplied by -1 if $\mathcal{N}(-\mathbf{W}_{\cdot j}|\mathbf{m}_j, \mathbf{\Sigma}_j) > \mathcal{N}(\mathbf{W}_{\cdot j}|\mathbf{m}_j, \mathbf{\Sigma}_j)$, correcting any reflection invariance. If $\mathbf{m}_j$, and $\mathbf{\Sigma}_j$ are not known a priori, then they can be estimated from posterior samples, either selecting a reference sample and setting $\mathbf{\Sigma}_j = \mathbf{I}_D$, or by an iterative updating scheme. When this is the case, correcting reflection invariance is left until after the full Markov chain has been sampled.

The second block draws samples for ℓ_w using an Ancillarity-Sufficiency Interweaving Strategy (ASIS) [Yu and Meng, 2011], a popular approach to hyperparameter inference for GPs [Murray and Adams, 2010; Filippone et al., 2013]. This involves making two sub-updates for ℓ_w within each full updating step. Firstly, rather than defining target densities with respect to $\ell_w \in \mathbb{R}^+$, consider instead $\log \ell_w \in \mathbb{R}$. This allows an Adaptive Metropolis (AM) sampling scheme to be employed (see Algorithm 2), which tunes proposal densities automatically while preserving the detailed balance condition in its limit [Haario et al., 2001; Roberts and Rosenthal, 2009]. Thus equipped, the first sub-update, referred to as the Sufficient Augmentation by Yu and Meng [2011] ("unwhitened" in Murray and Adams [2010]), updates the Markov Chain according to the target distribution

$$\pi(\log \ell_w) \propto \mathcal{MN}(\mathbf{W}|\mathbf{0}, \mathbf{K}_{\mathcal{X}}, \mathbf{I}_Q)\, Gamma(\ell_w|a_w, b_w)\, \ell_w.$$

while the second, the Ancilliary Augmentation ("whitened"), updates with respect to

$$\pi(\log \ell_w) \propto \mathcal{L}(\mathbf{W} = \mathbf{L}\boldsymbol{\zeta}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}, Q)\, Gamma(\ell_w|a_w, b_w)\, \ell_w.$$

In the third block, the Markov Chain for $\theta_{\mathcal{T}}$ is considered as two sub-blocks, the first being ℓ conditional on $\{\mathbf{W}, \mathbf{\Lambda}, \kappa, \tau\}$. As was the case with ℓ_w, updates are performed within an AM sampling scheme where the target distribution is given by

$$\pi(\log \ell) \propto \mathcal{L}(\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}|\mathbf{Y}, Q)\, Gamma(\ell|a_\ell, b_\ell)\, \ell.$$

For the second sub-block, define the logit transform as

$$\operatorname{logit} x \equiv \log \frac{x}{1-x}, \tag{3.20}$$

such that logit : $[0,1] \to \mathbb{R}$. Then, the target distribution for an AM sampling

scheme for κ and τ is

$$\pi\left(\operatorname{logit}\kappa, \operatorname{logit}\tau\right) \propto \mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda} | \mathbf{Y}, Q\right) Beta\left(\kappa | a_{\kappa}, b_{\kappa}\right) Beta\left(\tau | a_{\tau}, b_{\tau}\right) \\ \kappa\left(1-\kappa\right)\tau\left(1-\tau\right).$$

The block-at-a-time MCMC inference scheme is then completed by a final AM step for $\boldsymbol{\Lambda}_i$ where

$$\pi\left(\log \boldsymbol{\Lambda}\right) \propto \mathcal{L}\left(\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda} | \mathbf{Y}, Q\right) \prod_{i=1}^{D} Gamma\left(\boldsymbol{\Lambda}_i | a_{\boldsymbol{\Lambda}}, b_{\boldsymbol{\Lambda}}\right) \boldsymbol{\Lambda}_i.$$

All that remains is to select the appropriate number of factors, Q, given the model evidence

$$p\left(\mathbf{Y} | Q\right) = \int_{\{\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}}\}} p\left(\mathbf{Y}, \mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}} | Q\right) d\left\{\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}}\right\}. \tag{3.21}$$

This is equivalent to the normalising constant of (3.19) and, given a Markov Chain with this stationary distribution, the evidence can be estimated by a Bridge Sampling scheme [Meng and Wong, 1996; Gronau et al., 2017a], which is straightforward to implement using the `bridgesampling` package in R [Gronau et al., 2017b; R Core Team, 2019]. With that, posterior inference for the PLVM is complete.

3.2.6 Ancestral Reconstruction

The PLVM developed above has been formulated for the ancestral reconstruction of a FVT, while allowing uncertainty about the reconstruction to be quantified. This amounts to obtaining

$$p\left(\mathbf{f}_* | \mathbf{Y}, Q\right) = \int_{\{\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}}\}} p\left(\mathbf{f}_*, \mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}} | \mathbf{Y}\right) d\left\{\mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda}, \theta_{\mathcal{X}}\right\} \tag{3.22}$$

where $\mathbf{f}_* = \left(f\left(\mathbf{t}_*, \mathbf{x}_1\right), \ldots, f\left(\mathbf{t}_*, \mathbf{x}_D\right)\right)$ for some ancestral position $\mathbf{t}_* \in \mathcal{T}$. This distribution can be sampled within the MCMC inference scheme.

Given the joint distribution

$$\begin{bmatrix} \operatorname{vec}\left(\mathbf{Y}\right) \\ \mathbf{f}_* \end{bmatrix} | \mathbf{W}, \theta_{\mathcal{T}}, \boldsymbol{\Lambda} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \left(\mathbf{K}_{\mathcal{X}} \otimes \mathbf{K}_{\mathcal{T}}\right) + \left(\boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_N\right) & \left(\mathbf{K}_{\mathcal{X}} \otimes \mathbf{k}_{\mathcal{T}*}\right) \\ \left(\mathbf{K}_{\mathcal{X}} \otimes \mathbf{k}_{\mathcal{T}*}\right)^{\top} & k_* \mathbf{K}_{\mathcal{X}} \end{bmatrix}\right),$$

where $\mathbf{K}_{\mathcal{X}} = \mathbf{W}\mathbf{W}^{\top}$, $\mathbf{k}_{\mathcal{T}*} = \left(k_{\mathcal{T}}\left(\mathbf{t}_1, \mathbf{t}_* | \theta_{\mathcal{X}}\right), \ldots, k_{\mathcal{T}}\left(\mathbf{t}_N, \mathbf{t}_* | \theta_{\mathcal{X}}\right)\right)^{\top}$, and $k_* = k_{\mathcal{T}}\left(\mathbf{t}_*, \mathbf{t}_* | \theta_{\mathcal{X}}\right)$,

the conditional distribution can be expressed as

$$p\left(\mathbf{f}_*|\mathbf{Y},\mathbf{W},\theta_{\mathcal{T}},\mathbf{\Lambda}\right)=\mathcal{N}\left(\mathbf{f}_*|\mathbf{m}_*,\mathbf{K}_*\right), \tag{3.23}$$

where

$$\begin{aligned}\mathbf{K}_* &= k_*\mathbf{K}_{\mathcal{X}}-\left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{k}_{\mathcal{T}*}\right)^{\top}\left(\left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{K}_{\mathcal{T}}\right)+\left(\mathbf{\Lambda}^{-1}\otimes\mathbf{I}_N\right)\right)^{-1}\left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{k}_{\mathcal{T}*}\right),\\ \mathbf{m}_* &= \left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{k}_{\mathcal{T}*}\right)^{\top}\left(\left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{K}_{\mathcal{T}}\right)+\left(\mathbf{\Lambda}^{-1}\otimes\mathbf{I}_N\right)\right)^{-1}\mathrm{vec}\left(\mathbf{Y}\right).\end{aligned}$$

While this distribution is analytically tractable, the $\mathcal{O}\left((ND)^3\right)$ cost of inverting $\left(\left(\mathbf{K}_{\mathcal{X}}\otimes\mathbf{K}_{\mathcal{T}}\right)+\left(\mathbf{\Lambda}^{-1}\otimes\mathbf{I}_N\right)\right)$ makes its computation infeasible. Fortunately, the principles that underpin the pruned likelihood can be extended to ancestral reconstruction.

Firstly, recall that $\{\mathbf{Y}\}_*^{post}$ denotes all the rows of $\mathbf{Y}$ descendant from and including $\mathbf{t}_*$, and let $\{\mathbf{Y}\}_*^{pre}=\mathbf{Y}/\{\mathbf{Y}\}_*^{post}$. Suppressing the notation of $(\mathbf{W},\theta_{\mathcal{T}},\mathbf{\Lambda},\mathcal{T})$ for clarity, it can be shown that

$$p\left(\mathbf{f}_*|\mathbf{Y}\right)\propto p\left(\{\mathbf{Y}\}_*^{post}|\mathbf{f}_*\right)p\left(\mathbf{f}_*|\{\mathbf{Y}\}_*^{pre}\right). \tag{3.24}$$

The quantity $p\left(\{\mathbf{Y}\}_i^{post}|\mathbf{f}_i\right)$ is computed for all $i=1,\ldots,N+S+M$ in the pruned likelihood algorithm presented in Appendix A.1, thus it remains only to find an expression for $p\left(\mathbf{f}_*|\{\mathbf{Y}\}_*^{pre}\right)$. The key point to note is that, for the Markov process over $\mathcal{T}$, $\mathbf{f}_*$ is independent of $\{\mathbf{Y}\}_*^{pre}$ given $\mathbf{f}_{\mathrm{pa}(*)}$, thus

$$\begin{aligned}&p(\mathbf{f}_*|\,\{\mathbf{Y}\}_*^{pre})\\ &=\int p\left(\mathbf{f}_*,\mathbf{f}_{\mathrm{pa}(*)}|\left\{\{\mathbf{Y}\}_{\mathrm{sib}(*)}^{post}\right\},\{\mathbf{Y}\}_{\mathrm{pa}(*)}^{pre}\right)d\mathbf{f}_{\mathrm{pa}(*)},\\ &\propto\int p\left(\mathbf{f}_*|\mathbf{f}_{\mathrm{pa}(*)}\right)\left(\prod_{j\in sib(*)}p\left(\{\mathbf{Y}\}_j^{post}|\mathbf{f}_{\mathrm{pa}(*)}\right)\right)p\left(\mathbf{f}_{\mathrm{pa}(*)}|\{\mathbf{Y}\}_{\mathrm{pa}(*)}^{pre}\right)d\mathbf{f}_{\mathrm{pa}(*)},\end{aligned} \tag{3.25}$$

where $\mathrm{sib}(*)$ denotes the siblings of $\mathbf{t}_*$ and $\left\{\{\mathbf{Y}\}_{\mathrm{sib}(*)}^{post}\right\}$ the set $\left\{\{\mathbf{Y}\}_i^{post}:i\in\mathrm{sib}(*)\right\}$. This expression defines a recursion which can be solved up to a normalising constant by traversing $\mathcal{T}$ from root to $\mathbf{t}_*$. This insight allowed Cybis et al. [2015] to efficiently compute closed-form conditional distributions for traits at terminal nodes of a bifurcating tree, under a BM model for trait evolution.

Substituting (3.25) into (3.24), the conditional distribution of a FVT can be efficiently computed at all internal nodes of $\mathcal{T}$ by following a post-order traversal of $\mathcal{T}$ with a pre-order traversal. The details of this algorithm, which applies to

Parameter:	κ	τ	ℓ	σ_w^2	ℓ_w	λ
Value:	0.95	0.05	2.5	1	0.5	10

Table 3.1: The assumed parameter values for a simulation study of the PLVM.

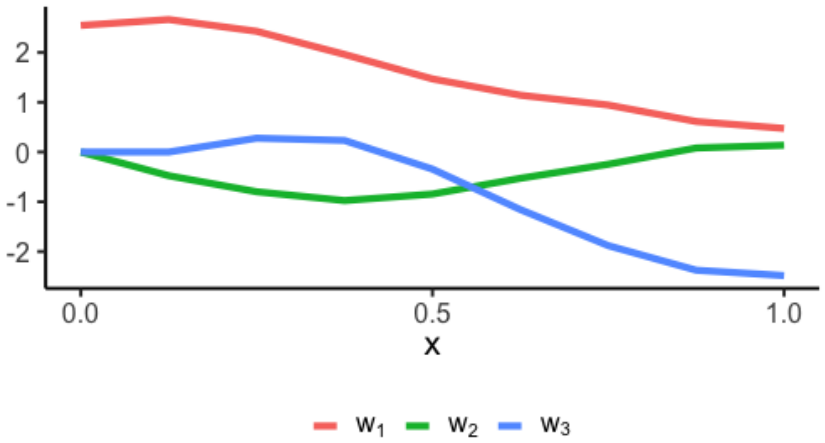
general Gauss-Markov models for trait evolution even when traits are subject to observation noise, is presented Appendix A.2. This allows samples be drawn from (3.23) in $\mathcal{O}\left(N^2\right)$ operations rather than $\mathcal{O}\left(N^3\right)$, which in turn allows (3.22) be sampled from efficiently.

3.3 Results for a Synthetic Example

Consider a dataset simulated from a PLVM, to which the MCMC inference scheme will be applied. Here, $Q = 3$ independent and identically distributed factors are sampled from an OU process over $\mathcal{T}$ where $S = 32$ and $N_i = 4$ for all $i = 1, \ldots, S$, yielding $N = 128$ samples from the model. Factors are then mapped to $D = 9$ manifest variables representing noisy observations of the FVT where $\mathbf{\Lambda} = \lambda \mathbf{I}_D$.

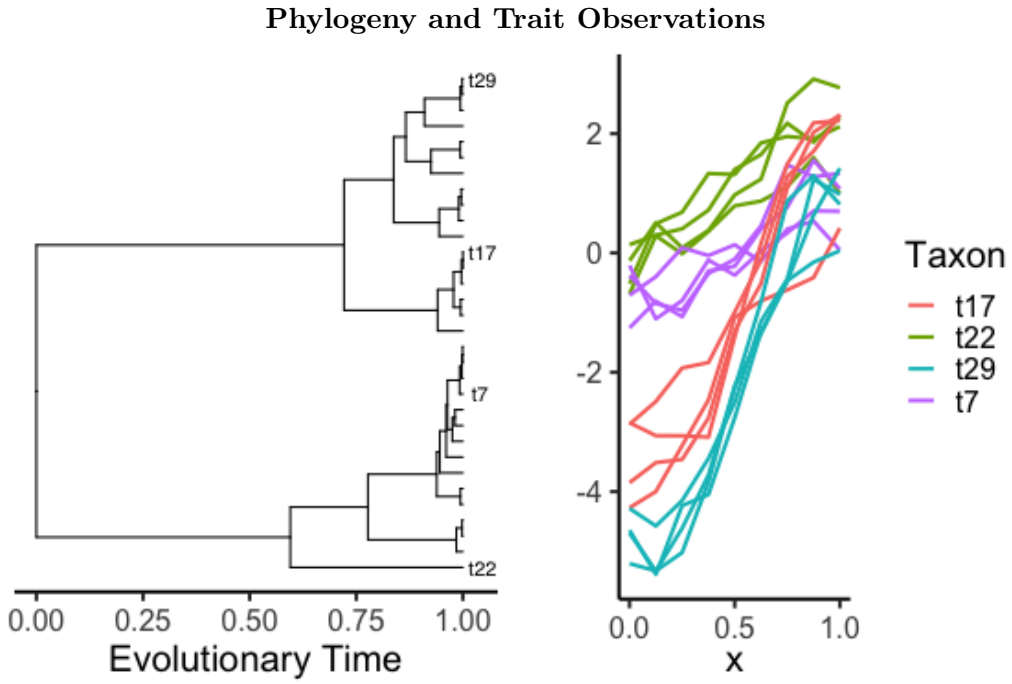
For the purposes of this experiment, $\mathcal{T}$ is set by first considering a phylogeny with 32 terminal nodes generated by a coalescent process using default parameters provided in `ape` [Paradis and Schliep, 2018], which is subsequently scaled such that the distance from the root to each tip is 1. This phylogeny is then extended to yield $\mathcal{T}$ by appending four nodes with zero edge weight to each terminal node. Phylogenetic hyper-parameters are fixed a-priori, where $\tau = 0.05$ reflects low intra-taxon variation, and $\kappa = 0.95$ implies that the process has strong heritability over $\mathcal{T}$. Setting $\ell = 2.5$ further implies that, for short time-scales, the process is more strongly correlated than BM with unit variance. The loading is fixed by sampling 3 independent zero-mean Matérn-$\frac{5}{2}$ GPs at 9 points spread uniformly over the unit interval, where $\sigma_w^2 = 1$ and $\ell_w = 0.5$, and rotating the result with a QR decomposition such that its upper triangular entries are 0. Finally, setting $\lambda = 10$ specifies the model. These parameter and hyper-parameter values are summarised in Table 3.1 and the loading in Figure 3.4. The phylogeny, along with some samples from the model are presented in Figure 3.5.

The first task is to check that bridge sampling identifies the number of latent factors correctly. To do this, five chains of length 20,000 are sampled from the posterior distribution (3.19) for $Q = 1, \ldots, 4$, given $a_\kappa = b_\kappa = a_\tau = b_\tau = 1$ and $a_w = b_w = a_{\mathbf{\Lambda}} = b_{\mathbf{\Lambda}} = 0.1$. Chains are initialised by sampling the prior at random with the first 1000 samples discarded as warm-up samples and the remainder being used for posterior inference, leaving 95,000 samples in total. These are used to



Loadings

Figure 3.4: Loadings are sampled from zero-mean Matérn-$\frac{5}{2}$ GPs at $D = 9$ points spread uniformly over the unit interval, and rotated by the QR decomposition such that upper-triangular entries are 0.



(a) Phylogenetic Tree

(b) Observed FVTs

Figure 3.5: Sub-figure (a) illustrates the taxon-level phylogeny simulated according to a coalescent process and scaled such that the distance from root to each tip is one, while sub-figure (b) illustrate the sampled manifest variables for selected taxa. The position of each taxon on $\mathcal{T}$ is noted in (a).

Q:	1	2	3	4
$\log p(\mathbf{Y}\|Q=3) - \log p(\mathbf{Y}\|Q)$:	686.53	109.81	0	20.34

Table 3.2: The log Bayes factor under $p(Q) \propto 1$ for each models considered, where $Q = 3$ is the null model.

compute the model evidence (3.21) for each of the models sampled. Assuming a uniform prior for Q, Bayes factors are then computed [Jarosz and Wiley, 2014], letting $Q = 3$ serve as the null model. Log Bayes factors for each model are presented in Table 3.2. According to Jeffreys [1939], such Bayes factors can be interpreted as decisive evidence that the number of latent factors is, in fact, three.

Having correctly identified $Q = 3$ as the most probable model for the data, it is the FVT distribution at internal nodes of $\mathcal{T}$ that is the primary object of interest. These nodes fall into two broad categories. Taxon-level nodes, those that are parents of nodes corresponding to observations, allow the definition of the FVT distribution for each observed taxon, while nodes corresponding to the unobserved ancestral taxa yield the ancestral reconstruction of the FVT. Samples from the FVT distribution at one of each node type, that is the node labelled `t17` in Figure 3.5 and the root of $\mathcal{T}$, are presented in Figure 3.6. Comparing samples to the true conditional distribution of the FVT given $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$, it can be seen that they match very closely, with strikingly similar regions of high density, even as samples integrate over uncertainty on $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}, \ell_w\}$. This is a very satisfying result, demonstrating the accuracy of MCMC inference for the ancestral reconstruction of the FVT within a PLVM.

Given that the primary goal, the ancestral reconstruction of a FVT, has been achieved, the convergence of Markov chains sampling from the model parameter and hyper-parameter posterior distribution is also of interest. Consider first the loading $\mathbf{W}$, illustrated in Figure 3.7, for which rotation invariance has been corrected by a QR decomposition and reflection invariance via the relabelling algorithm outlined above. The correlation structure of the FVT over $\mathcal{X}$ is being identified accurately, in that samples all have a very similar shape to the true loading, however, the magnitude, i.e. $\sqrt{\mathrm{tr}(\mathbf{W}^\top \mathbf{W})}$, is being underestimated. As would be expected, this inflates the implied values for latent factors, which in turn inflates the intra-taxon variation parameter τ and reduces the phylogenetic length-scale ℓ. These effects are manifest in Figure 3.8, although in no case does the true hyper-parameter value lie outside the sampled posterior distribution. While this behaviour is somewhat disappointing, it is not entirely unexpected, given that variance and length-scale cannot be estimated consistently for Matérn covariance functions [Zhang, 2004]. Despite this, Markov chains do mix well and converge to a single posterior mode in

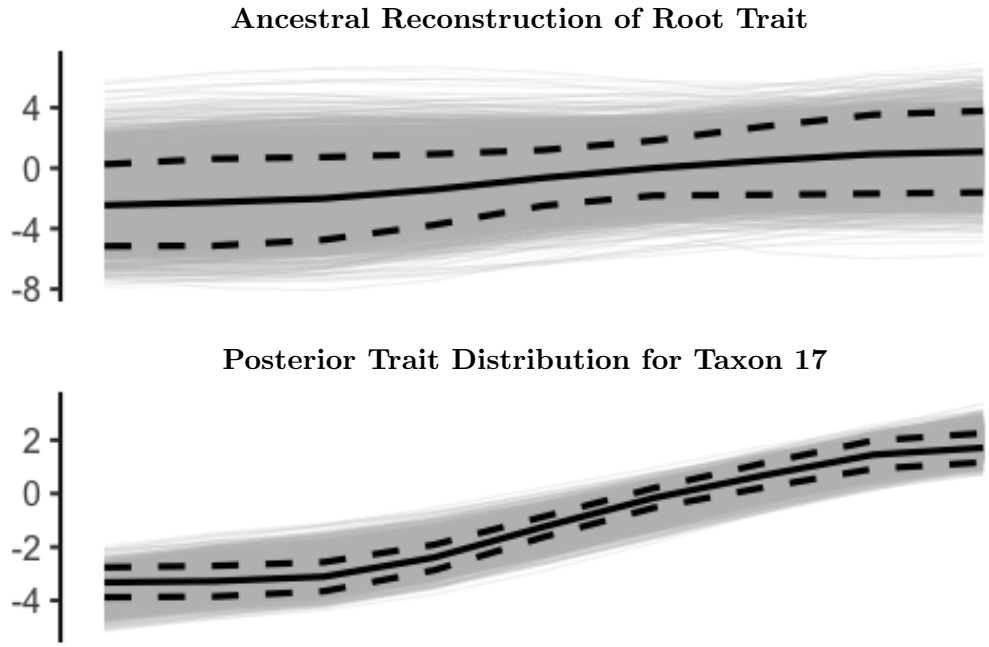


Figure 3.6: Comparison of the posterior distribution for a FVT at internal nodes of $\mathcal{T}$ for $Q = 3$, integrating over the uncertainty in $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}, \ell_w\}$ via the MCMC inference scheme described above, to the conditional distribution given $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$. Solid black lines represent the FVT conditional mean, dotted lines two standard deviations around the mean, and each opaque grey line represents a sample from the posterior.

each case.

The posterior distributions of the phylogenetic hyper-parameters indicate that care should be taken when attempting to interpret hyper-parameter values for the PLVM with respect to the heritability of a trait, however, it is knowledge about ancestral trait distributions that heritability seeks to define. Thus, given that the ancestral trait reconstruction matches the true ancestral distribution well, this is not considered a major cause for concern.

Analysis of the sampled Markov chains is completed in Figure 3.9, where chains for both ℓ_w and λ^{-1} can be seen to converge to a posterior mode. Samples for ℓ_w do reflect the fact that deflating the magnitude of $\mathbf{W}$ implies more strongly correlated functions however, and this is manifest in inflated values for ℓ_w given the fixed σ_w^2.

3.4 Discussion

This chapter has introduced a Phylogenetic Latent Variable Model (PLVM) for the ancestral reconstruction of function-valued traits (FVTs), describing a spatially inhomogeneous phylogenetic Gaussian process as a latent variable model within the Phylogenetic Gaussian Process Regression (PGPR) framework. Efficient algorithms computing the model likelihood and ancestral traits allow a Markov Chain Monte Carlo (MCMC) inference scheme to provide a Bayesian approach to estimation of model parameters and hyper-parameters, model selection, and ancestral reconstruction. Thus, it makes an important methodological contribution towards the study of FVTs in evolution.

Considering how this work builds upon that of Hadjipantelis et al. [2013], which had been state-of-the-art approach to inference within the PGPR framework, makes this contribution clear. Rather than simply assuming $\mathbf{Y} = \mathbf{Z}\mathbf{W}^\top$ and then breaking inference into two distinct steps, violating the assumption of dependence between taxa that is at the heart of all PCMs [Felsenstein, 1985], the PLVM allows for measurement error on traits and performs joint inference and uncertainty quantification for the PGPR phylogeny-trait covariance function. While the inference scheme proposed by Hadjipantelis et al. [2013] is computationally inexpensive and has been shown to perform well for synthetic datasets, its disregard of the phylogeny when inferring the trait covariance structure does run the risk of identifying spurious correlations within the data. Furthermore, choosing the number of latent variables is based on heuristics, such as the proportion of variance explained by principal components. This problem is also addressed in this work, where Q is selected after

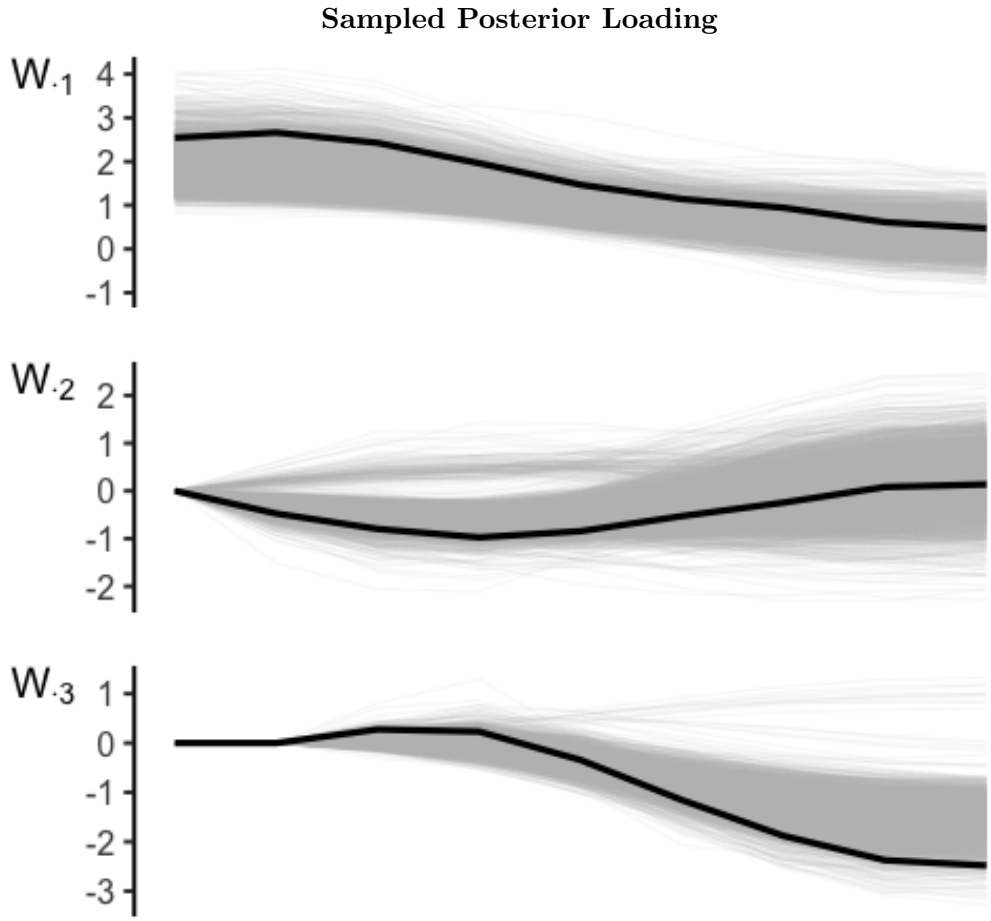


Figure 3.7: Samples from the posterior distribution of loading $\mathbf{W}$, mapped to a single mode for identifiability. Solid black lines represent the true loading, while each opaque grey line represents a sample.

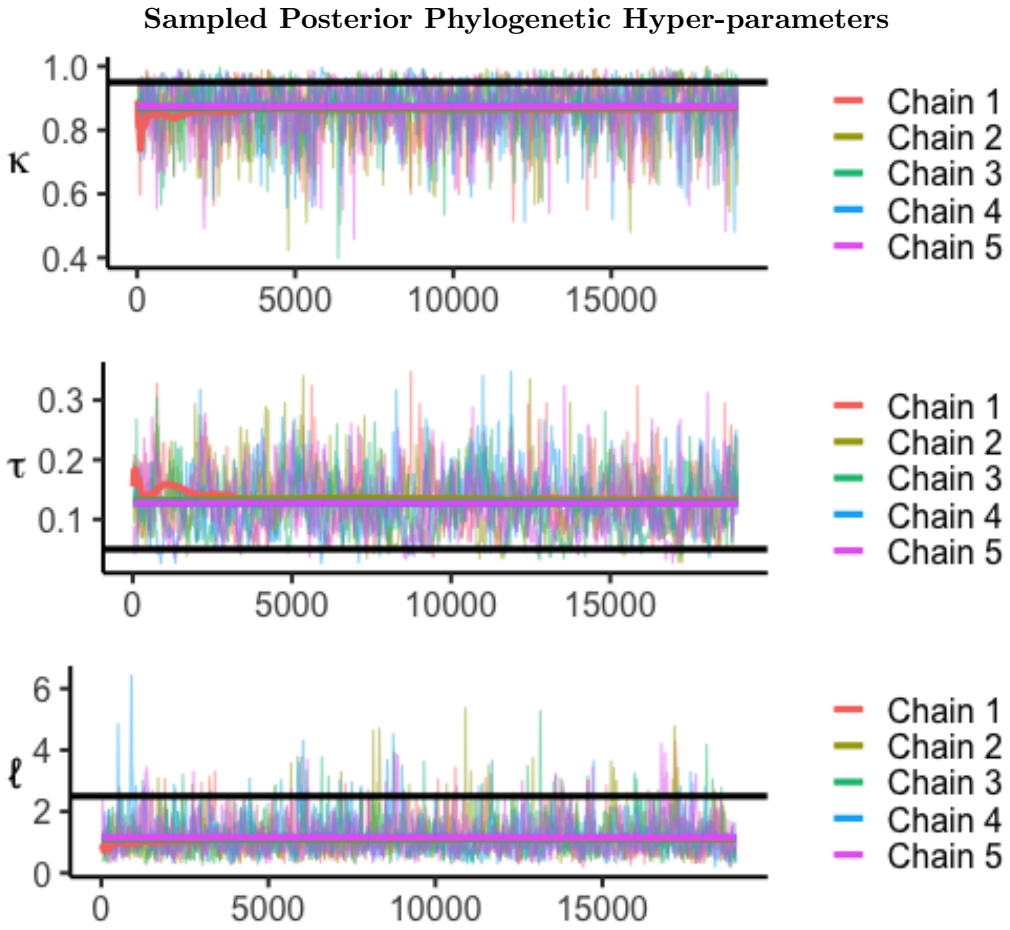


Figure 3.8: Trace plots of Markov chains sampling phylogenetic hyper-parameter posterior distributions. Solid black horizontal lines represent the true hyper-parameter value. MCMC chains converge to a single posterior mode, although the true hyper-parameter values lie in the tail of the posterior distribution.

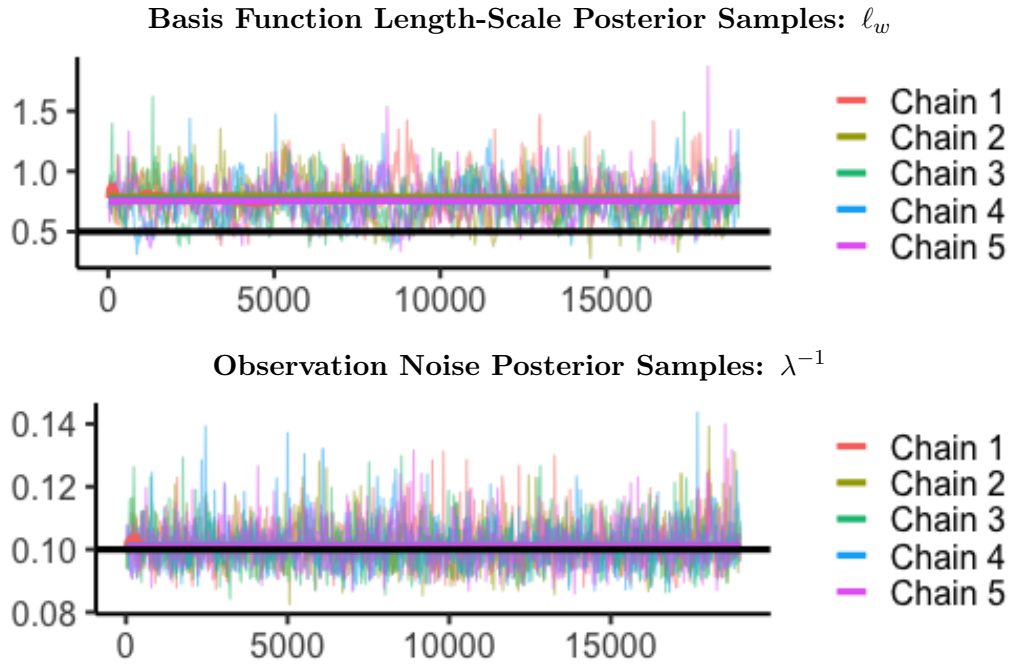


Figure 3.9: Trace plots of Markov chains sampling from the basis function length-scale and observation noise posterior distributions.

estimating the model evidence via Bridge sampling. Thus, the MCMC algorithm presented here offers a principled Bayesian approach to the ancestral reconstruction of FVTs which fits a probabilistic model for trait evolution to observed data.

It is worth noting that this work introduced a form of the Ornstein-Uhlenbeck (OU) phylogenetic covariance function which incorporates intra-taxon variation, offering a flexible approach to modelling repeated measurements of extant taxa. Doing so presented a number of challenges. In particular, it required the extension of efficient tree-traversal algorithms computing likelihood and conditional distributions for Brownian Motion over a phylogeny to the general Gauss-Markov case. This is a significant contribution in it's own right, offering a more flexible approach to modelling stochastic processes over a phylogeny. It also serves to highlight links between the PLVM and PCMs developed by Cybis et al. [2015] and Tolkoff et al. [2017], each of which rely on efficient tree-traversal algorithms for Bayesian inference. In the sense that they offer a model for the evolution of both discrete and continuous traits, these multivariate PCMs are more general than the PLVM presented here, however, neither method allows for intra-taxon variation, nor do they consider anything other than a BM model for trait evolution. In this respect, the PLVM generalises the models for trait evolution which underpin these PCMs.

MCMC inference for the PLVM has been shown to effectively reconstruct ancestral traits while providing uncertainty quantification for this reconstruction, the objective which motivated this work. As discussed above, this approach offers significant conceptual benefits over competing methods, however it does present some practical drawbacks. A first point to note is that care must be taken when drawing any conclusions from the absolute values of model parameters. While inference has been constrained to ensure MCMC chain convergence, scale invariance between the trait and phylogeny covariance functions remains problematic. This is due to well-known problems with the estimation of variance and length-scale in Matérn covariance functions [Zhang, 2004]. Thus, the PLVM is too flexible to make definitive statements about the absolute values of the model parameters and hyperparameters. Although this issue could be addressed by taking an approach similar to that of Tolkoff et al. [2017] and assuming a BM phylogenetic covariance function, this results in a more rigid model for trait evolution. This is undesirable when ancestral reconstruction is the objective.

A second, and altogether more serious problem with the method presented here is that MCMC inference scales with $\mathcal{O}\left(D^3\right)$, making the approach wholly impractical for high dimensional data. While it may be argued that this is simply the price that must be paid in order to model FVTs, there is no denying that this is a major limitation of the method. This is not an issue problem for Phylogenetic Factor Analysis [Tolkoff et al., 2017], for which inference scales with $\mathcal{O}\left(D\right)$. Within this framework, the trait space of a FVT could be sampled much more densely. That is to say, a much larger value for D may be chosen. It is also true that the PLVM is formulated for the evolution a single FVT over the fixed phylogeny $\mathcal{T}$. While incorporating a distribution over phylogenies would be relatively straightforward, such a distribution could be sampled within the MCMC inference, generalising the model to a collection of discrete and continuous valued traits is not so elementary. As discussed above, Phylogenetic Factor Analysis has been developed for collections of continuous and discrete traits, although it does not incorporate FVTs.

Thus, even though work presented in this chapter does address important issues, incorporating intra-taxon variation and scaling inference linearly with N, there do remain avenues for further development. One option would be to impose stronger constraints on the model, i.e. assume a BM model for trait evolution over the phylogeny, and link the result to well established methods for assessing phylogenetic signal, such as Pagel's λ [Pagel, 1999b], Blombergs's K [Blomberg et al., 2003], or the Phylogenetic Mixed Model [Housworth et al., 2004]. Alternatively, an even more flexible model could be developed, one which is unconcerned with

identifiability and instead focusses solely on ancestral reconstruction for some set of traits. It is the second option that will be tackled in the next chapter of this thesis. Although measures of phylogenetic signal and heritability are well understood and widely reported, their purpose is to describe knowledge of ancestral traits. Thus, developing a method which tackles this question directly may provide even more valuable insight. Furthermore, by addressing the issue of computational efficiency and generalising the model to collections of traits, it is hoped that an effective, practical method for ancestral reconstruction will be developed.

Chapter 4

A Generalised Phylogenetic Latent Variable Model

4.1 Introduction

In the context of evolutionary biology and phylogenetic comparative analysis, bat echolocation represents a particularly fascinating characteristic. The call production and signal processing system represents a particularly intricate natural phenomenon. While echolocation in bats is well-studied [Fenton et al., 2016], describing the developmental pathways leading to the diversity seen in the call structures of extant bats has proven challenging [Simmons and Stein, 1980; Schnitzler et al., 2004; Eick et al., 2005; Collen, 2012; Meagher et al., 2018a,b]. Echolocation calls are complex, multi-harmonic acoustic signals [Fenton et al., 2016], and obtaining a parsimonious representation of such objects is a challenging task in and of itself [Cohen, 1995; Oppenheim and Schafer, 2014]. Furthermore, there are over a thousand species of bat currently recognised [Simmons, 2005], making the ancestral reconstruction of bat echolocation calls, the objective of this thesis, a problem requiring the implementation of techniques for "*big data*". This chapter presents a novel method developed specifically for this task, making an important contribution to the field of evolutionary biology.

As has been discussed in earlier chapters, a bats echolocation call can be thought of as a Function-Valued Trait (FVT). As such, the PGPR framework proposed by Jones and Moriarty [2013] offers a suitable probabilistic model for its evolution. A Markov Chain Monte Carlo (MCMC) sampling scheme offering a Bayesian approach to the ancestral reconstruction of FVTs was proposed in Chapter 3. This method addressed many of the limitations associated with Phylogenetic

Comparative Methods (PCMs) proposed by Hadjipantelis et al. [2012], Cybis et al. [2015], and Tolkoff et al. [2017]. In particular, allowing joint inference over the full phylogeny-trait covariance function meant that trait evolution could be flexibly modelled as any Gauss-Markov process over a phylogeny. Despite this, the method suffers from limitations of its own.

The first of these issues was the computational expense associated with the inference scheme. MCMC methods are an inherently time-consuming approach to Bayesian inference, best suited to small, expensive datasets [Blei et al., 2017]. The algorithm presented in Chapter 3 was particularly problematic as, although it scaled linearly with the number of extant taxa, it scaled cubically with the number of measurements of each FVT. A second problem arose from the fact that constraints needed to be placed on the PLVM in order to ensure convergence of MCMC chains. This meant that latent variables over the phylogeny were assumed to be independent and identically distributed, imposing a certain rigidity on the model. This may not reflect reality, as it is entirely possible that some aspects of a FVT are strongly correlated over the phylogeny while others are not. The final, and possibly most pressing, limitation identified was the lack of generality for the PGPR framework, which has been formulated for evolutionary inference on a single FVT only [Jones and Moriarty, 2013; Hadjipantelis et al., 2013; Goolsby, 2015; Mariñas-Collado et al., 2019]. Thus, PGPR is ineffective if the characterisation of a trait requires a combination of discrete and continuous characters, or if correlation over a set of traits is to be explored, limiting its scope.

Each of these issues will be addressed in this chapter. Firstly, PGPR is generalised to include collections of discrete and continuous traits, such that ordinal, categorical, and continuous scalar-valued traits can all be modelled alongside FVTs. To this end, the approach proposed by Cybis et al. [2015] is implemented, which extends the probit likelihood of Albert and Chib [1993] to phylogenetic comparative analysis by augmenting manifest (observed) traits with a set of real-valued auxiliary variables. The evolution of these auxiliary variables over the phylogeny is then modelled as a PLVM, where the assumption of independent and identically distributed latent variables is relaxed such that latent variables are only assumed to be independent. This relaxation provides a more flexible model, one which can fit observed data closely. A Co-ordinate Ascent Variational Inference (CAVI) scheme allows efficient approximate Bayesian inference to be performed for the model.

Variational Inference (VI) describes a set of techniques for approximating intractable posterior distributions [Jordan et al., 1999; Bishop, 2006; Blei et al., 2017]. Rather than sampling from the distribution of interest, VI proposes a variational

family of distributions, governed by some variational parameters. The inference is then treated as an optimisation problem, whereby variational parameters are chosen to minimise the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] between the variational family and the true model posterior. This approach has been used to derive efficient approximate solutions to many Bayesian inference problems, including the classification of binary [Csató et al., 2000; Opper and Winther, 2000] and multinomial [Girolami and Rogers, 2006; Damoulas and Girolami, 2008] random variables, principal components and factor analysis [Bishop, 1999; Ghahramani and Beal, 2000], and Gaussian process latent variable models [Titsias and Lawrence, 2010]. More recently, VI has been applied to phylogenetics with both Dang and Kishino [2019] and Zhang and Matsen IV [2018] developing variational approaches to inferring phylogenies from molecular sequences.

This chapter presents the generalised Phylogenetic Latent Variable Model, a PCM for the ancestral reconstruction of collections of ordinal, categorical, continuous, or function-valued traits. Based on the threshold model for trait evolution [Wright, 1934; Felsenstein, 2011; Cybis et al., 2015; Tolkoff et al., 2017], the Phylogenetic Gaussian Process Regression (PGPR) framework is extended from a single FVTs to any collection of traits via the probit likelihood [Albert and Chib, 1993]. A CAVI scheme for approximate Bayesian inference is derived, the performance of which is assessed for a synthetic dataset, based on that presented by Hadjipantelis et al. [2013].

4.2 Methods

4.2.1 Data Augmentation

Consider a set of P discrete and continuous traits, observed for N related individuals, belonging to $S \leq N$ separate taxa. Discrete traits may be categorical- or ordinal-valued, while continuous traits are scalar- or function-valued. Given that each FVT is a multivariate object, let $\mathbf{Y}_{n\cdot} = (y_{n1}, \ldots, y_{nD})^\top$ for $D \geq P$ denote the *manifest traits* such that $\mathbf{Y} = (\mathbf{Y}_{1\cdot}, \ldots, \mathbf{Y}_{N\cdot})^\top$, where $\mathcal{O}_{\mathbf{Y}}$ is the set of ordinal trait indices for $\mathbf{Y}_{n\cdot}$, with $\mathcal{C}_{\mathbf{Y}}$ and $\mathcal{R}_{\mathbf{Y}}$ being analogously defined for categorical and continuous traits respectively. Furthermore, the shared ancestry between individuals is given by the phylogeny $\mathcal{T}$, which is known and of the form described in sub-section 3.2.1, although, for notational ease, it is assumed in this chapter that the taxon level phylogeny is a bifurcating tree such that $M = S - 1$.

As described by Albert and Chib [1993] and Cybis et al. [2015], assume that there exists a set of continuous random variables $\mathbf{X}_{n\cdot} = (x_{n1}, \ldots, x_{nD'})^\top$ for

$D' \geq D$ that govern the behaviour of $\mathbf{Y}_{n\cdot}$ via the deterministic mapping function $g : \mathbf{X}_{n\cdot} \rightarrow \mathbf{Y}_{n\cdot}$. In this case, manifest traits $\mathbf{Y}$ are a function of *auxiliary traits*, denoted $\mathbf{X} = (\mathbf{X}_{1\cdot}, \ldots, \mathbf{X}_{N\cdot})^\top$. When y_{ni} is an ordinal trait, which is to say that it takes one of K_i ordered values and $i \in \mathcal{O}_\mathbf{Y}$, then

$$y_{ni} = g\left(x_{ni'}\right) = k, \text{ if } \gamma_{i,k-1} \leq x_{ni'} < \gamma_{i,k},$$

is a one-to-one map from $x_{ni'}$ to y_{ni}, with i' indexing the auxiliary trait corresponding to the i^{th} manifest trait throughout this chapter, and $\boldsymbol{\gamma}_i = (\gamma_{i,0}, \ldots \gamma_{i,K_i})$ being the set of cut-off points where $\gamma_{i,0} = -\infty, \gamma_{i,1} = 0$, and $\gamma_{i,K_i} = \infty$.

If y_{ni} is a categorical trait, in that it falls into to one of K_i unordered states (categories), labelled $c_{i,k}$ for $k = 1, \ldots, K_i$, and $i \in \mathcal{C}_\mathbf{Y}$, then

$$\begin{aligned} y_{ni} &= g\left(x_{ni'}, \ldots, x_{n,i'+K_i-1}\right), \\ &= c_{i,k-1} \text{ if } x_{n,i'+k} = \sup\left(x_{ni'}, \ldots, x_{n,i'+K_i-1}\right), \end{aligned}$$

defines the K_i-to-one map where, without any loss of generality, $x_{n,i'+k} = 0$ when $y_{ni} = c_{i,k-1}$.

In order to complete the mapping, consider the continuous or function-valued manifest traits, that is, $\mathbf{Y}_{\cdot i}$ for all $i \in \mathcal{R}_\mathbf{Y}$. In this case, any monotonic function from $\mathbb{R}$ to the manifest traits will suffice. For example,

$$y_{ni} = g\left(x_{ni'}\right) = x_{ni'},$$

is appropriate for $y_{ni} \in \mathbb{R}$.

4.2.2 A Generalised Phylogenetic Latent Variable Model

In order to develop a generalised model for $\mathbf{Y}$, the auxiliary traits $\mathbf{X}$ are modelled as a PLVM. That is to say, $\mathbf{X}_{n\cdot} = \mathbf{W}\mathbf{Z}^\star_{n\cdot} + \boldsymbol{\epsilon}_{n\cdot}$, such that

$$p\left(\mathbf{X}_{n\cdot}|\mathbf{W}, \mathbf{Z}^\star, \boldsymbol{\Lambda}\right) = \mathcal{N}\left(\mathbf{X}_{n\cdot}|\mathbf{W}\mathbf{Z}^\star_{n\cdot}, \boldsymbol{\Lambda}^{-1}\right),$$

where $\mathbf{W} = (\mathbf{W}_{1\cdot}, \ldots \mathbf{W}_{D'\cdot})^\top$ is the $D' \times Q$ loading matrix, $\mathbf{Z}^\star_{n\cdot}$ is the Q-dimensional vector of factors associated with position $\mathbf{t}_n \in \mathcal{T}$ such that $\mathbf{Z}^\star = (\mathbf{Z}^\star_{1\cdot}, \ldots, \mathbf{Z}^\star_{N\cdot})^\top$, and $\boldsymbol{\epsilon}_{n\cdot} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right)$ is a D'-dimensional error vector with diagonal precision matrix $\boldsymbol{\Lambda}$, This allows the auxiliary likelihood to be defined as

$$p\left(\mathbf{Y}, \mathbf{X}|\boldsymbol{\gamma}, \mathbf{W}, \mathbf{Z}^\star, \boldsymbol{\Lambda}\right) = p\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma}\right) p\left(\mathbf{X}|\mathbf{W}, \mathbf{Z}^\star, \boldsymbol{\Lambda}\right),$$

$$= \prod_{n=1}^{N} \delta\left(\mathbf{Y}_{n\cdot} = g\left(\mathbf{X}_{n\cdot}\right)\right) \mathcal{N}\left(\mathbf{X}_{n\cdot} | \mathbf{W}\mathbf{Z}^{\star}_{n\cdot}, \mathbf{\Lambda}^{-1}\right), \tag{4.1}$$

where $\boldsymbol{\gamma}$ is the set of cut-off points associated with any ordinal traits and $\delta\left(\cdot\right)$ is an indicator function.

In order to fully specify the model, prior distributions for the model loading, factors, error precision, and cut-off points must be defined. Consider factors $\mathbf{Z}^{\star}_{n\cdot} = \left(z_1\left(\mathbf{t}_n\right), \ldots, z_Q\left(\mathbf{t}_n\right)\right)^{\top}$, which are modelled as independent, zero-mean Gauss-Markov processes over $\mathcal{T}$, such that

$$z_j\left(\mathbf{t}_n\right) \sim \mathcal{GP}\left(0, k_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_m | \kappa_j, \tau_j, \ell_j\right)\right), \tag{4.2}$$

where the phylogenetic covariance function is of the form

$$\begin{aligned} k_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_m | \kappa_j, \tau_j, \ell_j\right) = \left(1 - \tau_j\right) \Bigg(& \kappa_j \, \exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_m\right)}{\ell_j}\right) + \\ & \left(1 - \kappa_j\right) \, \delta\left(\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_m\right) = 0\right) \delta\left(n \leq N + S\right) \Bigg) + \\ & \tau_j \, \delta\left(n = m\right) \delta\left(n \leq N\right). \end{aligned} \tag{4.3}$$

As discussed in section 3.2.4, given that $\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_m\right)$ defines the patristic distance between $\mathbf{t}_n$ and $\mathbf{t}_m$ over $\mathcal{T}$, such a prior distribution assumes that factors are generated by an Ornstein-Uhlenbeck (OU) process over $\mathcal{T}$ with phylogenetic length-scale $\ell_j \in \mathbb{R}^{+}$, within a Phylogenetic Mixed Model with heritability $\kappa_j \in (0, 1)$ [Housworth et al., 2004], and intra-taxon variation $\tau_j \in (0, 1)$, for $j = 1, \ldots, Q$. Furthermore, this prior ensures that $k_{\mathcal{T}}\left(\mathbf{t}_n, \mathbf{t}_n | \kappa_j, \tau_j, \ell_j\right) = 1$ for all $n = 1, \ldots, N$ and $j = 1, \ldots, Q$, fixing the scale of factors, a modelling choice which will force loadings to account for the magnitude of $\mathbf{X}$, easing the interpretation of model parameters and hyperparameters.

Note that factors are also implied at internal nodes of $\mathcal{T}$ under this prior distribution for $\mathbf{Z}^{*}_{n\cdot}$, which is to say that there exists $\mathbf{Z}^{*}_{m\cdot} = \left(z_1\left(\mathbf{t}_{N+m}\right), \ldots, z_Q\left(\mathbf{t}_{N+m}\right)\right)^{\top}$ for $m = 1, \ldots, 2S - 1$, such that $\mathbf{Z}^{*} = \left(\mathbf{Z}^{;}_{1\cdot} \ldots, \mathbf{Z}^{*}_{2S-1,\cdot}\right)^{\top}$. Thus, given

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}^{\star} \\ \mathbf{Z}^{*} \end{bmatrix},$$

the matrix of factors at all nodes of $\mathcal{T}$, the Gauss-Markov structure of (4.2) allows

the definition of

$$p\left(\mathbf{Z}|\boldsymbol{\kappa},\boldsymbol{\tau},\boldsymbol{\ell}\right)=\prod_{j=1}^{Q}\prod_{n=1}^{N+2S-2}\mathcal{N}\left(\mathbf{Z}_{nj}|\phi_{n,j}\mathbf{Z}_{\mathrm{pa}(n),j},\eta_{n,j}\right)\mathcal{N}(\mathbf{Z}_{R,j}|0,\eta_{R,j})$$

where $\boldsymbol{\kappa}=(\kappa_1,\ldots,\kappa_Q)^\top$, $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_Q)^\top$, $\boldsymbol{\ell}=(\ell_1,\ldots,\ell_Q)^\top$, and

$$\begin{aligned}\phi_{n,j}&=k_{\mathcal{T}}\left(\mathbf{t}_n,\mathbf{t}_{\mathrm{pa}(n)}|\kappa_j,\tau_j,\ell_j\right)k_{\mathcal{T}}\left(\mathbf{t}_{\mathrm{pa}(n)},\mathbf{t}_{\mathrm{pa}(n)}|\kappa_j,\tau_j,\ell_j\right)^{-1},\\ \eta_{n,j}&=k_{\mathcal{T}}\left(\mathbf{t}_n,\mathbf{t}_n|\kappa_j,\tau_j,\ell_j\right)-k_{\mathcal{T}}\left(\mathbf{t}_n,\mathbf{t}_{\mathrm{pa}(n)}|\kappa_j,\tau_j,\ell_j\right)^2k_{\mathcal{T}}\left(\mathbf{t}_{\mathrm{pa}(n)},\mathbf{t}_{\mathrm{pa}(n)}|\kappa_j,\tau_j,\ell_j\right)^{-1},\end{aligned}$$

for $n=1,\ldots,N+2S-2$, with $\phi_{R,j}\equiv 1$ and $\eta_{R,j}\equiv k_j\left(\mathbf{t}_R,\mathbf{t}_R|\mathcal{T}\right)$, where $R\equiv N+2S-1$ denotes the root node of $\mathcal{T}$. The prior distribution for $\mathbf{Z}$ is completed by defining $p\left(\kappa_j\right)=Beta\left(\kappa_j|a_\kappa,b_\kappa\right)$, $p\left(\kappa_j\right)=Beta\left(\tau_j|a_\tau,b_\tau\right)$, and $p\left(\ell_j\right)=Gamma\left(\ell_j|2,1\right)$. The Beta prior is a natural choice for random variable defined on the unit interval, while the Gamma prior reflects an approximation of the assumption that the OU process over $\mathcal{T}$, which is scaled such that $\max\left\{\mathrm{d}_{\mathcal{T}}\left(\mathbf{t}_n,\mathbf{t}_R\right)\right\}_{n=1}^{N}=1$, is similar to Brownian Motion with unit variance over short time scales, as discussed in subsection 3.2.4.

An independent Gaussian prior is chosen for each column of the loading matrix, denoted $\mathbf{W}_{\cdot j}$, which is to say that

$$p\left(\mathbf{W}_{\cdot j}|\alpha_j\right)=\mathcal{N}\left(\mathbf{W}_{\cdot j}|\mathbf{0},\alpha_j^{-1}\mathbf{K}_{\mathbf{W}}\right),\tag{4.4}$$

for $j=1,\ldots,Q$, where α_j is an Automatic Relevance Determination (ARD) [Neal, 2012] hyper-parameter and $\mathbf{K}_{\mathbf{W}}$ is the prior loading covariance matrix. Letting $p\left(\alpha_j\right)=Gamma\left(\alpha_j|a_\alpha,b_\alpha\right)$, ARD hyper-parameters tune the prior distribution for each column such that it can flexibly adjust to the magnitude of $\mathbf{X}$. Furthermore, large values of α_j indicate that $\mathbf{W}_{\cdot j}$ is close $\mathbf{0}$, allowing unnecessary columns of $\mathbf{W}$ to be deflated away to irrelevance. This means that, when fitting the model to data, some large value for Q can be selected, with superfluous factors being effectively pruned away without any further user input.

The covariance matrix $\mathbf{K}_{\mathbf{W}}$ has a block diagonal structure, with non-zero off diagonal entries occurring only in those blocks corresponding to FVTs. As in sub-section 3.2.4, each FVT is assumed to be a twice mean square differentiable function observed over $\mathcal{X}=\mathbb{R}^d$ such that the corresponding block of $\mathbf{K}^{\mathbf{W}}$ is given

by the Gram matrix of the Matérn-$\frac{5}{2}$ kernel

$$k(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right), \tag{4.5}$$

where the length-scale ℓ has been fixed, and $r = |\mathbf{x} - \mathbf{x}'|$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. In this case, fixing ℓ is not a restrictive assumption, due to the flexibility provided by α_j, and doing so will simplify inference for the model significantly.

At this point it is worth defining $\mathcal{O}_{\mathbf{X}}$, $\mathcal{C}_{\mathbf{X}}$, and $\mathcal{R}_{\mathbf{X}}$, which index ordinal, categorical, and continuous trait columns of $\mathbf{X}$ respectively, and are analogous to $\mathcal{O}_{\mathbf{Y}}$, $\mathcal{C}_{\mathbf{Y}}$, and $\mathcal{R}_{\mathbf{Y}}$. Thus, following the approach of Albert and Chib [1993] a prior distribution for $\mathbf{\Lambda}$ is then given by fixing $\mathbf{\Lambda}_{i'} \equiv 1$ for $i' \in \{\mathcal{O}_{\mathbf{X}}, \mathcal{C}_{\mathbf{X}}\}$, ensuring identifiability in the model for discrete traits, and setting $p(\mathbf{\Lambda}_{i'}) = Gamma(\mathbf{\Lambda}_{i'} | a_{\mathbf{\Lambda}}, b_{\mathbf{\Lambda}})$ for $i' \in \mathcal{R}_{\mathbf{X}}$.

Finally, defining a uniform prior for free ordinal cut-off points, that is

$$\begin{aligned} p(\gamma_{i,k} | \gamma_{i,k-1}) &= \begin{cases} \frac{1}{b_\gamma - \gamma_{i,k-1}}, & \text{for } \gamma_{i,k} \in [\gamma_{i,k-1}, \gamma_{i,k-1} + b], \\ 0, & \text{otherwise,} \end{cases} \\ &= \mathcal{U}(\gamma_{i,k} | \gamma_{i,k-1}, \gamma_{i,k-1} + b_\gamma), \end{aligned} \tag{4.6}$$

for $i \in \mathcal{O}_{\mathbf{Y}}$ and $k \in \{2, \dots K_i - 1\}$, completes the model specification, a graphical representation of which is presented in Figure 4.1.

4.2.3 Approximate Posterior Inference

Fitting a generalised PLVM to some set of manifest traits $\mathbf{Y}$, given the phylogeny $\mathcal{T}$, involves learning about the posterior distribution over model parameters and hyper-parameters, which can be expressed as

$$\begin{aligned} &p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{\Lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell} | \mathbf{Y}) \\ &\propto p(\mathbf{Y}, \mathbf{X} | \mathbf{Z}, \mathbf{W}, \mathbf{\Lambda}, \boldsymbol{\gamma}) \, p(\mathbf{Z} | \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}) \, p(\boldsymbol{\kappa}) \, p(\boldsymbol{\tau}) \, p(\boldsymbol{\ell}) \, p(\mathbf{W} | \boldsymbol{\alpha}) \, p(\boldsymbol{\alpha}) \, p(\mathbf{\Lambda}) p(\boldsymbol{\gamma}), \end{aligned} \tag{4.7}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_Q)^\top$. Letting $\mathbf{\Psi} = \{\mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{\Lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}\}$, the model defined in sub-section 4.2.2 implies that this posterior is a multi-modal distribution. For any set of parameter values $\mathbf{\Psi}^*$, there exist $2 \times Q! - 1$ equivalent parametrisations due to permutations and reflections of $\mathbf{W}$ and $\mathbf{Z}$. Furthermore, the model's flexibility suggests that there are likely to be multiple local optima. Thus, Markov Chain Monte Carlo methods are unsuitable for inference on this object.

The Generalised Phylogenetic Latent Variable Model

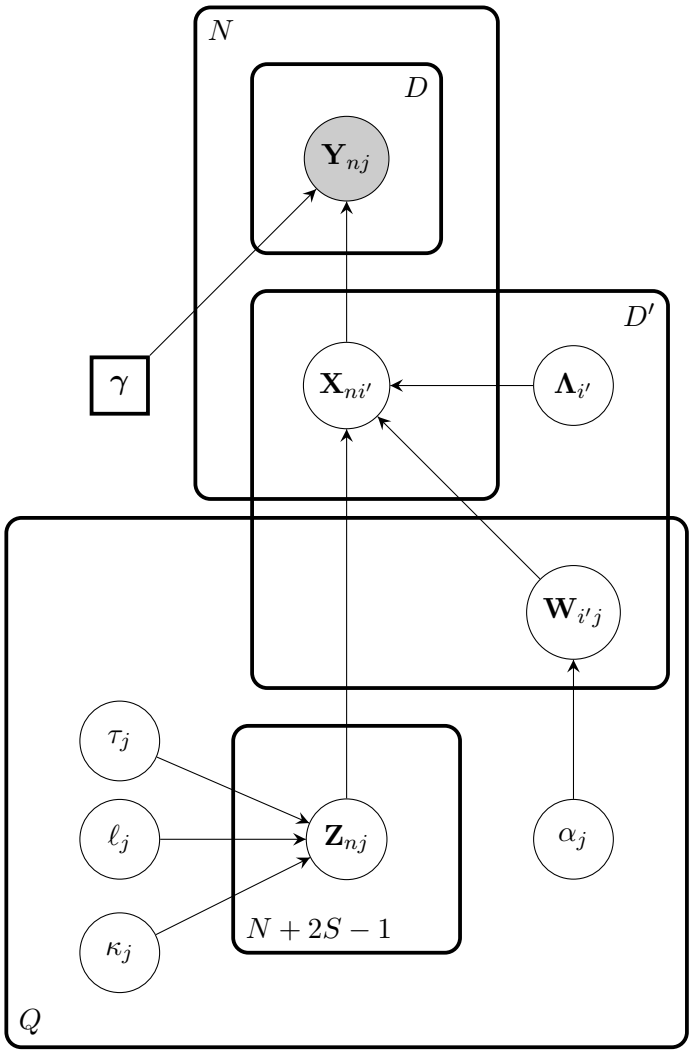


Figure 4.1: A graphical representation of the Generalised Phylogenetic Latent Variable Model presented in section 4.2.2. As with Figure 3.3, each circle represents a random variable, where those that are shaded grey have been observed. Boxes around circles are plates, denoting the number of existing random variables of that type. The box around $\boldsymbol{\gamma}$ represents the fact that there is a variable number of ordinal cut off parameters within the model, depending on the number of ordinal traits observed and the values each trait can take. Note also that the phylogeny $\mathcal{T}$, the space on which factors $\mathbf{Z}_{nj}$ are observed, has been omitted for clarity of presentation.

While this multi-modality is a cause for concern, the more constrained PLVM for FVTs in Chapter 3 was developed to address this. A latent variable model, such as Factor Analysis, is only ever picking one explanation for an observed dataset from an infinitum of possibilities, all of which effectively describe the same model for data [Lopes, 2014]. Thus, given that the objective of this analysis is an ancestral reconstruction for some collection of traits, these concerns are set aside while some "optimal" expression for (4.7) is found. One approach to problems of this nature, popular in Machine Learning, is Variational Inference (VI) [Jordan et al., 1999; Bishop, 2006; Blei et al., 2017].

Given the variational family of distributions, denoted $\mathcal{Q}$, parametrised by $\boldsymbol{\xi}$, VI approximates the posterior distribution by finding

$$q^{*}\left(\boldsymbol{\Psi}\right) = \arg\min_{q(\boldsymbol{\Psi})\in\mathcal{Q}} \mathrm{KL}\left(q\left(\boldsymbol{\Psi}\right) || \, p\left(\boldsymbol{\Psi}|\mathbf{Y}\right)\right), \tag{4.8}$$

where $\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951]. The approximate posterior $q^{*}\left(\boldsymbol{\Psi}\right)$ is optimal, not only in the sense that it minimises the KL divergence, but also in that it is the best approximation within $\mathcal{Q}$ to the Bayesian posterior [Knoblauch et al., 2019].

From the definition of KL divergence

$$\begin{aligned}\mathrm{KL}\left(q\left(\boldsymbol{\Psi}\right) || \, p\left(\boldsymbol{\Psi}|\mathbf{Y}\right)\right) &= \mathbb{E}_{q}\left[\log q\left(\boldsymbol{\Psi}\right)\right] - \mathbb{E}_{q}\left[\log p\left(\boldsymbol{\Psi}|\mathbf{Y}\right)\right], \\ &= \mathbb{E}_{q}\left[\log q\left(\boldsymbol{\Psi}\right)\right] - \mathbb{E}_{q}\left[\log p\left(\boldsymbol{\Psi},\mathbf{Y}\right)\right] + \log p\left(\mathbf{Y}\right), \end{aligned} \tag{4.9}$$

where $\mathbb{E}_{q}[\cdot]$ denotes the expectation with respect to $q\left(\boldsymbol{\Psi}\right)$. This demonstrates the dependence of the KL divergence in (4.8) on $p\left(\mathbf{Y}\right)$, the model evidence, a quantity which cannot be computed, but is constant with respect to $q\left(\cdot\right)$. However, as $\mathrm{KL}(\cdot||\cdot) \geq 0$ by definition, an equivalent objective function can be defined, that is

$$\begin{aligned}\mathrm{ELBO}\left(q\right) &= \mathbb{E}_{q}\left[\log p\left(\boldsymbol{\Psi},\mathbf{Y}\right)\right] - \mathbb{E}_{q}\left[\log q\left(\boldsymbol{\Psi}\right)\right], \\ &= \mathbb{E}_{q}\left[\log p\left(\mathbf{Y}|\boldsymbol{\Psi}\right)\right] + \mathbb{E}_{q}\left[\log p\left(\boldsymbol{\Psi}\right)\right] - \mathbb{E}_{q}\left[\log q\left(\boldsymbol{\Psi}\right)\right], \end{aligned} \tag{4.10}$$

the log Evidence Lower Bound (ELBO), which, as the name suggests, bounds the logarithm of the model evidence from below.

The approach to VI taken here is to assume a mean-field variational family for $\mathcal{Q}$, and implement a Co-ordinate Ascent Variational Inference (CAVI) algorithm which maximises the ELBO in (4.10) [Bishop, 2006; Blei et al., 2017]. That is to

say

$$q(\boldsymbol{\Psi}) = \prod_{i=1}^{M} q_i(\boldsymbol{\Psi}_i),$$

where each set within a partition of $\boldsymbol{\Psi}$ has its own variational factor $q_i(\boldsymbol{\Psi}_i)$. In this case

$$q_i^*(\boldsymbol{\Psi}_i) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\Psi}/\boldsymbol{\Psi}_i)}\left[\log p(\boldsymbol{\Psi}, \mathbf{Y})\right]\right), \tag{4.11}$$

which is to say that the optimal distribution over $\boldsymbol{\Psi}_i$ within the mean-field variational family is proportional to the exponentiated expectation of the log joint distribution over traits, model parameters, and model hyper-parameters, where every other variational factor has been held fixed. By iteratively finding $q_i^*(\boldsymbol{\Psi}_i)$ for each variational factor, where ELBO (q) increases at every iteration, the CAVI algorithm finds a locally optimal solution for (4.8).

The mean-field variational family approximating (4.7) is given by

$$\begin{aligned} q(\boldsymbol{\Psi}) = q(\mathbf{X})\, q(\boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \\ \left(\prod_{i'=1}^{D'} q(\mathbf{W}_{i'\cdot})\right)\left(\prod_{n=1}^{N+2S-1} q(\mathbf{Z}_{n\cdot})\right)\left(\prod_{i\in\mathcal{O}_{\mathbf{Y}}}\prod_{k=2}^{K_i-1} q(\gamma_{i,k})\right), \end{aligned} \tag{4.12}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}\}$, and subscripts on variational factors have been suppressed for clarity of exposition. Here, the approximate posterior distribution factorises over the auxiliary traits, free auxiliary precision parameters and model hyper-parameters, the factors at each node of $\mathcal{T}$, the loading matrix's rows, and each of the ordinal trait cut-off points.

Deriving variational parameters for each variational factor is a somewhat involved process, and as such is relegated to Appendix B.1, however, the resulting approximate posterior distribution is presented here. Employing the notation $\langle\boldsymbol{\Psi}_i\rangle \equiv \mathbb{E}_{q(\boldsymbol{\Psi})}[\boldsymbol{\Psi}_i]$ it is shown that

$$q^*(\mathbf{W}_{i'\cdot}) = \mathcal{N}\left(\mathbf{W}_{i'\cdot}|\langle\mathbf{W}_{i'\cdot}\rangle, \mathbf{S}_{i'}^{\mathbf{W}}\right),$$

and

$$q^*(\mathbf{Z}_{n\cdot}) = \mathcal{N}\left(\mathbf{Z}_{n\cdot}|\langle\mathbf{Z}_{n\cdot}\rangle, \mathbf{S}_{n}^{\mathbf{Z}}\right),$$

where the variational means and covariances are defined by Equations (B.6), (B.7), (B.9), and (B.10) of Appendix B.1.

Interaction between the variational family and the true model posterior in-

duces further factorisation such that

$$q\left(\boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) = \prod_{i' \in \mathcal{R}_{\mathbf{X}}} q\left(\boldsymbol{\Lambda}_{i'}\right) \prod_{j=1}^{Q} q\left(\alpha_j\right) q\left(\theta_j\right),$$

where $\theta_j = \{\kappa_j, \tau_j, \ell_j\}$. It is then shown that

$$q^*\left(\boldsymbol{\Lambda}_{i'}\right) = Gamma\left(\boldsymbol{\Lambda}_{i'} | \tilde{a}_{\boldsymbol{\Lambda}_{i'}}, \tilde{b}_{\boldsymbol{\Lambda}_{i'}}\right)$$

with shape $\tilde{a}_{\boldsymbol{\Lambda}}^{i'}$ and rate $\tilde{b}_{\boldsymbol{\Lambda}}^{i'}$ defined by (B.14) and (B.14) respectively, and

$$q^*\left(\alpha_j\right) = Gamma\left(\alpha_j | \tilde{a}_{\alpha}^{j}, \tilde{b}_{\alpha}^{j}\right)$$

with $\tilde{a}_{\alpha}^{j}$ and $\tilde{b}_{\alpha}^{j}$ defined by (B.18) and (B.19).

A problem arises for $q^*\left(\theta_j\right)$, in that no closed form solution for this variational factor exists. This could be addressed by drawing a Monte Carlo sample from the optimal mean-field variational family distribution, presented up to a normalising constant in (B.23), and then estimating the required expectations. This approach is computationally expensive, however, particularly when CAVI requires a large number of iterations to converge. Instead, given that θ_j are hyper-parameters for the phylogenetic Gaussian process prior over factors $\mathbf{Z}$, it is deemed appropriate to simply optimise ELBO (q) with respect to θ_j. This is equivalent to setting $q\left(\theta_j\right) = \delta\left(\theta_j = \langle\theta_j\rangle\right)$, and offers a computationally efficient approach, which does not require any difficult to compute expectations, as could be the case if some other parametric form was chosen for the variational factor $q\left(\theta_j\right)$. Furthermore, given that no closed form solution exists for the variational factors of the free ordinal trait cut off points, the same arguments apply in setting $q\left(\gamma_{i,l}\right) = \delta\left(\gamma_{i,l} = \langle\gamma_{i,l}\rangle\right)$.

The final set of variational factors to be considered are those for auxiliary traits, also subject to an induced factorisation, such that

$$q\left(\mathbf{X}\right) = \prod_{n=1}^{N} \prod_{i'=1}^{D'} q\left(\mathbf{X}_{ni'}\right).$$

Three separate cases must be considered. The first is for continuous and function-valued traits, that is when $i' \in \mathcal{R}_{\mathbf{X}}$, where the optimal approximate posterior is simply

$$q^*\left(\mathbf{X}_{ni'}\right) = \mathcal{N}\left(\mathbf{X}_{ni'} | \langle\mathbf{W}_{i'.}\rangle^{\top} \langle\mathbf{Z}_{n.}\rangle, \langle\boldsymbol{\Lambda}_{i'}\rangle\right).$$

Secondly, ordinal traits, for which $i' \in \mathcal{O}_{\mathbf{X}}$, imply that

$$q^*\left(\mathbf{X}_{ni'}|\mathbf{Y}_{ni}=k\right)=\mathcal{TN}\left(\mathbf{X}_{ni'}|\langle\mathbf{W}_{i'\cdot}\rangle^\top\langle\mathbf{Z}_{n\cdot}\rangle,1,\langle\gamma_{i,k-1}\rangle,\langle\gamma_{i,k}\rangle\right),$$

which is to say that $\mathbf{X}_{ni'}$ follows a truncated Gaussian distribution with unit variance bounded below by $\langle\gamma_{i,k-1}\rangle$ and above by $\langle\gamma_{i,k}\rangle$. Finally, for categorical traits, that is when $i \in \mathcal{C}_{\mathbf{Y}}$ and $\mathbf{Y}_{ni}=c_{i,k}$

$$q^*\left(\mathbf{X}_{n,i'+k-1}\right)=\delta\left(\mathbf{X}_{n,i'+k-1}=0\right)$$

and the optimal approximate posterior for the remaining auxiliary traits associated with $\mathbf{Y}_{ni}$ can be expressed as

$$\prod_{j\neq k}q^*\left(\mathbf{X}_{n,i'+j-1}|\mathbf{Y}_{ni}=c_{i,k}\right)=\prod_{j\neq k}\mathcal{TN}\left(\mathbf{X}_{n,i'+j-1}|\langle\mathbf{W}_{i'+j-1,\cdot}\rangle^\top\langle\mathbf{Z}_{n\cdot}\rangle,1,-\infty,0\right).$$

Iteratively updating the variational parameters for each of these variational factors in CAVI will then optimise $\mathrm{ELBO}\left(q\right)$ defined in (4.10), the derivation of which is presented in Appendix B.2.

4.2.4 Ancestral Reconstruction

For the generalised PLVM presented above, the problem of ancestral reconstruction is equivalent to finding the predictive distribution at some new position on the phylogeny $\mathbf{t}_* \in \mathcal{T}$. A variational approximation to this predictive distribution is given by

$$\begin{aligned}p\left(\mathbf{Y}_{*\cdot}|\mathbf{t}_*,\mathbf{Y}\right)&=\int p\left(\mathbf{Y}_{*\cdot}|\boldsymbol{\Psi},\mathbf{t}_*,\mathbf{Y}\right)p\left(\boldsymbol{\Psi}|\mathbf{Y}\right)d\boldsymbol{\Psi}\\&\approx\int p\left(\mathbf{Y}_{*\cdot}|\boldsymbol{\Psi},\mathbf{t}_*,\mathbf{Y}\right)q^*\left(\boldsymbol{\Psi}|\mathbf{Y}\right)d\boldsymbol{\Psi},\end{aligned}$$

however, integrating over the variational distributions of both the loading matrix and phylogenetic factors is an intractable problem. While this integral could be evaluated by Monte Carlo simulation, a more appealing approach is to simply obtain an approximate predictive distribution by integrating over the phylogenetic factors and auxiliary traits only, which yields

$$\begin{aligned}&p\left(\mathbf{Y}_{*\cdot}|\mathbf{t}_*,\mathbf{Y},\mathbf{W},\boldsymbol{\Lambda},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\kappa},\boldsymbol{\tau},\boldsymbol{\ell}\right)\\&\approx\int\delta\left(g\left(\mathbf{X}_{*\cdot}\right)=\mathbf{Y}_{*\cdot}\right)\mathcal{N}\left(\mathbf{X}_{*\cdot}|\langle\mathbf{W}\rangle\langle\mathbf{Z}_{*\cdot}\rangle,\langle\Lambda\rangle^{-1}+\langle\mathbf{W}\rangle\mathbf{S}_*^{\mathbf{Z}}\langle\mathbf{W}\rangle^\top\right)d\mathbf{X}_{*\cdot}.\end{aligned}\tag{4.13}$$

Consider the marginal predictive distribution at internal nodes of $\mathcal{T}$ for each type of manifest trait in turn, detailed derivations for which are included in Appendix B.3. It is can be shown that, for $i \in \mathcal{O}_{\mathbf{Y}}$

$$\begin{aligned} &p\left(\mathbf{Y}_{*i} = k | \mathbf{t}_*, \mathbf{Y}, \mathbf{W}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}\right) \\ &\approx F_{\mathcal{N}}\left(\frac{\langle \gamma_{i,k} \rangle - \langle \mathbf{W}_{i'\cdot} \rangle \langle \mathbf{Z}_{*\cdot} \rangle}{\nu^*_{i'}}\right) - F_{\mathcal{N}}\left(\frac{\langle \gamma_{i,k-1} \rangle - \langle \mathbf{W}_{i'\cdot} \rangle \langle \mathbf{Z}_{*\cdot} \rangle}{\nu^*_{i'}}\right), \end{aligned}$$

where $F_{\mathcal{N}}(\cdot)$ denotes a standard normal cumulative density function and $\nu^*_{i'} = \sqrt{1 + \langle \mathbf{W}_{i'\cdot} \rangle^\top \mathbf{S}^{\mathbf{Z}}_* \langle \mathbf{W}_{i'\cdot} \rangle}$. For $i \in \mathcal{C}_{\mathbf{Y}}$, a generalisation of the multinomial probit regression predictive distribution derived by Girolami and Rogers [2006] yields

$$p\left(\mathbf{Y}_{*i} = c_{i,k} | \mathbf{t}_*, \mathbf{Y}, \mathbf{W}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}\right) \approx \mathbb{E}_{p(u)}\left[\prod_{l=1}^{K_i - 1} F_{\mathcal{N}}\left(\mathbf{u}_l^{ki*}\right)\right],$$

where $\mathbf{u}^{ki*}$ is a function of $u \sim \mathcal{N}(0,1)$ and the variational parameters.

Finally, for $i \in \mathcal{R}_{\mathbf{Y}}$ the predictive distribution is given by

$$\begin{aligned} &p\left(\mathbf{Y}_{*i} | \mathbf{t}_*, \mathbf{Y}, \mathbf{W}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\ell}\right) \\ &\approx \mathcal{N}\left(\mathbf{Y}_{*i} | \langle \mathbf{W}_{i'\cdot} \rangle \langle \mathbf{Z}_{*\cdot} \rangle, \langle \Lambda_{i'} \rangle^{-1} + \langle \mathbf{W}_{i'\cdot} \rangle^\top \mathbf{S}^{\mathbf{Z}}_* \langle \mathbf{W}_{i'\cdot} \rangle\right), \end{aligned}$$

completing the ancestral reconstruction.

It is worth noting that manifest traits at internal nodes of $\mathcal{T}$ have been denoted $\mathbf{Y}_{*\cdot}$ rather than $\mathbf{f}_*$ as in sub-section 3.2.6. This is because, although intra-taxon variation and, in the case of ancestral nodes, non-phylogenetic noise effects on the factors have been stripped away, the observation noise has been included in the predictive distribution presented here, which was not the case for (3.22).

4.3 Results for a Synthetic Example

Performance of the CAVI algorithm for ancestral reconstruction of traits modelled with a generalised PLVM is investigated for a synthetic dataset, based on that studied by Hadjipantelis et al. [2013]. Given the phylogeny with $S = 128$ terminal nodes presented in Figure 4.2a, three observations are made for each extant taxon such that $N = 384$ and the full phylogeny with N terminal nodes is denoted $\mathcal{T}$.

A collection of $P = 4$ traits are considered in this analysis, an ordinal trait with three ordered categories labelled as $\{1, 2, 3\}$, a categorical trait made up of three unordered categories labelled $\{0, 1, 2\}$, a continuous trait, and a function valued trait

Q	ℓ	κ	τ
1	2.50	0.850	0.050
2	1.00	0.100	0.050
3	1.75	0.500	0.010
4	2.00	0.950	0.025

Table 4.1: Phylogenetic hyper-parameter values for independent Ornstein-Uhlenbeck processes over $\mathcal{T}$.

which has been observed at 32 points spread uniformly over the interval $[0, 1]$.

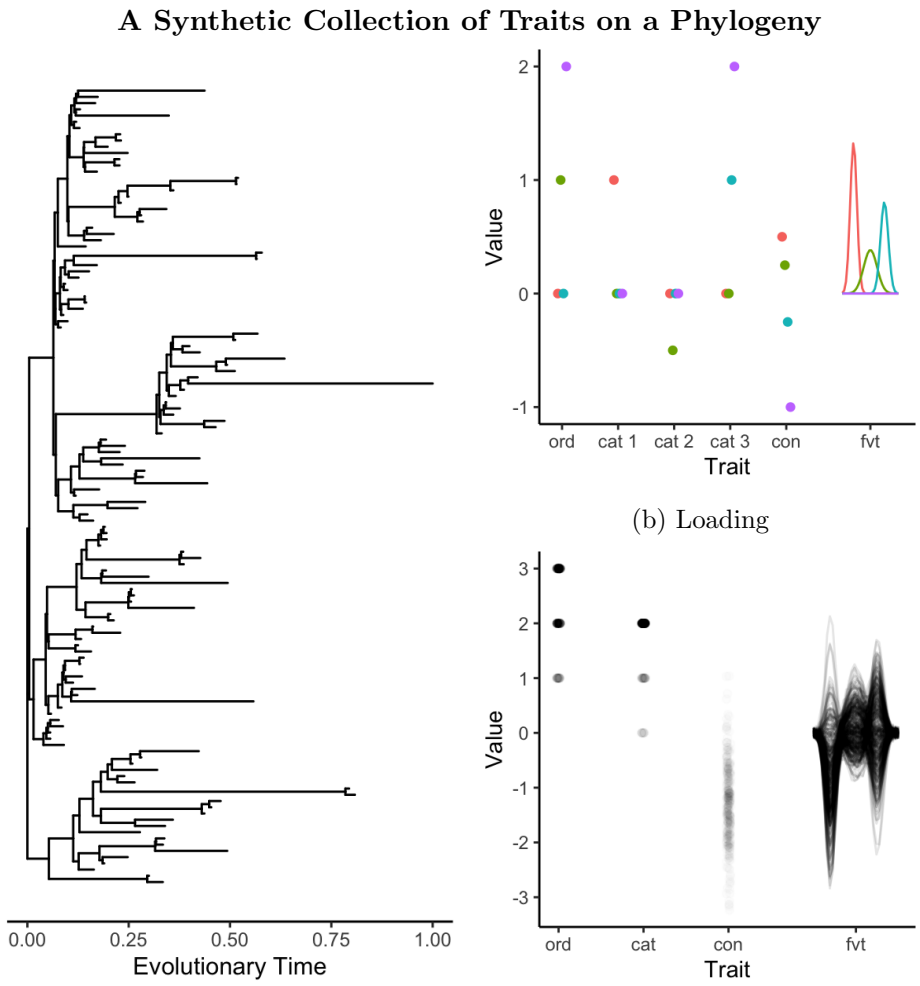
The evolution of these traits over the phylogeny is driven by $Q = 4$ independent OU processes, as defined by (4.2) and (4.3). That is to say, trait evolution over $\mathcal{T}$ is driven by four independent factors. Hyper-parameters governing each of these processes are presented in Table 4.1. To gain some intuition on an interpretation of hyper-parameter values, consider the first independent factor in some detail. In this case, $\ell_1 = 2.5$ indicates that on short time scales the OU process is more slowly varying than a Brownian Motion with unit variance over the same interval, implying a strong phylogenetic signal for this factor. The heritability $\kappa_1 = 0.85$ can be interpreted as saying 85% of inter-taxon variation is due to the phylogeny as opposed to independent environmental effects [Housworth et al., 2004]. Finally, $\tau_1 = 0.05$ is the within-taxon variation and indicates that variability of factors within each taxon is low.

Factors are mapped to auxiliary traits given the loading matrix $\mathbf{W}$ and diagonal trait precision matrix $\mathbf{\Lambda}$, where the precision for discrete traits is always fixed to 1, and for continuous and function valued traits is set to 10 and 500 respectively. Mapping auxiliary to manifest traits requires the definition of ordinal trait cut-off points, where $\gamma = (-\infty, 0, 2, \infty)$. The phylogeny, loadings, and manifest traits are all presented in Figure 4.2.

4.3.1 Model Fitting

In order to fit the model, a length-scale for the Matérn-$\frac{5}{2}$ prior covariance function on the FVT loading, that is, ℓ in (4.5), must be set. Here, ℓ is chosen by Type-II Maximum Likelihood estimation [Rasmussen and Williams, 2006], under the assumption that manifest FVTs are themselves independent and identically distributed Matérn-$\frac{5}{2}$ processes with unit variance, subject to some observation noise, allowing the choice of ℓ to be informed by the data. Thus, the value $\ell = 0.13$ is set prior to fitting a generalised PLVM for $\mathbf{Y}$.

The second choice to be made when fitting a generalised PLVM is the number



(a) Phylogenetic Tree

(b) Loading

(c) Manifest Traits

Figure 4.2: The phylogeny of evolutionary relationships between taxa for the simulated data (a), along with the loading mapping factors to the auxiliary traits (b) and the set of synthetic manifest traits (c), such that each trait is represented by opaque points and lines.

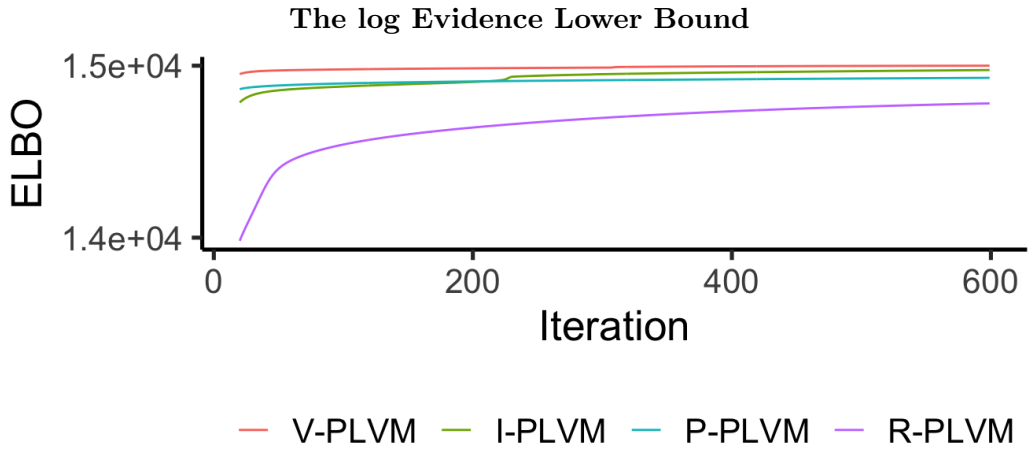


Figure 4.3: Variational Inference from four different initialisation values of $\mathbf{W}$, where the first 20 iterations in each case have been omitted for clarity. The Varimax initialisation (V-PLVM) performs uniformly better than any other candidate model maximises the ELBO, converging after approximately 600 iterations.

of latent factors, Q. One option is to simply choose some value for Q that is greater than the anticipated number of factors required and trust that the ARD precision parameters $\boldsymbol{\alpha}$ and phylogenetic length-scales $\boldsymbol{\ell}$ will effectively prune away superfluous factors without overfitting the data. An alternative approach, the one taken here, is first to perform Principal Components Analysis (PCA) [Tipping and Bishop, 1999] for a random sample of auxiliary traits given $\mathbf{Y}$ and randomly selected ordinal cut off points. Given the results of this analysis, Q can be selected such that the first Q principal components capture some proportion of variation in the data. For the synthetic dataset, 90% of the variation in a random set of auxiliary traits is explained by the first five principal components.

Fixing $Q = 5$, the generalised PLVM is fitted to the data for multiple initialisation values, with the model that maximises ELBO (q) being selected for ancestral reconstruction. Four strategies for initialising the CAVI algorithm are considered here. The first strategy is simply to initialise inference at random, producing a model which is referred to as R-PLVM. The three alternative strategies are very closely related and they are: initialising inference at the first Q principal components (P-PLVM); the Varimax rotation of the first Q principal components (V-PLVM) [Kaiser, 1958]; and initialising with Q independent components [Blaschke and Wiskott, 2002]. In each case, CAVI is said to have converged when ELBO (q) increases by less than 10^{-2} from one iteration to the next, or after 1000 iterations have been completed.

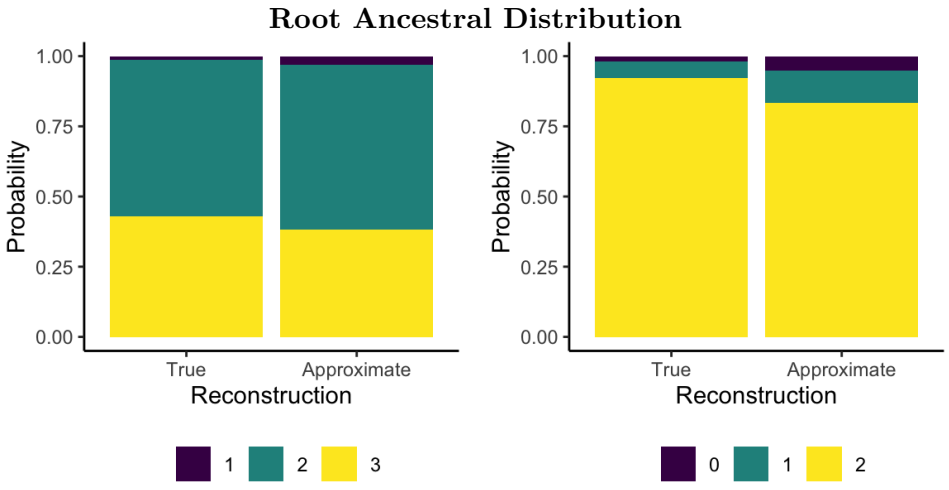
On inspection of Figure 4.3, it can be seen that initialisations informed by an exploratory analysis of the data resulted in uniformly better performance than R-PLVM. The V-PLVM converged in approximately 600 iterations[1], at which point it was found to maximise ELBO (q), even after allowing all other candidate models to run for the full 1000 iterations. Given that it is the best performing model, the V-PLVM is considered for further analysis.

4.3.2 Ancestral Reconstruction

Selecting the V-PLVM as a model for trait evolution allows a distribution for traits at each internal node of $\mathcal{T}$ to be defined. For those nodes that are parents of terminal nodes, that is $\mathbf{t}_n \in \mathcal{T}$ for $n \in \{N+1, \ldots, N+S\}$, this is the extant taxon trait distribution, while at all other internal nodes, i.e. $\mathbf{t}_n \in \mathcal{T}$ for $n \in \{N+S+1, \ldots, N+2S-1\}$, this distribution yields an ancestral reconstruction.

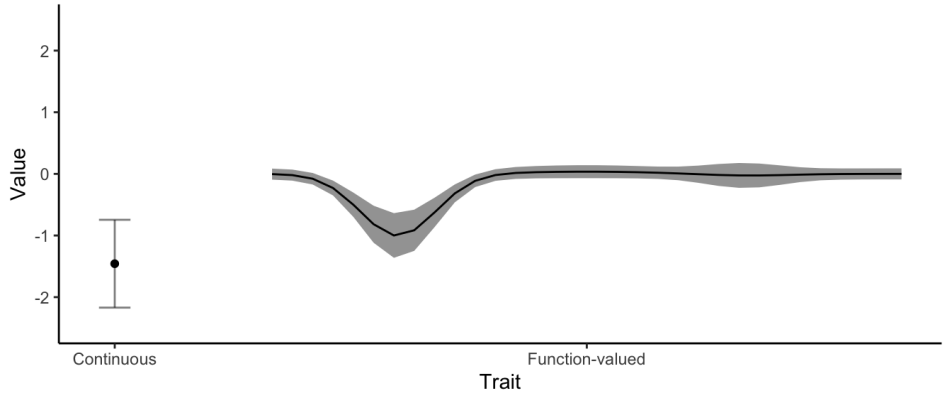
In order to investigate the performance of the V-PLVM for ancestral reconstruction, consider the trait distribution at $\mathbf{t}_R \in \mathcal{T}$, that is the root of $\mathcal{T}$. The ancestral reconstruction, as defined by (4.13), is presented in Figure 4.4. An examination of this figure reveals that the inferred ancestral distribution captures the salient features of the true ancestral distribution. The probabilities for discrete traits are well approximated and distributions for continuous traits closely matched. For the ordinal trait, presented in Figure 4.4a, both ancestral distributions have state 2 as most probable, followed closely by state 3, with state 1 being very improbable. The categorical trait (Figure 4.4b) has state 2 as the overwhelmingly most probable, with state 1 being slightly less improbable than state 0 in both cases. This represents a remarkably faithful ancestral reconstruction by the V-PLVM, particularly as it is based only on knowledge of the phylogeny $\mathcal{T}$ and manifest traits $\mathbf{Y}$, while the true distribution is based on full knowledge of the model including factors at terminal nodes of $\mathcal{T}$. The means and standard deviations are nearly identical in both ancestral distributions for the continuous trait, while for the FVT, the trait mean function is reproduced faithfully, including a negative bump early in the interval, and only slight differences exist between regions of high density over the interval. These differences can be attributed to uncertainty in the inferred factors, which may inflate the variance of ancestral reconstruction, as compared to the case when factors at the terminal nodes are known. Thus, using the V-PLVM for ancestral reconstruction provides a very satisfactory result.

[1]This took approximately 20 minutes on a Mac Book Pro with a 2.3 GHz Intel Core i5 processor, and an 8 GB 2133 MHz LPDDR3 memory

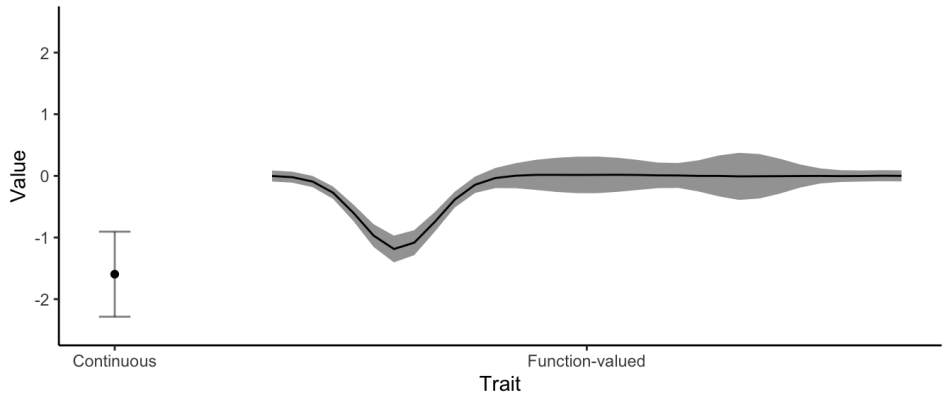


(a) Ordinal Trait

(b) Categorical Trait



(c) Ancestral Distribution



(d) Approximate Ancestral Distribution

Figure 4.4: A comparison of the true ancestral distribution at the root of $\mathcal{T}$, with approximate ancestral distribution given by V-PLVM. In (a) and (b) each colour in the bars represent the probability that the trait was of that particular state, while in (d) and (c), grey error markers represent two standard deviations from the mean.

4.3.3 Parameter Inference

While ancestral reconstruction is the primary objective of the generalised PLVM, a task for which it has been demonstrated to perform well, the accuracy of approximate inference for model parameters and hyper-parameters, in particular, the loading $\mathbf{W}$ and phylogenetic hyper-parameters θ, are also of interest.

Figure 4.5 presents a comparison of the true loading $\mathbf{W}$ and posterior expected loading $\langle \mathbf{W} \rangle$, which have been reordered and reflected to match $\mathbf{W}$. It can be seen that the shape of $\langle \mathbf{W} \rangle$ matches the truth, indicating that the correlation structure between traits has been modelled faithfully. $\mathbf{W}$ does not always lie within the region of high density around $\langle \mathbf{W} \rangle$ however, and while it is well known that variational inference underestimates the posterior distribution of correlated variables [Bishop, 2006], affecting $\langle \mathbf{W} \rangle$ with respect to the FVT in particular, this may also be attributed to scale invariance in the model, as discussed in section 3.3. Furthermore, the large bias in the categorical weights for the fourth loading can be explained by the link function defined in sub-section 4.2.1, where $\mathbf{X}_{n,i'+k} = 0$ when $\mathbf{Y}_{ni} = c_{i,k-1}$. This causes the large positive values in this loading to be shifted towards zero. The fifth loading demonstrates the effect of including superfluous latent factors in the inference and is close to $\mathbf{0}$, as desired. The results represent remarkably good performance, given the flexibility of the model.

Within the CAVI algorithm, $\langle \boldsymbol{\theta} \rangle$ is updated via optimisation of ELBO (q), i.e. assuming that $q(\theta_j) = \delta(\theta_j = \langle \theta_j \rangle)$, which does not provide any uncertainty quantification for the phylogenetic hyper-parameters. An expression for the mean-field variational family approximation to the posterior distribution of $\boldsymbol{\theta}$, up to a normalising constant, has been derived and is presented in Appendix B.2. Sampling from this distribution after convergence of the CAVI algorithm allows some insight into the uncertainty on $\langle \boldsymbol{\theta} \rangle$. For the V-PLVM, this sample is presented in Figure 4.6.

Consider the phylogenetic length-scale $\boldsymbol{\ell}$. At first glance, it appears to be poorly estimated, in no case is the true value within the region of high posterior density. On closer inspection, however, factors 2 and 3 have low heritability, and so inference for ℓ_j is somewhat irrelevant. For factor 4, which has high heritability, the approximation is much closer to the truth. As regards factor 1, although ℓ_1 is underestimated, this corresponds to the first bump on the FVT interval and would appear to be well modelled over the phylogeny, given that it is faithfully reconstructed at the ancestral node. Thus, this underestimation does not have a significant impact on the conclusions drawn from the model. If anything, these observations simply indicate that caution should be exercised when attempting to

draw conclusions from the inferred values for phylogenetic hyper-parameters. As a final note on $\langle \boldsymbol{\ell} \rangle$, the superfluous fifth factor has a very large length-scale, indicating that these factors are near-constant over $\mathcal{T}$. In this respect it is behaving as an ARD hyper-parameter [Neal, 2012]. While this may be undesirable, in that it obscures the interpretation of $\langle \boldsymbol{\alpha} \rangle$, it does indicate that unnecessary factors will automatically become irrelevant within the model.

Next, examine the approximate posterior for the heritability of each factor $\boldsymbol{\kappa}$. These approximations perform much better than those for $\boldsymbol{\ell}$, in that, with factor 3 being an exception, they all lie close to the true hyper-parameter value. A possible explanation for the poor estimation in factor 3 is that the intra-taxon variation is so large as to completely dominate the inference on $\langle \kappa_3 \rangle$. Interestingly, $\langle \kappa_5 \rangle$ is very close to 1, indicating that those factors are indeed constant over $\mathcal{T}$.

Finally, note that in each case intra-taxon variation $\boldsymbol{\tau}$ is estimated remarkably well, with $\langle \tau_5 \rangle$ being close to 0, indicating that the V-PLVM will provide very accurate estimates of extant-taxon trait distributions. Thus, this examination of Figure 4.6 suggests a hierarchical approach to interpreting phylogenetic hyper-parameters. That is, $\langle \boldsymbol{\tau} \rangle$ can be trusted to reflect the underlying process, then, when $\langle \tau_j \rangle$ is low, the corresponding $\langle \kappa_j \rangle$ will reflect the heritability of the process over $\mathcal{T}$. Finally, care must be taken when interpreting $\langle \ell_j \rangle$ as even relatively small values may result in an important phylogenetic signal.

4.4 Discussion

This chapter presents a generalised Phylogenetic Latent Variable Model (PLVM) for ancestral trait reconstruction. It extends the Phylogenetic Gaussian Process Regression (PGPR) framework for function-valued traits (FVTs) to include scalar-valued ordinal, categorical, and continuous traits with Co-ordinate Ascent Variational Inference (CAVI) providing a computationally efficient method for approximate Bayesian inference. In doing so, the generalised PLVM offers a novel, flexible tool for evolutionary inference on any set of phenotypes, designed for the analysis of thousands of data points.

The generalised PLVM and its CAVI scheme represents an important methodological contribution of this thesis, building on the work presented in Chapter 3. It retains advantages of the PLVM, incorporating repeated measurements for extant taxa and joint inference of the phylogeny-trait covariance function, while addressing its shortcomings. Firstly, by linking PGPR to the threshold model for discrete trait evolution [Wright, 1934; Felsenstein, 2011], the generalised PLVM allows the

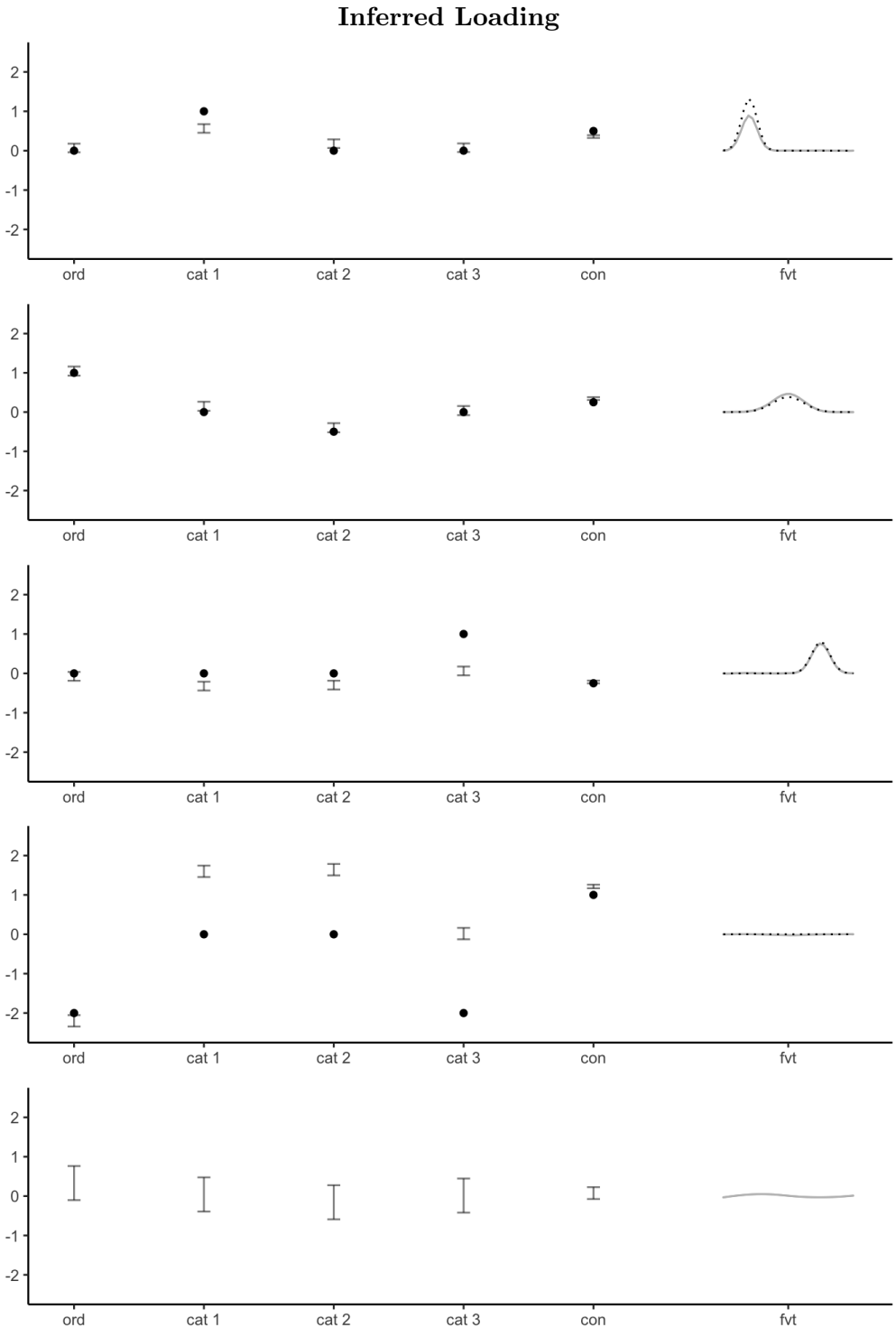
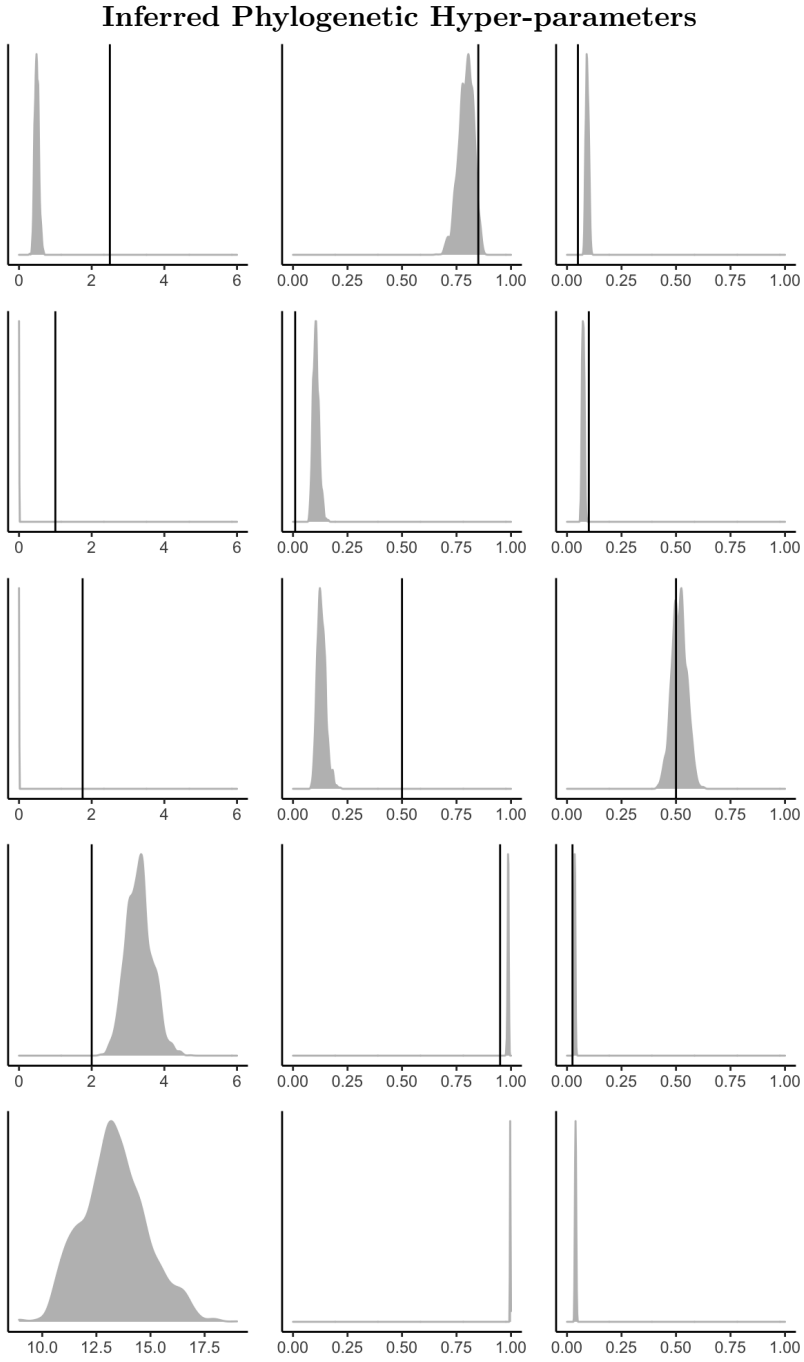


Figure 4.5: A comparison of the true loading (black points / dotted line) to the approximate posterior (grey error bars / ribbons) inferred by V-PLVM. Error bars and ribbons represent two standard deviations around the approximate posterior mean. See sub-section 4.3.3 for a discussion of the results presented in this figure.



(a) ℓ (b) κ (c) τ

Figure 4.6: A comparison of the true phylogenetic hyper-parameters (horizontal black lines) to the approximate posterior (grey density plots) inferred at the final iteration of V-PLVM. Note that rows are ordered such that the hyper-parameters in the i^{th} row correspond with the loading in the i^{th} row of Figure 4.5. Note also the fifth phylogenetic length-scale has been plotted on a different scale. See sub-section 4.3.3 for a discussion of the results presented in this figure.

ancestral reconstruction of ordinal and categorical traits alongside FVTs. As such, it offers a model for the evolution of a far richer class of phenotypes than that considered in Chapter 3. In addition, relaxing the assumption of independent and identically distributed latent variables in the PLVM provides a more flexible model for observed data, allowing for correlations between taxa and traits that depend on the phylogeny to varying degrees. Finally, CAVI performs approximate Bayesian inference for the phylogeny-trait covariance function in minutes where the MCMC algorithm proposed in Chapter 3 would take days, if not weeks. Thus, the generalised PLVM offers a general approach to ancestral trait reconstruction and is a practical tool for the phylogenetic comparative analysis of big data.

The Phylogenetic Comparative Method (PCM) introduced in this chapter offers a novel approach which develops and extends those proposed by Hadjipantelis et al. [2013], Cybis et al. [2015], and Tolkoff et al. [2017]. Setting aside the fact that a generalised PLVM models discrete and continuous scalar-valued traits alongside FVTs, it builds upon the method proposed by Hadjipantelis et al. [2013] in much the same way as the PLVM presented in Chapter 3. CAVI for the generalised PLVM clarifies that inference within the PGPR framework should be cognisant of dependence between taxa due to the phylogeny and performs joint inference for the phylogeny-trait covariance function, rather than separating inference into two distinct steps. While dimension reduction is relevant to selecting suitable initialisation values, this iterative inference scheme updates the parameters and hyper-parameters to account for the shared evolutionary history and patterns within the observed data. Furthermore, the problem of model selection with respect to the number of latent variables Q is addressed via automatic relevance determination [Neal, 2012].

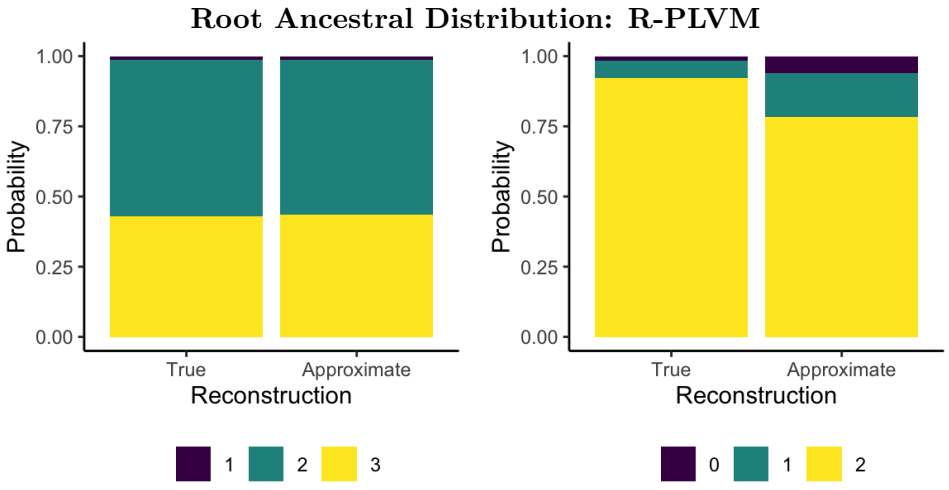
It is worth noting that the objective of this work differs from that of Cybis et al. [2015] and Tolkoff et al. [2017], in that it has been explicitly formulated for ancestral reconstruction rather than focussing on elucidating relationships between traits for related taxa. Although this is a subtle distinction, indeed these alternative approaches do imply a set of reconstructed ancestral traits, it does manifest itself in several key differences. The most important of these is that the generalised PLVM provides a far more flexible model for trait evolution than the multivariate phylogenetic latent liability model [Cybis et al., 2015] or Phylogenetic Factor Analysis (PFA) [Tolkoff et al., 2017]. In fact, each of these methods lie within the extended PGPR framework and can be thought of as special cases of the generalised PLVM where $\mathbf{X} = \mathbf{Z}\mathbf{W}^\top$ and $\mathbf{X} = \mathbf{Z}\mathbf{W}^\top + \boldsymbol{\epsilon}$ respectively. For each of these models, latent variables $\mathbf{Z}$ are assumed to be fixed, independent, and identically distributed Brownian Motion processes over the phylogeny. This leads to the second key differ-

ence, that is, these models do not include intra-taxon variation and as such do not accommodate repeated trait measurements for extant taxa. Finally, optimising the Evidence Lower Bound via CAVI provides a scalable approach to inference, while the MCMC schemes proposed by Cybis et al. [2015] and Tolkoff et al. [2017] scale with $\mathcal{O}\left(N^2\right)$ and repeated sampling from their respective posterior distributions.

A potential issue for this generalised PLVM is rooted in its flexibility, which is a result of non-identifiable parameters and hyper-parameters. It is possible that this model fits to observed data without necessarily providing a sensible model for trait evolution and that conclusions are heavily dependant on the initialisation of CAVI. Given that phylogenetic comparative analyses rarely posses validation datasets (it is often impossible to measure traits of long extinct ancestors), this would call into question any resulting ancestral trait reconstruction These concerns are allayed by results presented in Appendix C, where the ancestral trait reconstruction at the phylogeny's root is presented for each of the generalised PLVMs discarded in favour of the V-PLVM. These figures shows that each model results in remarkably similar ancestral trait distributions, even the R-PLVM (see Figure 4.7), which was initialised at random. Thus, with respect to the ancestral reconstruction of synthetic data at least, is seems that CAVI for a generalised PLVM results in broadly similar conclusions, irrespective of initialisation values.

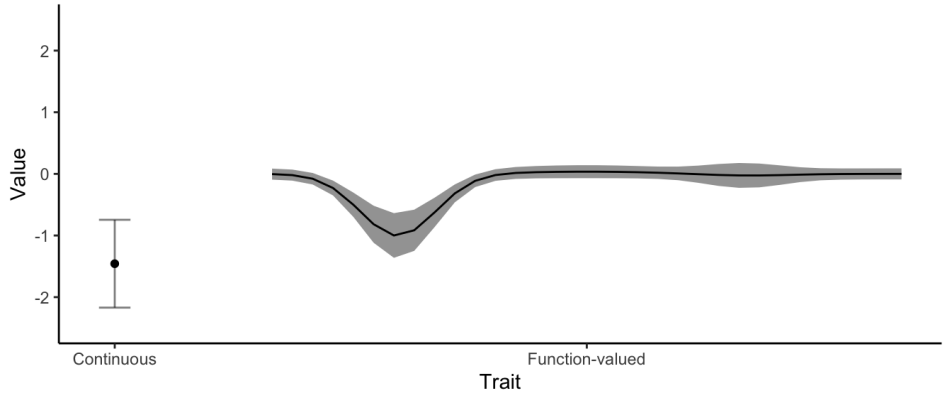
There do remain some extensions to the generalised PLVM that could be considered. Firstly, a more structured Variational inference scheme could be implemented, perhaps assuming that the mean-field variational family factorised over columns of $\mathbf{W}$ and $\mathbf{Z}$ rather than rows. This would likely result in improved uncertainty quantification; however, it would come at the cost of greater computational expense. A second extension may be to consider a deep latent variable model, linking phylogenetic Gaussian processes and deep Gaussian processes [Damianou and Lawrence, 2013]. This would result in an even more flexible model that the generalised PLVM, potentially allowing non-Gaussian traits to be modelled. Finally, extending the model to allow latent factors to be distributed according to a stable process [Elliot and Mooers, 2014] may offer an alternative approach to modelling evolution.

In conclusion, a generalised PLVM for ancestral reconstruction has been developed which can be fitted flexibly and efficiently to datasets containing thousands of observations. It is hoped that over the coming years this will prove to be an invaluable tool in the evolutionary biologists' toolbox. All that remains now is to apply this model to a set of bat echolocation calls.



(a) Ordinal Trait

(b) Categorical Trait



(c) Ancestral Distribution

(d) Approximate Ancestral Distribution

Figure 4.7: A comparison of the true ancestral distribution at the root of $\mathcal{T}$, with approximate ancestral distribution given by R-PLVM. In (a) and (b) each colour in the bars represent the probability that the trait was of that particular state, while in (d) and (c), grey error markers represent two standard deviations from the mean.

Chapter 5

Ancestral Reconstruction of the Bat Echolocation Call

5.1 Introduction

Ancestral reconstruction can be understood as an interpolation between the traits of existing populations to those of their common ancestors, providing insight into the behaviour and reproductive strategies of organisms living millions of years ago. Such trait reconstructions are inherently uncertain however, and only as reliable as the model for evolution which underpins them. The Phylogenetic Latent Variable Model (PLVM) developed in Chapters 3 & 4 represents the current state-of-the-art approach to this problem, although applying these methods to the evolution of bat echolocation is not straightforward. Echolocation calls are acoustic signals, precisely structured in both time and frequency, best described by a time-frequency representation [Cohen, 1995; Hlawatsch and Auger, 2008]. Ancestral reconstruction of the bat echolocation call must be based on manifest traits derived from such a representation, however, care must be taken to ensure that the PLVM offers a coherent model for the evolution of these features, providing a sensible distribution for ancestral traits. Thus, by identifying an appropriate time-frequency representation for a bat's echolocation call and modelling the evolution of these traits as a PLVM, this chapter presents the first ancestral reconstruction of bat echolocation to allow playback of ancestral bat echolocation calls, with results presented in this web application.[1]

A bats echolocation call is best described as a multi-harmonic signal of duration $T \in \mathbb{R}^+$ [Hopp et al., 2012; Fenton et al., 2016], subject to both frequency

[1] https://jpmeagher.shinyapps.io/test_reconstruction/

and amplitude modulation. This is to say that the signal

$$y(t) = \sum_{k=1}^{K} A_k(t) \cos(\phi_k(t)) + \epsilon(t), \tag{5.1}$$

is defined by its harmonic order K, its amplitude envelope $A_k(t) \in \mathbb{R}^+$ and instantaneous phase $\phi_k(t)$ for $k = 1, \ldots, K$, and an independent noise process, assumed to be of the form $\epsilon(t) \sim \mathcal{N}(0, \sigma^2(t))$, when $t \in [0, T]$. The harmonic structure of the call is then defined by

$$\phi_k(t) = 2\pi k \left(\int_0^t f(s)\, ds \right) + \varphi_k, \tag{5.2}$$

where the instantaneous fundamental frequency $f(t) \in \mathbb{R}^+$ must be a slowly varying function and phase shift $\varphi_k \in [0, 2\pi]$ is constant.

When characterising echolocation calls it is the fundamental frequency curve $f(\cdot)$, harmonic order K, and duration T that are of most interest [Fenton et al., 2016]. While amplitude envelope curves $A_k(\cdot)$ are dependent on the conditions in which the signal was recorded [Hopp et al., 2012], the dominant harmonic, the component carrying most energy in the signal, is an important feature [Fenton et al., 2016]. Thus, estimating $A_k(\cdot)$, at least relative to $A_{k'}(\cdot)$ for $k \neq k'$, is also relevant for the characterisation of bat echolocation calls. Finally, the phase shift φ_k and variance $\sigma^2(t)$ are considered to be irrelevant and are treated as nuisance parameters.

It is assumed that by estimating the relevant parameters for call characterisation and subsequently modelling their evolution over a phylogeny, the ancestral reconstruction of bat echolocation calls may be performed. Unfortunately, problems with this approach are immediately apparent. In general, the model described by (5.1) and (5.2) is ill defined. There exist infinite combinations of $\{A_k(t), \phi_k(t)\}$ pairs which will yield a signal equivalent to $y(t)$ [Cohen, 1995; Hlawatsch and Auger, 2008]. For mono-component signals, this can be addressed by defining an analytic signal $y_a(\cdot)$ such that $y(t) = \Re(y_a(t))$, for which an instantaneous frequency can be estimated [Gabor, 1946; Boashash, 1992; Huang et al., 2009]. Some attempts have been made at extending this approach to multi-component signals, interestingly both Olhede and Walden [2005] and DiCecco et al. [2013] used bat echolocation calls as a motivating example. Each of these methods requires the definition of frequency bands within which components lie a priori, however, making their application to large datasets difficult.

An alternative approach to parameter estimation is to consider fundamental

frequency extraction, also known as pitch tracking, which is an important problem in speech processing [Gerhard, 2003]. Given that no set of signal and fundamental frequency curve pairs for bat echolocation are known a priori, an unsupervised approach is required. Unsupervised pitch tracking algorithms can be separated into two categories, parametric and non-parametric methods. Non-parametric approaches, which include Cepstrum pitch determination [Noll, 1967] along with the RAPT [Talkin, 1995], YIN [De Cheveigné and Kawahara, 2002], SWIPE [Camacho and Harris, 2008] and PEFAC [Gonzalez and Brookes, 2014] algorithms, provide a computationally efficient estimate for the fundamental frequency curve. Despite this efficiency, however, these methods are not appropriate for the problem at hand. Fundamental frequency curve estimates can be error-prone, particularly when the signal-to-noise ratio (SNR) is low, and hyper-parameters within an algorithm may require careful manual tuning. Furthermore, the harmonic order and amplitude envelopes are not estimated by these methods. Thus, a parametric approach to fundamental frequency extraction must be considered.

Parametric pitch tracking is based on a harmonic model for acoustic signals [Quinn and Thomson, 1991; Shi et al., 2019], similar in spirit to the Spectrogram [Cohen, 1995]. Defining a set of (overlapping) frames of a signal and assuming each frame to be stationary allows a harmonic model approximating that defined by (5.1) and (5.2), to be fit for each frame. This approach provides estimates for $f(\cdot)$, K, and $A_k(\cdot)$. A major benefit of the parametric approach is that it allows for Bayesian inference of these parameters [Davy and Godsill, 2003; Nielsen et al., 2013; Shi et al., 2019], providing a coherent approach to model fitting and selection. Prior to this work, such methods had not been applied to bat echolocation, despite offering a parsimonious call characterisation and a promising approach to the comparative analysis of calls.

This chapter is laid out as follows. In section 5.2 a harmonic model for bat echolocation is formulated, for which a maximum-a-posteriori inference scheme is derived. Selected results from the fitting of this model to bat echolocation call recordings are presented and discussed. Subsequently, section 5.3 presents the problem of ancestral call reconstruction. A set of echolocation call features is then inferred from a post-hoc correction of the raw harmonic model output. One of those features, the fundamental frequency curve, represents a function-valued trait and as such is subject to a functional data analysis prior to performing evolutionary inference [Meyer and Kirkpatrick, 2005; Srivastava and Klassen, 2016]. Given this set traits representing echolocation calls and a phylogeny describing the shared evolutionary history between the observed bat taxa, the generalised Phylogenetic

Latent Variable Model developed in Chapter 4 is employed for ancestral trait reconstruction, allowing ancestral bat echolocation calls to be estimated. The chapter concludes with a discussion of these results and signposts directions for future research.

5.2 A Harmonic Model for Bat Echolocation

Consider the sinusoidal signal $y(t) \in \mathbb{R}$ for $t \in [0, T]$, that is a bats echolocation call. Firstly, define a rectangular window function of size ρ, that is

$$w_\rho(t) \equiv \delta\left(-\frac{\rho}{2} < t \leq \frac{\rho}{2}\right)$$

where $\delta(\cdot)$ is the indicator function. Then, let $t_n \in [0, T]$ for $n = 1, \ldots, N$ and $t_n < t_{n+1}$ define a set of (possibly overlapping) *frames* spanning $[0, T]$ such that

$$x_n(t) \equiv w_\rho(t - t_n)\, y(t).$$

A harmonic model for each frame is then defined as

$$x_n(t) = w_\rho(t - t_n)\, z_n(t),$$

for

$$z_n(t) = \sum_{k=1}^{K_n} \beta_{n,k}^{(1)} \cos(2\pi\, k\, f_n t) + \beta_{n,k}^{(2)} \sin(2\pi\, k\, f_n t) + \epsilon_n(t),$$

where K_n is the harmonic order, f_n is the fundamental frequency, and $\boldsymbol{\beta}_n(K_n) \equiv \left(\beta_{n,1}^{(1)}, \beta_{n,1}^{(2)}, \ldots, \beta_{n,K_n}^{(1)}, \beta_{n,K_n}^{(2)}\right)^\top$ is the set of sinusoidal basis coefficients, for which the shorthand $\boldsymbol{\beta}_{K_n} \equiv \boldsymbol{\beta}_n(K_n)$ is used when appropriate. The independent noise process is then assumed to be Gaussian with constant variance, which is to say that $\epsilon_n(t) \sim \mathcal{N}\left(0, \sigma_n^2\right)$.

The intuition which underpins the model described above is exactly that which motivates the short-time Fourier transform (STFT) [Cohen, 1995]. Firstly, the window function defines an interval of size ρ over which the signal is assumed to be locally stationary [Dahlhaus, 1996]. Then, defining frames of $y(\cdot)$ via $w_\rho(\cdot)$, rather than applying a Fourier transform, as would be the case with a STFT, each frame is described by a harmonic model. It is worth noting that a rectangular window, rather than the Hamming or Gaussian windows typically used in the STFT [Hlawatsch and Auger, 2008], should be employed when modelling frames as stationary multi-harmonic signals. This is due to the assumption of constant amplitude

in the harmonic model. Furthermore, the model defined by (5.2) is simply a re-parametrisation of (5.1), where

$$\begin{aligned} f(\cdot) &= f_n, \\ A_k(\cdot) &= \sqrt{\left(\beta_{n,k}^{(1)}\right)^2 + \left(\beta_{n,k}^{(2)}\right)^2}, \\ \varphi_k &= \operatorname{atan2}\left(\beta_{n,k}^{(2)}, \beta_{n,k}^{(1)}\right), \\ \sigma^2(\cdot) &= \sigma_n^2, \end{aligned} \tag{5.3}$$

and $K = K_n$. As will be seen in the following, estimation of the model parameters becomes much more straightforward for the sinusoidal basis defined under this parametrisation.

This harmonic model has been defined as a process that is continuous in time. While conceptually useful, the signal will, in fact, be observed at discrete time points. Therefore, letting $t_{n,m} \in \left[t_n - \frac{\rho}{2}, t_n + \frac{\rho}{2}\right]$ for $t_{n,m} < t_{n,m+1}$ and $m = 1, \ldots, M$ define a uniform sampling over the interval, which indexes all non-zero observations of $x_n(\cdot)$, the n^{th} frame can be defined as

$$\mathbf{x}_n \equiv \left(x_n(t_{n,1}), \ldots, x_n(t_{n,M})\right)^\top,$$

where $x_n(\cdot)$ has been sampled at rate $\frac{1}{t_{n,m+1} - t_{n,m}}$. Furthermore, defining the sinusoidal basis matrix

$$\mathbf{W}(K_n, f_n) \equiv \left(\mathbf{w}(K_n, f_n, t_{n,1}), \ldots, \mathbf{w}(K_n, f_n, t_{n,M})\right)^\top,$$

given the sinusoidal basis functions

$$\begin{aligned} \mathbf{w}(K_n, f_n, t_{n,m}) \equiv \big(&\cos(2\pi\, f_n\, t_{n,m}), \sin(2\pi\, f_n\, t_{n,m}), \ldots, \\ &\cos(2\pi\, K_n\, f_n\, t_{n,m}), \sin(2\pi\, K_n\, f_n\, t_{n,m}) \big)^\top, \end{aligned}$$

and $\boldsymbol{\epsilon}_n \equiv \left(\epsilon_n(t_{n,1}), \ldots, \epsilon_n(t_{n,1})\right)^\top$, (5.2) implies that

$$\mathbf{x}_n = \mathbf{W}(K_n, f_n)\, \boldsymbol{\beta}_{K_n} + \boldsymbol{\epsilon}_n.$$

And so, the likelihood associated with the n^{th} frame is

$$\mathcal{L}(\theta_n | \mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n | \mathbf{W}(K_n, f_n)\, \boldsymbol{\beta}_{K_n}, \sigma_n^2 \mathbf{I}\right), \tag{5.4}$$

where $\theta_n \equiv \left\{K_n, f_n, \boldsymbol{\beta}_{K_n}, \sigma_n^2\right\}$. With that, a harmonic model for bat echolocation

calls has been defined.

5.2.1 Prior Specification

In order to complete the specification of this model, a prior distribution for $\Theta = \{\theta_1, \ldots, \theta_N\}$ is required. In this respect, an approach similar to that of Shi et al. [2019], which builds on methods for Bayesian signal processing developed by Nielsen et al. [2013], is adopted.

The first point to note is that frames are not independent of one another. If they overlap, then the intersection between adjacent frames is a non-empty set, and even if $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ partitions the observed signal, dependence between frames is implied by the assumption that $f(\cdot)$ in (5.1) is a slowly varying function. A practical approach to such problem is to assume a first order Markov prior for Θ, which is to say that

$$p(\Theta) = p(\theta_1) \prod_{n=1}^{N} p(\theta_n|\theta_{n-1}), \tag{5.5}$$

preserving dependence between frames. Thus, by defining $p(\theta_1)$ and $p(\theta_n|\theta_{n-1})$, a Hidden Markov Model for $\mathbf{X}$ is developed [Rabiner, 1989; Bishop, 2006]

Consider first $K_n \in \mathbb{N}$, the harmonic order of $\mathbf{x}_n$. In (5.1), K is assumed to be constant, however, allowing each frame its own harmonic order defines a far more flexible model, one which fits to data in a straightforward manner. Thus, it is assumed that K_n is dependent on K_{n-1}, such that

$$\begin{aligned} &p(K_n|K_{n-1}, n_K, K_{max}) \\ &\propto \delta(1 \leq K_n \leq K_{max}) \binom{n_K}{K_n} \left(\frac{K_{n-1}}{n_K}\right)^{K_n} \left(1 - \frac{K_{n-1}}{n_K}\right)^{n_K - K_n}, \\ &= \delta(1 \leq K_n \leq K_{max}) \, \mathcal{B}\left(K_n|n_K, \frac{K_{n-1}}{n_K}\right) \end{aligned} \tag{5.6}$$

where K_{max} and n_K are the maximum harmonic order and number of prior Bernoulli trials respectively, both of which must be fixed a-priori. This truncated Binomial prior for K_n, which implies that $\max\{p(K_n|K_{n-1})\}_{K_n=1}^{K_{max}} = K_{n-1}$, encourages stability in the harmonic order between frames, while still allowing the model to fit observed data well. This prior is completed by setting $p(K_1) \propto \delta(1 \leq K_n \leq K_{max})$.

Given K_n, a prior distribution for the fundamental frequency f_n must now be defined. Firstly, let $f_{1/2}$ be the Nyquist frequency, defined as half the sampling rate, which is the highest frequency component of $\mathbf{x}_n$ that can be detected [Oppenheim

and Schafer, 2014]. Then,

$$\begin{aligned} &p\left(f_n|f_{n-1}, K_n, f_{min}, f_{max}, \sigma_f^2\right) \\ &\propto \delta\left(f_{min} \leq f_n < \min\left\{f_{max}, \frac{f_{1/2}}{K_n}\right\}\right)\mathcal{N}\left(f_n|f_{n-1}, \sigma_f^2\right) \end{aligned} \tag{5.7}$$

where the fundamental frequency variance σ_f^2, along with minimum and maximum fundamental frequencies, f_{min} and f_{max} respectively, all of which are fixed a-priori, defines a truncated Gaussian prior for f_n such that $\max_{f_n}\{p\left(f_n|f_{n-1}\right)\} = f_{n-1}$. In order to complete this prior, simply define $p\left(f_1\right) \propto \delta\left(f_{min} \leq f_1 < \min\left\{f_{max}, \frac{f_{1/2}}{K_n}\right\}\right)$.

The next parameter to be considered is the independent noise process variance, σ_n^2. Strictly speaking, σ_n^2 should be dependent on σ_{n-1}^2, either through observations being shared for adjacent frames, or by making an assumption that $\sigma^2\left(\cdot\right)$ in (5.1) is some slowly varying process. Encoding this dependence will result in a more complex model, however, and such an effort is deemed unnecessary for what is a nuisance parameter [Shi et al., 2019]. Thus, Jeffreys' prior is assumed [Jeffreys, 1946], such that

$$p\left(\sigma_n^2\right) \propto \frac{1}{\sigma_n^2}. \tag{5.8}$$

The final set of model parameters for which a prior distribution must be defined is the set of sinusoidal basis coefficients, $\boldsymbol{\beta}_{K_n}$. These parameters relate to both the amplitude $A_k\left(\cdot\right)$ and phase shift φ_k in (5.1). As such, defining a dependence between $\boldsymbol{\beta}_{K_n}$ and $\boldsymbol{\beta}_{K_{n-1}}$ would require very careful consideration. Similarly to σ_n^2 however, these parameters are not of particular interest in and of themselves, and given that dependency between frames has already been encoded in $p\left(K_n, f_n|K_{n-1}, f_{n-1}\right)$, the conditionally independent prior distribution described by Nielsen et al. [2013] is employed here. In order to motivate this choice, first consider

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{K_n} &= \arg\max_{\boldsymbol{\beta}_{K_n}} \mathcal{L}(\theta_n|\mathbf{x}_n), \\ &= \left(\mathbf{W}\left(K_n, f_n\right)^\top \mathbf{W}\left(K_n, f_n\right)\right)^{-1} \mathbf{W}\left(K_n, f_n\right)^\top \mathbf{x}_n, \end{aligned} \tag{5.9}$$

which states that, given K_n and f_n, an analytic expression for the maximum likelihood estimate of $\boldsymbol{\beta}_{K_n}$ exists. Zellner's g-prior [Zellner, 1986] then states that

$$p\left(\boldsymbol{\beta}_{K_n}|\sigma_n^2, K_n, f_n, g_n\right) = \mathcal{N}\left(\boldsymbol{\beta}_{K_n}|\mathbf{0}, g_n\sigma_n^2\left(\mathbf{W}\left(K_n, f_n\right)^\top \mathbf{W}\left(K_n, f_n\right)\right)^{-1}\right), \tag{5.10}$$

where

$$p(g_n|\zeta) = \frac{\zeta - 2}{2}\left(1 + g_n\right)^{-\frac{\zeta}{2}}, \tag{5.11}$$

is the hyper-prior distribution for the hyper-parameter g_n, with $2 < \zeta \leq 4$ such that (5.11) is a special case of the Beta prime distribution [Liang et al., 2008]. Zellner's g-prior can be interpreted as the posterior distribution for $\boldsymbol{\beta}_{K_n}$ that results from the analysis of a sample $\mathbf{x}_0 = \mathbf{0}$, given the basis $\mathbf{W}\left(K_n, f_n\right)$, a uniform prior on $\boldsymbol{\beta}_{K_n}$, and the scaled variance $g_n\sigma_n^2$ [Bové et al., 2011]. The g-prior covariance, which is the scaled inverse Fisher information matrix, implies that a large prior variance is assigned when $\boldsymbol{\beta}_{K_n}$ is difficult to estimate.

With that, a harmonic model for bat echolocation calls has been fully specified, a graphical representation of which is presented in Figure 5.1. Before considering inference for this model, however, there remains an important point to note. Firstly, let

$$\textit{Inv-Gamma}\left(x|a,b\right) \equiv \frac{b^a}{\Gamma\left(a\right)}x^{-a-1}\exp\left(-\frac{b}{x}\right), \tag{5.12}$$

be the pdf of an inverse Gamma random variable. Then, given the prior defined by (5.8) and (5.10), the joint distribution for $\mathbf{x}_n$, $\boldsymbol{\beta}_{K_n}$, and σ_n^2, conditional on K_n, f_n, and g_n can be expressed as

$$\begin{aligned}
&p\left(\mathbf{x}_n, \boldsymbol{\beta}_{K_n}, \sigma_n^2 | K_n, f_n, g_n\right) \\
&= \mathcal{L}(\theta_n|\mathbf{x}_n, \mathbf{p}_n)\, p\left(\boldsymbol{\beta}_{K_n}|\sigma_n^2, K_n, f_n, g_n\right) p\left(\sigma_n^2\right), \\
&\propto \mathcal{N}\left(\mathbf{x}_n|\mathbf{W}\left(K_n, f_n\right)\boldsymbol{\beta}_{K_n}, \sigma_n^2\mathbf{I}\right) \\
&\qquad \mathcal{N}\left(\boldsymbol{\beta}_{K_n}|\mathbf{0}, g_n\sigma_n^2\left(\mathbf{W}\left(K_n, f_n\right)^\top \mathbf{W}\left(K_n, f_n\right)\right)^{-1}\right)\sigma_n^{-2}, \\
&= \mathcal{N}\left(\boldsymbol{\beta}_{K_n}|\frac{g_n}{1+g_n}\hat{\boldsymbol{\beta}}_{K_n}, \frac{g_n}{1+g_n}\sigma_n^2\left(\mathbf{W}\left(K_n, f_n\right)^\top \mathbf{W}\left(K_n, f_n\right)\right)^{-1}\right) \\
&\qquad \textit{Inv-Gamma}\left(\sigma_n^2|\frac{M}{2}, \frac{M\hat{\sigma}_n^2}{2}\right)\Gamma\left(\frac{M}{2}\right)\left(\frac{1}{\pi M\hat{\sigma}_n^2}\right)^{\frac{M}{2}}\left(\frac{1}{1+g_n}\right)^{\frac{2K_n}{2}},
\end{aligned} \tag{5.13}$$

where

$$\hat{\sigma}_n^2 \equiv \frac{\mathbf{x}_n^\top\left(\mathbf{I} - \frac{g_n}{1+g_n}\mathbf{P}_\mathbf{W}\right)\mathbf{x}_n}{M},$$

and

$$\mathbf{P}_\mathbf{W} \equiv \mathbf{W}\left(K_n, f_n\right)\left(\mathbf{W}\left(K_n, f_n\right)^\top \mathbf{W}\left(K_n, f_n\right)\right)^{-1}\mathbf{W}\left(K_n, f_n\right)^\top.$$

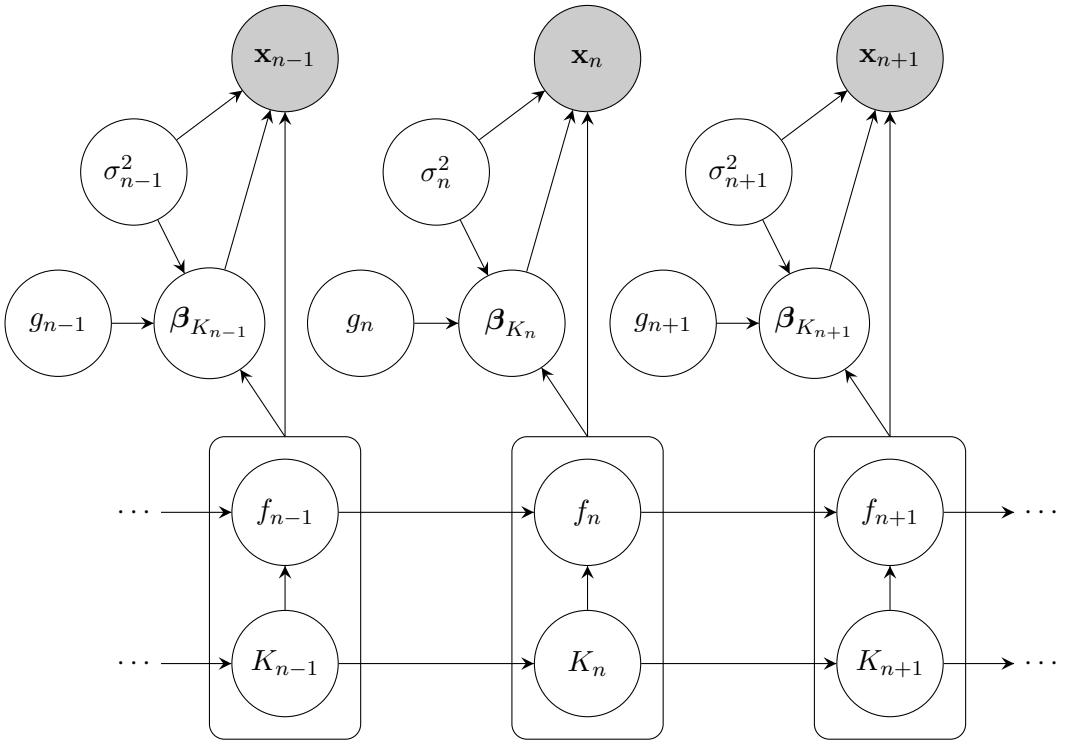


Figure 5.1: A graphical representation of the harmonic model for bat echolocation calls.

Integrating over the basis coefficients and variance implies that

$$\begin{aligned} p\left(\mathbf{x}_n | K_n, f_n, g_n\right) &\propto \Gamma\left(\frac{M}{2}\right)\left(\frac{\mathbf{x}_n^\top \mathbf{x}_n}{\pi M\, \hat{\sigma}_n^2\, \mathbf{x}_n^\top \mathbf{x}_n}\right)^{\frac{M}{2}}\left(\frac{1}{1+g_n}\right)^{\frac{2K_n}{2}} \\ &\propto \frac{\left(1+g_n\right)^{\frac{M-2K_n}{2}}}{\left(1+g_n\left(1-\mathbf{R}^2\left(K_n, f_n\right)\right)\right)^{\frac{M}{2}}}, \end{aligned} \tag{5.14}$$

where

$$\mathbf{R}^2\left(K_n, f_n\right) \equiv \frac{\mathbf{x}_n^\top \mathbf{P}_{\mathbf{W}} \mathbf{x}_n}{\mathbf{x}_n^\top \mathbf{x}_n}.$$

Thus, a marginal likelihood is defined for each frame. This expression implies that an inference scheme need only estimate K_n, f_n, and g_n for each frame.

5.2.2 Maximum-a-Posteriori Inference

Performing inference for the harmonic model defined above involves learning about the posterior distribution

$$\begin{aligned} p\left(\mathbf{k}, \mathbf{f}, \boldsymbol{g} | \mathbf{X}, \boldsymbol{\alpha}\right) &\propto p\left(\mathbf{X} | \mathbf{k}, \mathbf{f}, \boldsymbol{g}\right) p\left(\mathbf{f} | \mathbf{k}, f_{min}, f_{max}, \sigma_f^2\right) p\left(\mathbf{k} | n_K, K_{max}\right) p\left(\boldsymbol{g} | \zeta\right), \\ &= p\left(\mathbf{x}_1 | K_1, f_1, g_1\right) p\left(f_1 | K_1, f_{min}, f_{max}\right) p\left(K_1 | K_{max}\right) p(g_1 | \zeta) \\ &\quad \prod_{n=2}^{N} p\left(\mathbf{x}_n | K_n, f_n, , g_n\right) p\left(f_n | f_{n-1}, K_n, f_{min}, f_{max}, \sigma_f^2\right) \\ &\quad p\left(K_n | K_{n-1}, n_K, K_{max}\right) p(g_n | \zeta), \end{aligned} \tag{5.15}$$

where $\mathbf{k} = (K_1, \ldots, K_N)$, $\mathbf{f} = (f_1, \ldots, f_N)^\top$, $\boldsymbol{g} = (g_1, \ldots, g_N)^\top$, and the set of quantities to be specified a-priori is denoted $\boldsymbol{\alpha} = \left\{f_{min}, f_{max}, \sigma_f^2, n_K, K_{max}, \zeta\right\}$. In order to complete this task, a forward-backwards algorithm for maximum-a-posteriori (MAP) inference is developed [Rabiner, 1989; Bishop, 2006], which shall be presented as a Variational Inference (VI) scheme where the approximating distribution is a product of indicator functions.

As discussed in sub-section 4.2.3, rather than attempting inference for a posterior distribution directly, the objective of VI is to find the parametrisation for a variational family $\mathcal{Q}$ which maximises the log evidence lower bound. For (5.15), that is

$$\begin{aligned} \text{ELBO}\left(q\right) \equiv \mathbb{E}_q &\left[\log p\left(\mathbf{X} | \mathbf{k}, \mathbf{f}, \boldsymbol{g}\right) p\left(\mathbf{f} | \mathbf{k}, f_{min}, f_{max}, \sigma_f^2\right) p\left(\mathbf{k} | n_K, K_{max}\right) p\left(\boldsymbol{g} | \zeta\right)\right] \\ &- \mathbb{E}_q\left[\log q\left(\mathbf{k}, \mathbf{f}, \boldsymbol{g}\right)\right], \end{aligned}$$

which must be maximised with respect to the approximate posterior $q\left(\mathbf{k}, \mathbf{f}, \boldsymbol{g}\right) \in \mathcal{Q}$.

Firstly, it is assumed that the variational family consists of distributions that factorise according to

$$q\left(\mathbf{k}, \mathbf{f}, \boldsymbol{g}\right) = \prod_{n=1}^{N} q\left(K_n, f_n, g_n\right),$$

then, interaction between this variational family and (5.15) induces a further factorisation such that

$$q\left(K_n, f_n, g_n\right) = q\left(K_n, f_n\right) q\left(g_n\right).$$

Finally, the variational family is fully specified assuming that

$$q\left(K_n, f_n\right) = \delta\left(K_n = \langle K_n \rangle, f_n = \langle f_n \rangle\right),$$

and

$$q\left(g_n\right) = \delta\left(g_n = \langle g_n \rangle\right).$$

This implies that $\mathbb{E}_q\left[\log q\left(\mathbf{k}, \mathbf{f}, \boldsymbol{g}\right)\right] = 0$ and optimising $\text{ELBO}\left(q\right)$ with respect to variational parameters $\xi_n \equiv \{\langle K_n \rangle, \langle f_n \rangle, \langle g_n \rangle\}$ for $n = 1, \ldots, N$ is equivalent to MAP estimation.

Consider first the optimisation of $\text{ELBO}\left(q\right)$ with respect to $\langle g_n \rangle$. It can be shown that

$$\frac{\partial \,\text{ELBO}\left(q\right)}{\partial g_n} = \frac{\partial \mathbb{E}_q\left[\log p\left(\mathbf{x}_n | K_n, f_n, g_n\right) p\left(g_n | \zeta\right)\right]}{\partial g_n},$$

which allows the definition of

$$\langle g_n \rangle = \max\left\{\frac{M\mathbf{R}^2\left(\langle K_n \rangle, \langle f_n \rangle\right) - \left(2\langle K_n \rangle + \zeta\right)}{\left(2\langle K_n \rangle + \zeta\right)\left(1 - \mathbf{R}^2\left(\langle K_n \rangle, \langle f_n \rangle\right)\right)}, 0\right\}, \tag{5.16}$$

depending only on the n^{th} frame.

In order to complete the inference scheme, define the respective forward and backward framewise objective functions as

$$\begin{aligned}
\text{ELBO}_{n+1}^{f}\left(q\right) \equiv\, & \mathbb{E}_q\left[\log p\left(\mathbf{x}_{n+1} | K_{n+1}, f_{n+1}, g_{n+1}\right)\right] \\
& + \mathbb{E}_q\left[\log p\left(f_{n+1} | f_n, K_{n+1}, f_{min}, f_{max}, \sigma_f^2\right)\right] \\
& + \mathbb{E}_q\left[\log p\left(K_{n+1} | K_n, n_K, K_{max}\right)\right] + \mathbb{E}_q\left[\log p(g_{n+1} | \zeta)\right], \qquad (5.17) \\
\text{ELBO}_{n}^{b}\left(q\right) \equiv\, & \text{ELBO}_{n}^{f}\left(q\right) + \mathbb{E}_q\left[\log p\left(f_{n+1} | f_n, K_{n+1}, f_{min}, f_{max}, \sigma_f^2\right)\right] \\
& + \mathbb{E}_q\left[\log p\left(K_{n+1} | K_n, n_K, K_{max}\right)\right], \qquad (5.18)
\end{aligned}$$

for $n = 1, \ldots, N-1$, and let

$$\begin{aligned}
\text{ELBO}_{1}^{f}\left(q\right) \equiv\, & \mathbb{E}_q\left[\log p\left(\mathbf{x}_1 | K_1, f_1, g_1\right)\right] + \mathbb{E}_q\left[\log p\left(f_1 | K_1, f_{min}, f_{max}\right)\right] \\
& + \mathbb{E}_q\left[\log p\left(K_1 | K_{max}\right)\right] + \mathbb{E}_q\left[\log p(g_1 | \zeta)\right] \qquad (5.19)
\end{aligned}$$

be the initialisation objective. Each of these quantities can be computed up to a normalising constant given (5.6), (5.7), (5.14), and (5.11). This allows $\left(\langle K_n \rangle, \langle f_n \rangle\right) \in \{1, \ldots, K_{max}\} \times \left[f_{min}, \min\left\{f_{max}, \frac{f_{1/2}}{\langle K_n \rangle}\right\}\right]$ to be found by a grid search, given g_0, an initial value for the g-hyperprior, tolerance ε, and maximum number of iterations for each frame I_{max}. A description of the forward-backwards inference scheme is presented in Algorithm 4.

Algorithm 4: A Harmonic Model for Bat Echolocation Calls

Data: $\mathbf{X}, \boldsymbol{\alpha}, g_0, \varepsilon, I_{max}$
Result: MAP estimation of $\mathbf{k}$, $\mathbf{f}$ and $\boldsymbol{g}$
`/* Forward Pass */`
for $n \in \{1, \ldots, N\}$ **do**
  $g_0 \rightarrow \langle g_n \rangle$;
  $\max_{(\langle K_n \rangle, \langle f_n \rangle)} \left\{ \mathrm{ELBO}_n^f(q) \right\} \rightarrow (\langle K_n \rangle, \langle f_n \rangle)$;
  $\mathrm{ELBO}_n(q) \rightarrow \ell_n$;
  for $i \in \{1, \ldots, I_{max}\}$ **do**
    update $\langle g_n \rangle$ according to (5.16);
    $\max_{(\langle K_n \rangle, \langle f_n \rangle)} \left\{ \mathrm{ELBO}_n^f(q) \right\} \rightarrow (\langle K_n \rangle, \langle f_n \rangle)$;
    if $\mathrm{ELBO}_n^f(q) - \ell_n < \varepsilon$ **then**
      break
    else
      $\mathrm{ELBO}_n^f(q) \rightarrow \ell_n$;
    end
  end
end
`/* Backward Pass */`
for $n \in \{N-1, \ldots, 1\}$ **do**
  $\max_{(\langle K_n \rangle, \langle f_n \rangle)} \left\{ \mathrm{ELBO}_1^b(q) \right\} \rightarrow (\langle K_n \rangle, \langle f_n \rangle)$;
  $\mathrm{ELBO}_n(q) \rightarrow \ell_n$;
  for $i \in \{1, \ldots, I_{max}\}$ **do**
    update $\langle g_n \rangle$ according to (5.16);
    $\max_{(\langle K_n \rangle, \langle f_n \rangle)} \left\{ \mathrm{ELBO}_n^b(q) \right\} \rightarrow (\langle K_n \rangle, \langle f_n \rangle)$;
    if $\mathrm{ELBO}_n^b(q) - \ell_n < \varepsilon$ **then**
      break
    else
      $\mathrm{ELBO}_n^b(q) \rightarrow \ell_n$;
    end
  end
end

5.2.3 Fitting the Harmonic Model

The harmonic model described above was fit to a sample of 1816 bat echolocation calls recorded at a sampling rate of 500 kHz. This dataset has been made publicly available by Stathopoulos et al. [2018] and will be discussed in more detail in subsection 5.3.1. The fundamental frequency for calls in the sample is assumed to lie over the interval $[f_{min} = 15 \text{ kHz}, f_{max} = 212 \text{ kHz}]$. While the range $[9 \text{ kHz}, 212 \text{ kHz}]$ was discussed in section 2.1, this value of f_{min} better reflects the properties of calls observed in this particular dataset. Furthermore, raising f_{min} can significantly improve the fit of model. It forces low frequency background noise to be ignored and reduces incidence of "pitch halving", which refers to the phenomenon of a fitted fundamental frequency being half the true value, as identified by visual inspection of the signals spectrogram. Standard deviation for the change in fundamental frequency from one frame to the next is then assumed to be $\sigma_f = 5$ kHz. This represents a somewhat balanced prior, that should discourage pitch halving or doubling between frames without prohibiting it. It is also worth noting that the Nyquist frequency $f_{1/2} = 250$ kHz

The maximum harmonic order takes some relatively large value, which is to say that it is greater than the anticipated number of harmonics for any call in the sample. In this case $K_n = 8$ is deemed appropriate, while the number of trials in the Binomial prior on harmonic order n_K must be greater than K_{max} to prevent $\mathcal{B}(K_n|n_K, 1)$ from occurring. Here $n_K = 2K_{max}$ is chosen. Fixing the g-hyper-prior parameter $\zeta = 3$, all the required quantities have been specified such that

$$\boldsymbol{\alpha} = \left\{ f_{min} = 15 \text{ kHz}, f_{max} = 212 \text{ kHz}, \sigma_f^2 = (5 \text{ kHz})^2, n_K = 16, K_{max} = 8, \zeta = 3 \right\}.$$

The harmonic model is then fit the set of bat echolocation calls. After first passing recordings through a Butterworth bandpass filter of order 10 defined by $[f_{min}, f_{max}]$ [Butterworth et al., 1930], frames are defined by a window of size $\rho \equiv 0.512$ ms, implying that $M \equiv 256$, with 75% overlap of adjacent frames.

For each harmonic order, maximisation of the fundamental frequency is performed by a grid search. Given a coarse uniform grid over permissible frequencies, defined by intervals of size 0.25 kHz, an intermediate frequency, f_{tmp}, maximising the objective is identified. Then a finer grid search, defined by the interval of size 0.01 kHz, is performed over

$$\left[\max \left\{ f_{min}, f_{tmp} - 0.25 \text{ kHz} \right\}, \min \left\{ f_{max}, f_{tmp} + 0.25 \text{ kHz} \right\} \right].$$

The tolerance and maximum number of iterations selected for the inference scheme is $\varepsilon \equiv 10^{-2}$ and $I_{max} \equiv 10$.

The model fit is assessed by a visual comparison of call spectrograms and fitted values for the fundamental frequency and harmonic order in each frame. A representative selection of echolocation calls are presented in Figure 5.2.

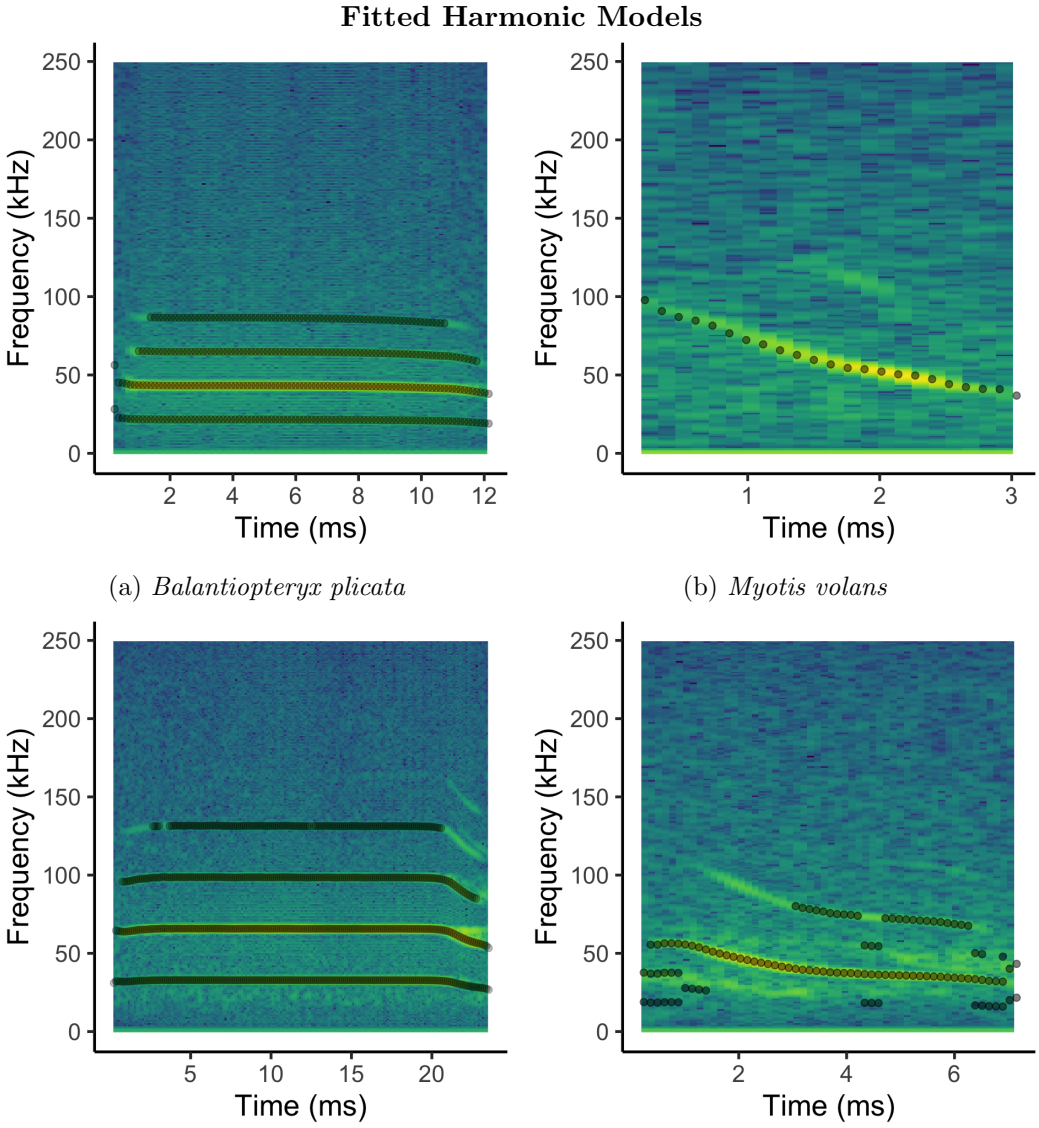
Figure 5.2a shows the model fit and call spectrogram for a bat from the species *Balantiopteryx plicata* within the Emballonuridae family. This call can be best described as a constant frequency, multi-harmonic signal, for which the second component is dominant. The fitted model agrees with the call spectrogram, yielding a harmonic order of 4 and frequency component curves that lie along peaks of the spectrogram. However, on close inspection of Figure 5.2a, there appears to be a discontinuity in the frequency components from the first to the second frame. This error occurs due to an absence of any frequency component in the first frame, i.e. there is a low signal-to-noise ratio (SNR).

Similarly the model fits well to the single-component, short-duration, broadband sweep of *Myotis volans* (Vespertilionidae, Figure 5.2b), and the multi-harmonic constant frequency to broadband sweep call of *Pteronotus parnellii* (Mormoopidae, Figure 5.2c). In the case of *Tadarida brasiliensis* (Molossidae, Figure 5.2d) however, the model fit is poor. This multi-harmonic, broadband call appears to have been recorded with a low SNR, as evidenced by seemingly unstructured areas of high energy density on the spectrogram. Thus, the fitted model suffers from order errors in a number of frames, resulting in pitch halving, although some frequency component of the model does identify the dominant component throughout.

Although the model does not provide a perfect fit for the data, fundamental frequency extraction is a very challenging problem, as evidenced by the plethora of algorithms that have been developed for this purpose [Noll, 1967; Talkin, 1995; De Cheveigné and Kawahara, 2002; Gerhard, 2003; Camacho and Harris, 2008; Gonzalez and Brookes, 2014]. Thus, such performance is to be expected when applying any fundamental frequency extraction algorithm to recordings sampled in the field. In reality, Figure 5.2 presents reasonably satisfactory results, which will allow the desired feature representation of echolocation calls to be obtained after a post-hoc correction of raw output from the model, as will be discussed in sub-section 5.3.3.

5.2.4 A Brief Discussion of the Harmonic Model

The objective in developing this harmonic model for bat echolocation was to define a set of features for which ancestral reconstruction could be performed. In many respects, this has been successful. As will be discussed in detail in the next section,



(a) *Balantiopteryx plicata*

(b) *Myotis volans*

(c) *Pteronotus parnellii*

(d) *Tadarida brasiliensis*

Figure 5.2: The harmonic model fit to a sample of bat echolocation calls, overlaid on the call spectrogram. Frequency components of the echolocation call are identified by black points in each frame. Thus, at any particular point in time, the lowest frequency point is the fundamental frequency, while the number of points corresponds to the harmonic order. Discussion of each case presented above is provided in sub-section 5.2.3.

this model allows the harmonic order, fundamental frequency curve, and dominant harmonic of a given call recording to be defined. Figure 5.2 demonstrates that, for the most part, the model provides a very good fit for the data. This represents an important contribution towards the comparative analysis of bat echolocation.

The model described here is not without some significant shortcomings, however. Firstly, it is computationally expensive to implement, taking an average of three minutes to fit each call in the sample, although this issue could be addressed by implementing an algorithm for efficient likelihood computation derived by Nielsen et al. [2017]. This would represent the first step in any further development of this method for fundamental frequency estimation.

A second issue is that the model does not include the case where there is no periodic signal in a frame. This causes discontinuities in the fundamental frequency curve to occur over frames at the beginning and end of calls (see Figure 5.2a). A potential solution to this problem would be to replace the model defined by (5.2) with

$$z_n(t) = u_n \left(\sum_{k=1}^{K_n} \beta_{n,k}^{(1)} \cos(2\pi\, k\, f_n t) + \beta_{n,k}^{(2)} \sin(2\pi\, k\, f_n t) \right) + \epsilon_n(t), \tag{5.20}$$

which introduces the variable $u_n \in \{0, 1\}$ indicating the presence or absence of a periodic signal. In this respect, the model and inference scheme developed here is less sophisticated than that presented by Shi et al. [2019]. Inclusion of such an indicator variable may allow the model to be adapted for problems such as call identification and classification [Stathopoulos et al., 2018; Mac Aodha et al., 2018].

The third problem identified here is that order errors do occur for some calls within the sample (Figure 5.2d). When these are due to unstructured noise corrupting the signal, this may prove an impossible problem to solve completely, however, a more sophisticated Bayesian inference scheme may mitigate such issues. While it may be computationally expensive, a Reversible Jump Markov Chain Monte Carlo inference scheme inferring the harmonic order and fundamental frequency for each frame, similar to the approach for non-stationary periodic signals proposed by Hadj-Amar et al. [2019], could provide an interesting avenue for future research. Alternatively, a more carefully considered Variational inference scheme than that presented above, perhaps employing a Variational family where $q(f_n) = \mathcal{N}\left(f_n | \mu_n, \lambda_n^{-1}\right)$, could offer similar advantages. Such an approach would allow some uncertainty quantification without a massive computational expense

Despite these issues and potential directions for further work, when a rich set of echolocation call features is required, the parametric approach to fundamen-

tal frequency extraction proposed here is undoubtedly superior to alternative, non-parametric methods. While the YIN algorithm is based on the autocorrelation function for the signal in question De Boor [1972], methods such as RAPT [Talkin, 1995], SWIPE [Camacho and Harris, 2008] and PEFAC [Gonzalez and Brookes, 2014] are not based on any explicit model for periodic signals. Instead, they rely on frequency-domain representations of a signal, attempting to identify peaks in the power spectral density associated with the fundamental frequency. Although these methods represent the standard approach to pitch determination, they do not allow estimation of the harmonic order and amplitude envelope in a straightforward manner. As such, they would be unsuitable for obtaining feature representations of bat echolocation calls. Furthermore, they can be error prone and require careful parameter tuning for effective performance. While the harmonic model developed here does result in some errors, as judged by a comparison of its raw output with call spectrograms, these occur in a clear and systematic manner. Thus, a post-hoc correction procedure offers a pragmatic solution to this issue, defining a feature representation for bat echolocation call, which in turn allows their ancestral reconstruction. This analysis will be presented in the following section.

5.3 Echolocation Call Reconstruction

Ancestral reconstruction for a set of echolocation call features is presented here. Features are defined by a post-hoc correction of raw output from the harmonic model presented in section 5.2 and will subsequently be modelled as a generalised Phylogenetic Latent Variable Model (PLVM), for which Variational Inference will be performed, as described in Chapter 4.

5.3.1 Echolocation Call Data

Bat echolocation call recordings gathered across north and central Mexico have been made publicly available by Stathopoulos et al. [2018]. These calls were collected from June to November 2012 and from February to May 2013. Bats were captured in 10 mist nets placed at ground level, that is 0-3 metres high, and identified to species level using field keys. Echolocation calls were recorded by two methods: bats were released from the hand in open areas away from vegetation, between 6 and 10 metres away from a bat detector, or; bats were attached to a zip line and recorded as they flew along the zip line path. Calls were recorded by a Pettersson 1000x bat detector, set to record calls manually in realtime, full-spectrum, at a sampling rate of 500 kHz. Calls were selected from recordings of 449 individual bats

Species	Key	Individuals	Calls
Family: Emballonuridae			
Balantiopteryx plicata	Bapl	16	100
Family: Molossidae			
Nyctinomops femorosaccus	Nyfe	16	100
Tadarida brasiliensis	Tabr	49	100
Family: Vespertilionidae			
Antrozous pallidus	Anpa	58	100
Eptesicus fuscus	Epfu	74	100
Idionycteris phyllotis	Idph	6	100
Lasiurus blossevillii	Labl	10	90
Lasiurus cinereus	Laci	5	42
Lasiurus xanthinus	Laxa	8	100
Myotis volans	Myvo	8	100
Myotis yumanensis	Myyu	5	89
Pipistrellus hesperus	Pihe	85	100
Family: Mormoopidae			
Mormoops megalophylla	Mome	10	100
Pteronotus davyi	Ptda	8	100
Pteronotus parnellii	Ptpa	23	100
Pteronotus personatus	Ptpe	7	51
Family: Phyllostomidae			
Artibeus jamaicensis	Arja	11	82
Desmodus rotundus	Dero	6	38
Leptonycteris yerbabuenae	Leye	26	100
Macrotus californicus	Maca	6	53
Sturnira ludovici	Stlu	8	51
Sturnira lilium	Stli	4	20

Table 5.1: Mexican Bat Echolocation Call Dataset

from $S^b = 22$ species across five families, each consisting of multiple calls. For each species, as many calls as possible, up to a maximum of 100, were selected from the recordings. This resulted in $N^b = 1816$ sample calls, with the number of calls for each species denoted N_i^b for $i = 1, \ldots, S^b$. Species have been assigned a four-letter identifying key based on their binomen (the scientific name for the species), made up of the first two letters of the genus and species names respectively. This dataset is summarised in Table 5.1.

5.3.2 Bat Phylogeny

It is assumed that evolutionary relationships between the $S^b = 22$ species of bat are accurately described by the bat supertree of Collen [2012], based on studies

conducted between 1970 and 2009, which dates the most recent common ancestor for all bats in the sample to 52.5 million years ago. This phylogeny, denoted $\mathcal{P}_S$, is presented in Figure 5.3.

As described in section 3.2.1, multiple observations per species are accommodated by defining $\mathcal{P}$, which retains the inter-taxon structure of $\mathcal{P}_S$, but now has N^b observation (terminal) nodes, each of which has an edge weight of zero with its parent, one of the S^b extant taxon nodes. Let $\mathbf{p}_n \in \mathcal{P}$ for $n = 1, \ldots, N^b$ denote the positions on $\mathcal{P}$ of the observation nodes, $\mathbf{p}_n \in \mathcal{P}$ for $n = N^b + 1, \ldots, N^b + S^b$ be extant taxon nodes, and $\mathbf{p}_n \in \mathcal{P}$ for $n = N^b + S^b + 1, \ldots, N^b + S^b + M^b$ be the ancestral nodes corresponding to the $M^b = 16$ internal nodes in $\mathcal{P}_S$. Furthermore, given the patristic distance operator $\mathrm{d}_{\mathcal{P}}(\cdot, \cdot)$, which computes the distance between positions on $\mathcal{P}$, the phylogeny $\mathcal{P}$ is scaled such that $\mathrm{d}_{\mathcal{P}}\left(\mathbf{p}_n, \mathbf{p}_{N^b+S^b+M^b}\right) = 1$ for $n = 1, \ldots, N$, which is to say that the patristic distance between each of the observation nodes and the root of $\mathcal{P}$ is 1.

5.3.3 Echolocation Call Features

It is assumed that echolocation calls are well characterised by a set of four features. They are the harmonic order, fundamental frequency curve, call duration, and dominant component. Fitting the harmonic model described in section 5.2 to each echolocation call recording produces the raw output $\hat{\mathbf{f}}_n = \left(\langle f_1^n \rangle, \ldots, \langle f_{N_n^c}^n \rangle\right)^\top$, $\hat{\mathbf{k}}_n = \left(\langle K_1^n \rangle, \ldots, \langle K_{N_n^c}^n \rangle\right)^\top$, and $\hat{\boldsymbol{g}}_n = \left(\langle g_1^n \rangle, \ldots, \langle g_{N_n^c}^n \rangle\right)^\top$, where N_n^c denotes the number of frames for the n^{th} call recording for $n = 1, \ldots, N^b$. As demonstrated in Figure 5.2, this raw output may contain order errors, and so a post-hoc correction of the data is performed before defining the feature set.

Before proceeding, define the operator

$$\text{which}\max\left(x_1, \ldots, x_N\right) = i,$$

which identifies the index i such that

$$\max\left(x_1, \ldots, x_N\right) = x_i$$

for some $N \in \mathbb{N}$. Then, note that for any $\{f_i^n, K_i^n, g_i^n\}$, (5.3), (5.9) and (5.13) imply the values $\hat{\mathbf{a}}_{n,i} = \left(\langle A_{i,1}^n \rangle, \ldots, \langle A_{i,K_i^n}^n \rangle\right)^\top$ and let $\hat{\mathbf{a}}_{n,i}^2 = \left(\langle A_{i,1}^n \rangle^2, \ldots, \langle A_{i,K_i^n}^n \rangle^2\right)^\top$.

Consider first the fundamental frequency curve. In cases such as those presented in Figures 5.2a-5.2c, any errors can be corrected simply by pruning away components for which the estimated squared amplitude envelope, denoted $\hat{\mathbf{a}}_{n,i}^2$, is

A Phylogeny for Sampled Mexican Bats

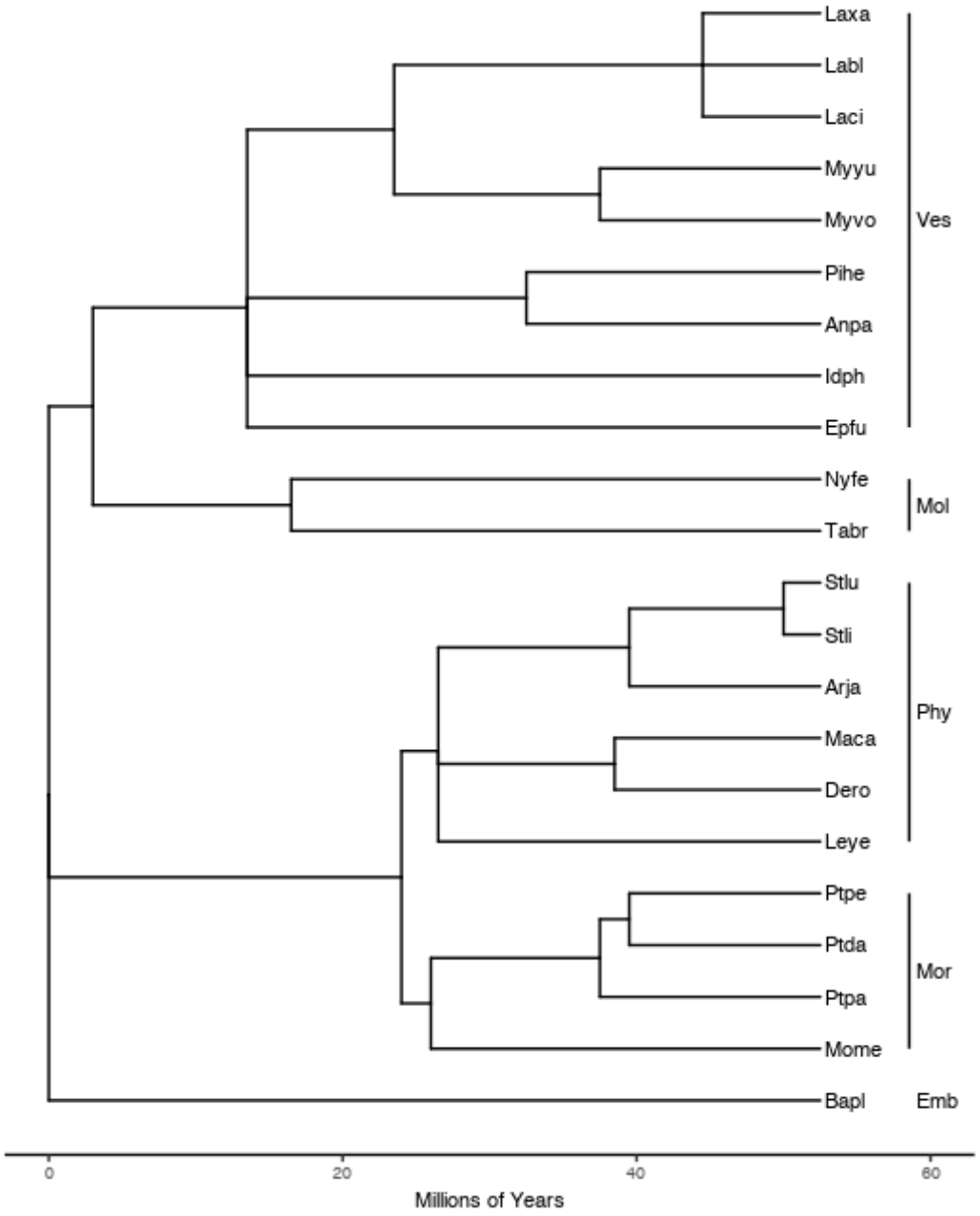


Figure 5.3: The phylogenetic tree assumed to model evolutionary relationships between observed bat species, as described by Collen [2012].

below a defined threshold, $a_{n,0}^2$, typically set at $0.005 \max \left\{ \hat{\mathbf{a}}_{n,i}^2 \right\}_{i=1}^{N_n^c}$, although there is variation from one call to the next. In Figure 5.2d, however, the fundamental frequency has been correctly identified in some frames, but order errors occur in others, resulting in discontinuous jumps in fundamental frequency curve. These errors are addressed by identifying a reference frame $i_{n,0}$ and implementing a procedure iteratively matching frequency components with those of an adjacent frame. In most cases $i_{n,0} = \text{which} \max \left\{ \hat{\mathbf{a}}_{n,i}^\top \hat{\mathbf{a}}_{n,i} \right\}_{i=1}^{N_n^c}$, which is approximately the frame for which the SNR is highest, is an appropriate choice. Furthermore, in a very small proportion of calls, pitch halving occurs in every frame. For such cases, the fundamental frequency curve is simply doubled. If after executing each of these steps the model fit remains unsatisfactory, fitting the model for new values for f_{min} and f_{max} was attempted. If this proved unsuccessful, as happened for some recordings with a particularly low SNR, the call was omitted from the analysis. In total 1805 of the 1816 calls were deemed suitable for comparative analysis. Examples of corrected fundamental frequency vector, denoted $\tilde{\mathbf{f}}_n$, are presented in Figure 5.4.

Before completing the definition of fundamental frequency curves, it is worth considering the harmonic order, duration and dominant component for each call. A straightforward definition of the harmonic order for the call would be to set it as $\max \hat{\mathbf{k}}_n$ simply. Such a definition could result in the harmonic order being misspecified, however, given that order errors may have occurred. A similar definition, which accounts for corrections made to the fundamental frequency, is to let $\tilde{\mathbf{k}}_n$ be the vector of harmonic orders corresponding to $\tilde{\mathbf{f}}_n$ such that $\left(\tilde{\mathbf{k}}_n \right)_i = \tilde{K}_i^n$ where

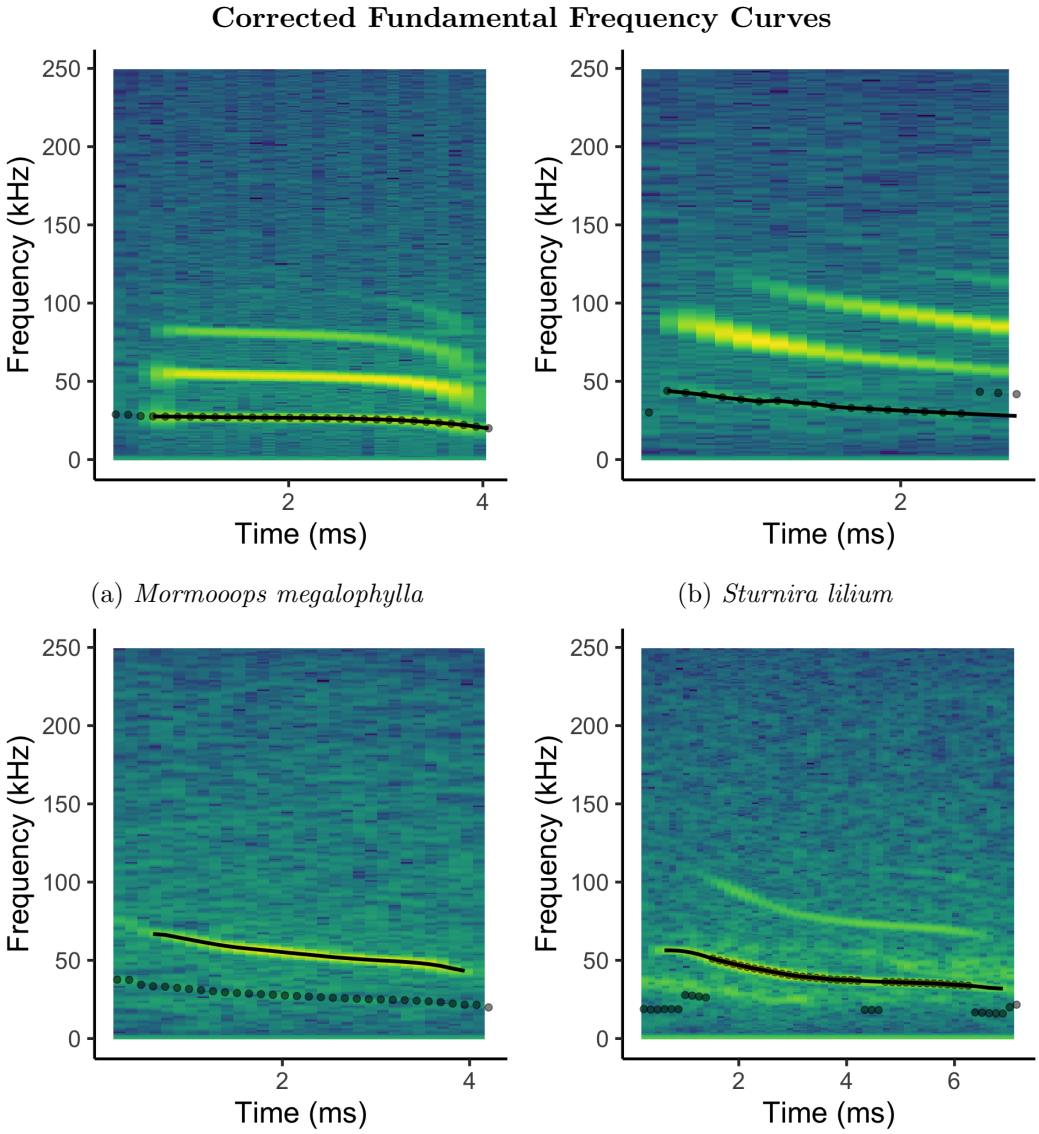
$$\tilde{K}_i^n = \max \left\{ k | \left(\tilde{A}_{i,k}^n \right)^2 > a_{n,0}^2 \right\}_{k=1}^{K_{max}}$$

and $\tilde{A}_{i,k}^n$ is analogous to $\langle A_{i,k}^n \rangle$. Then, the final estimate for the harmonic order of the call is given by

$$\bar{K}_n = \max \tilde{\mathbf{k}}_n. \tag{5.21}$$

The duration of the call is then estimated given $\tilde{t}_{n,1}$ and $\tilde{t}_{n,N_n^{c\prime}}$, the first and last time index associated with $\tilde{\mathbf{f}}_n$ respectively, by

$$\bar{T}_n = \tilde{t}_{n,N_n^{c\prime}} - \tilde{t}_{n,1} + \rho, \tag{5.22}$$



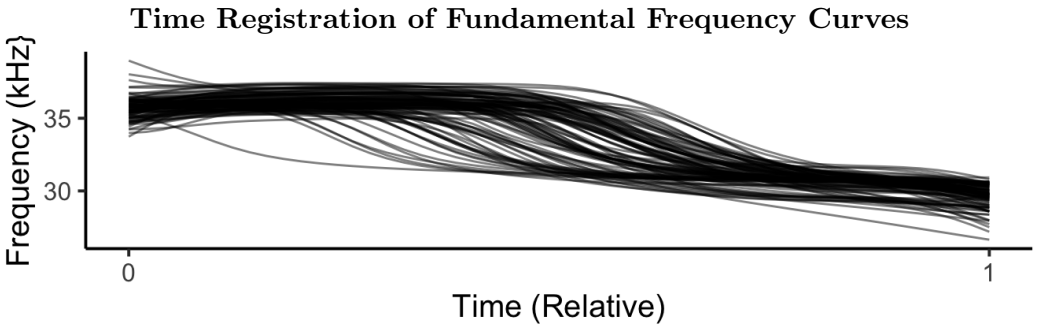
(a) *Mormooops megalophylla*

(b) *Sturnira lilium*

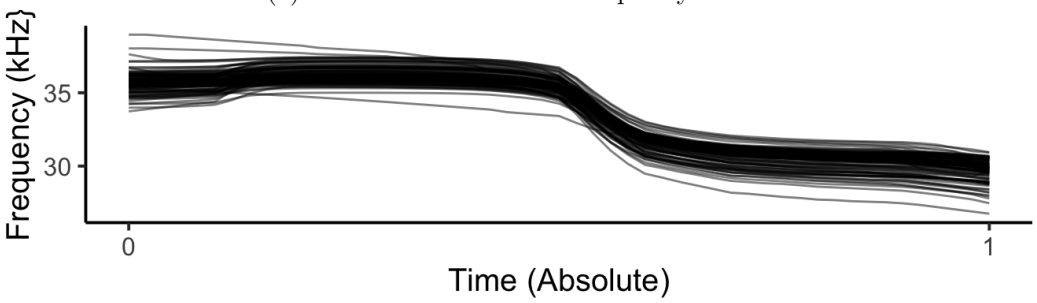
(c) *Myotis yumanensis*

(d) *Tadarida brasiliensis*

Figure 5.4: The corrected fundamental frequency fit for a sample of bat echolocation calls, overlaid on the call spectrogram. The fitted fundamental frequency component of the echolocation call is identified by black points in each frame, while the corrected fundamental frequency curve is presented as a solid black line. Figure 5.4a presents a case where fundamental frequency estimates for the first four frames have been discarded due to the absence of any signal; Figure 5.4b corrects for misidentification of the fundamental frequency in the final three frames of the call; Figure 5.4c illustrates the effect of correcting for pitch halving throughout the call recording; and Figure 5.4d demonstrates the correction when pitch halving occurs in a few frames only.



(a) Observed fundamental frequency curves



(b) Registered fundamental frequency curves

Figure 5.5: A comparison of observed and registered fundamental frequency curves for sampled *Pteronotus parnellii* calls.

with the dominant component given by

$$\bar{d}_n = \text{which max} \left\{ \sum_{i=1}^{N_n^{c\prime}} \left(\tilde{A}_{i,k}^n \right) \right\}_{k=1}^{\bar{K}_n}.$$

All that remains now, is the definition of fundamental frequency curves given $\tilde{\mathbf{f}}_n$ for $n = 1, \ldots, N^b$. The fundamental frequency curve, as defined by $f(\cdot)$ in (5.2), is a slowly varying function of time, Therefore, it is assumed that $\tilde{\mathbf{f}}_n$, representing discrete observations of the fundamental frequency curve for the n^{th} echolocation call which are associated with time indices $\left\{\tilde{t}_{n,1}, \ldots, \tilde{t}_{n,N_n^{c\prime}}\right\}$, is a function-valued trait (FVT) for which Functional Data Analysis (FDA) must be performed. [Ramsay, 2004; Meyer and Kirkpatrick, 2005] Thus, in order to define the set of fundamental frequency curves for ancestral reconstruction, each $\tilde{\mathbf{f}}_n$ must be smoothed and registered appropriately.

The first step in the FDA is to map $\tilde{\mathbf{f}}_n$ to the unit interval. This is simply a

case of defining a new set of time indices where

$$t_{n,i} = \frac{\tilde{t}_{n,i} - \tilde{t}_{n,1}}{\tilde{t}_{n,N_n^{c\prime}} - \tilde{t}_{n,1}}.$$

Then, assuming that $\tilde{f}_n$ represents noisy observations of a twice differentiable function, a smoothing spline can be fitted [Friedman et al., 2001; R Core Team, 2019], such that $\tilde{f}_n(t)$ is defined for all $t \in [0, 1]$. Fundamental frequency curves must then be time registered [Ramsay and Li, 1998]. This describes the process of stripping out phase variation in the curves such that $f_n(t)$ is meaningfully comparable with $f_m(t)$ for $n \neq m$. Here, fundamental frequency curves are registered within each taxon, though no between-taxon registration is performed. The Bayesian registration algorithm of Cheng et al. [2016], which registers the square-root velocity function of the fundamental frequency curves and is implemented in the `fdasrvf` R package [Tucker, 2019; R Core Team, 2019], is employed for this task, with the effect of time registration on the fundamental frequency curves for *Pteronotus parnellii* presented in Figure 5.5. Finally, let N^f-dimensional vector $\bar{\mathbf{f}}_n$ represent the smoothed, registered fundamental frequency curve sampled on a regular grid over the unit interval for $n = 1, \ldots, N^b$, where $N^f = 51$.

5.3.4 Ancestral Reconstruction

The set of echolocation call features defined above are: the harmonic order $\bar{K}_n \in \{1, \ldots, 6\}$, which is considered to be an ordinal trait; the time registered fundamental frequency curve $\bar{\mathbf{f}}_n \in (\mathbb{R}^+)^{N^f}$, a FVT; the dominant component $\bar{d}_n \in \{1, 2, 3\}$, a categorical trait; and the call duration $\bar{T}_n \in \mathbb{R}^+$; all of which are defined for $n = 1, \ldots, N^{b\prime}$, where $N^{b\prime} = 1805$ given that the harmonic model did not fit successfully to all $N^b = 1816$ recordings. Note that the maximum value of $\bar{K}_n$ inferred during feature extraction was 6, indicating that setting $K_{max} = 8$ was appropriate.

The evolution of these traits over the phylogeny $\mathcal{P}$ is modelled as a generalised PLVM, as presented in Chapter 4, with $P = 4$ manifest traits such that

$$\mathbf{Y}_{n\cdot} = \begin{bmatrix} \bar{K}_n \\ \bar{d}_n \\ \left(\log \bar{T}_n - \frac{1}{N^{b\prime}} \sum_{n=1}^{N^{b\prime}} \log \bar{T}_n\right) / \hat{\sigma}_{\bar{T}} \\ \log \bar{\mathbf{f}}_n - \frac{1}{N^{b\prime}} \sum_{n=1}^{N^{b\prime}} \log \bar{\mathbf{f}}_n \end{bmatrix},$$

where

$$\hat{\sigma}_{\bar{T}} = \sqrt{\frac{\left(\log \bar{T}_n - \frac{1}{N^{b\prime}} \sum_{n=1}^{N^{b\prime}} \log \bar{T}_n\right)^2}{N^{b\prime} - 1}},$$

and $D = N^f + 3$. That is to say, the logarithm of call duration has been centred and scaled to have variance 1, while the fundamental frequency curves logarithm has been centred. Taking the logarithm of variables that have been defined for the positive real numbers allows them to be modelled over the real number line, as specified for the generalised PLVM.

Manifest traits are then modelled by auxiliary traits $\mathbf{X}_{n\cdot} \in \mathbb{R}^{D'}$ given the map $g : \mathbf{X}_{n\cdot} \to \mathbf{Y}_{n\cdot}$, defined in sub-section 4.2.1, and ordinal cut off points $\boldsymbol{\gamma}$ where

$$\mathbf{X}_{n\cdot} = \mathbf{W}\mathbf{Z}_{n\cdot} + \boldsymbol{\epsilon},$$

for $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right)$, with diagonal precision matrix $\boldsymbol{\Lambda}$ and $D' = N^f + 5$.

The factors $\mathbf{Z}_{nj} = z_j\left(\mathbf{p}_n\right)$ for $j = 1, \ldots, Q$ are assumed to follow univariate phylogenetic Gaussian processes, that is

$$z_j\left(\mathbf{p}_n\right) \sim \mathcal{GP}\left(0, k_{\mathcal{P}}\left(\mathbf{t}_n, \mathbf{p}_m | \kappa_j, \tau_j, \ell_j\right)\right),$$

where the phylogenetic covariance function is of the form

$$\begin{aligned} k_{\mathcal{T}}\left(\mathbf{p}_n, \mathbf{p}_m | \kappa_j, \tau_j, \ell_j\right) = {} & \left(1 - \tau_j\right)\left(\kappa_j \ \exp\left(-\frac{\mathrm{d}_{\mathcal{T}}\left(\mathbf{p}_n, \mathbf{p}_m\right)}{\ell_j}\right) + \right. \\ & \left. \left(1 - \kappa_j\right) \ \delta\left(\mathrm{d}_{\mathcal{P}}\left(\mathbf{p}_n, \mathbf{p}_m\right) = 0\right) \delta\left(n \leq N + S\right)\right) + \\ & \tau_j \, \delta\left(n = m\right) \delta\left(n \leq N\right), \end{aligned}$$

where heritability κ_j, intra-taxon variation τ_j, phylogenetic length-scale ℓ_j, and the patristic distance operator $\mathrm{d}_{\mathcal{P}}\left(\cdot, \cdot\right)$ define an Ornstein-Uhlenbeck Phylogenetic Mixed Model with intra-taxon variation.

The prior distribution for loading $\mathbf{W}$ is defined for each of the Q columns with

$$p\left(\mathbf{W}_{\cdot j} | \alpha_j\right) = \mathcal{N}\left(\mathbf{W}_{\cdot j} | \mathbf{0}, \alpha_j^{-1} \mathbf{K}_{\mathbf{W}}\right), \tag{5.23}$$

where $\mathbf{K}_{\mathbf{W}}$ is a block diagonal matrix. Non-zero off diagonal elements of $\mathbf{K}_{\mathbf{W}}$ occur only in the block corresponding to the fundamental frequency curve, which is assumed to be twice mean square differentiable, given by the Gram matrix of the

Matérn-$\frac{5}{2}$ kernel

$$k(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right), \tag{5.24}$$

for $r = |t_i - t_{i'}|$, when $t_i \in [0, 1]$ indexes the time registered fundamental frequency curves. The length-scale is fixed a-priori such that $\ell = 0.5$. This value chosen after Type II maximum likelihood estimation for a zero-mean GP, given covariance function (5.24), fitted to the sample of manifest traits corresponding to fundamental frequency curves [Rasmussen and Williams, 2006].

In order to define the prior for $\mathbf{\Lambda}$, note that those diagonal elements corresponding to discrete traits are fixed to 1 and let λ_T, λ_f be the precision parameters for the call duration and fundamental frequency curve respectively. Then set $p(\lambda_T) = Gamma(\lambda_T|1, 1)$ and $p(\lambda_f) = Gamma(\lambda_f|1, 1)$.

The model specification is completed given hyper-prior distributions for the model hyper-parameters. They are $p(\kappa_j) = Beta(\kappa_j|1, 1)$; $p(\tau_j) = Beta(\tau_j|1, 1)$; $p(\ell_j) = Gamma(\ell_j|2, 1)$; and $p(\alpha_j) = Gamma(\alpha_j|0.001, 0.001)$ for $j = 1, \ldots, Q$.

This model is then fit for $\mathbf{Y} = \left(\mathbf{Y}_{1\cdot}, \ldots, \mathbf{Y}_{N^{b\prime}\cdot}\right)^\top$, given the phylogeny $\mathcal{P}$, by the Co-ordinate Ascent Variational Inference (CAVI) scheme presented in Chapter 4. In order to initialise this algorithm, the number of factors Q must be selected, and so a set of auxiliary traits were generated at random and were subject to Principal Components Analysis (PCA) [Tipping and Bishop, 1999]. This PCA indicated that the first four principal components capture 94% of variance in the auxiliary dataset, while eight principal components capture 99.9%. Given that ancestral reconstruction is the objective, and that the Automatic Relevance Determination (ARD) hyper-parameters α_j will automatically deflate superfluous factors to insignificance, $Q = 8$ was chosen for the model.

Two alternative initialisations for CAVI were then considered, initialising the loading at the first Q principal components, referred to as P-PLVM, and at the VARIMAX rotation of those components (V-PLVM), with the model maximising the log evidence lower bound ELBO(q) (see Appendix B.2) at convergence being selected for ancestral reconstruction. Here, CAVI is adjudged to have converged when ELBO(q) increases by less than 10^{-2} from one iteration to the next. As can be seen in Figure 5.6, it is V-PLVM which maximises ELBO(q) after approximately 6000 iterations, and so this is the model selected for the ancestral reconstruction of bat echolocation calls

A reconstruction of the echolocation calls of bats Most Recent Common Ancestor (MRCA), based on the sample of Mexican bat echolocation call recordings and

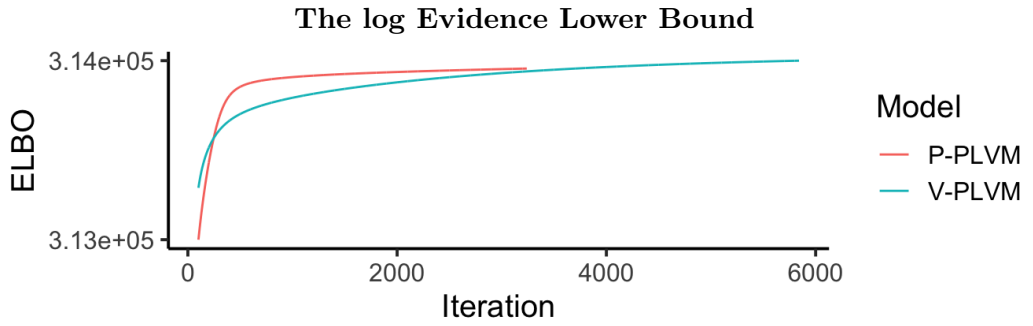
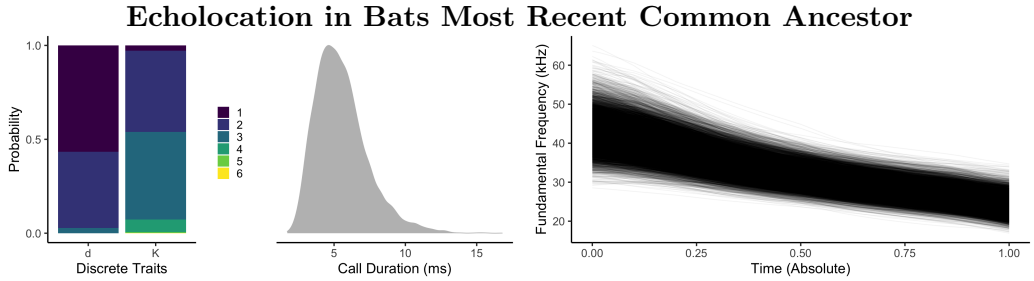


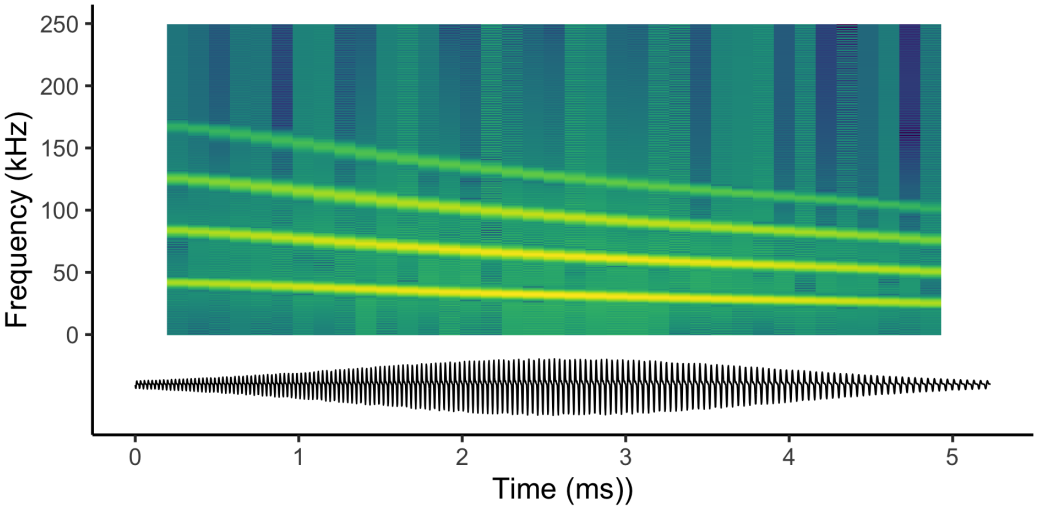
Figure 5.6: A comparison of log Evidence Lower Bounds (ELBO (q)) for the P-PLVM and V-PLVM models for the evolution of bat echolocation. It is V-PLVM that maximises the ELBO.

the phylogeny $\mathcal{P}$ is presented in Figure 5.7. This analysis suggests that the ancestral bat echolocation call was a multi-harmonic, broadband sweep from approximately 40 to 30 kHz, lasting 3-8 ms. Finally, the fundamental frequency component was dominant with a probability greater than 0.5.

The V-PLVM is not restricted to analysing the ancestral trait at the root of $\mathcal{P}$ only. Trait distributions have been defined at every internal node of $\mathcal{P}$, such that the distributions of traits for all extant and ancestral taxa can be explored. This allows the identification of intermediate echolocation calls, those that may have existed as bats evolved from using one call structure to another. Furthermore, it provides a sense check for the model. Should a reconstructed echolocation call be unreasonable (i.e. physically impossible for the larynx to produce) for any node on $\mathcal{P}$, this would call any conclusions based on the model into question. A representation of the implied echolocation call parameters at each internal node of $\mathcal{P}$ under the V-PLVM is presented in Figure 5.8. Here, the shape and position of frequency components are given by the MAP fundamental frequency, while the length of components along the x-axis is proportional to MAP duration. The probability that component k is present in the call, that is $p\left(K(\mathbf{p}_i) \geq k\right)$, is proportional to the opacity of the line used to represent that component, while line width is proportional to the probability of the component being dominant. Call representations at terminal nodes can be thought of as the representative call for that species, with edges (which are in no way representative of evolutionary time between calls) illustrating the evolutionary path taken by each call. Those internal calls that are labelled represent MRCA for that particular family. Furthermore, an interactive web application allowing the exploration and playback of ancestral echolocation calls throughout bats life history can be found at https://jpmeagher.shinyapps.io/test_reconstruction/.



(a) Ancestral Reconstruction of Echolocation Call Parameters



(b) Reconstructed Call Recording and Spectrogram

Figure 5.7: The ancestral reconstruction of echolocation in the most recent common ancestor for the sample of Mexican bat species. Sub-plot (a) presents the posterior distribution of call parameters while (b) presents a hypothetical call recording and spectrogram for this bat. The call was simulated given the MAP call duration and fundamental frequency, assuming a Gaussian amplitude envelope and random phase for each component where $A_k(t) \propto (1 + p\,(d = k))\,p\,(K = k)$.

Q	ℓ_j	κ_j	τ_j
1	2.55 (1.52, 3.62)	1.00 (0.88, 1.00)	0.040 (0.038, 0.043)
2	4.90 (3.41, 7.82)	1.00 (0.95, 1.00)	0.006 (0.006, 0.006)
3	3.59 (2.14, 5.42)	0.96 (0.81, 1.00)	0.005 (0.004, 0.005)
4	3.34 (2.09, 5.35)	1.00 (0.89, 1.00)	0.003 (0.002, 0.003)
5	7.95 (4.94, 11.77)	1.00 (0.97, 1.00)	0.000 (0.000, 0.000)
6	3.48 (2.27, 5.78)	1.00 (0.88, 1.00)	0.004 (0.003, 0.004)
7	5.70 (3.44, 8.38)	1.00 (0.94, 1.00)	0.002 (0.002, 0.002)
8	3.04 (1.76, 4.81)	1.00 (0.89, 1.00)	0.029 (0.028, 0.031)

Table 5.2: MAP estimates and intervals of 90% posterior density for phylogenetic hyper-parameters of the V-PLVM.

An examination of the two unobserved descendants of the sample MRCA reveals a point of particular interest. As may be expected, these echolocation calls are very similar; however, the most probable dominant component for the MRCA of Vespertilionidae and Molossidae is the first, while for the MRCA of Mormoopidae and Phyllistomidae it is the second. This suggests that early bat species separated based on a preference for one frequency component over the other.

As discussed in Chapter 4, interpretation of phylogenetic hyper-parameters and loading can be challenging, particularly when signal over the phylogeny is low. Nonetheless, for V-PLVM, the inferred phylogenetic hyper-parameters and intervals of 90% posterior density are presented in Table 5.2, while the loading is illustrated in Figure 5.9.

Consider first the intra-taxon variation τ_j. In every case, this is less than 0.05, indicating that the intra-taxon variation for echolocation calls is low and so variation in the factors over the phylogeny must be described by heritability κ_j and phylogenetic length-scale ℓ_j. Given that $\kappa_j \approx 1$ in every case bar one ($\kappa_3 \approx 0.96$), there is strong heritability for factors over the phylogeny. Finally, consider ℓ_j. Although care must be taken when interpreting the value of the phylogenetic length-scale, in this case, intra-taxon variation is low while heritability is high. Therefore, any variation over the phylogeny must be modelled by ℓ_j. Thus, more weight can be placed on the interpretation of its value. To this end, note that for short time scales, Brownian Motion with unit variance is well approximated by a unit variance Ornstein-Uhlenbeck process with a length scale of 2. Therefore, large values for ℓ_j (that is $\ell_j > 2$), indicate the presence of a strong phylogenetic signal in the factors over $\mathcal{P}$. That is to say, factors for closely related taxa are more strongly correlated than would be expected under a Brownian Motion model for factor evolution.

An examination of the inferred loading may shed further light on the workings

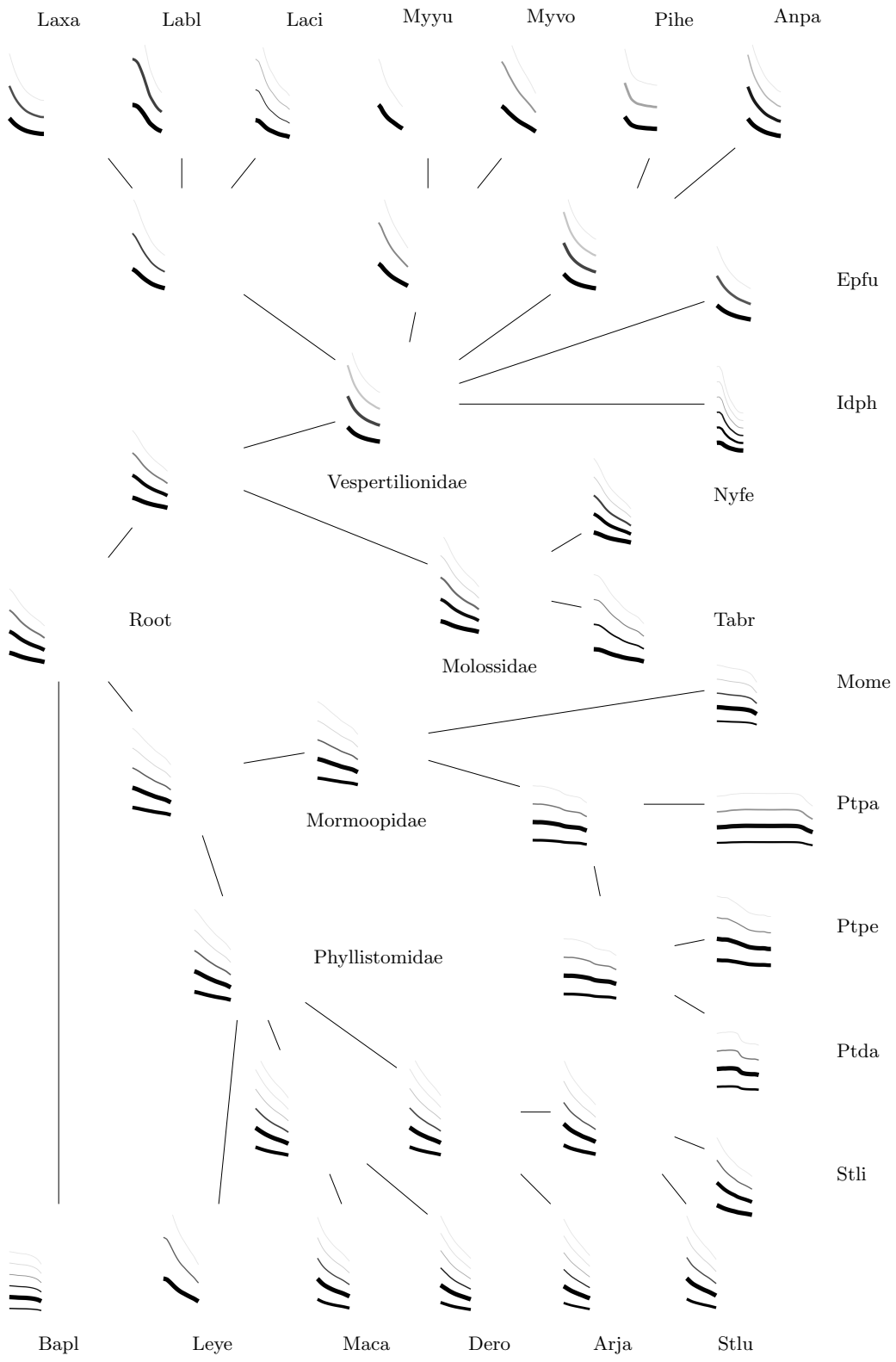


Figure 5.8: The Evolution of Bat Echolocation.

of this model, illustrating the structure of independent evolutionary features. For example, the first factor links a flattening and raising of the fundamental frequency with a reduction in the harmonic order, the second factor links the introduction of a broadband sweep in the latter part of an otherwise constant frequency call to a lengthening in call duration, and the eighth factor links flattening and lowering of the fundamental frequency to a decreased probability of the first component being dominant.

Finally, the free observation noise parameters associated with the scaled duration and fundamental frequency traits are $\lambda_T \approx 2.06$ and $\lambda_f \approx 3 \times 10^5$ respectively. This indicates that approximately half the variation in call duration is independent of the factors, while fundamental frequency curves are modelled with very high precision. These results imply that call duration is variable, even within species, while high precision measurements of the fundamental frequency are to be expected for traits that have been smoothed and registered.

5.4 Discussion

This chapter has presented a harmonic model for bat echolocation calls and performed ancestral reconstruction for a set of call features, given the phylogeny $\mathcal{P}$. This represents a novel application of the generalised PLVM, allowing a previously unattainable insight into the evolutionary dynamics of bat echolocation. Based on an analysis of 1816 echolocation call recordings sampled from 22 species of extant bat, conclusions regarding the structure of ancestral bat echolocation calls over $\mathcal{P}$ can be drawn and hypothesised call recordings synthesised, allowing the playback ancestral bat calls.

The Most Recent Common Ancestor (MRCA) for this sample of Mexican bats employed multi-harmonic, broadband sweep from approximately 40 to 30 kHz, lasting 3-8 ms. In the MRCA of the Vespertilionidae and Molossidae families, The first harmonic was most probably dominant, while in the MRCA of Mormoopidae and Phyllistomidae, it was most probably the second. These conclusions are in broad agreement with those of Collen [2012] and Schnitzler et al. [2004], however, results presented here are based on statistical models for both echolocation and trait evolution.

There are several particularly pleasing aspects to these results. The first is the clarity with which ancestral echolocation calls are reconstructed. The structure of frequency components over the phylogeny is being modelled directly, meaning that the output can be interpreted without any post-processing, a feature that

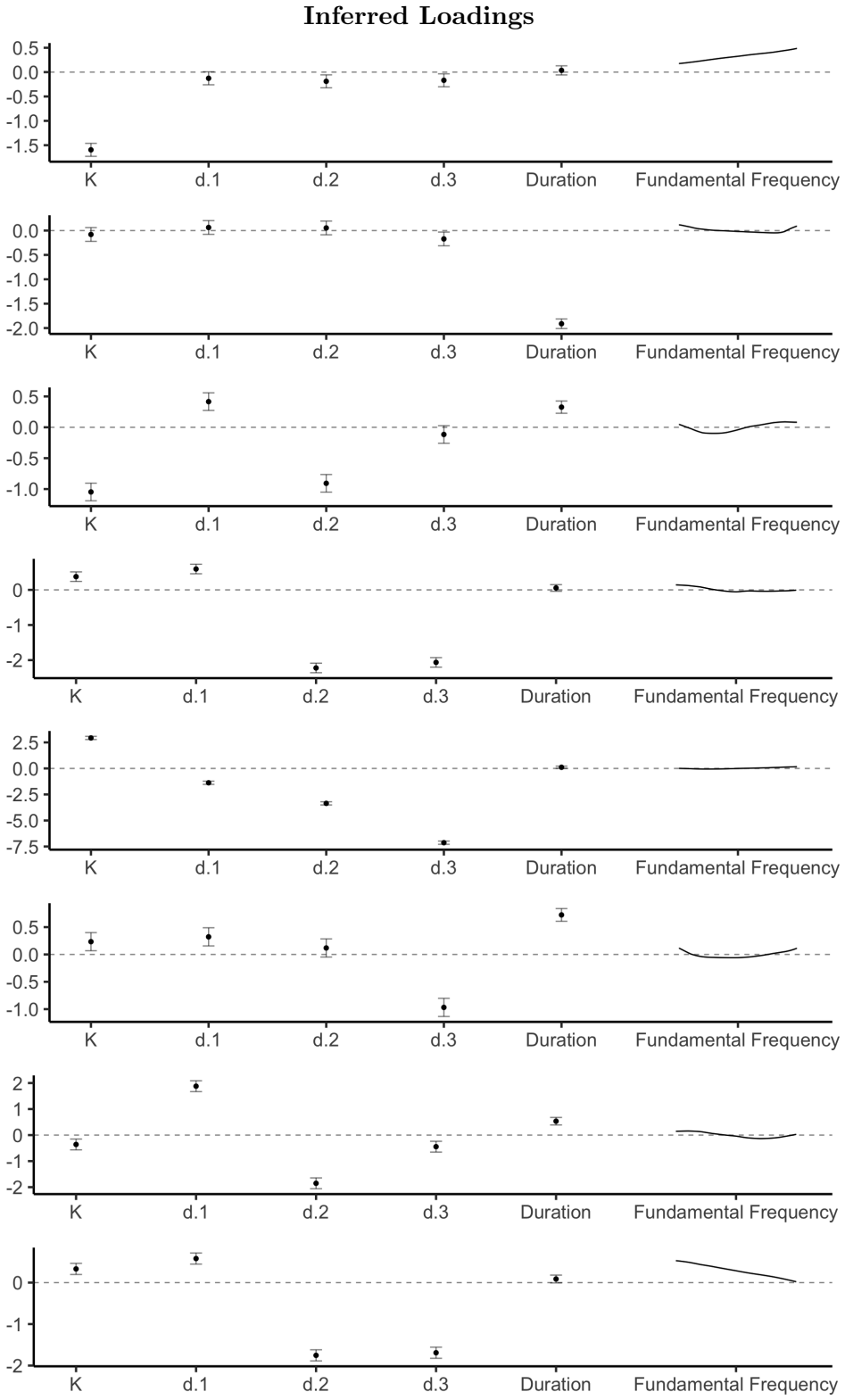


Figure 5.9: V-PLVM inferred loadings.

Collen [2012] could not achieve. Secondly, using the reconstructed echolocation call features to propose ancestral echolocation recordings, which can then be listened to, is straightforward given the model defined by (5.1) and (5.2). Finally, inference, for what could be considered a large and unwieldy dataset in the context of phylogenetic comparative analysis, was performed by CAVI for a generalised PLVM in the order of tens of minutes. This demonstrates that the true potential of generalised PLVM lies in the analysis of much larger datasets than are studied here. Applying this method to the Echobank of bat echolocation calls [Collen, 2012], which contains echolocation call recordings for 410 species of extant bat, may allow new insight into the development of echolocation in Chiroptera.

Some areas for further research do remain. As discussed in sub-section 5.2.4, the raw output of the harmonic model for echolocation does require a post-hoc correction to be applied prior to analysis. Developing an approach to fundamental frequency extraction for which this correction is not required represents an avenue worth exploring. Furthermore, linking this model to the estimation of instantaneous frequency in multi-component signals, as described by Olhede and Walden [2005] and DiCecco et al. [2013], may provide a general approach to precise time-frequency analysis.

For the ancestral reconstruction of bat echolocation, however, this model appears to perform as well as could be hoped, given the dataset available. This presents the most exciting avenue for future work stemming from this thesis, the application of a generalised PLVM, as described here, to a much more diverse set of calls, allowing scientific conclusions on bats ancestral echolocation call to be drawn. This remains a priority for future work.

Chapter 6

Final Remarks

The development and application of novel statistical methodology have driven advances in both Phylogenetics and Phylogenetic Comparative Methods (PCMs) for decades [Felsenstein, 1973, 1985; Hansen, 1997; Drummond et al., 2002; Suchard et al., 2018]. Models for the mutation of molecular sequences elucidate shared ancestry with a degree of certainty that was previously unattainable [Suchard et al., 2018; Amador et al., 2018], while PCMs allow the evolution of ever more complex phenotypes to be studied [Hadjipantelis et al., 2012; Cybis et al., 2015; Tolkoff et al., 2017]. This thesis makes a significant contribution to the latter of these endeavours. The methods developed here may yet prove widely applicable across the disciplines of Evolutionary Biology, Morphometrics, and Bioacoustics, providing new and profound insights into the development of life on earth.

From a human's perspective, echolocation represents a fascinating natural phenomenon, being so far removed from our own lived experiences as to be near incomprehensible. Despite this, the principles which underpin the process have come to be well-understood [Denny, 2007; Fenton et al., 2016]. What has remained much more mysterious, is the path by which this characteristic developed in Chiroptera [Simmons and Stein, 1980; Schnitzler et al., 2004; Collen, 2012], despite the consensus that has emerged on the structure and timing of the order's ancestral relationships [Teeling et al., 2000, 2005; Eick et al., 2005; Tsagkogeorga et al., 2013; Amador et al., 2018]. Thus, the objective of this thesis was to shed light on this mystery through the development of novel techniques for the phylogenetic comparative analysis acoustic signals, and by so doing, reconstruct the calls of ancestral bats with a degree of certainty that was previously unavailable.

Given that echolocation is a continuous process in time and, as such, is a Function-Valued Trait (FVT) [Kirkpatrick and Heckman, 1989; Meyer and Kirk-

patrick, 2005], the Phylogenetic Gaussian Process Regression (PGPR) framework provided a useful, though limited, approach to modelling its evolution [Jones and Moriarty, 2013]. The framework suffered from several shortcomings. It was formulated for a single FVT, assuming a separable phylogeny-trait covariance structure. This implied trait measurements were free of independent observation noise. Furthermore, existing inference schemes for models of trait evolution either approximated PGPR [Hadjipantelis et al., 2013], or failed to take a fully Bayesian approach to inferring the phylogenetic covariance structure [Cybis et al., 2015; Tolkoff et al., 2017]. These limitations made such methods unsuitable for ancestral reconstruction. Thus, the first contribution of this thesis was to introduce the Phylogenetic Latent Variable Model (PLVM), offering a new perspective on the PGPR framework which addressed each of these challenges.

The PLVM provides a flexible approach to modelling the evolution of a FVT over a known phylogeny. Its construction, which is similar to Factor Analysis [Bartholomew et al., 2011; Lopes, 2014] and Phylogenetic Factor Analysis [Tolkoff et al., 2017] allowed separability of the phylogeny-trait covariance function to be relaxed. Not only that, but it also facilitated the development of the first fully Bayesian approach to PGPR, which inferred both the trait and phylogeny covariance functions and incorporated repeated measurements for extant taxa. A Markov Chain Monte Carlo (MCMC) inference scheme for doing so was developed in Chapter 3, based on state-of-the-art sampling techniques for Gaussian processes [Murray et al., 2010; Murray and Adams, 2010; Yu and Meng, 2011; Filippone et al., 2013]. This inference scheme relied on the efficient computation of both the pruned likelihood and conditional distributions for general Gauss-Markov processes over a phylogeny, each a novel contribution in its own right. This algorithm allowed the extension of both Brownian Motion (BM) and Ornstein-Uhlenbeck (OU) models for trait evolution to the Phylogenetic Mixed Model (PMM) [Housworth et al., 2004] while also modelling intra-taxon variation. The planned release of a statistical software package implementing this likelihood computation will aid its broader dissemination, allowing more researchers to fit flexible models for trait evolution and maximising the impact of this contribution. While the approach presented in Chapter 3 did offer excellent ancestral reconstruction and uncertainty quantification, it suffered from significant shortcomings. The PLVM considered the evolution of a single FVT only and while inference scaled linearly with observed individuals, it scaled cubically with the number of trait measurements. It was these issues that motivated development of the practical approach to ancestral reconstruction that followed.

The generalised PLVM, presented in Chapter 4, modelled the evolution of

any collection of traits taking discrete or continuous values within a single framework. Furthermore, the development of a Co-ordinate Ascent Variational Inference (CAVI) scheme provided a flexible and efficient approach to approximate Bayesian inference for the model [Jordan et al., 1999; Bishop, 2006; Blei et al., 2017]. This approach offers significant theoretical and practical advantages over the PCMs proposed by Hadjipantelis et al. [2013], Cybis et al. [2015], and Tolkoff et al. [2017], particularly with respect to ancestral reconstruction. It represents the first method for evolutionary inference on a collection of traits that includes FVTs alongside scalar-valued discrete and continuous traits, allows further relaxation of separability for the phylogeny-trait covariance function governing PGPR, and easily accommodates repeated measurements of extant taxa. Finally, CAVI performs approximate Bayesian inference in a fraction of the time required by MCMC inference schemes, allowing the method scale to datasets consisting of thousands of observations.

Though the generalised PLVM represents a complete solution for phylogenetic comparative analysis, there do remain many opportunities for further research. For example, the method could be adapted to the case when some or all extant taxa are missing one or more trait measurements. Given that the generalised PLVM defines a probabilistic model for evolution, it provides a natural approach to this problem, to which CAVI could be adapted in a reasonably straightforward manner. In a similar vein, the model could also include trait measurements for ancestral taxa, should they be available. Because the generalised PLVM infers correlation structure over a set of traits, including some elements of this set for ancestral taxa may allow the reconstruction of remaining traits with a much higher degree of certainty than would otherwise be possible. Furthermore, this would provide valuable data for the model to fit. As an illustrative example, consider the relationship between a bats body mass and its echolocation call. The echolocation calls of more massive bats tend to be at lower frequencies than for those with less body mass [Collen, 2012]. Fitting a generalised PLVM which includes a measurement of body mass alongside the feature representation of bat echolocation should allow this correlation to be quantified. When this is the case, estimating the body mass of ancestral bats from the fossil record and including this ancestral trait in the generalised PLVM may result in more certainty on the reconstruction of ancestral bat echolocation calls.

A more speculative direction to explore would be the development of a robust approach to evolutionary inference with the PLVM. Biological data is often non-Gaussian, with empirical trait distributions often having heavy tails [Elliot and Mooers, 2014]. In such situations, a small number of outlying observations may severely bias any ancestral reconstruction based on Gaussian models for trait evo-

lution. Modelling data as a stable process presents one approach to this problem, applied to the ancestral reconstruction of eutherian mammal's body size [Elliot and Mooers, 2014]. Such models allow mean square continuity for the stochastic process describing trait evolution over the phylogeny to be relaxed, and as such offer methods that are robust to outliers [Nolan, 2012]. Assuming that latent variables in the PLVM come from such a distribution may provide an even more flexible class of models for trait evolution, although inference may prove challenging. Alternatively, the application of recently proposed methods for robust Bayesian inference to the PLVM may yield a similar effect [Futami et al., 2017; Knoblauch et al., 2019; Nakagawa and Hashimoto, 2020].

As a final remark on the PLVM, note that this PCM is conditional on a phylogeny, as are those proposed by Hadjipantelis et al. [2013], Cybis et al. [2015], and Tolkoff et al. [2017]. There is no doubt that such an approach is justified. Phenotypes result from the interaction between an organism's genotype and its environment, which is to say that they are a function of both the genes and environmental conditions [Campbell et al., 1997]. This plasticity means that closely related species may exhibit vastly different phenotypes, making phylogenetic inference challenging. On the other hand, the genotype is passed directly from one generation to the next, preserving far more of the phylogenetic signal. Thus, Bayesian inference for models of gene mutation represents the current state-of-the-art approach to inferring phylogenies [Suchard et al., 2018; Amador et al., 2018], which can then provide the structure required for phylogenetic comparative analysis, as described by both Cybis et al. [2015] and Tolkoff et al. [2017]. Despite this clear justification for separating phylogenetic inference from the phenotype, future research may challenge this reasoning. Even after molecular analysis, there does remain uncertainty on the phylogeny. The logic of phylogenetic comparative analysis and ancestral reconstruction assumes that traits do carry some phylogenetic signal; otherwise, each individual would represent an independent observation, making the reconstruction of common ancestors impossible. Although the phylogenetic signal-to-noise ratio may be low for any given phenotype, including multiple phenotypes within a single analysis may allow this ratio to be improved. Thus, it is possible that the generalised PLVM, which models evolution for extensive collections of traits, may be adapted to provide rigorous methods for phylogenetic inference. Furthermore, the pruned likelihood for generalised Gauss-Markov processes could allow such methods to look beyond a BM model for trait evolution. Adapting the MCMC and Variational inference schemes proposed here to this problem even opens up the possibility of developing a unified approach to phylogenetic inference, incorporating both molecular and phenotypic

data.

Applying the PLVM to bat echolocation posed a particularly unusual set of challenges for evolutionary inference. Echolocation calls are non-stationary periodic signals. As such, their comparative analysis required a time-frequency representation of the call. Furthermore, this representation had to cohere with a linear model for evolution. Although echolocation calls are well known to be multi-harmonic signals [Fenton et al., 2016], exploiting this structure for their comparative analysis had proven challenging. A harmonic model for bat echolocation, such as that developed in Chapter 5, provides a straightforward characterisation of the echolocation call, based on models for human speech [Shi et al., 2019]. Though inference is challenging, as it is for time-frequency analysis in general [Olhede and Walden, 2005; Hlawatsch and Auger, 2008; Huang et al., 2009; DiCecco et al., 2013], this model offers a new perspective on bat echolocation and rigorously defines a feature representation of the echolocation call that is straightforward to interpret. As discussed in sub-section 5.2.4, there remain many strategies for developing this harmonic model further and doing so may well facilitate the development of new methods for echolocation call classification [Redgwell et al., 2009; Stathopoulos et al., 2018; Mac Aodha et al., 2018]. Given that bats have been identified as a bioindicator species [Jones et al., 2009], bat call classification may offer a low-cost approach to biodiversity monitoring. As such, methods for doing so accurately and efficiently are essential. Existing methods do not model the echolocation call itself, thus, classifiers based on the harmonic model may prove to generalise more easily to data from many different sources.

A single overarching goal motivated each of the methodological contributions outlined above, that was, the ancestral reconstruction of bat echolocation. Their application to this problem, which considered a sample of 1816 call recordings for 22 species of Mexican bat [Stathopoulos et al., 2018], concluded that extant bats most recent common ancestor employed a multi-harmonic call with at least two frequency components. These components consisted of a broadband sweep from approximately 40 to 30 kHz, lasting between 3 and 8 ms. Either the first or second harmonic was dominant, although there was a greater than 50% chance that it was the first. These findings contradict the conclusions of Simmons and Stein [1980], who posited that the ancestral bat call was a narrowband multi-harmonic signal; however, they are in broad agreement with those of Schnitzler et al. [2004] and Collen [2012]. An exceptionally satisfying aspect of the analysis presented here is that hypothetical call recordings are easy to produce, allowing playback of the estimated calls for long-extinct bats. Though this analysis does offer new insight

into the evolution of bat echolocation, it was somewhat limited in its scope. The study considered 22 species spread over 5 families [Stathopoulos et al., 2018], though over 1000 extant species and 21 families are currently recognised [Simmons, 2005]. Thus, the conclusions drawn are subject to the caveat that a more representative sample may offer a far greater degree of certainty on the structure of ancestral bat echolocation calls. Despite this being the case, these results do concur with those of [Schnitzler et al., 2004] and [Collen, 2012]. Furthermore, the rigorous approach to uncertainty quantification taken here makes it reasonable to expect that a more extensive study would still produce results that lie within the regions of high posterior density identified here. Such a study is currently in preparation, applying the methods developed here to the Echobank, a database containing call recordings for the 410 species of extant bat [Collen, 2012]. Thus, while the Bayesian approach to ancestral reconstruction has already provided novel insight into the evolution of bat echolocation, there remains the promise of yet more to come.

Appendix A

Tree Traversal Algorithms

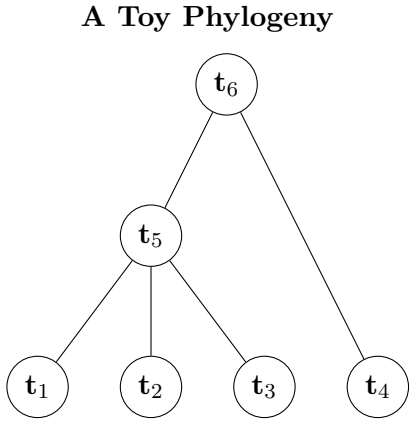


Figure A.1: A phylogeny with $N = 4$ terminal nodes and $M = 2$ internal nodes, the positions for which are denoted $\mathbf{t}_i$ for $i = 1, \ldots, N + M$.

A.1 Pruned Likelihood Calculation

Consider the trait-phylogeny separable Gaussian process over $\mathcal{T} \times \mathcal{X}$ such that

$$f(\mathbf{t}, \mathbf{x}) \sim \mathcal{GP}\left(0, k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right) k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{A.1}$$

where $\mathcal{T}$ denotes a phylogeny with N terminal and M internal nodes, and $k_{\mathcal{T}}(\cdot, \cdot)$ is the covariance function for a first order Markov process. The model for observations at terminal nodes is given by

$$\mathbf{Y}_{n\cdot} \sim \mathcal{N}\left(\mathbf{f}_n, \mathbf{\Psi}\right),$$

for $n = 1, \ldots, N$, where $\mathbf{f}_n = \left(f\left(\mathbf{t}_n, \mathbf{x}_1\right), \ldots, f\left(\mathbf{t}_n, \mathbf{x}_D\right)\right)^\top$ for registered indices $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \ldots, D$, and $\boldsymbol{\Psi}$ is a diagonal covariance matrix.

Let $k_{i,j} \equiv k_{\mathcal{T}}\left(\mathbf{t}_i, \mathbf{t}_j\right)$, and $k_i \equiv k_{i,i}$, then define

$$\phi_i \equiv k_{i,\mathrm{pa}(i)} k_{\mathrm{pa}(i)}^{-1}, \tag{A.2}$$

$$\eta_i \equiv k_i - k_{i,\mathrm{pa}(i)} k_{\mathrm{pa}(i)}^{-1} k_{\mathrm{pa}(i),i}, \tag{A.3}$$

the conditional weighted mean and variance due to the process over $\mathcal{T}$, such that

$$\mathbf{f}_i | \mathbf{f}_{\mathrm{pa}(i)} \sim \mathcal{N}(\phi_i \mathbf{f}_{\mathrm{pa}(i)}, \eta_i \mathbf{K}_{\mathcal{X}}), \tag{A.4}$$

for $i = 1, \ldots, N + M - 1$, where $\mathbf{K}_{\mathcal{X}}$ is the Gram matrix of $k_{\mathcal{X}}\left(\cdot, \cdot\right)$.

It can be shown that

$$\mathbf{Y}_n | \mathbf{f}_{\mathrm{pa}(n)} \sim \mathcal{N}\left(\phi_n \mathbf{f}_{\mathrm{pa}(n)}, \eta_n \mathbf{K}_{\mathcal{X}} + \boldsymbol{\Psi}\right), \tag{A.5}$$

for $n = 1, \ldots, N$, and the distribution of the root node is assumed to be

$$\mathbf{f}_{N+M} \sim \mathcal{N}(\mathbf{0}, k_{N+M} \mathbf{K}_{\mathcal{X}}). \tag{A.6}$$

Let $\boldsymbol{\Lambda}_n = \left(\eta_n \mathbf{K}_{\mathcal{X}} + \boldsymbol{\Psi}\right)^{-1}$ for $n = 1, \ldots, N$, and $\boldsymbol{\Lambda}_i = \eta_i^{-1} \mathbf{K}_{\mathcal{X}}^{-1}$ for $i = N + 1, \ldots, N + M - 1$. If $\mathbf{K}_{\mathcal{X}} = \mathbf{W}\mathbf{W}^\top + \epsilon \mathbf{I}_D$, where $\epsilon > 0$ is some small constant which prevents kx from being ill conditioned. $\boldsymbol{\Lambda}_n$ can be calculated efficiently by the Woodbury identity [Petersen and Pedersen, 2012, sec 3.2.2]. Reformulating Equation (3.8) here for clarity, the joint distribution over observed trait values and the latent phylogenetic GP can be expressed as

$$p\left(\mathbf{Y}, \mathbf{f}_{N+1}, \ldots \mathbf{f}_{N+M}\right) = \left(\prod_{n=1}^{N} p\left(\mathbf{Y}_{n\cdot} | \mathbf{f}_{\mathrm{pa}(n)}\right)\right) \left(\prod_{i=N+1}^{N+M} p\left(\mathbf{f}_i | \mathbf{f}_{\mathrm{pa}(i)}\right)\right) p\left(\mathbf{f}_{N+M}\right),$$

Given (3.8), consider the marginal density in the case of the toy example in figure 3.1, that is

$$\begin{aligned} p(\mathbf{Y}_{1\cdot}, \mathbf{Y}_{2\cdot}, \mathbf{Y}_{3\cdot}, \mathbf{Y}_{4\cdot}) &= \int \left(\int \left(\prod_{i=1}^{3} p(\mathbf{Y}_{i\cdot} | \mathbf{f}_5) \right) p(\mathbf{f}_5 | \mathbf{f}_6) d\mathbf{f}_5 \right) p(\mathbf{Y}_{4\cdot} | \mathbf{f}_6) p(\mathbf{f}_6) d\mathbf{f}_6, \\ &= \int p(\mathbf{Y}_{1\cdot}, \mathbf{Y}_{2\cdot}, \mathbf{Y}_{3\cdot} | \mathbf{f}_6) p(\mathbf{Y}_{4\cdot} | \mathbf{f}_6) p(\mathbf{f}_6) d\mathbf{f}_6. \end{aligned}$$

From this example, (3.10) can be deduced, that is

$$p(\{\mathbf{Y}\}_h^{post}|\mathbf{f}_{\mathrm{pa}(h)}) = \int p(\{\mathbf{Y}\}_h^{post}|\mathbf{f}_h)p(\mathbf{f}_h|\mathbf{f}_{\mathrm{pa}(h)})d\mathbf{f}_h,$$
$$= \int \left(\prod_{i\in \mathrm{ch}(h)} p(\{\mathbf{Y}\}_i^{post}|\mathbf{f}_h) \right) p(\mathbf{f}_h|\mathbf{f}_{\mathrm{pa}(h)})d\mathbf{f}_h, \qquad \text{(A.7)}$$

where $\{\mathbf{Y}\}_h^{post}$ denotes the set of observed traits descendant from and including $\mathbf{t}_h$, and $\mathrm{ch}\,(h)$ is the set of children for $\mathbf{t}_h$. Thus, the marginal density $p(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ can be calculated by a post-order traversal of $\mathcal{T}$, which proceeds from terminal nodes to the root, calculating the partial mean vector, $\mathbf{m}_i^{post}$; precision matrix, $\mathbf{\Lambda}_i^{post}$; and scaling constant, c_i; associated with distribution at each $\mathbf{t}_i \in \mathcal{T}$; for $i = 1, \ldots, N+M$, "pruning" away descendant nodes. The marginal density calculated by this algorithm is called the *pruned likelihood*.

Initialise the algorithm by setting $\mathbf{m}_n^{post} = \mathbf{Y}_{n\cdot}, \mathbf{\Lambda}_n^{post} = \mathbf{\Lambda}_n$, and $c_n = 1$ for $n = 1, \ldots, N$. For $h = N+1, \ldots, N+M$ and $i \in \mathrm{ch}\,(h)$

$$p(\{\mathbf{Y}\}_i^{post}|\mathbf{f}_h) = \mathcal{N}\left(\mathbf{m}_i^{post}|\phi_i\mathbf{f}_h, \left(\mathbf{\Lambda}_i^{post}\right)^{-1}\right),$$
$$= \phi_i^{-D}\mathcal{N}\left(\mathbf{f}_h|\phi_i^{-1}\mathbf{m}_i^{post}, \left(\phi_i^2\mathbf{\Lambda}_i^{post}\right)^{-1}\right),$$
$$= \phi_i^{-D}\mathcal{N}\left(\mathbf{f}_h|\tilde{\mathbf{m}}_i, \tilde{\mathbf{\Lambda}}_i^{-1}\right),$$

where mean of the Gaussian distribution has been rearranged [Petersen and Pedersen, 2012, eq 357] and

$$\tilde{\mathbf{m}}_i = \phi_i^{-1}\mathbf{m}_i^{post},$$
$$\tilde{\mathbf{\Lambda}}_i = \phi_i^2\mathbf{\Lambda}_i^{post}.$$

It can be shown that

$$\prod_{i\in \mathrm{ch}(h)} p(\{\mathbf{Y}\}_i^{post}|\mathbf{f}_h) = \prod_{i\in \mathrm{ch}(h)} \phi_i^{-D}\mathcal{N}\left(\mathbf{f}_h|\tilde{\mathbf{m}}_i, \tilde{\mathbf{\Lambda}}_i^{-1}\right),$$
$$= c_h \left(\prod_{i\in \mathrm{ch}(h)} \phi_i^{-D} \right) \mathcal{N}\left(\mathbf{f}_h|\mathbf{m}_h^{post}, \left(\sum_{i\in \mathrm{ch}(h)} \tilde{\mathbf{\Lambda}}_i\right)^{-1}\right) \qquad \text{(A.8)}$$

by considering the product of Gaussian densities [Petersen and Pedersen, 2012, eq 371].

In this case

$$\mathbf{m}_h^{post} = \left(\sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \right)^{-1} \left(\sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \tilde{\mathbf{m}}_i \right), \tag{A.9}$$

is the partial mean vector for $\mathbf{f}_h$ given $\{\mathbf{f}\}_h^{post}$.

The associated scaling constant is given by

$$c_h = \prod_{i=2}^{d_j - 1} \mathcal{N}\left(\tilde{\mathbf{m}}_{\text{ch}(h)_i} | \mathbf{m}_i^c, \Sigma_i^c \right),$$

where

$$\begin{aligned}
\mathbf{m}_i^c &= \left(\sum_{j=1}^{i-1} \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_j} \right)^{-1} \left(\sum_{j=1}^{i-1} \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_j} \tilde{\mathbf{m}}_{\text{ch}(h)_j} \right), \\
\Sigma_i^c &= \left(\sum_{j=1}^{i-1} \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_j} \right)^{-1} + \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_i}^{-1}, \\
&= \left(\sum_{j=1}^{i-1} \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_j} \right)^{-1} \left(\sum_{j=1}^{i-1} \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_j} + \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_i} \right) \tilde{\mathbf{\Lambda}}_{\text{ch}(h)_i}^{-1}.
\end{aligned}$$

Substituting (A.8) into (A.7) yields

$$\begin{aligned}
p(\{\mathbf{Y}\}_h^{post} | \mathbf{f}_{\text{pa}(h)}) &= \int \mathcal{N}\left(\mathbf{f}_h | \phi_h \mathbf{f}_{\text{pa}(h)}, (\mathbf{\Lambda}_h)^{-1} \right) \times \\
&\qquad c_h \left(\prod_{i \in \text{ch}(h)} \phi_i^{-D} \right) \mathcal{N}\left(\mathbf{f}_h | \mathbf{m}_h^{post}, \left(\sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \right)^{-1} \right) d\mathbf{f}_h, \\
&= c_h \left(\prod_{i \in \text{ch}(h)} \phi_i^{-D} \right) \mathcal{N}\left(\mathbf{m}_h^{post} | \phi_h \mathbf{f}_{\text{pa}(h)}, \left(\mathbf{\Lambda}_h^{post} \right)^{-1} \right),
\end{aligned} \tag{A.10}$$

where

$$\begin{aligned}
\mathbf{\Lambda}_h^{post} &= \left(\mathbf{\Lambda}_h^{-1} + \left(\sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \right)^{-1} \right)^{-1}, \\
&= \mathbf{\Lambda}_h \left(\mathbf{\Lambda}_h + \sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \right)^{-1} \left(\sum_{i \in \text{ch}(h)} \tilde{\mathbf{\Lambda}}_i \right),
\end{aligned}$$

The algorithm is completed by noting that $\phi_{N+M} = 1$ and $\eta_{N+M} = k_{N+M}$ and so the pruned likelihood is expressed as

$$p(\mathbf{Y}) = \left(\prod_{h=1}^{N+M} \phi_h^{-D} c_h \right) \mathcal{N}\left(\mathbf{m}_{N+M}^{post} | \mathbf{0}, \left(\mathbf{\Lambda}_{N+M}^{post} \right)^{-1} \right).$$

A.2 Pruned Conditional Distribution

Given that

$$p(\mathbf{f}_* | \mathbf{Y}, \mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}) = \mathcal{N}(\mathbf{f}_* | \mathbf{m}_*, \mathbf{K}_*), \tag{A.11}$$

for $\mathbf{f}_* = (f(\mathbf{t}_*, \mathbf{x}_1), \ldots, f(\mathbf{t}_*, \mathbf{x}_D))$ for some ancestral position $\mathbf{t}_* \in \mathcal{T}$, where $\mathcal{T}$ denotes a phylogeny with N terminal and M internal nodes, and $\mathbf{x}_i \in \mathcal{X}$ indexes the registered trait space, the following describes an algorithm computing $\mathbf{m}_*$ and $\mathbf{K}_*$ in $\mathcal{O}(N^2)$ operations.

Let $\mathbf{Y}_{-n\cdot}$ denote $\mathbf{Y}$ less it's n^{th} row. Let $\{\mathbf{Y}\}_*^{pre} = \mathbf{Y} / \{\mathbf{Y}\}_*^{post}$, where $\{\mathbf{Y}\}_h^{post}$ denotes the set of observed traits descendant from and including $\mathbf{t}_h$. This is to say that $\{\mathbf{Y}\}_*^{pre} \cup \{\mathbf{Y}\}_*^{post} = \mathbf{Y}$, $\{\mathbf{Y}\}_{N+M}^{pre} = \varnothing$, and $\{\mathbf{Y}\}_n^{pre} = \mathbf{Y}_{-n\cdot}$ for $n = 1, \ldots, N$. Suppressing notation for the dependence of $\mathbf{f}_*$ on $\{\mathbf{W}, \theta_{\mathcal{T}}, \mathbf{\Lambda}\}$, rewriting (A.11) yields

$$\begin{aligned} p(\mathbf{f}_* | \mathbf{Y}) &= p\left(\mathbf{f}_* | \{\mathbf{Y}\}_*^{post}, \{\mathbf{Y}\}_*^{pre} \right), \\ &\propto p\left(\mathbf{f}_*, \{\mathbf{Y}\}_*^{post} | \{\mathbf{Y}\}_*^{pre} \right), \\ &= p\left(\{\mathbf{Y}\}_*^{post} | \mathbf{f}_* \right) p(\mathbf{f}_* | \{\mathbf{Y}\}_*^{pre}). \end{aligned} \tag{A.12}$$

It has already been shown in (A.8) that

$$\begin{aligned} p(\{\mathbf{Y}\}_*^{post} | \mathbf{f}_*) &= \prod_{i \in \mathrm{ch}(*)} p(\{\mathbf{Y}\}_i^{post} | \mathbf{f}_*), \\ &\propto \mathcal{N}\left(\mathbf{f}_* | \mathbf{m}_*^{post}, \left(\sum_{i \in \mathrm{ch}(*)} \tilde{\Lambda}_i \right)^{-1} \right), \end{aligned}$$

and so it remains to find $p(\mathbf{f}_* | \{\mathbf{Y}\}_*^{pre})$.

Consider the expression

$$\begin{aligned} p(\mathbf{f}_* | \{\mathbf{Y}\}_*^{pre}) &= \int p\left(\mathbf{f}_*, \mathbf{f}_{\mathrm{pa}(*)} | \left\{ \{\mathbf{Y}\}_{\mathrm{sib}(*)}^{post} \right\}, \{\mathbf{Y}\}_{\mathrm{pa}(*)}^{pre} \right) d\mathbf{f}_{\mathrm{pa}(*)}, \\ &= \int p\left(\mathbf{f}_* | \mathbf{f}_{\mathrm{pa}(*)} \right) p\left(\mathbf{f}_{\mathrm{pa}(*)} | \left\{ \{\mathbf{Y}\}_{\mathrm{sib}(*)}^{post} \right\}, \{\mathbf{Y}\}_{\mathrm{pa}(*)}^{pre} \right) d\mathbf{f}_{\mathrm{pa}(*)}. \end{aligned} \tag{A.13}$$

where sib(∗) denotes the siblings of $\mathbf{t}_*$ and $\left\{\{\mathbf{Y}\}^{post}_{\text{sib}(*)}\right\}$ is the set $\left\{\{\mathbf{Y}\}^{post}_{i} : i \in \text{sib}(*)\right\}$.

The first term of (A.13) is defined in (A.4). The second term is more involved and, similarly to (A.12), can be expressed as

$$\begin{aligned} p\left(\mathbf{f}_{\text{pa}(*)}|\{\{\mathbf{Y}\}^{post}_{\text{sib}(*)}\},\{\mathbf{Y}\}^{pre}_{\text{pa}(*)}\right) &\propto p\left(\mathbf{f}_{\text{pa}(*)},\{\{\mathbf{Y}\}^{post}_{\text{sib}(*)}\}|\{\mathbf{Y}\}^{pre}_{\text{pa}(*)}\right), \\ &= p\left(\{\{\mathbf{Y}\}^{post}_{\text{sib}(*)}\}|\mathbf{f}_{\text{pa}(*)}\right) p\left(\mathbf{f}_{\text{pa}(*)},|\{\mathbf{Y}\}^{pre}_{\text{pa}(*)}\right) \\ &= \left(\prod_{j\in sib(*)} p\left(\{\mathbf{Y}\}^{post}_{j}|\mathbf{f}_{\text{pa}(*)}\right)\right) p\left(\mathbf{f}_{\text{pa}(*)}|\{\mathbf{Y}\}^{pre}_{\text{pa}(*)}\right). \end{aligned} \tag{A.14}$$

From (A.8) it can be seen that that

$$\prod_{j\in sib(*)} p\left(\{\mathbf{Y}\}^{post}_{j}|\mathbf{f}_{\text{pa}(*)}\right) \propto \prod_{j\in sib(*)} \mathcal{N}\left(\mathbf{f}_{\text{pa}(*)}|\tilde{\mathbf{m}}_j,\tilde{\Lambda}_j^{-1}\right), \tag{A.15}$$

and so, substituting (A.15) into (A.14), which in turn is substituted into (A.13),

$$\begin{aligned} p(\mathbf{f}_*|\{\mathbf{Y}\}^{pre}_{*}) \propto \int &\mathcal{N}\left(\mathbf{f}_*|\phi_*\mathbf{f}_{\text{pa}(*)},\Lambda_*^{-1}\right)\times \\ &\left(\prod_{j\in sib(*)} \mathcal{N}\left(\mathbf{f}_{\text{pa}(*)}|\tilde{\mathbf{m}}_j,\tilde{\Lambda}_j^{-1}\right)\right)\times \\ &\mathcal{N}\left(\mathbf{f}_{\text{pa}(*)}|\mathbf{m}^{pre}_{\text{pa}(*)},\left(\Lambda^{pre}_{\text{pa}(*)}\right)^{-1}\right) d\mathbf{f}_{\text{pa}(*)}, \\ \propto \int &\mathcal{N}\left(\mathbf{f}_{\text{pa}(*)}|\phi_*^{-1}\mathbf{f}_*,\left(\phi_*^2\Lambda_*\right)^{-1}\right)\mathcal{N}\left(\mathbf{f}_{\text{pa}(*)}|\hat{\mathbf{m}}_*,\hat{\Lambda}_*^{-1}\right) d\mathbf{f}_{\text{pa}(*)}, \\ \propto &\mathcal{N}\left(\phi_*^{-1}\mathbf{f}_*|\hat{\mathbf{m}}_*,\left(\phi_*^2\Lambda_*\right)^{-1}+\hat{\Lambda}_*^{-1}\right) \\ \propto &\mathcal{N}\left(\mathbf{f}_*|\mathbf{m}^{pre}_*,(\Lambda^{pre}_*)^{-1}\right), \end{aligned}$$

where the new temporary variables

$$\begin{aligned} \hat{\Lambda}_* &= \Lambda^{pre}_{\text{pa}(*)} + \sum_{j\in\text{sib}(*)} \tilde{\Lambda}_j, \\ \hat{\mathbf{m}}_* &= \hat{\Lambda}_*^{-1}\left(\Lambda^{pre}_{\text{pa}(*)}\mathbf{m}^{pre}_{\text{pa}(*)} + \sum_{j\in\text{sib}(*)} \tilde{\Lambda}_j\tilde{\mathbf{m}}_j\right), \end{aligned}$$

allow a convenient expression of the pre-order traversal partial mean vector and

precision scalar, that is

$$\mathbf{m}_*^{pre} = \phi_* \hat{\mathbf{m}}_*. \tag{A.16}$$

$$\begin{aligned} \Lambda_*^{pre} &= \left((\Lambda_*)^{-1} + \left(\phi_*^{-2} \hat{\Lambda}_* \right)^{-1} \right)^{-1}, \\ &= \Lambda_* \left(\Lambda_* + \phi_*^{-2} \hat{\Lambda}_* \right)^{-1} \phi_*^{-2} \hat{\Lambda}_* \end{aligned} \tag{A.17}$$

The quantities (A.17) and (A.16) can be calculated by traversing $\mathcal{T}$ the root to $\mathbf{t}_*$, given that $\mathbf{m}_{N+M}^{pre} = \mathbf{0}$ and $\Lambda_{N+M}^{pre} = (k_{N+M} \mathbf{K}_{\mathcal{X}})^{-1}$.

Therefore

$$\begin{aligned} p\left(\mathbf{f}_* | \mathbf{Y}\right) &\propto p\left(\{\mathbf{Y}\}_*^{post} | \mathbf{f}_* \right) p\left(\mathbf{f}_* | \{\mathbf{Y}\}_*^{pre} \right), \\ &= \mathcal{N}\left(\mathbf{f}_* | \mathbf{m}_*^{post}, \left(\sum_{i \in \mathrm{ch}(*)} \tilde{\Lambda}_i \right)^{-1} \right) \mathcal{N}\left(\mathbf{f}_* | \mathbf{m}_*^{pre}, (\Lambda_*^{pre})^{-1} \right), \\ &\propto \mathcal{N}\left(\mathbf{f}_* | \mathbf{m}_*, \mathbf{K}_* \right), \end{aligned}$$

where

$$\begin{aligned} \mathbf{K}_* &= \left(\Lambda_*^{pre} + \sum_{i \in \mathrm{ch}(*)} \tilde{\Lambda}_i \right)^{-1}, \\ \mathbf{m}_* &= \mathbf{K}_* \left(\Lambda_*^{pre} \mathbf{m}_*^{pre} + \sum_{i \in \mathrm{ch}(*)} \tilde{\Lambda}_i \tilde{\mathbf{m}}_i \right). \end{aligned}$$

Appendix B

Derivations for Variational Inference

B.1 Co-ordinate Ascent Variational Inference Updates

The posterior distribution at (4.7) is to be approximated by the mean-field variational family which factorises according to (4.12). Given that the optimal distribution over the variational factors is $q^*(\boldsymbol{\Psi}_i) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\Psi}/\boldsymbol{\Psi}_i)}\left[\log p\left(\boldsymbol{\Psi}, \mathbf{Y}|\mathcal{T}\right)\right]\right)$, each of the variational parameters required to implement CAVI can be found in turn where the shorthand $\mathbb{E}_{-\boldsymbol{\Psi}_i}[\cdot] \equiv \mathbb{E}_{q(\boldsymbol{\Psi}/\boldsymbol{\Psi}_i)}[\cdot]$ and $\langle \boldsymbol{\Psi}_i \rangle \equiv \mathbb{E}_{q(\boldsymbol{\Psi})}[\boldsymbol{\Psi}_i]$ is used to provide a less cluttered notation and dependence on fixed model parameters has been supressed.

$q^*(\mathbf{X})$

Interaction between the variational and true joint distributions induces the factorisation

$$q^*(\mathbf{X}) \propto \prod_{n=1}^{N}\prod_{i=1}^{D} q^*(\{\mathbf{X}\}_{ni}),$$

where $\{\mathbf{X}\}_{ni}$ denotes the auxiliary traits associated with $\mathbf{Y}_{ni}$ and so $q^*(\{\mathbf{X}\}_{ni})$ can be considered for ordinal, categorical, and continuous traits independently.

For ordinal traits, that is for $i \in \mathcal{O}_{\mathbf{Y}}$,

$$\begin{aligned}
&q^*(\mathbf{X}_{ni'}|\mathbf{Y}_{ni} = k) \\
&\propto \exp\left(\mathbb{E}_{-\mathbf{X}_{ni'}}\left[\log p\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma}\right) p\left(\mathbf{X}|\mathbf{Z}, \mathbf{W}\right)\right]\right) \\
&\propto \exp\left(\mathbb{E}_{-\mathbf{X}_{ni'}}\left[\log \delta\left(\gamma_{i,k-1} \leq \mathbf{X}_{ni'} < \gamma_{i,k}\right) \mathcal{N}\left(\mathbf{X}_{ni'}|\mathbf{W}_{i'\cdot}^{\top}\mathbf{Z}_{n\cdot}, 1\right)\right]\right),
\end{aligned}$$

$$
\begin{aligned}
&\propto \delta\left(\langle\gamma_{i,k-1}\rangle \leq \mathbf{X}_{ni'} < \langle\gamma_{i,k}\rangle\right)\ \mathcal{Z}_{ni'}^{-1}\,\mathcal{N}\left(\mathbf{X}_{ni'}|\langle\mathbf{W}_{i'\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle, 1\right),\\
&= \mathcal{TN}\left(\mathbf{X}_{ni'}|\langle\mathbf{W}_{i'\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle, 1, \langle\gamma_{i,k-1}\rangle, \langle\gamma_{i,k}\rangle\right)
\end{aligned}
$$

a truncated normal distribution, where setting $a_{ni'}^{\mathbf{X}} \equiv \langle\gamma_{i,k-1}\rangle - \langle\mathbf{W}_{i'\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle$ and $b_{ni'}^{\mathbf{X}} \equiv \langle\gamma_{i,k}\rangle - \langle\mathbf{W}_{i'\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle$ implies that

$$
\mathcal{Z}_{ni'} = F_{\mathcal{N}}\left(b_{ni'}^{\mathbf{X}}\right) - F_{\mathcal{N}}\left(a_{ni'}^{\mathbf{X}}\right) \tag{B.1}
$$

$$
\langle\mathbf{X}_{ni'}\rangle = \langle\mathbf{W}_{i'\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle + \mathcal{Z}_{ni'}^{-1}\left(\mathcal{N}\left(a_{ni'}^{\mathbf{X}}\right) - \mathcal{N}\left(b_{ni'}^{\mathbf{X}}\right)\right), \tag{B.2}
$$

where $\mathcal{Z}_{ni'}$ is a normalising constant, $\mathcal{N}(\cdot) \equiv \mathcal{N}(\cdot|0,1)$ and $F_{\mathcal{N}}(\cdot) \equiv \int_{-\infty}^{\cdot}\mathcal{N}(x|0,1)\,dx$.

For categorical traits, when $i \in \mathcal{C}_{\mathbf{Y}}$, the optimal variational distribution is

$$
\begin{aligned}
&q^{*}(\{\mathbf{X}\}_{ni}|\mathbf{Y}_{ni} = c_{i,k})\\
&\propto \exp\left(\mathbb{E}_{-\{\mathbf{X}\}_{ni}}\left[\log p\left(\mathbf{Y}|\mathbf{X}\right)p\left(\mathbf{X}|\mathbf{Z},\mathbf{W}\right)\right]\right)\\
&\propto \exp\left(\mathbb{E}_{-\{\mathbf{X}\}_{ni}}\left[\log\delta\left(\mathbf{X}_{n,i'+k-1} = 0\right)\right]\right)\\
&\qquad \exp\left(\mathbb{E}_{-\{\mathbf{X}\}_{ni}}\left[\prod_{j\neq k}\delta\left(\mathbf{X}_{n,i'+j-1} < 0\right)\mathcal{N}\left(\mathbf{X}_{n,i'+j-1}|\mathbf{W}_{i'+j-1,\cdot}^{\top}\mathbf{Z}_{n\cdot}, 1\right)\right]\right),\\
&\propto \prod_{j\neq k}\delta\left(\mathbf{X}_{n,i'+j-1} < 0\right)\mathcal{Z}_{n,i'+j-1}^{-1}\,\mathcal{N}\left(\mathbf{X}_{n,i'+j-1}|\langle\mathbf{W}_{i'+j-1,\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle, 1\right),\\
&= \prod_{j\neq k}\mathcal{TN}\left(\mathbf{X}_{n,i'+j-1}|\langle\mathbf{W}_{i'+j-1\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle, 1, -\infty, 0\right)
\end{aligned}
$$

where for $b_{n,i'+j-1}^{\mathbf{X}} \equiv -\langle\mathbf{W}_{i'+j-1,\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}\rangle$

$$
\mathcal{Z}_{n,i'+j-1} = F_{\mathcal{N}}\left(b_{n,i'+j-1}^{\mathbf{X}}\right) \tag{B.3}
$$

$$
\langle\mathbf{X}_{n,i'+j-1}\rangle = \langle\mathbf{W}_{i+j-1,\cdot}\rangle^{\top}\langle\mathbf{Z}_{n\cdot}^{\star}\rangle - \mathcal{Z}_{n,i'+j-1}^{-1}\mathcal{N}\left(b_{n,i'+j-1}^{\mathbf{X}}\right). \tag{B.4}
$$

Finally, for continuous and function-valued traits, that is for all $i \in \mathcal{R}_{\mathbf{Y}}$, the auxiliary and manifest traits are equivalent and so $q^{*}(\mathbf{X}_{ni'}|\mathbf{Y}_{ni}) = \delta\left(\mathbf{X}_{ni'} = \mathbf{Y}_{ni}\right)$ and

$$
\langle\mathbf{X}_{ni'}\rangle = \mathbf{Y}_{ni}. \tag{B.5}
$$

.

$q^*(\mathbf{W}_{i'\cdot})$

The optimal variational distribution for the loading matrix rows is given by

$$
\begin{aligned}
&q^*(\mathbf{W}_{i'\cdot}) \\
&\propto \exp\left(\mathbb{E}_{-\mathbf{W}_{i'\cdot}}\left[\log p(\mathbf{X}|\mathbf{W},\mathbf{Z},\Lambda)\, p(\mathbf{W}|\mathbf{W},\boldsymbol{\alpha})\right]\right), \\
&\propto \exp\left(\mathbb{E}_{-\mathbf{W}_{i'\cdot}}\left[\log \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{X}_{ni'}|\mathbf{W}_{i'\cdot}^{\top}\mathbf{Z}_{n\cdot}, \Lambda_{i'}^{-1}\right) \mathcal{N}\left(\mathbf{W}_{i'\cdot}|\mathbf{W}^*_{i'\cdot}, \left(\mathbf{A}^{i'}\right)^{-1}\right)\right]\right), \\
&\propto \mathcal{N}\left(\mathbf{W}_{\cdot i'}|\langle\mathbf{W}_{i'\cdot}\rangle, \mathbf{S}_{i'}^{\mathbf{W}}\right),
\end{aligned}
$$

where the variational parameters are defined as

$$\mathbf{S}_{i'}^{\mathbf{W}} = \left(\langle\Lambda_{i'}\rangle\langle\mathbf{Z}^{\star\top}\mathbf{Z}^{\star}\rangle + \langle\mathbf{A}^{i'}\rangle\right)^{-1}, \tag{B.6}$$

$$\langle\mathbf{W}_{i'\cdot}\rangle = \mathbf{S}_{i'}^{\mathbf{W}}\left(\langle\Lambda_{i'}\rangle\langle\mathbf{Z}^{\star}\rangle^{\top}\langle\mathbf{X}_{\cdot i'}\rangle + \langle\mathbf{A}^{i'}\rangle\langle\mathbf{W}^*_{i'\cdot}\rangle\right) \tag{B.7}$$

which are functions of

$$
\begin{aligned}
\langle\mathbf{Z}^{\star\top}\mathbf{Z}^{\star}\rangle &= \sum_{n=1}^{N}\langle\mathbf{Z}_{n\cdot}\mathbf{Z}_{n\cdot}^{\top}\rangle, \\
&= \sum_{n=1}^{N}\mathbf{S}_n^{\mathbf{Z}} + \langle\mathbf{Z}_{n\cdot}\rangle\langle\mathbf{Z}_{n\cdot}\rangle^{\top}, \\
\langle\mathbf{W}^*_{i'\cdot}\rangle &= \langle\mathbf{W}_{-i',\cdot}\rangle^{\top}\left(\mathbf{K}^{\mathbf{W}}_{-i',-i'}\right)^{-1}\mathbf{K}^{\mathbf{W}}_{-i',i'}
\end{aligned}
\tag{B.8}
$$

and the Q-dimensional diagonal matrix $\mathbf{A}^i$ with entries

$$\langle\mathbf{A}_j^{i'}\rangle = \langle\alpha_j\rangle\left(\mathbf{K}^{\mathbf{W}}_{i'i'} - \mathbf{K}^{\mathbf{W}}_{i',-i'}\left(\mathbf{K}^{\mathbf{W}}_{-i',-i'}\right)^{-1}\mathbf{K}^{\mathbf{W}}_{-i',i'}\right)^{-1},$$

for the fixed prior covariance matrix $\mathbf{K}_{\mathbf{W}}$.

$q^*(\mathbf{Z}_{n\cdot})$

When deriving the approximate posterior for the factors over $\mathcal{T}$, the terminal and internal nodes must be considered as separate cases. Starting with the terminal nodes, that is $\mathbf{Z}_{n\cdot}$ for $n = 1, \ldots, N$

$$
\begin{aligned}
&q^*(\mathbf{Z}_{n\cdot}) \\
&\propto \exp\left(\mathbb{E}_{-\mathbf{Z}_{n\cdot}}\left[\log p(\mathbf{X}|\mathbf{W},\mathbf{Z},\Lambda)\, p(\mathbf{Z}|\boldsymbol{\theta})\right]\right),
\end{aligned}
$$

$$
\begin{aligned}
&\propto \exp\left(\mathbb{E}_{-\mathbf{Z}_{n\cdot}}\left[\log \mathcal{N}\left(\mathbf{X}_{n\cdot}|\mathbf{W}\mathbf{Z}_{n\cdot}, \Lambda^{-1}\right)\prod_{j=1}^{Q}\mathcal{N}\left(\mathbf{Z}_{nj}|\phi_{n,j}\mathbf{Z}_{\mathrm{pa}(n),j}, \eta_{n,j}\right)\right]\right),\\
&\propto \mathcal{N}\left(\mathbf{Z}_{n\cdot}|\langle\mathbf{Z}_{n\cdot}\rangle, \mathbf{S}_n^{\mathbf{Z}}\right),
\end{aligned}
$$

where

$$
\mathbf{S}_n^{\mathbf{Z}} = \left(\langle\mathbf{W}^\top\Lambda\mathbf{W}\rangle + \langle\mathbf{E}^n\rangle\right)^{-1}, \tag{B.9}
$$

$$
\langle\mathbf{Z}_{n\cdot}\rangle = \mathbf{S}_n^{\mathbf{Z}}\left(\langle\mathbf{W}\rangle^\top\langle\Lambda\rangle\langle\mathbf{X}_{n\cdot}\rangle + \langle\mathbf{E}^n\Phi^n\rangle\langle\mathbf{Z}_{\mathrm{pa}(n),\cdot}\rangle\right). \tag{B.10}
$$

Note that $\mathrm{pa}\,(n)$ denotes the parent node of node n, Φ^n and $\mathbf{E}^n$ are Q-dimensional diagonal matrices with entries $\Phi_j^n = \phi_{n,j}$ and $\mathbf{E}_j^n = \eta_{n,j}^{-1}$, and

$$
\begin{aligned}
\langle\mathbf{W}^\top\Lambda\mathbf{W}\rangle &= \sum_{i=1}^{D'}\langle\Lambda_{i'}\rangle\langle\mathbf{W}_{i'\cdot}\mathbf{W}_{i'\cdot}\rangle^\top,\\
&= \sum_{i'=1}^{D'}\langle\Lambda_{i'}\rangle\left(\mathbf{S}_i^{\mathbf{W}} + \langle\mathbf{W}_{i'\cdot}\rangle\langle\mathbf{W}_{i'\cdot}\rangle^\top\right).
\end{aligned} \tag{B.11}
$$

For internal nodes, that is $\mathbf{Z}_{n\cdot}$ for $n = 1, \ldots, N + 2S - 1$, the optimal approximate posterior is given by

$$
\begin{aligned}
&q^*\left(\mathbf{Z}_{n\cdot}\right)\\
&\propto \exp\left(\mathbb{E}_{-\mathbf{Z}_{n\cdot}}\left[\log p\left(\mathbf{Z}|\boldsymbol{\theta}\right)\right]\right),\\
&\propto \exp\left(\mathbb{E}_{-\mathbf{Z}_{n\cdot}}\left[\log\prod_{j=1}^{Q}\mathcal{N}\left(\mathbf{Z}_{nj}|\phi_{n,j}\mathbf{Z}_{\mathrm{pa}(n),j}, \eta_{n,j}\right)\prod_{k\in\{\mathrm{ch}(n)\}}\mathcal{N}\left(\mathbf{Z}_{k,j}|\phi_{k,j}\mathbf{Z}_{nj}, \eta_{k,j}\right)\right]\right),\\
&\propto \prod_{j=1}^{Q}\mathcal{N}\left(\mathbf{Z}_{nj}|\langle\mathbf{Z}_{nj}\rangle, \left(\mathbf{S}_n^{\mathbf{Z}}\right)_j\right),
\end{aligned}
$$

where $\mathbf{S}_n^{\mathbf{Z}}$ is a diagonal matrix with elements

$$
\left(\mathbf{S}_n^{\mathbf{Z}}\right)_j = \left(\langle\eta_{n,j}^{-1}\rangle + \sum_{k\in\{\mathrm{ch}(n)\}}\langle\phi_{n,k}^2\eta_{n,k}^{-1}\rangle\right)^{-1}, \tag{B.12}
$$

$$
\langle\mathbf{Z}_{nj}\rangle = \left(\mathbf{S}_n^{\mathbf{Z}}\right)_j\left(\langle\eta_{n,j}^{-1}\phi_{n,j}\rangle\langle\mathbf{Z}_{\mathrm{pa}(n),j}\rangle + \sum_{k\in\{\mathrm{ch}(n)\}}\langle\phi_{n,k}\eta_{n,k}^{-1}\rangle\langle\mathbf{Z}_{kj}\rangle\right), \tag{B.13}
$$

and at the root node, that is $R \equiv N + 2S - 1$, $\phi_{R,j} \equiv 1, \eta_{R,j} \equiv k_j\left(\mathbf{t}_R, \mathbf{t}_R|\mathcal{T}\right)$, and

$\mathbf{Z}_{\mathrm{pa}(R)\cdot} \equiv \mathbf{0}$.

$q^*\left(\Lambda, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$

For the approximate distribution over auxiliary trait precision parameters, ARD precision hyper-parameters on the loading, and hyper-parameters of the Gaussian process over $\mathcal{T}$, interaction between the true joint distribution and the mean-field variational family results in further factorisation of the approximate posterior such that

$$\begin{aligned} q\left(\Lambda, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) &= q\left(\Lambda\right) q\left(\boldsymbol{\alpha}\right) q\left(\boldsymbol{\theta}\right), \\ &= \prod_{i' \in \mathcal{R}_{\mathbf{X}}} q\left(\Lambda_{i'}\right) \prod_{j=1}^{Q} q\left(\alpha_j\right) q\left(\theta_j\right), \end{aligned}$$

and each of the approximate posterior distributions can be derived independently.

The optimal distribution for the $i^{'th}$ diagonal element of the model precision matrix, when $i' \in \mathcal{R}_{\mathbf{X}}$ and so is a free parameter, is given by

$$\begin{aligned} q^*\left(\Lambda_i\right) &\propto \exp\left(\mathbb{E}_{-\Lambda_{i'}}\left[\log p\left(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \Lambda\right) p(\Lambda)\right]\right), \\ &\propto \exp\left(\mathbb{E}_{-\Lambda_{i'}}\left[\log \mathcal{N}\left(\mathbf{X}_{\cdot i'}|\mathbf{Z}\mathbf{W}_{i'\cdot}, \Lambda_{i'}^{-1}\right) Gamma\left(\Lambda_{i'}|\, a_\Lambda, b_\Lambda\right)\right]\right), \\ &\propto Gamma\left(\Lambda_{i'}|\hat{a}_\Lambda^{i'}, \hat{b}_\Lambda^{i'}\right), \end{aligned}$$

for

$$\hat{a}_\Lambda^{i'} = a_\Lambda + \frac{N}{2}, \tag{B.14}$$

$$\hat{b}_\Lambda^{i'} = b_\Lambda + \frac{1}{2}\left(\langle \mathbf{X}_{\cdot i'}\rangle^\top \langle \mathbf{X}_{\cdot i'}\rangle - 2\langle \mathbf{X}_{\cdot i'}\rangle^\top \langle \mathbf{Z}\rangle \langle \mathbf{W}_{i'\cdot}\rangle + \mathrm{tr}\left(\langle \mathbf{W}_{i'\cdot}\mathbf{W}_{i'\cdot}^\top\rangle \langle \mathbf{Z}^\top\mathbf{Z}\rangle\right)\right), \tag{B.15}$$

which implies that

$$\langle \Lambda_{i'}\rangle = \frac{\hat{a}_\Lambda^{i'}}{\hat{b}_\Lambda^{i'}}, \tag{B.16}$$

$$\langle \log \Lambda_{i'}\rangle = \psi\left(\hat{a}_\Lambda^{i'}\right) - \log \hat{b}_\Lambda^{i}, \tag{B.17}$$

where $\psi\left(\cdot\right)$ is the digamma function. When $i \notin \mathcal{R}_{\mathbf{X}}$, $\langle \Lambda_{i'}\rangle \equiv 1$ and $\langle \log \Lambda_{i'}\rangle = 0$. If it is assumed that any $\Lambda_{i'} = \Lambda_{k'}$ for some $i', k' \in \mathcal{R}_{\mathbf{X}}, i' \neq k'$, then $q^*\left(\Lambda_{i'}\right)$ can be obtained by summing over all relevant indices.

Similarly, the optimal approximate posterior for the ARD precision para-

mater of the j^{th} column of the loading matrix is

$$\begin{aligned} q^*(\alpha_j) &\propto \exp\left(\mathbb{E}_{-\alpha_j}\left[\log p(\mathbf{W}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\right]\right), \\ &\propto \exp\left(\mathbb{E}_{-\alpha_j}\left[\log \mathcal{N}\left(\mathbf{W}_{\cdot j}|\mathbf{0}, \alpha_j^{-1}\mathbf{K}^{\mathbf{W}}\right) Gamma\left(\alpha_j|\, a_\alpha, b_\alpha\right)\right]\right), \\ &\propto Gamma\left(\alpha_j|\hat{a}_\alpha^j, \hat{b}_\alpha^j\right), \end{aligned}$$

for

$$\hat{a}_\alpha^j = \frac{D'}{2} + a_\alpha \tag{B.18}$$

$$\hat{b}_\alpha^j = \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{K}^{\mathbf{W}}\right)^{-1}\langle \mathbf{W}_{\cdot j}\mathbf{W}_{\cdot j}^\top\rangle\right) + b_\alpha \tag{B.19}$$

where

$$\langle \mathbf{W}_{\cdot j}\mathbf{W}_{\cdot j}^\top\rangle = \mathbf{V}_j^{\mathbf{W}} + \langle \mathbf{W}_{\cdot j}\rangle\langle \mathbf{W}_{\cdot j}\rangle^\top, \tag{B.20}$$

with diagonal matrix $\left(\mathbf{V}_j^{\mathbf{W}}\right)_{ii} \equiv \left(\mathbf{S}_i^{\mathbf{W}}\right)_{jj}$. This implies that

$$\langle \alpha_i\rangle = \frac{\hat{a}_\alpha^j}{\hat{b}_\alpha^j}, \tag{B.21}$$

$$\langle \log \alpha_j\rangle = \psi\left(\hat{a}_\alpha^j\right) - \log \hat{b}_\alpha^j., \tag{B.22}$$

Finally, the variational distribution for θ_j is given by

$$\begin{aligned} q^*(\theta_j) &\propto \exp\left(\mathbb{E}_{-\theta_j}\left[\log p(\mathbf{Z}_{\cdot j}|\theta_j)\, p(\theta_j)\right]\right) \\ &\propto \exp\left(\mathbb{E}_{-\theta_j}\left[\prod_{n=1}^{N+2S-1} \log \mathcal{N}\left(\mathbf{Z}_{nj}|\phi_{n,j}\mathbf{Z}_{\mathrm{pa}(n)}, \eta_{n,j}\right)\right.\right. \\ &\qquad \left.\left. Gamma\left(\ell_j|2,1\right) Beta(\kappa_j|1,1) Beta(\tau_j|1,1)\right]\right) \\ &\propto \exp\left(\sum_{n=1}^{N+2S-1} -\frac{1}{2}\log \eta_{n,j} - \frac{\langle \mathbf{Z}_{nj}^2\rangle - 2\langle \mathbf{Z}_{nj}\rangle\langle \mathbf{Z}_{\mathrm{pa}(n),j}\rangle + \phi_{n,j}^2\langle \mathbf{Z}_{\mathrm{pa}(n),j}^2\rangle}{2\eta_{n,j}}\right. \\ &\qquad \left. + \log \tau_j + \log\left(1-\tau_j\right) + \log \kappa_j + \log\left(1-\kappa_j\right) + \log \ell_j - \ell_j\right) \end{aligned} \tag{B.23}$$

where

$$\langle \mathbf{Z}_{nj}^2\rangle = \left(\mathbf{S}_n^{\mathbf{Z}}\right)_{jj} + \langle \mathbf{Z}_{nj}\rangle^2.$$

This distribution does not yield an analytic solution for θ_j, however it's computation does scale linearly with N. If full uncertainty quantification for these hyperparameters is required then a Monte Carlo method (i.e. Adaptive Metropolis [Haario et al., 2001; Roberts and Rosenthal, 2009]) can be employed to draw S_{MC} samples from this distribution. The required expectations can then be estimated by

$$\langle \varphi_j \rangle = \sum_{s=1}^{S_{MC}} \varphi_j^{(s)} \tag{B.24}$$

where

$$\varphi_j = \left\{ \ell_j, \kappa_j, \tau_j, \log p\left(\theta_j\right), \left\{ \phi_{n,j}, \eta_{n,j}, \log \eta_{n,j}, \phi_{n,ji}\eta_{n,j}^{-1}, \phi_{n,j}^2 \eta_{n,j}^{-1} \right\}_{n=1}^{N+2S-1} \right\}$$

A less computationally expensive approach is to set the approximating distribution to be an indicator function, such that $q\left(\theta_j\right) \equiv \delta\left(\theta_j = \langle \theta_j \rangle\right)$, and optimise the ELBO directly. In this case $\left\{ \langle \phi_{n,j} \rangle, \langle \eta_{n,j} \rangle, \langle \log \eta_{n,j} \rangle \langle \phi_{n,ji}\eta_{n,j}^{-1} \rangle, \langle \phi_{n,j}^2 \eta_{n,j}^{-1} \rangle \right\}_{n=1}^{N+2S-1}$ can be computed directly from $\langle \theta_j \rangle$.

$q^*\left(\gamma_{i,k}\right)$

The optimal mean-field variational family approximate posterior distribution for free ordinal cut-off points, that is when $i \in \mathcal{O}_{\mathbf{Y}}$ and $k = 2, \ldots, K_i - 1$, is given by

$$\begin{aligned} & q^*\left(\gamma_{i,k}\right) \\ & \propto \exp\left(\mathbb{E}_{-\mathbf{X}_{ni'}}\left[\log p\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right)\right]\right) \\ & \propto \exp\left(\mathbb{E}_{-\gamma_{i,k}}\left[\log \prod_{n=1}^{N} \left(\sum_{j=1}^{K_i} \delta\left(\mathbf{Y}_{ni} = j\right) \delta\left(\gamma_{i,j-1} \leq \mathbf{X}_{ni'} < \gamma_{i,j}\right) \right) \right] \right), \end{aligned} \tag{B.25}$$

for which no closed form solution exists. In this case however, uncertainty quantification for $\gamma_{i,k}$ is not a priority. Thus rather than attempting to find $\langle \gamma_{i,k} \rangle$ under (B.25), simply set $q\left(\gamma_{i,k}\right) = \delta\left(\gamma_{i,k} = \langle \gamma_{i,k} \rangle\right)$ and optimse with respect to the ELBO , noting that $\langle \gamma_{i,k} \rangle \in \left[\lfloor \gamma_{i,k} \rfloor, \lceil \gamma_{i,k} \rceil \right)$ where $\lfloor \gamma_{i,k} \rfloor = \max\{\max\{\langle \mathbf{X}_{ni'} \rangle | \mathbf{Y}_{ni} = k\}, \langle \gamma_{i,k-1} \rangle\}$ and $\lceil \gamma_{i,k} \rceil = \min\{\min\{\langle \mathbf{X}_{ni'} \rangle | \mathbf{Y}_{ni} = k+1\}, \langle \gamma_{i,k+1} \rangle, \langle \gamma_{i,k-1} \rangle + b_\gamma\}$.

B.2 The Evidence Lower Bound

The Evidence Lower Bound (ELBO) is the objective function being maximised in Variational Inference. For the generalised PLVM the ELBO is of the form

$$\begin{aligned}
\text{ELBO}\left(q\right) = &\mathbb{E}_q\left[\log p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\mathbf{X}|\mathbf{Z},\mathbf{W},\Lambda\right)\right] - \mathbb{E}_q\left[\log q\left(\mathbf{X}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\mathbf{Z}|\theta,\mathcal{T}\right)\right] - \mathbb{E}_q\left[\log q\left(\mathbf{Z}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\mathbf{W}|\boldsymbol{\alpha}\right)\right] - \mathbb{E}_q\left[\log q\left(\mathbf{W}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\Lambda\right)\right] - \mathbb{E}_q\left[\log q\left(\Lambda\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\boldsymbol{\gamma}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\gamma}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\boldsymbol{\alpha}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\alpha}\right)\right] \\
&+ \mathbb{E}_q\left[\log p\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\theta}\right)\right]. \qquad \text{(B.26)}
\end{aligned}$$

It is worth noting that, for a continuous random variable X, the differential entropy is defined by $-\mathbb{E}_{p(x)}\left[\log p\left(X\right)\right]$ [Bishop, 2006]. Thus if $\mathbb{E}_q\left[\log p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}\right)\right]$ is considered to be the expected likelihood of the manifest traits with respect to the variational distribution $q\left(\Theta\right)$, then each subsequent each line of the ELBO presented above can be in terms of an entropy term $-\mathbb{E}_q\left[\log q\left(\boldsymbol{\Psi}_i\right)\right]$ less a cross-entropy term $-\mathbb{E}_q\left[\log p\left(\boldsymbol{\Psi}_i|\boldsymbol{\Psi}_{-i}\right)\right]$. Each line of ELBO $\left(q\right)$ is considered in turn below and by substituting each of these into (B.26), ELBO $\left(q\right)$ can be computed.

$\mathbb{E}_q\left[\log p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}\right)\right]$

The expected log likelihood of the manifest traits given auxiliary traits and ordinal cut-off points can be expressd as $\mathbb{E}_q\left[\log p\left(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}\right)\right] = \mathbb{E}_q\left[\log \delta\left(\mathbf{Y}_{n\cdot} = g\left(\mathbf{X}_{n\cdot}\right)\right)\right]$, which is to say that ELBO $\left(q\right)$ is undefined when the conditions set out by the auxiliary to manifest mapping are not satisfied.

$\mathbb{E}_q\left[\log p\left(\mathbf{X}|\mathbf{Z},\mathbf{W},\Lambda\right)\right] - \mathbb{E}_q\left[\log q\left(\mathbf{X}\right)\right]$

To compute the contribution of auxiliary traits to ELBO $\left(q\right)$ consider each type of trait in turn. Given that the entropy of an indicator function is 0, for function-valued and continuous traits the auxiliary contribution is given by

$$\sum_{i'\in\mathcal{R}_{\mathbf{X}}}\sum_{n=1}^{N}\mathbb{E}_q\left[\log p\left(\mathbf{X}_{ni'}|\mathbf{Z}_{n\cdot},\mathbf{W}_{i'\cdot},\Lambda_{i'}\right)\right] - \mathbb{E}_q\left[\log q\left(\mathbf{X}_{ni'}\right)\right]$$

$$
\begin{aligned}
&= \sum_{i' \in \mathcal{R}_{\mathbf{X}}} \sum_{n=1}^{N} \mathbb{E}_q \left[\log \mathcal{N} \left(\mathbf{X}_{ni'} | \mathbf{W}_{i'\cdot} \mathbf{Z}_{n\cdot}, \Lambda_{i'}^{-1} \right) \right] - \mathbb{E}_q \left[\log \delta \left(\mathbf{X}_{ni'} = \mathbf{Y}_{ni'} \right) \right], \\
&= \sum_{i' \in \mathcal{R}_{\mathbf{X}}} -\frac{N}{2} \log 2\pi + \frac{N}{2} \langle \log \Lambda_i \rangle \\
&\qquad - \frac{\langle \Lambda_i \rangle}{2} \left(\langle \mathbf{X}_{\cdot i'} \rangle^\top \langle \mathbf{X}_{\cdot i'} \rangle - 2 \langle \mathbf{X}_{\cdot i} \rangle^\top \langle \mathbf{Z}^\star \rangle \langle \mathbf{W}_{i'\cdot} \rangle + \operatorname{tr} \left(\langle \mathbf{Z}^{\star\top} \mathbf{Z}^\star \rangle \langle \mathbf{W}_{i'\cdot} \mathbf{W}_{i'\cdot}^\top \rangle \right) \right),
\end{aligned} \tag{B.27}
$$

where the required quantities are defined in (B.17), (B.16), (B.5), (B.10), (B.7), (B.8), and (B.11).

For continuous traits, when $\mathbf{Y}_{ni} = c_{i,k_n}$ for $k_n \in \{1, \ldots, K_i\}$, the following holds

$$
\begin{aligned}
&\sum_{i' \in \mathcal{C}_{\mathbf{X}}} \sum_{n=1}^{N} \sum_{l=1}^{K_i} \mathbb{E}_q \left[\log p \left(\mathbf{X}_{n,i'+l-1} | \mathbf{Z}_{n\cdot}, \mathbf{W}_{i'+l-1,\cdot} \right) \right] - \mathbb{E}_q \left[\log q \left(\mathbf{X}_{n,i'+l-1} \right) \right] \\
&= \sum_{i' \in \mathcal{C}_{\mathbf{X}}} \sum_{n=1}^{N} \mathbb{E}_q \left[\log \mathcal{N} \left(0 | \mathbf{W}_{i'+k_n-1,\cdot} \mathbf{Z}_{n\cdot}, 1 \right) \right] - \mathbb{E}_q \left[\log \delta \left(\mathbf{X}_{n,i+k_n-1} = 0 \right) \right] \\
&\qquad + \sum_{l \neq k_n} \mathbb{E}_q \left[\log \mathcal{N} \left(\mathbf{X}_{n,i'+l-1} | \mathbf{W}_{i'+l-1,\cdot} \mathbf{Z}_{n\cdot}, 1 \right) \right] \\
&\qquad - \mathbb{E}_q \left[\log \delta \left(\mathbf{X}_{n,i'+l-1} < 0 \right) \mathcal{Z}_{n,i'+l-1}^{-1} \mathcal{N} \left(\mathbf{X}_{n,i'+l-1} | \langle \mathbf{W}_{i'+l-1,\cdot} \rangle^\top \langle \mathbf{Z}_{n\cdot} \rangle, 1 \right) \right], \\
&= \sum_{i' \in \mathcal{C}_{\mathbf{X}}} \sum_{n=1}^{N} -\frac{1}{2} \log 2\pi - \frac{1}{2} \operatorname{tr} \left(\langle \mathbf{Z}_{n\cdot} \mathbf{Z}_{n\cdot}^\top \rangle \langle \mathbf{W}_{i'+k_n-1,\cdot} \mathbf{W}_{i'+k_n-1,\cdot}^\top \rangle \right) \\
&\qquad + \sum_{l \neq k_n} -\frac{1}{2} \log 2\pi - \frac{1}{2} \Big(\langle \mathbf{X}_{n,i'+l-1}^2 \rangle - 2 \langle \mathbf{X}_{n,i'+l-1} \rangle \langle \mathbf{W}_{i'+l-1,\cdot} \rangle^\top \langle \mathbf{Z}_{n\cdot} \rangle \\
&\qquad\qquad + \operatorname{tr} \left(\langle \mathbf{Z}_{n\cdot} \mathbf{Z}_{n\cdot}^\top \rangle \langle \mathbf{W}_{i'+l-1,\cdot} \mathbf{W}_{i'+l-1,\cdot}^\top \rangle \right) \Big) \\
&\qquad + \log \mathcal{Z}_{n,i'+l-1} + \frac{1}{2} \log 2\pi + \frac{1}{2} \Big(\langle \mathbf{X}_{n,i'+l-1}^2 \rangle \\
&\qquad\qquad - 2 \langle \mathbf{X}_{n,i'+l-1} \rangle \langle \mathbf{W}_{i'+l-1,\cdot} \rangle^\top \langle \mathbf{Z}_{n\cdot} \rangle + \left(\langle \mathbf{W}_{i'+l-1,\cdot} \rangle^\top \langle \mathbf{Z}_{n\cdot} \rangle \right)^2 \Big), \\
&= \sum_{i' \in \mathcal{C}_{\mathbf{X}}} -\frac{N}{2} \log 2\pi + \sum_{n=1}^{N} -\frac{1}{2} \operatorname{tr} \left(\langle \mathbf{Z}_{n\cdot} \mathbf{Z}_{n\cdot}^\top \rangle \langle \mathbf{W}_{i'+k_n-1,\cdot} \mathbf{W}_{i'+k_n-1,\cdot}^\top \rangle \right) \\
&\qquad + \sum_{l \neq k_n} -\frac{1}{2} \operatorname{tr} \left(\langle \mathbf{Z}_{n\cdot} \mathbf{Z}_{n\cdot}^\top \rangle \langle \mathbf{W}_{i'+l-1,\cdot} \mathbf{W}_{i'+l-1,\cdot}^\top \rangle \right) + \frac{1}{2} \left(\langle \mathbf{W}_{i'+l-1,\cdot} \rangle^\top \langle \mathbf{Z}_{n\cdot} \rangle \right)^2 \\
&\qquad\qquad + \log \mathcal{Z}_{n,i'+l-1},
\end{aligned} \tag{B.28}
$$

where the categorical normalising constant $\mathcal{Z}_{n,i'+l-1}$ is defined in (B.3).

Finally, the contribution of ordinal traits to ELBO (q) when $\mathbf{Y}_{ni} = k_n$ with $k_n \in \{1, \ldots, K_i\}$ can be expressed as

$$
\begin{aligned}
&\sum_{i' \in \mathcal{O}_{\mathbf{X}}} \sum_{n=1}^{N} \mathbb{E}_q \left[\log p\left(\mathbf{X}_{n,i'} | \mathbf{Z}_{n\cdot}, \mathbf{W}_{i'\cdot}\right)\right] - \mathbb{E}_q \left[\log q\left(\mathbf{X}_{ni'}\right)\right] \\
&= \sum_{i' \in \mathcal{R}_{\mathbf{X}}} \sum_{n=1}^{N} \mathbb{E}_q \left[\log \mathcal{N}\left(\mathbf{X}_{ni'} | \mathbf{W}_{i'\cdot} \mathbf{Z}_{n\cdot}, 1\right)\right] \\
&\qquad - \mathbb{E}_q \left[\log \delta\left(\langle \gamma_{i,k_n - 1} \rangle \leq \mathbf{X}_{ni'} < \langle \gamma_{i,k_n} \rangle\right) \, \mathcal{Z}_{ni'}^{-1} \, \mathcal{N}\left(\mathbf{X}_{ni'} | \langle \mathbf{W}_{i'\cdot} \rangle^{\top} \langle \mathbf{Z}_{n\cdot} \rangle, 1\right)\right], \\
&= \sum_{i' \in \mathcal{O}_{\mathbf{X}}} -\frac{N}{2} \log 2\pi - \frac{1}{2}\Big(\langle \mathbf{X}_{\cdot i'}^{\top} \mathbf{X}_{\cdot i'} \rangle - 2 \langle \mathbf{X}_{\cdot i} \rangle^{\top} \langle \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \rangle \\
&\qquad\qquad + \operatorname{tr}\left(\langle \mathbf{Z}^{\star \top} \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \mathbf{W}_{i'\cdot}^{\top} \rangle \right) \Big) \\
&\qquad + \frac{N}{2} \log 2\pi + \frac{1}{2}\Big(\langle \mathbf{X}_{\cdot i'}^{\top} \mathbf{X}_{\cdot i'} \rangle - 2 \langle \mathbf{X}_{\cdot i} \rangle^{\top} \langle \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \rangle \\
&\qquad\qquad + \operatorname{tr}\left(\langle \mathbf{Z}^{\star} \rangle^{\top} \langle \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \rangle \langle \mathbf{W}_{i'\cdot}^{\top} \rangle \right) \Big) + \sum_{n=1}^{N} \log \mathcal{Z}_{ni'}, \\
&= \sum_{i' \in \mathcal{O}_{\mathbf{X}}} -\frac{1}{2} \operatorname{tr}\left(\langle \mathbf{Z}^{\star \top} \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \mathbf{W}_{i'\cdot}^{\top} \rangle \right) + \frac{1}{2} \operatorname{tr}\left(\langle \mathbf{Z}^{\star} \rangle^{\top} \langle \mathbf{Z}^{\star} \rangle \langle \mathbf{W}_{i'\cdot} \rangle \langle \mathbf{W}_{i'\cdot}^{\top} \rangle \right) + \sum_{n=1}^{N} \log \mathcal{Z}_{ni'},
\end{aligned}
\tag{B.29}
$$

where the ordinal normalising constant $\mathcal{Z}_{n,i'}$ is given in (B.1). This completes the contribution of the auxiliary traits to ELBO (q)

$\mathbb{E}_q \left[\log p\left(\mathbf{Z} | \theta, \mathcal{T}\right)\right] - \mathbb{E}_q \left[\log q\left(\mathbf{Z}\right)\right]$

The contribution of all factors over $\mathcal{T}$ to ELBO (q) is given by

$$
\begin{aligned}
&\mathbb{E}_q \left[\log p\left(\mathbf{Z} | \theta, \mathcal{T}\right)\right] - \mathbb{E}_q \left[\log q\left(\mathbf{Z}\right)\right] \\
&= \sum_{n=1}^{N+2S-1} -\mathbb{E}_q \left[\log q\left(\mathbf{Z}_{n\cdot}\right)\right] + \sum_{j=1}^{Q} \mathbb{E}_q \left[\log \mathcal{N}\left(\mathbf{Z}_{nj} | \phi_{n,j} \mathbf{Z}_{\mathrm{pa}(n),j}, \eta_{n,j}\right)\right], \\
&= \sum_{n=1}^{N+2S-1} \frac{Q}{2} \log 2\pi + \frac{1}{2} \log |e \mathbf{S}_n^{\mathbf{Z}}| + \sum_{j=1}^{Q} -\frac{1}{2} \log 2\pi - \frac{1}{2} \langle \log \eta_{n,j} \rangle \\
&\qquad - \frac{\langle \mathbf{Z}_{nj}^2 \rangle \langle \eta_{n,j}^{-1} \rangle}{2} + \langle \mathbf{Z}_{nj} \rangle \langle \phi_{n,j} \eta_{n,j}^{-1} \rangle \langle \mathbf{Z}_{\mathrm{pa}(n),j} \rangle - \frac{\langle \phi_{n,j}^2 \eta_{n,j}^{-1} \rangle \langle \mathbf{Z}_{\mathrm{pa}(n)j}^2 \rangle}{2}, \\
&= \sum_{n=1}^{N+2S-1} \frac{1}{2} \log |e \mathbf{S}_n^{\mathbf{Z}}| + \sum_{j=1}^{Q} -\frac{1}{2} \langle \log \eta_{n,j} \rangle
\end{aligned}
$$

$$- \frac{\langle \mathbf{Z}_{nj}^2 \rangle \langle \eta_{n,j}^{-1} \rangle}{2} + \langle \mathbf{Z}_{nj} \rangle \langle \phi_{n,j} \eta_{n,j}^{-1} \rangle \langle \mathbf{Z}_{\mathrm{pa}(n),j} \rangle - \frac{\langle \phi_{n,j}^2 \eta_{n,j}^{-1} \rangle \langle \mathbf{Z}_{\mathrm{pa}(n)j}^2 \rangle}{2},$$

where $\mathbf{S}_n^{\mathbf{Z}}$ is defined in (B.9) and (B.12) and the functions of θ_j can be estimated by (B.24) or computed directly when $q(\theta_j) = \delta(\theta_j = \langle \theta_j \rangle)$.

$\mathbb{E}_q[\log p(\mathbf{W}|\boldsymbol{\alpha})] - \mathbb{E}_q[\log q(\mathbf{W})]$

The contribution to ELBO (q) from the loading matrix is given by

$$\begin{aligned}
&\mathbb{E}_q[\log p(\mathbf{W}|\boldsymbol{\alpha})] - \mathbb{E}_q[\log q(\mathbf{W})] \\
&= \sum_{j=1}^{Q} \mathbb{E}_q\left[\log \mathcal{N}\left(\mathbf{W}_{\cdot j}|\mathbf{0}, \alpha_j^{-1}\mathbf{K}^{\mathbf{Z}}\right)\right] - \sum_{i'=1}^{D'} \mathbb{E}_q[\log q(\mathbf{W}_{i'\cdot})], \\
&= -\frac{Q}{2}\log|\mathbf{K}^{\mathbf{W}}| + \sum_{j=1}^{Q}\frac{D'}{2}\langle \log \alpha_j \rangle - \frac{\langle \alpha_j \rangle}{2}\operatorname{tr}\left(\langle \mathbf{W}_{\cdot j}\mathbf{W}_{\cdot j}^{\top}\rangle \left(\mathbf{K}^{\mathbf{W}}\right)^{-1}\right) \\
&\quad + \sum_{i'=1}^{D'}\frac{1}{2}\log|e\mathbf{S}_{i'}^{\mathbf{W}}|,
\end{aligned} \tag{B.30}$$

where the required quantities are defined at (B.22), (B.21), (B.20), and (B.6).

$\mathbb{E}_q[\log p(\Lambda)] - \mathbb{E}_q[\log q(\Lambda)] + \mathbb{E}_q[\log p(\boldsymbol{\alpha})] - \mathbb{E}_q[\log q(\boldsymbol{\alpha})]$

Auxiliary and ARD precision parameters both have Gamma prior and approximate posterior distributions and so are presented together. Note that this presentation assumes that $\Lambda_{i'}$ is a free parameter for all $i' \in \mathcal{R}_{\mathbf{X}}$. If this does not apply then the summation for Λ should be over free parameters only. Their contribution to ELBO (q) is

$$\begin{aligned}
&\mathbb{E}_q[\log p(\Lambda)] - \mathbb{E}_q[\log q(\Lambda)] + \mathbb{E}_q[\log p(\boldsymbol{\alpha})] - \mathbb{E}_q[\log q(\boldsymbol{\alpha})] \\
&= \sum_{i' \in \mathcal{R}_{\mathbf{X}}} \mathbb{E}_q[\log Gamma(\Lambda_{i'}|a_\Lambda, b_\Lambda)] - \mathbb{E}_q\left[\log Gamma\left(\Lambda_{i'}|\hat{a}_\Lambda^{i'}, \hat{b}_\Lambda^{i'}\right)\right] \\
&\qquad + \sum_{j=1}^{Q}\mathbb{E}_q[\log Gamma(\alpha_j|a_\alpha, b_\alpha)] - \mathbb{E}_q\left[\log Gamma\left(\alpha_j|\hat{a}_\alpha^{j}, \hat{b}_\alpha^{j}\right)\right], \\
&= \sum_{i' \in \mathcal{R}_{\mathbf{X}}} a_\Lambda \log b_\Lambda - \hat{a}_\Lambda^{i'}\log \hat{b}_\Lambda^{i'} - \log\Gamma(a_\Lambda) + \log\Gamma(\hat{a}_\Lambda^{i'}) \\
&\qquad\qquad + \left(a_\Lambda - \hat{a}_\Lambda^{i'}\right)\langle \log \Lambda_i \rangle - \left(b_\Lambda - \hat{b}_\Lambda^{i'}\right)\langle \Lambda_i \rangle
\end{aligned} \tag{B.31}$$

$$+\sum_{j=1}^{Q} a_\alpha \log b_\alpha - \hat{a}_\alpha^j \log \hat{b}_\alpha^j - \log\Gamma(a_\alpha) + \log\Gamma(\hat{a}_\alpha^j)$$
$$+ \left(a_\alpha - \hat{a}_\alpha^j\right) \langle \log \alpha_j \rangle - \left(b_\alpha - \hat{b}_\alpha^j\right) \langle \alpha_j \rangle \tag{B.32}$$

where the relevant quantities are defined at (B.14), (B.15), (B.17), (B.16), (B.18), (B.19), (B.22), and(B.21).

$\mathbb{E}_q\left[\log p\left(\boldsymbol{\gamma}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\gamma}\right)\right]$

The contribution of the ordinal trait cut-offs is relatively straightforward to compute, and is given by

$$\begin{aligned}
&\mathbb{E}_q\left[\log p\left(\boldsymbol{\gamma}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\gamma}\right)\right] \\
&= \sum_{i \in \mathcal{O}_{\mathbf{Y}}} \sum_{k \in \{2,\ldots.K_i - 1\}} \mathbb{E}_q\left[\log \mathcal{U}\left(\gamma_{i,k} | \gamma_{i,k-1}, \gamma_{i,k-1} + b_\gamma\right)\right] - 0, \\
&= \sum_{i \in \mathcal{O}_{\mathbf{Y}}} \sum_{k \in \{2,\ldots.K_i - 1\}} - \log b_\gamma,
\end{aligned} \tag{B.33}$$

where $\lfloor \gamma_{i,k} \rfloor$ and $\lceil \gamma_{i,k} \rceil$ have been defined along with (B.25).

$\mathbb{E}_q\left[\log p\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\theta}\right)\right]$

The final contribution to ELBO (q) comes from the phylogenetic hyperparameters $\boldsymbol{\theta}$. Two approaches to optimising $\boldsymbol{\theta}$ have been outlined, sampling from the optimal mean field variational family approximate posterior and setting $q\left(\boldsymbol{\theta}\right) = \delta\left(\boldsymbol{\theta} = \langle \boldsymbol{\theta} \rangle\right)$. Each of these approximations require different approaches to calculating ELBO (q). When the sampling approach has been taken

$$\mathbb{E}_q\left[\log p\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\theta}\right)\right] = \sum_{j=1}^{Q} \langle \log p\left(\theta_j\right) \rangle - \langle \log q\left(\theta_j\right) \rangle,$$

where $\langle \log p\left(\theta_j\right) \rangle$ is estimated in (B.24) and $\langle \log q\left(\theta_j\right) \rangle$ must be estimated using some multivariate estimation technique such as that provided by the "IndepTest" package in R [Berrett et al., 2018, 2019].

Alternatively, when approximating the posterior for $\boldsymbol{\theta}$ with an indicator function,

$$\mathbb{E}_q\left[\log p\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_q\left[\log q\left(\boldsymbol{\theta}\right)\right] = \sum_{j=1}^{Q} \log p\left(\langle \theta_j \rangle\right).$$

B.3 Predictive Distribution

Consider the variational predictive distribution over manifest traits at some unobserved position $\mathbf{t}_* \in \mathcal{T}$ approximating the true predictive distribution given the observed manifest trait, a known phylogeny, loading matrix, auxiliary precision matrix, ordinal trait cut-off points, hyper-parameters for the Gaussian processes over $\mathcal{T}$, and the ARD precision parameters. Assuming that each parameter takes it's expectation unter the optimal approximate posterior yields

$$
\begin{aligned}
&p\left(\mathbf{Y}_{*\cdot}|\mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \mathbf{W}, \Lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\alpha}\right) \\
&\approx q\left(\mathbf{Y}_{*\cdot}|\mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \langle\mathbf{W}\rangle, \langle\Lambda\rangle, \langle\boldsymbol{\gamma}\rangle, \langle\boldsymbol{\theta}\rangle, \langle\boldsymbol{\alpha}\rangle\right), \\
&= \int q\left(\mathbf{Y}_{*\cdot}\mathbf{X}_{*\cdot}, \mathbf{Z}_{*\cdot}|\mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \langle\mathbf{W}\rangle, \langle\Lambda\rangle, \langle\boldsymbol{\gamma}\rangle, \langle\boldsymbol{\theta}\rangle, \langle\boldsymbol{\alpha}\rangle\right) d\mathbf{X}_{*\cdot}\, d\mathbf{Z}_{*\cdot}, \\
&= \int p\left(\mathbf{Y}_{*\cdot}|\mathbf{X}_{*\cdot}, \langle\boldsymbol{\gamma}\rangle\right) p\left(\mathbf{X}_{*\cdot}|\mathbf{Z}_{*\cdot}, \langle\mathbf{W}\rangle, \langle\Lambda\rangle\right) q\left(\mathbf{Z}_{*\cdot}\right) d\mathbf{X}_{*\cdot}\, d\mathbf{Z}_{*\cdot}, \\
&= \int \delta\left(\mathbf{Y}_{*\cdot} = g\left(\mathbf{X}_{*\cdot}\right)\right) \mathcal{N}\left(\mathbf{X}_{*\cdot}|\langle\mathbf{W}\rangle\mathbf{Z}_{*\cdot}, \langle\Lambda\rangle^{-1}\right) \mathcal{N}\left(\mathbf{Z}_{*\cdot}|\langle\mathbf{Z}_{*\cdot}\rangle, \mathbf{S}_*^{\mathbf{Z}}\right) d\mathbf{X}_{*\cdot}\, d\mathbf{Z}_{*\cdot}, \\
&= \int \delta\left(\mathbf{Y}_{*\cdot} = g\left(\mathbf{X}_{*\cdot}\right)\right) \mathcal{N}\left(\mathbf{X}_{*\cdot}|\langle\mathbf{W}\rangle\langle\mathbf{Z}_{*\cdot}\rangle, \langle\Lambda\rangle^{-1} + \langle\mathbf{W}\rangle\mathbf{S}_*^{\mathbf{Z}}\langle\mathbf{W}\rangle^\top\right) d\mathbf{X}_{*\cdot}.
\end{aligned}
$$

To obtain the marginal predictive distribution for each manifest trait, recall that $\Lambda_{i'} = 1$ for all $i' \in \{\mathcal{O}_{\mathbf{X}}, \mathcal{C}_{\mathbf{X}}\}$ and set $\nu_{i'}^* = \sqrt{1 + \langle\mathbf{W}_{i'\cdot}\rangle^\top \mathbf{S}_*^{\mathbf{Z}} \langle\mathbf{W}_{i'\cdot}\rangle}$, then for $i \in \mathcal{O}_{\mathbf{Y}}$

$$
\begin{aligned}
&p\left(\mathbf{Y}_{*i} = k|\mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \mathbf{W}, \Lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\alpha}\right) \\
&\approx \int \delta\left(\langle\gamma_{i,k-1}\rangle \leq \mathbf{X}_{*i'} < \langle\gamma_{i,k}\rangle\right) \mathcal{N}\left(\mathbf{X}_{*i'}|\langle\mathbf{W}_{i'\cdot}\rangle^\top\langle\mathbf{Z}_{*\cdot}\rangle, \left(\nu_{i'}^*\right)^2\right) d\mathbf{X}_{*i'}. \\
&= \int_{\langle\gamma_{i,k-1}\rangle}^{\langle\gamma_{i,k}\rangle} \mathcal{N}\left(\frac{\mathbf{X}_{*i'} - \langle\mathbf{W}_{i'\cdot}\rangle^\top\langle\mathbf{Z}_{*\cdot}\rangle}{\nu_{i'}^*}|0, 1\right) d\mathbf{X}_{*i'}. \\
&= F_{\mathcal{N}}\left(\frac{\langle\gamma_{i,k}\rangle - \langle\mathbf{W}_{i'\cdot}\rangle^\top\langle\mathbf{Z}_{*\cdot}\rangle}{\nu_{i'}^*}\right) - F_{\mathcal{N}}\left(\frac{\langle\gamma_{i,k-1}\rangle - \langle\mathbf{W}_{i'\cdot}\rangle^\top\langle\mathbf{Z}_{*\cdot}\rangle}{\nu_{i'}^*}\right)
\end{aligned}
$$

where $F_{\mathcal{N}}\left(\cdot\right)$ is the standard Gaussian CDF.

Deriving the predictive distribution for categorical traits requires the definition of further notation. Let $\{\mathbf{W}\}_{*i}$ and $\{\mathbf{X}\}_{*i}$ denote the loading and auxiliary traits associated with manifest trait $\mathbf{Y}_{*i}$, while $\{\mathbf{W}_{-k}\}_{*i}$ and $\{\mathbf{X}_{-k}\}_{*i}$ are the same auxiliary traits and loadings less those associated with $c_{i,k}$. For notational ease, define $\mathbf{m}^{i*} \equiv \langle\{\mathbf{W}\}_{*i}\rangle\langle\mathbf{Z}_{*\cdot}\rangle$, $\mathbf{N}^{i*} \equiv \mathbf{I} + \langle\{\mathbf{W}\}_{*i}\rangle\mathbf{S}_*^{\mathbf{Z}}\langle\{\mathbf{W}\}_{*i}\rangle$ and $\nu_{i'+k-1}^* \equiv \nu_k^*$, which can be extended to $\tilde{\mathbf{m}}_{-k}^{i*} = \mathbf{m}_{-k}^{i*} + \mathbf{N}_{-k,k}^{i*}\left(\nu_k^*\right)^{-2}\left(\mathbf{X}_{*i'+k-1} - \mathbf{m}_k^{i*}\right)$and $\mathbf{N}_{-k}^{i*} = \mathbf{N}_{-k,-k}^{i*} - \mathbf{N}_{-k,k}^{i*}\left(\nu_k^*\right)^{-2}\mathbf{N}_{k,-k}^{i*}$, where k again refers to $c_{i,k}$. Let $\mathbf{L}_{-k}^i\left(\mathbf{L}_{-k}^i\right)^\top = \mathbf{N}_{-k}^{i*}$,

then, when $i \in \mathcal{C}_{\mathbf{Y}}$

$$
\begin{aligned}
&p\left(\mathbf{Y}_{*i} = c_{i,k} | \mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \mathbf{W}, \Lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\alpha}\right) \\
&\approx \int \delta\left(\mathbf{X}_{*i'+k-1} > \mathbf{X}_{*i'+l-1} \, \forall \, l \neq k\right) \mathcal{N}\left(\{\mathbf{X}\}_{*i} | \mathbf{m}^{i*}, \mathbf{N}^{i*}\right) d\{\mathbf{X}\}_{*i}, \\
&= \int_{-\infty}^{\infty} \mathcal{N}\left(\mathbf{X}_{*i'+k-1} | \mathbf{m}_k^{i*}, \left(\nu_k^*\right)^2\right) \int_{-\infty}^{\mathbf{X}_{*i'+k-1}} \mathcal{N}\left(\{\mathbf{X}_{-k}\}_{*i} | \tilde{\mathbf{m}}_{-k}^{i*}, \mathbf{N}_{-k}^{i*}\right) d\{\mathbf{X}\}_{*i} \, d\mathbf{X}_{*i'+k-1}, \\
&= \int_{-\infty}^{\infty} \mathcal{N}\left(u | 0, 1\right) \\
&\qquad \int_{-\infty}^{u\nu_k^* + \mathbf{m}_k^{i*}} \mathcal{N}\left(\{\mathbf{X}_{-k}\}_{*i} | \mathbf{m}_{-k}^{i*} + \mathbf{N}_{-k,k}^{i*} \left(\nu_k^*\right)^{-1} u, \mathbf{L}_{-k}^{i} \left(\mathbf{L}_{-k}^{i}\right)^{\top}\right) d\{\mathbf{X}\}_{*i} \, du, \\
&= \mathbb{E}_{p(u)}\left[\prod_{l=1}^{K_i - 1} \int_{-\infty}^{\mathbf{u}_l^{ki*}} \mathcal{N}\left(v_l | 0, 1\right) dv_l\right], \\
&= \mathbb{E}_{p(u)}\left[\prod_{l=1}^{K_i - 1} F_{\mathcal{N}}\left(\mathbf{u}_l^{ki*}\right)\right],
\end{aligned}
$$

where $\mathbf{u}^{ki*}$ has been defined such that $\mathbf{L}_{-k}^{i} \mathbf{u}^{ki*} = \left(u\nu_k^* + \mathbf{m}_k^{i*}\right) \mathbf{1} - \left(\mathbf{m}_{-k}^{i*} + \mathbf{N}_{-k,k}^{i*} \left(\nu_k^*\right)^{-1} u\right)$, where $\mathbf{1}$ is a vector of 1's.

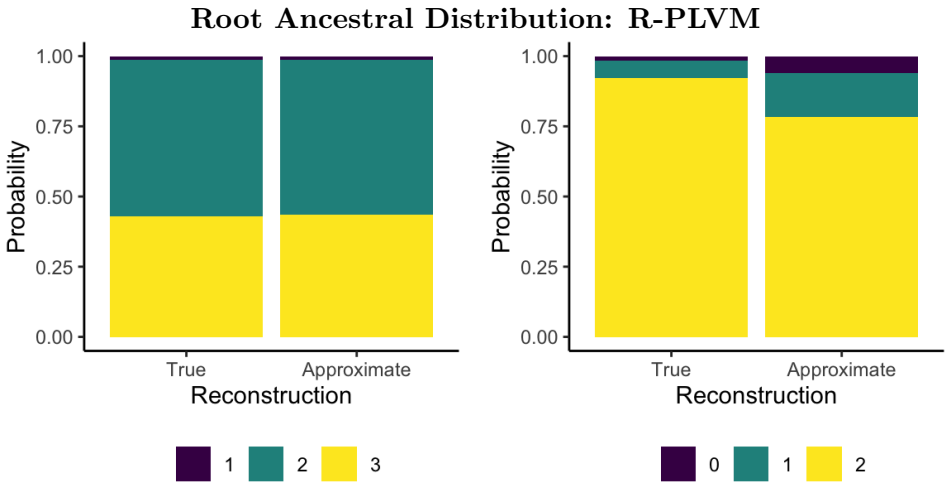
Finally, when $i \in \mathcal{R}_{\mathbf{Y}}$,

$$
\begin{aligned}
&p\left(\mathbf{Y}_{*i} = c_{i,k} | \mathbf{t}_*, \mathbf{Y}, \mathcal{T}, \mathbf{W}, \Lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\alpha}\right) \\
&\approx \int \delta\left(\mathbf{X}_{*i'} = \mathbf{Y}_{*i}\right) \mathcal{N}\left(\mathbf{X}_{*i'} | \langle \mathbf{W}_{i'\cdot} \rangle^{\top} \langle \mathbf{Z}_{*\cdot} \rangle, \left(\nu_{i'}^*\right)^2\right) d\{\mathbf{X}\}_{*i}, \\
&= \mathcal{N}\left(\mathbf{Y}_{*i} | \langle \mathbf{W}_{i'\cdot} \rangle^{\top} \langle \mathbf{Z}_{*\cdot} \rangle, \left(\nu_{i'}^*\right)^2\right).
\end{aligned}
$$

Appendix C

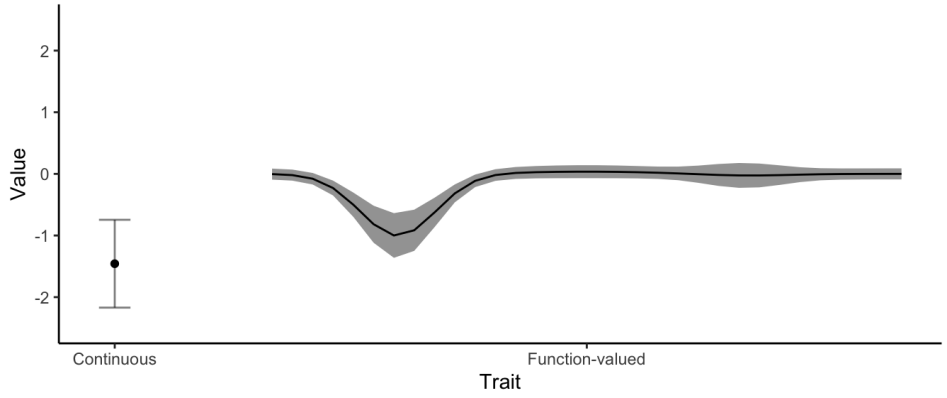
Alternative generalised PLVMs

The ancestral distributions implied by the R-PLVM, P-PLVM, and I-PLVM models fit in Chapter 4 are presented in the following.



(a) Ordinal Trait

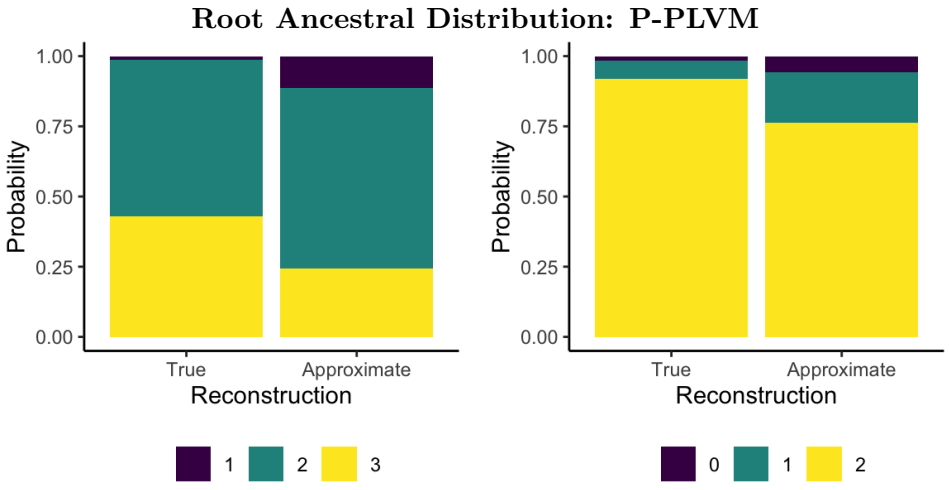
(b) Categorical Trait



(c) Ancestral Distribution

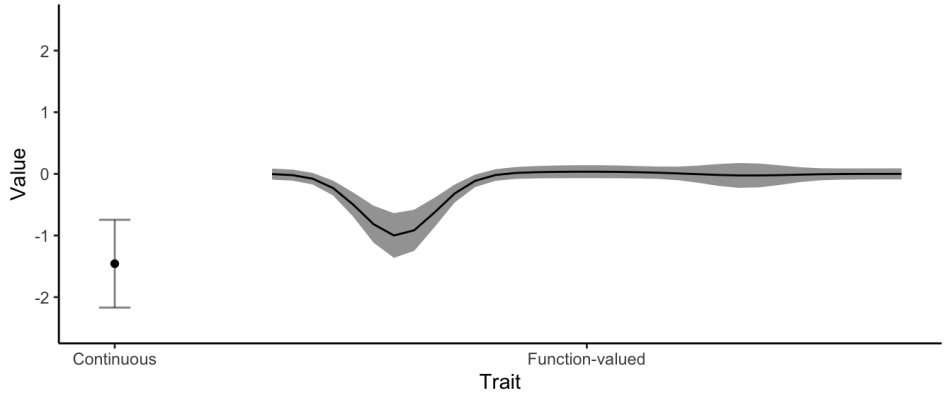
(d) Approximate Ancestral Distribution

Figure C.1: A comparison of the true ancestral distribution at the root of $\mathcal{T}$, with approximate ancestral distribution given by R-PLVM. In (a) and (b) each colour in the bars represent the probability that the trait was of that particular state, while in (d) and (c), grey error markers represent two standard deviations from the mean.

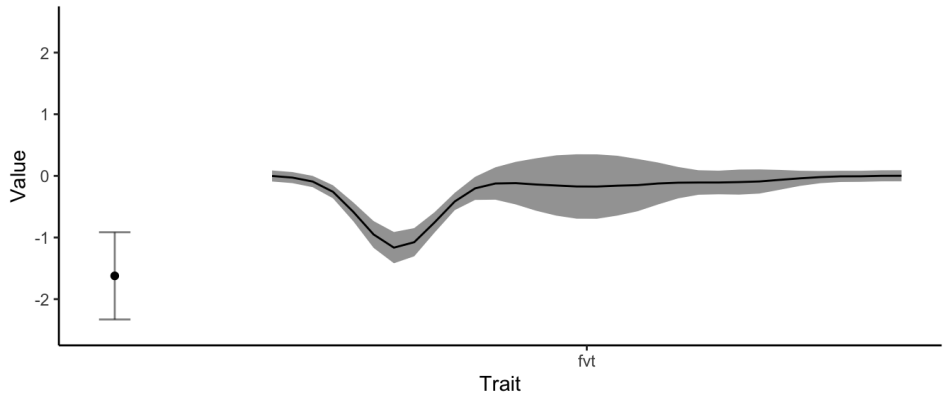


(a) Ordinal Trait

(b) Categorical Trait

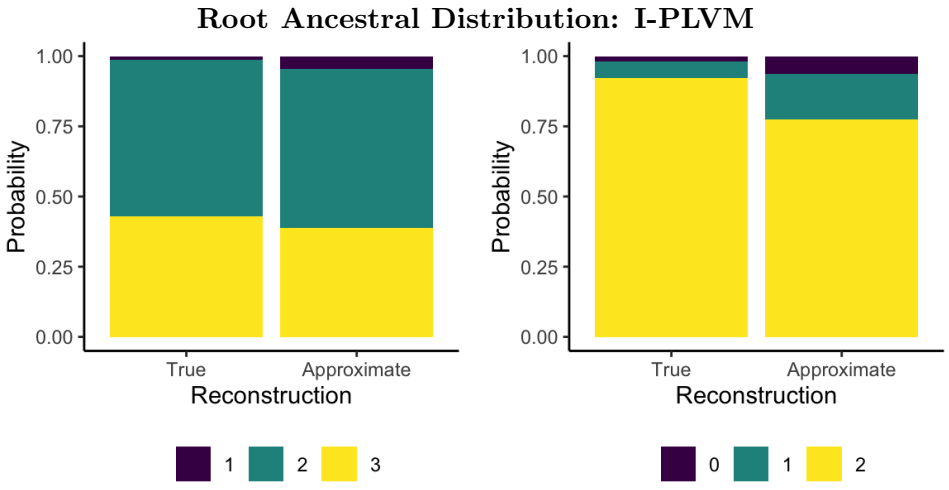


(c) Ancestral Distribution



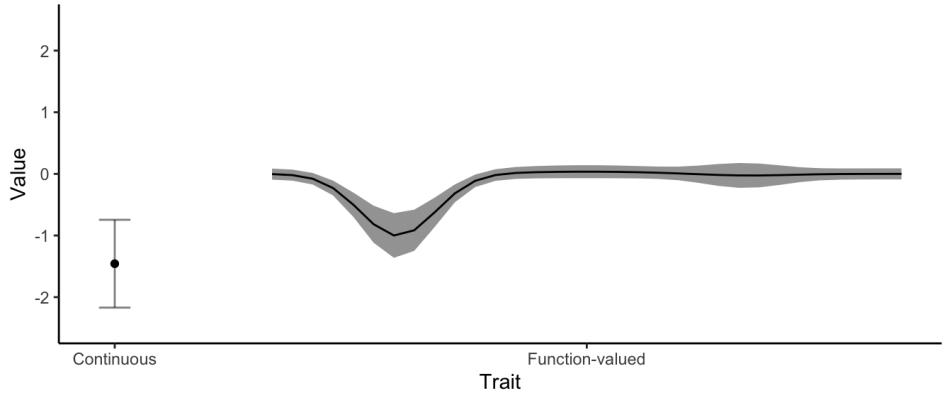
(d) Approximate Ancestral Distribution

Figure C.2: A comparison of the true ancestral distribution at the root of $\mathcal{T}$, with approximate ancestral distribution given by P-PLVM. In (a) and (b) each colour in the bars represent the probability that the trait was of that particular state, while in (d) and (c), grey error markers represent two standard deviations from the mean.

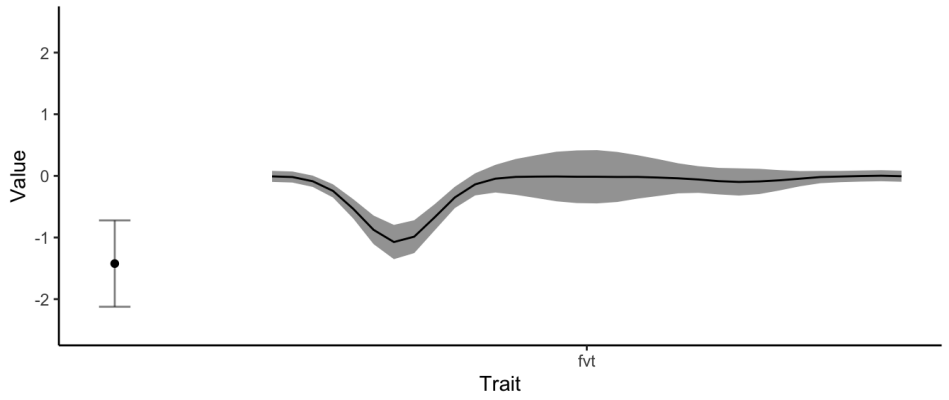


(a) Ordinal Trait

(b) Categorical Trait



(c) Ancestral Distribution



(d) Approximate Ancestral Distribution

Figure C.3: A comparison of the true ancestral distribution at the root of $\mathcal{T}$, with approximate ancestral distribution given by I-PLVM. In (a) and (b) each colour in the bars represent the probability that the trait was of that particular state, while in (d) and (c), grey error markers represent two standard deviations from the mean.

Bibliography

Dean C Adams and Michael L Collyer. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. *Systematic biology*, 67(1): 14–31, 2017.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422): 669–679, 1993.

Jont B Allen and Lawrence R Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

Lucila I Amador, R Leticia Moyers Arévalo, Francisca C Almeida, Santiago A Catalano, and Norberto P Giannini. Bat systematics in the light of unconstrained analyses of a comprehensive molecular supermatrix. *Journal of Mammalian Evolution*, 25(1):37–70, 2018.

Loren K Ammerman and David M Hillis. A molecular test of bat relationships: monophyly or diphyly? *Systematic Biology*, 41(2):222–232, 1992.

S Banerjee and AE Gelfand. On smoothness properties of spatial processes. *Journal of Multivariate Analysis*, 84(1):85–100, 2003.

David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, 2011.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

Thomas B. Berrett, Daniel J. Grose, and Richard J. Samworth. *IndepTest: Nonparametric Independence Tests Based on Entropy Estimation*, 2018. URL `https://CRAN.R-project.org/package=IndepTest`. R package version 0.2.0.

Thomas B Berrett, Richard J Samworth, Ming Yuan, et al. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Christopher M Bishop. Variational principal components. 1999.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Tobias Blaschke and Laurenz Wiskott. An improved cumulant based method for independent component analysis. In *International Conference on Artificial Neural Networks*, pages 1087–1093. Springer, 2002.

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

Simon P Blomberg, Theodore Garland Jr, and Anthony R Ives. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003.

Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

Fred L Bookstein et al. Size and shape spaces for landmark data in two dimensions. *Statistical science*, 1(2):181–222, 1986.

Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

Daniel Sabanés Bové, Leonhard Held, et al. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 2011.

Signe Brinkløv, M Brock Fenton, and John Morgan Ratcliffe. Echolocation in oilbirds and swiftlets. *Frontiers in physiology*, 4:123, 2013.

Stephen Butterworth et al. On the theory of filter amplifiers. *Wireless Engineer*, 7 (6):536–541, 1930.

Arturo Camacho and John G Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3): 1638–1652, 2008.

Neil A Campbell, Lawrence G Mitchell, Jane B Reece, and Jacqueline Bishop. *Biology: concepts & connections*. Number QH308. 2 C35 1996. Benjamin Cummings Menlo Park, Calif., 1997.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76 (1), 2017.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Luigi L Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570, 1967.

Wen Cheng, Ian L Dryden, Xianzheng Huang, et al. Bayesian registration of functions and curves. *Bayesian Analysis*, 11(2):447–475, 2016.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Leon Cohen. *Time-frequency analysis*, volume 778. Prentice hall, 1995.

AL Collen. *The evolution of echolocation in bats: a comparative approach*. PhD thesis, UCL (University College London), 2012.

Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, and Ole Winther. Efficient approaches to gaussian process classification. In *Advances in neural information processing systems*, pages 251–257, 2000.

Gabriela B Cybis, Janet S Sinsheimer, Trevor Bedford, Alison E Mather, Philippe Lemey, and Marc A Suchard. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The annals of applied statistics*, 9 (2):969, 2015.

Rainer Dahlhaus. On the kullback-leibler information divergence of locally stationary processes. *Stochastic processes and their applications*, 62(1):139–168, 1996.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Theodoros Damoulas and Mark A Girolami. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 24(10):1264–1270, 2008.

Theodoros Damoulas, Samuel Henry, Andrew Farnsworth, Michael Lanzone, and Carla Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 424–429. IEEE, 2010.

Tung Dang and Hirohisa Kishino. Stochastic variational inference for bayesian phylogenetics: A case of cat model. *Molecular biology and evolution*, 36(4):825–833, 2019.

Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* Routledge, 1859.

Manuel Davy and SJ Godsill. Bayesian harmonic models for musical signal analysis. *Bayesian Statistics*, 7:105–124, 2003.

A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

Richard Dawkins. *The blind watchmaker: Why the evidence of evolution reveals a universe without design.* WW Norton & Company, 1996.

Carl De Boor. On calculating with b-splines. *Journal of Approximation theory*, 6 (1):50–62, 1972.

Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4): 1917–1930, 2002.

John R Deller Jr and John Hansen. Methods, models, and algorithms for modern speech processing. *The Electrical Engineering Handbook*, page 861, 2004.

Mark Denny. *Blip, Ping, and Buzz: Making Sense of Radar and Sonar*. JHU Press, 2007.

Annette Denzinger and Hans-Ulrich Schnitzler. Bat guilds, a concept to classify the highly diverse foraging and echolocation behaviors of microchiropteran bats. *Frontiers in physiology*, 4:164, 2013.

John DiCecco, Jason E Gaudette, and James A Simmons. Multi-component separation and analysis of bat echolocation calls. *The Journal of the Acoustical Society of America*, 133(1):538–546, 2013.

GE Dobson. Xlvii.—conspectus of the suborders, families, and genera of chiroptera arranged according to their natural affinities. *Annals and Magazine of Natural History*, 16(95):345–357, 1875.

Joseph L Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, pages 351–369, 1942.

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

Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29 (8):1969–1973, 2012.

Ian L Dryden and Kanti V Mardia. *Statistical Shape Analysis: With Applications in R*, volume 995. John Wiley & Sons, 2016.

Geeta N Eick, David S Jacobs, and Conrad A Matthee. A nuclear dna phylogenetic perspective on the evolution of echolocation and historical biogeography of extant bats (chiroptera). *Molecular Biology and Evolution*, 22(9):1869–1886, 2005.

Michael G Elliot and Arne Ø Mooers. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. *BMC evolutionary biology*, 14(1):226, 2014.

Elena A Erosheva and S McKay Curtis. Dealing with reflection invariance in bayesian factor analysis. *psychometrika*, 82(2):295–307, 2017.

Joseph Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics*, 25(5):471, 1973.

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

Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

Joseph Felsenstein. A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, 179(2):145–156, 2011.

M Brock Fenton and Gary P Bell. Recognition of species of insectivorous bats by their echolocation calls. *Journal of Mammalogy*, 62(2):233–243, 1981.

M Brock Fenton, Paul A Faure, and John M Ratcliffe. Evolution of high duty cycle echolocation in bats. *Journal of Experimental Biology*, 215(17):2935–2944, 2012.

M Brock Fenton, Alan D Grinnell, Arthur N Popper, and Richard R Fay. *Bat bioacoustics*, volume 54. Springer, 2016.

Maurizio Filippone, Mingjun Zhong, and Mark Girolami. A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning*, 93(1):93–114, 2013.

Walter M Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

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

Yanqing Fu and Laura N Kloepper. A systematic method for isolating, tracking and discriminating time-frequency components of bat echolocation calls. *The Journal of the Acoustical Society of America*, 143(2):716–726, 2018.

J Fullard and J Dawson. The echolocation calls of the spotted bat euderma maculatum are relatively inaudible to moths. *Journal of Experimental Biology*, 200(1): 129–137, 1997.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. *arXiv preprint arXiv:1710.06595*, 2017.

Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

Robert Galambos. The avoidance of obstacles by flying bats: Spallanzani's ideas (1794) and later theories. *Isis*, 34(2):132–140, 1942.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

David Gerhard. *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.

Daniel Gervini and Theo Gasser. Self-modelling warping functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):959–971, 2004.

John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587, 1996.

Zoubin Ghahramani and Matthew J Beal. Variational inference for bayesian mixtures of factor analysers. In *Advances in neural information processing systems*, pages 449–455, 2000.

Mark Girolami and Simon Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

Gene H Golub and Charles F Van Loan. *Matrix computations*. Johns Hopkins University Press, 4th edition, 2013.

Richard Gomulkiewicz, Joel G Kingsolver, Patrick A Carter, and Nancy Heckman. Variation and evolution of function-valued traits. *Annual Review of Ecology, Evolution, and Systematics*, 49:139–164, 2018.

Sira Gonzalez and Mike Brookes. Pefac-a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(2):518–530, 2014.

Eric W Goolsby. Phylogenetic comparative methods for evaluating the evolutionary history of function-valued traits. *Systematic biology*, 64(4):568–578, 2015.

Alan Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 326(1233):119–157, 1989.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Donald R Griffin. Echolocation by blind men, bats and radar. *Science*, 100(2609): 589–590, 1944.

Donald R Griffin and Robert Galambos. The sensory basis of obstacle avoidance by flying bats. *Journal of Experimental zoology*, 86(3):481–506, 1941.

Donald R Griffin, Frederic A Webster, and Charles R Michael. The echolocation of flying insects by bats. *Animal behaviour*, 8(3-4):141–154, 1960.

Quentin F Gronau, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S Leslie, Jonathan J Forster, Eric-Jan Wagenmakers, and Helen Steingroever. A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97, 2017a.

Quentin F Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. Bridgesampling: An r package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*, 2017b.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

Beniamino Hadj-Amar, Bärbel Finkenstädt, Mark Fiecas, Francis Lévi, and Robert Huckstepp. Bayesian model search for nonstationary periodic time series. *Journal of the American Statistical Association*, (just-accepted):1–36, 2019.

Pantelis Z Hadjipantelis, John AD Aston, and Jonathan P Evans. Characterizing fundamental frequency in mandarin: A functional principal component approach utilizing mixed effect models. *The Journal of the Acoustical Society of America*, 131(6):4651–4664, 2012.

Pantelis Z Hadjipantelis, Nick S Jones, John Moriarty, David A Springate, and Christopher G Knight. Function-valued traits in evolution. *Journal of The Royal Society Interface*, 10(82):20121032, 2013.

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

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Franz Hlawatsch and François Auger. *Time-frequency analysis*. Wiley Online Library, 2008.

Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical models with R*. Springer Science & Business Media, 2012.

Richard A Holland, Dean A Waters, and Jeremy MV Rayner. Echolocation signal structure in the megachiropteran bat rousettus aegyptiacus geoffroy 1810. *Journal of experimental biology*, 207(25):4361–4369, 2004.

Steven L Hopp, Michael J Owren, and Christopher S Evans. *Animal acoustic communication: Sound analysis and research methods*. Springer Science & Business Media, 2012.

Elizabeth A Housworth, Emília P Martins, and Michael Lynch. The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96, 2004.

Norden E Huang, Zhaohua Wu, Steven R Long, Kenneth C Arnold, Xianyao Chen, and Karin Blank. On instantaneous frequency. *Advances in adaptive data analysis*, 1(02):177–229, 2009.

Anthony R Ives and Theodore Garland Jr. Phylogenetic logistic regression for binary dependent variables. *Systematic biology*, 59(1):9–26, 2009.

Andrew F Jarosz and Jennifer Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, 7(1):2, 2014.

Harold Jeffreys. The theory of probability. *The Theory of Probability, 3rd Edition, by Harold Jeffreys. Oxford Classic Texts in the Physical Sciences. ISBN: 9780198503682. Oxford: Oxford University Press, 1939*, 1939.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Gareth Jones and Marc W Holderied. Bat echolocation calls: adaptation and convergent evolution. *Proceedings of the Royal Society B: Biological Sciences*, 274 (1612):905–912, 2007.

Gareth Jones and Emma C Teeling. The evolution of echolocation in bats. *Trends in Ecology & Evolution*, 21(3):149–156, 2006.

Gareth Jones, David S Jacobs, Thomas H Kunz, Michael R Willig, and Paul A Racey. Carpe noctem: the importance of bats as bioindicators. *Endangered species research*, 8(1-2):93–115, 2009.

Nick S Jones and John Moriarty. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of The Royal Society Interface*, 10(78):20120616, 2013.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Jeffrey B Joy, Richard H Liang, Rosemary M McCloskey, T Nguyen, and Art FY Poon. Ancestral reconstruction. *PLoS computational biology*, 12(7):e1004763, 2016.

Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

David G Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

Mark Kirkpatrick and Nancy Heckman. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of mathematical biology*, 27(4):429–450, 1989.

Mark Kirkpatrick, David Lofsvold, and Michael Bulmer. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, 124(4):979–993, 1990.

Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, pages 1266–1305, 1992.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC Press, 2017.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Russell Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334, 1976.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(Nov): 1783–1816, 2005.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

Henry George Liddell and Robert Scott. *A greek-english lexicon*. New York: American Book Company, 1897.

Hedibert Freitas Lopes. Modern bayesian factor analysis. *Bayesian Inference in the Social Sciences*, pages 115–153, 2014.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.

Patrick J Loughlin and Berkant Tacer. On the amplitude-and frequency-modulation decomposition of signals. *The Journal of the Acoustical Society of America*, 100 (3):1594–1601, 1996.

Oisin Mac Aodha, Rory Gibb, Kate E Barlow, Ella Browning, Michael Firman, Robin Freeman, Briana Harder, Libby Kinsey, Gary R Mead, Stuart E Newson, et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, 14(3):e1005995, 2018.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Alanna Maltby, Kate E Jones, and Gareth Jones. Understanding the evolutionary origin and diversification of bat echolocation calls. In *Handbook of Behavioral Neuroscience*, volume 19, pages 37–47. Elsevier, 2010.

Irene Mariñas-Collado, Adrian Bowman, and Vincent Macaulay. A phylogenetic gaussian process model for the evolution of curves embedded in d-dimensions. *Computational Statistics & Data Analysis*, 2019.

Emilia P Martins and Thomas F Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

JP Meagher, T Damoulas, KE Jones, and M Girolami. Discussion of "the statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages", by davide pigoli, pantelis z hadjipantelis, john s coleman, and john ad aston. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1132, 2018a.

JP Meagher, T Damoulas, KE Jones, and M Girolami. Phylogenetic gaussian processes for bat echolocation. *Statistical Data Science*, pages 111–122, 2018b. doi: 10.1142/q0159. URL `https://www.worldscientific.com/doi/abs/10.1142/q0159`.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Karin Meyer and Mark Kirkpatrick. Up hill, down dale: quantitative genetics of curvaceous traits. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1443–1455, 2005.

Frank Miller, Klaus Janson, Lynn Varley, John Costanza, and Bob Kane. *Batman: The dark knight returns*. Dc Comics New York, 2002.

Gerrit Smith Miller. *The families and genera of bats*. Number 57. US Government Printing Office, 1907.

Mary Ellen Miller and Karl Taube. *An illustrated dictionary of the gods and symbols of ancient Mexico and the Maya*. Thames and Hudson Londres, 1997.

Venelin Mitov and Tanja Stadler. Poumm: An r-package for bayesian inference of phylogenetic heritability. *bioRxiv*, page 115089, 2017.

Karla Monterrubio-Gómez, Lassi Roininen, Sara Wade, Theo Damoulas, and Mark Girolami. Posterior inference for sparse hierarchical non-stationary models. *arXiv preprint arXiv:1804.01431*, 2018.

Cynthia F Moss, Chen Chiu, and Annemarie Surlykke. Adaptive vocal behavior drives perception by echolocation in bats. *Current opinion in neurobiology*, 21(4): 645–652, 2011.

Tamara Münkemüller, Sebastien Lavergne, Bruno Bzeznik, Stephane Dray, Thibaut Jombart, Katja Schiffers, and Wilfried Thuiller. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756, 2012.

Iain Murray and Ryan P Adams. Slice sampling covariance hyperparameters of latent gaussian models. In *Advances in neural information processing systems*, pages 1732–1740, 2010.

Iain Murray, Ryan P Adams, and David JC MacKay. Elliptical slice sampling. *Journal of Machine Learning Research*, 9:541–548, 2010.

Cory S Myers and Lawrence R Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409, 1981.

Thomas Nagel. What is it like to be a bat? *The philosophical review*, 83(4):435–450, 1974.

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, 49(2):343–360, 2020.

Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.

Jesper Kjær Nielsen, Mads Græsbøll Christensen, Ali Taylan Cemgil, and Søren Holdt Jensen. Bayesian model comparison with the g-prior. *IEEE Transactions on Signal Processing*, 62(1):225–238, 2013.

Jesper Kjær Nielsen, Tobias Lindstrøm Jensen, Jesper Rindom Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient. *Signal Processing*, 135:188–197, 2017.

John P Nolan. *Stable distributions*, volume 1177108605. ISBN, 2012.

A Michael Noll. Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309, 1967.

Ronald M Nowak and Ernest P Walker. *Walker's bats of the world*. JHU Press, 1994.

S Olhede and Andrew T Walden. A generalized demodulation approach to time-frequency projections for multicomponent signals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 461(2059):2159–2179, 2005.

Alan V Oppenheim and Ronald W Schafer. *Discrete-time signal processing*. Pearson Education, 2014.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401 (6756):877, 1999a.

Mark Pagel. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic biology*, 48(3):612–622, 1999b.

E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2018.

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

Scott C Pedersen. Morphometric analysis of the chiropteran skull with regard to mode of echolocation. *Journal of Mammalogy*, 79(1):91–103, 1998.

Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge university press, 2006.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. *Technical University of Denmark*, November 2012.

John D Pettigrew. Flying primates? megabats have the advanced pathway from eye to midbrain. *Science*, 231(4743):1304–1306, 1986.

Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. The statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1103–1145, 2018.

M Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):55–69, 1980.

Oliver G Pybus, Marc A Suchard, Philippe Lemey, Flavien J Bernardin, Andrew Rambaut, Forrest W Crawford, Rebecca R Gray, Nimalan Arinaminpathy, Susan L Stramer, Michael P Busch, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071, 2012.

BG Quinn and PJ Thomson. Estimating the frequency of a periodic function. *Biometrika*, 78(1):65–74, 1991.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

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

Hafez A Radi and John O Rasmussen. *Principles of physics: for scientists and engineers*. Springer Science & Business Media, 2012.

James O Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4, 2004.

James O Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

JO Ramsay, RD Bock, and Theo Gasser. Comparison of height acceleration curves in the fels, zurich, and berkeley growth data. *Annals of Human Biology*, 22(5): 413–426, 1995.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian process for machine learning*. MIT press, 2006.

George P Rédei. *Encyclopedia of genetics, genomics, proteomics, and informatics*. Springer Netherlands, 2008.

Robert Redgwell, Joseph Szewczak, Gareth Jones, and Stuart Parsons. Classification of echolocation calls from 14 species of bat by support vector machines and ensembles of neural networks. *Algorithms*, 2(3):907–924, 2009.

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

WDJL Ride et al. *International code of zoological nomenclature*. International Trust for Zoological Nomenclature, 1999.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

HU Schnitzler, EKV Kalko, and A Denzinger. Evolution of echolocation and foraging behavior in bats. *Echolocation in bats and dolphins*, pages 331–339, 2004.

IJ Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. part a. on the problem of smoothing or graduation. a first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4 (1):45–99, 1946a.

Isaac Jacob Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. part b. on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141, 1946b.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

George AF Seber. *Multivariate observations*, volume 252. John Wiley & Sons, 2009.

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen, Max A Little, and Mads Graesboll Christensen. Bayesian pitch tracking based on the harmonic model. *arXiv preprint arXiv:1905.08557*, 2019.

James A Simmons and Roger A Stein. Acoustic imaging in bat sonar: echolocation signals and the evolution of echolocation. *Journal of comparative physiology*, 135 (1):61–84, 1980.

Nancy B Simmons. The case for chiropteran monophyly. *Amer. museum novitates*, 1994.

Nancy B Simmons. Order chiroptera. *Mammal species of the world: a taxonomic and geographic reference*, pages 312–529, 2005.

Nancy B Simmons, Kevin L Seymour, Jörg Habersetzer, and Gregg F Gunnell. Primitive early eocene bat from wyoming and the evolution of flight and echolocation. *Nature*, 451(7180):818, 2008.

Jim R Skinner. *Getting The Most From Medical VOC Data Using Bayesian Feature Learning*. PhD thesis, University of Warwick, 2019.

James Dale Smith. Chiropteran evolution. *Biology of the Bats of the New World Family Phylostomidae, Part I*, pages 49–69, 1976.

Charles Spearman. General intelligence objectively determined and measured. *American Journal of Psychology 15*, pages 202–292, 1904.

Anuj Srivastava and Eric P Klassen. *Functional and shape data analysis*. Springer, 2016.

Anuj Srivastava, Wei Wu, Sebastian Kurtek, Eric Klassen, and James Stephen Marron. Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

Vassilios Stathopoulos, Veronica Zamora-Gutierrez, Kate E Jones, and Mark Girolami. Bat echolocation call identification for biodiversity monitoring: A probabilistic approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):165–183, 2018.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

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

Bram Stoker. *Dracula*. Broadview Press, 1897.

Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016, 2018.

Vivien Sung. *Five-fold happiness: Chinese concepts of luck, prosperity, longevity, happiness, and wealth*. Chronicle Books, 2002.

Annemarie Surlykke, Paul E Nachtigall, Richard R Fay, Arthur N Popper, et al. *Biosonar*, volume 51. Springer, 2014.

Matthew RE Symonds and Simon P Blomberg. A primer on phylogenetic generalised least squares. In *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 105–130. Springer, 2014.

David Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518, 1995.

Rong Tang and Hans-Georg Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

Emma C Teeling, Mark Scally, Diana J Kao, Michael L Romagnoli, Mark S Springer, and Michael J Stanhope. Molecular evidence regarding the origin of echolocation and flight in bats. *Nature*, 403(6766):188, 2000.

Emma C Teeling, Mark S Springer, Ole Madsen, Paul Bates, Stephen J O'brien, and William J Murphy. A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, 307(5709):580–584, 2005.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

Max R Tolkoff, Michael E Alfaro, Guy Baele, Philippe Lemey, and Marc A Suchard. Phylogenetic factor analysis. *Systematic biology*, 67(3):384–399, 2017.

Georgia Tsagkogeorga, Joe Parker, Elia Stupka, James A Cotton, and Stephen J Rossiter. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Current Biology*, 23(22):2262–2267, 2013.

J. Derek Tucker. *fdasrvf: Elastic Functional Data Analysis*, 2019. URL `https://CRAN.R-project.org/package=fdasrvf`. R package version 1.9.2.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

Leigh M Van Valen. *The evolution of bats*. 1979.

Nina Veselka, David D McErlain, David W Holdsworth, Judith L Eger, Rethy K Chhem, Matthew J Mason, Kirsty L Brain, Paul A Faure, and M Brock Fenton. A bony connection signals laryngeal echolocation in bats. *Nature*, 463(7283):939, 2010.

Zhe Wang, Tengteng Zhu, Huiling Xue, Na Fang, Junpeng Zhang, Libiao Zhang, Jian Pang, Emma C Teeling, and Shuyi Zhang. Prenatal development supports a single origin of laryngeal echolocation in bats. *Nature ecology & evolution*, 1(2): 0021, 2017.

Sewall Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 19(6):506, 1934.

Fangzhou Yao, Jeff Coquery, and Kim-Anh Lê Cao. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, 13(1):24, 2012.

Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

Cheng Zhang and Frederick A Matsen IV. Variational bayesian phylogenetic inference. 2018.

Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99 (465):250–261, 2004.