CONFIDENCE PROCEDURES FOR PHYLOGENETIC TREES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Amy Donaldson Willis

May 2017

CONFIDENCE PROCEDURES FOR PHYLOGENETIC TREES

Amy Donaldson Willis, Ph.D.

Cornell University 2017

ABSTRACT

Inferring evolutionary histories, or *phylogenetic trees*, has important applications in biology, criminology and public health. However, phylogenetic trees are complex, non-Euclidean objects. While our mathematical, algorithmic, and probabilistic understanding of the behavior of trees in their metric space is mature, statistical infrastructure is relatively underdeveloped. This thesis proposes inferential and exploratory statistical methods for the analysis of tree-valued data. The inferential method is a confidence set for the Fréchet mean of a distribution with support on the metric space of phylogenetic trees. Two exploratory methods are proposed for visualizing collections of trees, which rely on similar tools to the confidence set procedure. Finally, some results relating to modeling estimates of trees are given, and related open problems are discussed.

BIOGRAPHICAL SKETCH

Amy Donaldson Willis grew up in Brisbane, Australia with parents Kathryn and Michael Willis and younger brother Hugh. Her grandfather, Ian Milford Wyne Wood, was an agricultural scientist and plant breeder at CSIRO, and encouraged Amy's early interest in science. Her particular interest in mathematics was inspired by her teachers at Somerville House, and in 2008 Amy won a National Undergraduate Scholarship to study at the Australian National University. She graduated with First Class Honours in Statistics in 2011. In 2012 Amy moved to Ithaca, New York to pursue a PhD in Statistics. A passionate environmentalist, she is particularly interested in methods development for biodiversity studies, and has developed statistical tools for analyzing species richness and phylogenetics.

In loving memory of Ian Milford Wyne Wood (1930 – 1999)

ACKNOWLEDGEMENTS

I don't know where to begin thanking my adviser, John Bunge. John's patience with me far exceeded any reasonable limit, and he has trusted my ideas and believed in my creativity since our first meeting. I cannot thank him enough.

Sidney Resnick has gently pushed me to higher and higher standards of mathematical rigor, while unconditionally supporting my research interests. He has been an amazing mentor, and his careful work and dedication continues to inspire me.

Louis BiIllera's enthusiasm for this project has repeatedly motivated me, and I'm grateful for his wisdom, encouragement and mathematical expertise. His support and Colombian connections ensured that 2016 was a productive year for me.

The faculty of Cornell's Department of Statistical Science have patiently supported me for 5 years, providing funding, advice and much instruction. I am especially grateful to Giles Hooker, Martin Wells and Paul Velleman. Jacob Bien, through his incredible example, inspires me to be a better scientist, instructor and colleague.

Rayna Bell's tolerance of my poor understanding of biology and assiduous explanations have been critical to work in this thesis that bridges statistics and biology. Her Emydid gene trees made Section 4.3 possible. I am grateful for her friendship and generosity with her time.

Tom Nye most kindly made available to me his TreeBase package, a critical component of the software used here. He also cheerfully gave helpful suggestions and much encouragement. Megan Owen generously provided constructive comments and excellent advice. Two anonymous referees gave very helpful feedback and suggestions on Chapter 3, vastly clarifying many aspects. Conver-

sations with Sarah Heaps, Susan Holmes, Bret Larget, Erick Matsen, Phil Spinks and Grady Weyenberg shaped important aspects of this work in its formative stages and I'm grateful for sharing their thoughts and comments. I thank every one of you for your feedback, suggestions, and criticisms.

Almost all of this thesis was dictated, not typed. Its completion would not have been possible without the hard-working team of VoiceCode. I truly do not know how I would I finished my graduate work without them.

My friends have been an amazing source of support and love throughout the last 5 years. Rayna Bell, Allison Boex, Erin Camp, Marsha Lampi, Johannes Lederer, Ezra Lencer, Catherine Navarrete, Joanna Upton and Thea Whitman have encouraged, commiserated and cajoled me into finishing this thesis.

Dan Kowal and David Sinclair were the best classmates imaginable, both always ready with witty conversations, inappropriate jokes, different perspectives, creative ideas, and great suggestions. I am incredibly grateful for the cheer they brought to my graduate school years.

Scott Henderson has supported me through every day of this thesis, and has kept me relatively healthy and grounded year after year. I thank him from the bottom of my heart.

Kathy and Mike Willis forgive me endless faults and have always encouraged my study. I love them very much and could not be in the U.S. pursuing my dreams without them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Phylogenetics is the study of evolutionary relationships using molecular sequencing data. Using phylogenetic analysis, Hillis and Huelsenbeck identified a perpetrator of wilful HIV infection [24], Scaduto *et al.* provided evidence against the claim that social workers introduced a hepatitis infection into children's hospital [53], and Ou *et al.* identified accidental disease transmission by a healthcare provider [44]. Estimating phylogenies is an important and practical problem.

Of equal importance as estimation is assessing uncertainty in the estimates of these relationships. Disagreements between different methods for constructing ancestral histories may be substantial [61], and failing to note uncertain relationships could lead to false conclusions of guilt, misdirected resources, or preventable disease spread.

In this thesis I argue that despite sophisticated techniques for estimation, infrastructure for assessing uncertainty in phylogenetic relationships is severely underdeveloped. I utilize recent developments in mathematics, algorithms and probability to construct a method for describing uncertainties in evolutionary relationships. While I use interdisciplinary tools, the approach that I take is statistical in that it concerns estimation of an unknown parameter and assessments of the variability in the estimate.

This thesis begins in Chapter 2 with an introduction to the mathematical framework for comparing and analyzing phylogenetic trees. Chapter 3 proposes the first confidence set procedure for a mean phylogenetic tree, and investigates coverage and two case studies. In Chapter 4 I propose an exploratory

procedure for visualizing uncertainty in phylogenetic trees. While this problem has an extensive literature, the procedure proposed here is unique because it can show the uncertainty of estimating trees in the correct dimension. I explore some related theoretical questions in Chapter 5 before discussing the context of this work in the broader literature in Chapter 6.

CHAPTER 2

TREES, TREE SPACE AND THE LOG MAP

One of the most common goals of a statistical analysis is the estimation of unknown parameters and assessments of their uncertainty. Phylogenetic trees have a unique structure that complicates assessments of estimation error. Here I review the metric space of phylogenetic trees and related tools, which is necessary to quantify tree uncertainty in following chapters.

2.1 Phylogenetic trees

A phylogenetic tree is a representation of the shared ancestry of a specific collection of organisms or taxa. The utility of representing ancestry as a tree is underpinned by the assumption that the taxa have descended from a common ancestor, and differ from the ancestor and each other in varying degrees. Trees can encode morphological, phenotypical, lifestyle and genetic differences. However, advances in biology over the past 100 years suggest that genetic data provide the most comprehensive information about true ancestry and the evolution of life.

Formally, a phylogenetic tree is an edge-weighted tree-graph (a connected acyclic graph with no vertices of degree 2). Vertices (also referred to as nodes) represent organisms: internal vertices (vertices with degree 3 or greater) represent organisms that existed at a point in history, and the leaves (vertices with degree one) represent modern organisms. The edges (often called branches) have lengths that represent the extent of divergence between nodes. In some contexts the graph will be directed and have a vertex of out-degree one repre-

senting the common ancestor. When this is the case, this vertex is called the root.

2.2 Tree space

It is often necessary to compare different estimates of a phylogeny. For example, different sections of genetic information may imply different phylogenies, or different estimation procedures may conflict. In order to compare collections of trees, Billera *et al.* [9] developed a metric space for the internal structure of phylogenetic trees with the same leaf set. *Tree space* is denoted by $(\mathcal{T}_m, \gamma)$ where $m \geq 4$ is the cardinality of the leaf set. The metric distance $\gamma(T_i, T_j)$ between two trees T_i and T_j accounts for differences with respect to both their topologies (branching structure) and branch lengths.

Tree space is constructed by first representing each of the $(2m-5)!! = (2m-5) \times (2m-7) \times \ldots \times 5 \times 3 \times 1$ possible tree topologies by a single non-negative Euclidean orthant of dimension $m-3$ (the largest possible number of internal branches). Then the orthants are "glued together" [9, p. 12] along nearest neighbor interchange (NNI) topologies, which are obtained by reducing the length of a single edge to zero and adding a new edge at the induced degree 4 vertex, lie in adjacent orthants along the boundary corresponding to the collapse of the relevant NNI edge (see Figure 2.1).

Orthants and orthant boundaries are also called *strata*. Trees with the largest possible number of internal branches are said to fall in *top-dimensional strata* while trees with less than the largest possible number of internal branches fall in *co-dimensional strata*. Topologies that correspond to top-dimensional strata

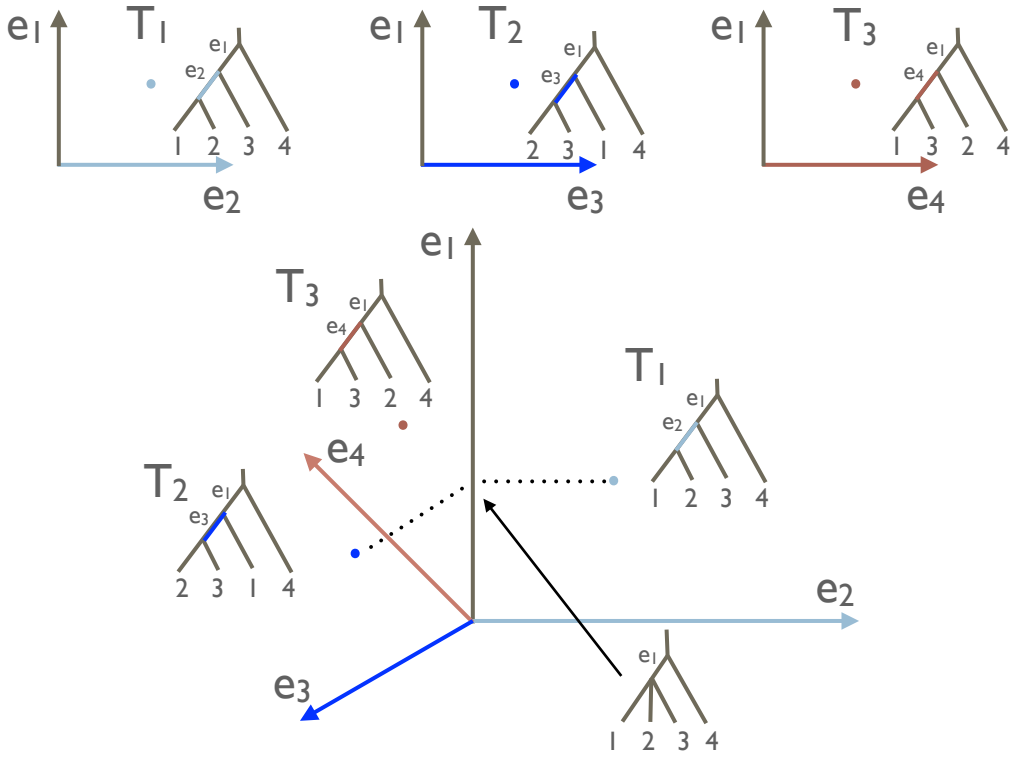

Figure 2.1: The structure of tree space with 5 leaves, $\mathcal{T}_5$, around a single co-dimension 1 stratum. Trees T_1, T_2 and T_3 mutually differ by a nearest neighbor interchange (NNI) move, and hence the orthants (or top-dimensional strata) associated with their topologies are connected along the co-dimension 1 stratum. The e_i values reflect the branch lengths, and the color coding connects the axes with the branch lengths. The dotted line between T_1 and T_2 is the unique shortest path between these trees.

are called *resolved*. Figure 2.1 shows the structure of tree space with 5 leaves, $\mathcal{T}_5$, around a single co-dimension 1 stratum, with the 3 associated NNI topologies.

The distance between two trees is defined to be the L_2-length of the shortest possible path between them, where the paths must pass through the orthants and their boundaries. The shortest path between any 2 trees is called the

geodesic path or the *geodesic*, and is necessarily piecewise linear. Geodesics are unique, a result which follows from [20] since the space is *non-positively curved* [9]. Non-positive curvature has also resulted in efficient algorithms for calculating geodesic paths [46], means [2, 37, 59, 57, 56] and principal paths [42, 43].

While the concept of path continuity drove the construction of the space, and completeness and separability follow naturally from the construction, other properties of the space emerged later. A proof of the Heine-Borel property can be found in Appendix 7. I proved and used this result for a consistency proof unrelated to the theme of this thesis [52].

In tree space, all leaves of the tree are treated equally, and so rooted trees can be analyzed using the methods described in this thesis by treating the root as another leaf. Therefore without loss of generality I will treat trees as unrooted for the remainder of this thesis. Furthermore, the metric space of external branches is $\mathbb{R}^m$, and the metric space of phylogenetic trees with external branches is the product space of tree space and $\mathbb{R}^m$. All methods in this thesis were built for tree space, but can be generalized to the product space.

2.2.1 Probability triples and spaces

In order to discuss limiting distributions of phylogenetic tree estimators, some tools from probability are necessary. In this section I briefly discuss construction of a complete probability space.

$\mathcal{T}_m$ is a metric space, so we can discuss the Borel σ-algebra. Enumerate the resolved tree topologies $j = 1, \ldots, (2m-5)!!$, and write any set in the Borel

algebra in the form $A = A_0 \cup \left(\bigcup_{j=1}^{(2m-5)!!} A_j\right)$ for A_j in the j-th topology, and A_0 the collection of trees in A with one or more internal vertices of degree 4 or greater. Then define $\nu(A) = \sum_{j=1}^{(2m-5)!!} \nu_B(A_j)$ for ν_B the Euclidean Borel measure of dimension $m-3$. This preserves σ-additivity because the Borel measure of any orthant boundary is zero, and there are only finitely many such boundaries. This construction does not lead to a complete measure space. However, its completion may be defined by appending all sets of Borel measure zero to give an analogue of Lebesgue-measurable sets in $\mathcal{T}_m$, which we call $\mathcal{L}(\mathcal{T}_m)$. We then have a complete measure space $(\Omega, \mathcal{L}(\mathcal{T}_m), \nu)$. Finally, for a probability measure $F : \mathcal{L}(\mathcal{T}_m) \to [0,1]$, defined with respect to the volume measure ν, we obtain a probability triple $(\Omega, \mathcal{L}(\mathcal{T}_m), F)$.

2.3 The log map

While tree space is not a manifold, the Euclidean-like structure in the interior of the orthants can be utilized to construct a surjection from tree space to Euclidean space. The log map, proposed by Barden *et al.* [5], captures both the distance and direction from a base tree T^* to a target tree T. Let T^* be a tree in a top-dimensional stratum. Define $\log_{T^*}(T) : \mathcal{T}_{m+3} \to \mathbb{R}^m$ to be

$$\log_{T^*}(T) = \gamma(T^*, T)\mathbf{v}_{T^*}(T), \tag{2.1}$$

where $\gamma(T^*, T)$ is the geodesic distance between T^* and T, and $\mathbf{v}_{T^*}(T)$ is a specifically chosen unit vector from T^* to T that reflects the direction of the first segment of the geodesic (details below). The function $\Phi_{T^*}(T)$ positions this vector to originate from the base tree,

$$\Phi_{T^*}(T) = \log_{T^*}(T) + \mathbf{t}^*, \tag{2.2}$$

for $\mathbf{t}^*$ the coordinates in $\mathbb{R}^m$ of T^*'s edge lengths. Throughout this thesis I refer to the function $\Phi_{T^*}(T)$ as the *log map*. The log map is surjection but not an injection.

The vector $\mathbf{v}_{T^*}(T)$ can be best illustrated via $\Phi_{T^*}(T)$. For a target tree T in the same orthant as the base tree T^* (identical topologies), $\Phi_{T^*}(T)$ coincides with a Euclidean representation of T, that is, is an m-vector with all positive components reflecting the lengths of the internal branches of T (Figure 2.2, bottom panel, red tree). For T in an adjacent orthant (NNI topology), $\Phi_{T^*}(T)$ has a single negative component with magnitude equal to the length of the branch present on T but not present on T^*, with the remaining components positive (adjusted to reflect the branch lengths of the T). If T is more topologically distinct than a NNI interchange from T^*, $\Phi_{T^*}(T)$ is found by continuing in the direction of the initial segment of the geodesic path (the segment contained in the same orthant as T^*) for the length of the geodesic across (potentially more than one) Euclidean orthant boundaries (Figure 2.2, bottom panel, blue tree).

Now equipped with the necessary mathematical infrastructure, I now define the sample and population mean of a distribution on tree space, and construct a confidence set for the population mean.

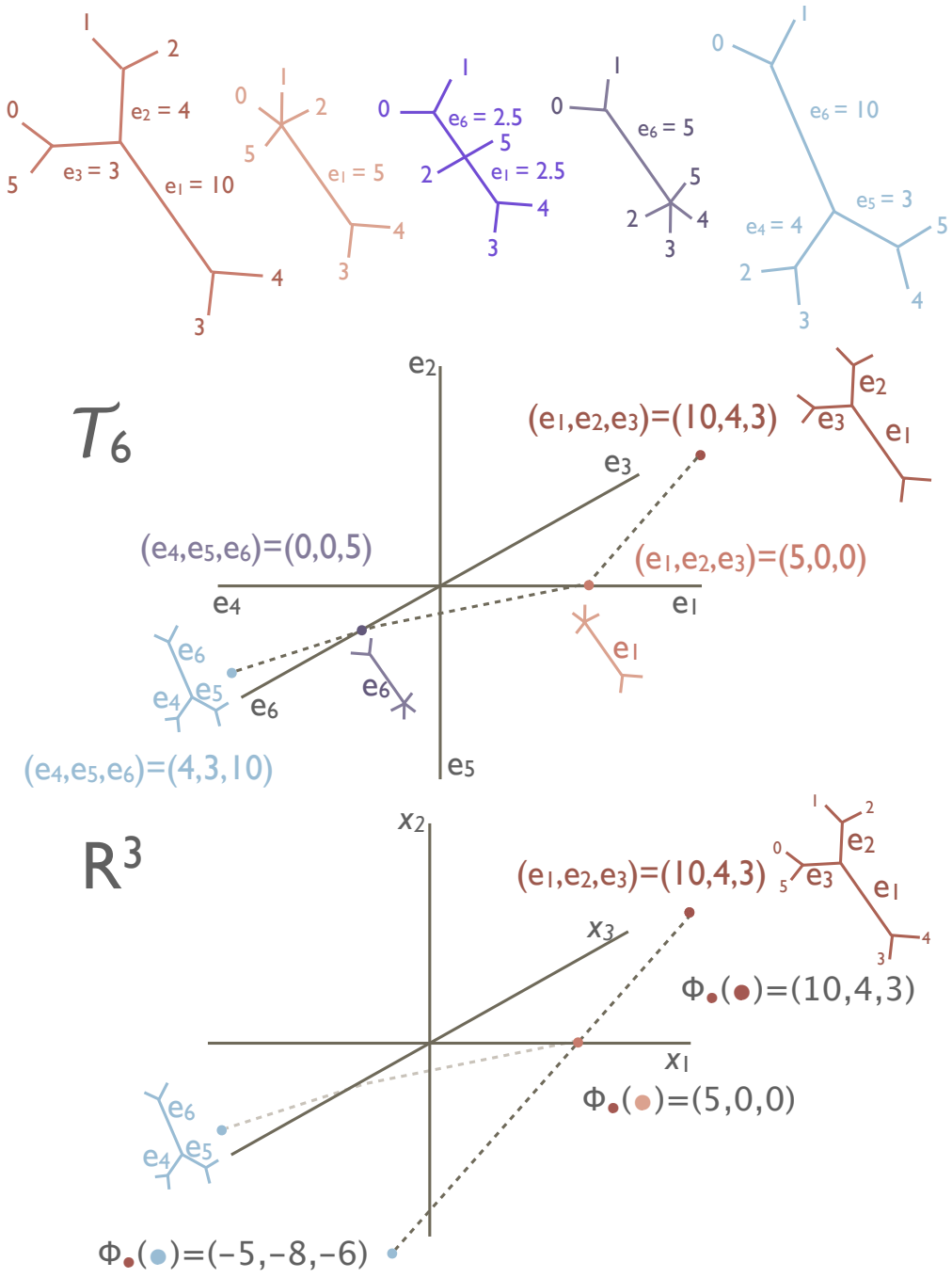

Figure 2.2: The geodesic path between 2 trees with 6 leaves (top panel), a representation of this path in $\mathcal{T}_6$ (middle panel), and the log map with respect to the $\{e_1, e_2, e_3\}$ (red) tree (bottom panel). The distance between the $\{e_1, e_2, e_3\}$ (red) tree and the $\{e_4, e_5, e_6\}$ (blue) tree is $15\sqrt{2}$, and so the log map of the blue tree based at the red tree is $(10, 4, 3) + 15\sqrt{2}\frac{(5,0,0)-(10,4,3)}{||(5,0,0)-(10,4,3)||} = (-5, -8, -6)$. The notation $e_i = j$ denotes that the length of edge i is j, where these edges are labelled on the diagrams.

CHAPTER 3

CONFIDENCE SETS FOR PHYLOGENETIC TREES

3.1 Means on metric spaces

For any probability space $(\Omega, \mathcal{B}, F)$ and metric space $(M, d(\cdot,\cdot))$, define the *Fréchet function* of the probability measure F to be

$$F(u) = \int d(q,u)^2 F(dq),$$

if it exists. Call any minimizer of the Fréchet function a *Fréchet mean* of F, noting that this the minimum of a least squares function. This definition arises naturally as the extension of Euclidean means to metric spaces, because

$$\arg\min_{u\in\mathbb{R}^n} \int |q-u|^2 G(dq) = \int qG(dq).$$

Thus the Fréchet mean is a centre of mass of a distribution on a metric space.

A sample Fréchet mean $\hat{M}_n$ of a collection of objects $M_1, \ldots, M_n \in M$ may be defined in the same way by replacing the probability measure F with the empirical distribution of the collection:

$$\hat{M}_n = \hat{M}_n(M_1, \ldots, M_n) := \arg\min_{m\in M} \sum_{i=1}^{n} d(M_i, m)^2.$$

I now turn to the specific case of the probability space $(\Omega, \mathcal{L}(\mathcal{T}_{m+3}), F)$ and the metric space of phylogenetic trees $(\mathcal{T}_{m+3}, \gamma)$. The nonpositive curvature of tree space [9] guarantees uniqueness of Fréchet means [63]. Throughout this thesis I refer to the true, or population, Fréchet mean of the distribution F:

$$\mu = \arg\min_{u\in\mathcal{T}_{m+3}} \int \gamma(q,u)^2 F(dq),$$

and the sample Fréchet mean of a collection of n trees as

$$\hat{T}_n = \hat{T}_n(T_1, \ldots, T_n) := \underset{u \in \mathcal{T}_{m+3}}{\arg\min} \sum_{i=1}^{n} \gamma(T_i, u)^2.$$

I am interested in the asymptotic behavior of the sample Fréchet mean as an estimator of the true Fréchet mean.

Central limit theorems (CLTs) for Fréchet means on general metric spaces have been developed [8]. However, this work relies on homeomorphisms to $\mathbb{R}^n$ from subsets of the space that are known to contain the true mean of the probability measure F. Because of the stratified structure of tree space, inverse functions will not exist for candidate homeomorphisms except restricted to subsets wholly contained in a single orthant. Thus without assuming the topology of the true mean *a priori*, general results for CLTs on manifolds or metric spaces are insufficient for tree mean inference.

3.2 Central limit theory on tree space

Key results regarding the asymptotic behavior of Fréchet means of phylogenetic trees were recently shown by Barden *et al.* [5]. I review some of their definitions and results here before presenting new results built on related ideas.

Definition 1. *The carrier of the geodesic between trees T_1 and T_2 is the sequence of orthants that the geodesic path traverses.*

Definition 2. *For a tree T^* in a top-dimensional stratum of tree space, a maximal cell is a set of trees T that share the same algebraic expression for $\log_{T^*}(T)$. Equivalently, trees T_1 and T_2 are in the same maximal cell if the geodesic from T^* to T_1 has the same carrier as the geodesic from T^* to T_2.*

Definition 3. *D_{T^*} is the set of trees that lie on the boundaries of the maximal cells of T^*.*

These definitions allow me to state the main theorem of Barden *et al.* [5].

Theorem 3.2.1. *[5, Theorem 2] Let F be a probability measure on $\mathcal{T}_{m+3}$ with finite Fréchet function and Fréchet mean T^* lying in a top-dimensional stratum. Assume that $F(\mathcal{D}_{T^*}) = 0$. Suppose that $\{T_i\}_{i\geq 1}$ is a sequence of iid random variables in $\mathcal{T}_{m+3}$ with probability measure F. Then*

$$\sqrt{n}(\hat{T}_n - T^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, A^T V A)$$

where V is the covariance matrix of $\Phi_{T^*}(T_1)$, and

$$A = \{I - E[M_{T^*}(T_1)]\}^{-1},$$

assuming that this inverse exists, and where $M_{T^*}(T)$ is the derivative of $\Phi_t(T)$, with respect to t, at T^*.

Remark. *The theorem above has not been modified from the original, but the method of proof makes it clear that the authors interpret $(\hat{T}_n - T^*)$ in the sense of Equation (2.2). In this sense, we might replace $\Phi_{\hat{T}_n}(\hat{T}_n)$ with $\hat{\mathbf{t}}_n$ and $\Phi_{T^*}(T^*)$ with $\hat{\mathbf{t}}^*$. However, when stated in this way, the ordering of the coordinates is not necessarily consistent across the two vectors. For this reason, it is more precise to say that the theorem describes the asymptotic behavior of $\sqrt{n}(\Phi_{\hat{T}^*}(\hat{T}^*) - \Phi_{\hat{T}^*}(\hat{T}_n))$, however, $\Phi_{\hat{T}^*}(\hat{T}_n)$ is an expression containing both population and sample quantities, and central limit theorems pertain to the behavior of sample quantities relative to population quantities. We therefore interpret $(\hat{T}_n - T^*)$ as $(\Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{T^*}(T^*))$, where the ordering of the coordinates of $\Phi_{\hat{T}_n}(\cdot)$ and $\Phi_{T^*}(\cdot)$ are identical. The LLN result of Ziezold [76] ensures that for large enough n, both quantities will correspond to the same orthant.*

While this theorem provides crucial insight into the behavior of tree-valued sample means, it is not usable for inference unless V and A are known. Knowing these values would involve very strong assumptions on F. I now discuss some sufficient conditions under which we can estimate the covariance matrix of the limiting distribution, and use these results to construct confidence sets for the true Fréchet mean of F.

3.3 Confidence sets for trees

The conditions of Theorem 3.2.1 are sufficient to describe the behavior of the sample Fréchet mean, but insufficient to describe the behavior of an estimate of the covariance matrix $\Sigma = A^T V A$. Much stronger conditions are required. Unfortunately, as we will see in Theorem 3.3.2, these stronger conditions imply the result of Theorem 3.2.1, and so Theorem 3.2.1 is ultimately not needed for constructing the confidence set. However, the following critical characterization of Fréchet means makes confidence set construction possible.

Lemma 3.3.1. *[5, Lemma 3] If T^*, the Fréchet mean of the distribution G, lies in a top-dimensional stratum, then*

$$\int \Phi_{T^*}(T) dG(T) = T^* \tag{3.1}$$

in the sense of Equation (2.2), that is, $\int \log_{T^*}(T) dG(T) = 0$.

I now combine results from Euclidean multivariate analysis with the above lemma to derive a pivoting distribution for the difference between the sample and true log-mapped Fréchet means after a rotation that reflects the covariance structure.

Theorem 3.3.2. *Let F be a probability measure on $\mathcal{T}_{m+3}$ with finite Fréchet function and Fréchet mean T^* lying in a top-dimensional stratum, and suppose that $\{T_i\}_{i\geq 1}$ is a sequence of iid random variables in $\mathcal{T}_{m+3}$ with probability measure F. Suppose that F satisfies $\Phi_{\hat{T}_n}(T_i) \sim \mathcal{N}(\tau, \nu)$. Define*

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\Phi_{\hat{T}_n}(T_i) - \Phi_{\hat{T}_n}(\hat{T}_n)\right)\left(\Phi_{\hat{T}_n}(T_i) - \Phi_{\hat{T}_n}(\hat{T}_n)\right)^T. \tag{3.2}$$

Then under the null hypothesis that $T^* = T_0$,

$$\frac{n(n-m)}{m(n-1)}\left(\hat{T}_n - T_0\right)^T S^{-1}\left(\hat{T}_n - T_0\right) \sim F_{m,n-m} \tag{3.3}$$

where $\hat{T}_n - T_0 := \Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{\hat{T}_n}(T_0)$, and F_{m_1,m_2} is the F-distribution with numerator degrees of freedom m_1 and denominator degrees of freedom m_2.

Proof. By applying Lemma 3.3.1 to the distribution $G(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{T_i\in A\}}$, I can write

$$\hat{T}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi_{\hat{T}_n}(T_i). \tag{3.4}$$

Therefore, $n^{1/2}(\hat{T}_n - T_0) \sim \mathcal{N}(0,\nu)$. By the normality assumption, $(n-1)S \sim W_p(n-1,\nu)$ and $\hat{T}_n$ and S are independent. The result then follows from the definition of Hotelling's T^2 distribution [29, 65]. □

Before discussing the strength of the assumptions of this theorem (final paragraph of Section 3.3), I state the following corollary.

Corollary 3.3.1. *Under the conditions of Theorem 3.3.2,*

$$A = \Bigg\{T \in \mathcal{T}_{m+3} : \tag{3.5}$$
$$\left(\Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{\hat{T}_n}(T)\right)^T S^{-1}\left(\Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{\hat{T}_n}(T)\right) < \frac{m(n-1)}{n(n-m)}F_{m,n-m}(1-\alpha)\Bigg\}$$

is a $100(1-\alpha)\%$ confidence set for T^*.

I believe this to be the first confidence set for a phylogenetic tree that accounts for uncertainty in estimation of both the topology and branch lengths of the tree.

While Theorem 3.2.1 was not necessary for Theorem 3.3.2, I argue that the development in [5] of Lemma 3.3.1 is the critical component to explaining the normality of log-mapped Fréchet means of trees. Under broad conditions, sums of random variables have normal limiting laws (e.g. [35]). However, minimizing arguments to convex functions, in general, do not. I believe that this is why the log map is a necessary tool for understanding the asymptotic behavior of tree means: by Lemma 3.3.1, log-mapped tree Fréchet means are, in fact, sums.

It is important to note that the assumption that $\Phi_{\hat{T}_n}(T_i)$ is normally distributed is very strong, and unlikely to be satisfied by broad classes of distributions on tree space. For this reason it is important to investigate the coverage of the proposed confidence set, and especially important to investigate this for phylogenetic models that are used in practice. I explore this topic in the following section.

3.4 Coverage

The probability that a confidence set contains the parameter to be estimated is called the *coverage*. The coverage depends on the data generating process and the sample size. I investigate different combinations of these components in order to gain a realistic picture of the coverage of the procedure described in Corollary 3.3.1, especially when the data generating process does not satisfy the requisite assumptions.

The first step of every coverage simulation that follows is fixing a data generating process. If the Fréchet mean of this process cannot be calculated exactly, it is approximated by simulating a very large number of observations and then calculating the sample Fréchet mean of these observations using the proximal point algorithm of Bačák [2] and confirming convergence. $1000 \times n$ trees are then simulated from the data generating process and grouped into 1000 sets of size n. For each set, I test the hypothesis that the Fréchet mean of the data generating process is the true mean based on the dataset that was simulated. This is equivalent to confirming that the confidence set contains the Fréchet mean. Finally, the proportion of the 1000 sets that did not reject the test at the α-level gives an estimate of the $100(1-\alpha)$%-level coverage.

3.4.1 Trivial case

To confirm that my implementation of the log map functions correctly, and that coverage decreases as the distribution of the log-mapped observations diverges from a multivariate normal distribution, I investigate coverage under two models that nearly satisfy the assumptions of Theorem 3.3.2. I arbitrarily chose an unrooted tree topology with 10 leaves, and drew a covariance matrix Σ_0 with eigenvalues drawn from a Uniform$(0.5, 2)$ distribution. I then simulated vectors from a $\mathcal{N}(\mu_0 \mathbb{1}_7, \Sigma_0)$ distribution, for $\mu_0 \in \{0.65, 2\}$, and discarded all observations with any negative coordinates. The data generating process for the trees is to assign these vectors as the length of the internal branches of the chosen topology, and assign length 1 to all external branches. This process was chosen to have a truncated multivariate normal distribution of the log-mapped observations. Because a (non-truncated) multivariate normal distribution satisfies the

Table 3.1: Estimated coverage of the confidence set procedure under a truncated multivariate normally-distributed tree-generating process with support in a single orthant. Larger values of μ reduce the level of truncation and result in a distribution closer to multivariate normal. The proportion of confidence sets containing the true tree is reported. Exact coverage would be signified by $(90, 95, 99)\%$.

μ_0: Mean parameter	n: Sample size	Coverage (%): $\alpha = (0.10, 0.05, 0.01)$
0.65	20	$(88.2, 93.0, 98.6)$
0.65	50	$(88.8, 94.1, 98.2)$
0.65	100	$(89.1, 93.4, 98.9)$
2	20	$(88.6, 93.9, 99.1)$
2	50	$(89.5, 94.4, 98.9)$
2	100	$(89.4, 93.5, 99.1)$

assumptions of Theorem 3.3.2, I can control the degree to which the assumption is violated with the choice of μ_0. The further μ_0 is from the orthant boundaries, the lower the proportion of trees that are truncated, since the covariance structure is unchanged. 88.8% and 21.3% of simulated trees were discarded due to truncation for $\mu_0 = 0.65$ and 2.

The results of the simulations can be found in Table 3.1. I first note that coverage is generally less than but very close to nominal, which provides some evidence that the programmed implementation of the procedure correctly reflects the described procedure. It also reflects that the data generating process does not satisfy the assumptions of Theorem 3.3.2, but that the disagreement is relatively small.

Coverage did not consistently increase with sample size. This is consistent with Theorem 3.3.2, which is an exact (non-asymptotic) result. However, coverage is smaller for smaller values of μ_0. This is as expected, because the distribution of the log-mapped observations less closely resembles a multivariate normal distribution due to truncating the distribution at the orthant boundaries.

However, coverage drops only slightly, which is because the distribution of the log-mapped observations is exactly multivariate normal close to the mean, and this is more impactful than matching this distribution near the tails.

Having found that the confidence procedure behaves as expected both with respect to sample size and distribution of the log-mapped observations, I now investigate coverage under a more realistic tree-generating process.

3.4.2 HKY

In contrast to the previous section, here I investigate coverage under a model for genetic sequences rather than a model on tree space. The HKY model [22] is a 5-parameter model for DNA mutations. Each location on the genome is modeled by an independent continuous-time Markov process with state space $\{A, T, G, C\}$. One parameter controls the transition ($A \leftrightarrow G$, $C \leftrightarrow T$) rate, one parameter controls the transversion ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ or $G \leftrightarrow T$) rate, and 4 parameters and 1 linear constraint control the stationary distribution. The transition matrix of the Markov chain is therefore

$$A = \begin{bmatrix} p_{TT} & p_{TC} & p_{TA} & p_{TG} \\ p_{CT} & p_{CC} & p_{CA} & p_{CG} \\ p_{AT} & p_{AC} & p_{AA} & p_{AG} \\ p_{GT} & p_{GC} & p_{GA} & p_{AA} \end{bmatrix} = \begin{bmatrix} - & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & - & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & - & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & - \end{bmatrix}$$

where the diagonals are such that the row totals are zero. The lengths of the branches of the tree represent the continuous time component of the model. Under this model, the parameters α, β, $\pi = \{\pi_T, \pi_C, \pi_A, \pi_G\}$ and the branch lengths can be estimated by maximum likelihood from the observed sequence data.

I define the following tree-generating-process based on this model. For a fixed tree, τ, I simulate 350 nucleotides under the HKY model using seq-gen [48]. I then use maximum likelihood to estimate the tree (also under a HKY model) using PhyML [21].

In contrast to the models of Section 3.4.1, I chose to investigate this data generating process because the resulting distribution of the log maps is unlikely to satisfy the assumptions of Theorem 3.3.2. However, this model is used in practice and so provides a useful test case for assessing the robustness of the confidence procedure to assumption violation.

The coverage of the confidence set procedure under the model described above is shown in Table 3.2 for two different choices of trees (see Section 3.5 for details). Coverage is lower than nominal. This is unsurprising, since the model does not satisfy the assumptions upon which the procedure is based. However, the simulations can be used to adjust for this when testing: for a true tree structurally similar to the Zika tree (Figure 3.1) and 20 data points, choosing $\alpha = 0.01$ should give a test that is conservative at the 10% level, since the coverage in this case is greater than 90%.

Interestingly, the coverage for the Zika tree simulations is stable across sample sizes, while the coverage for the HIV tree decreases with sample size. I believe that the coverage decreases because the statistical power to detect a difference between a normal distribution and the distribution of the log maps improves as more data are observed. This explanation is also consistent with the different patterns of coverage with sample size between the HIV and Zika trees. The HIV tree has 5 leaves, and log-mapped trees of dimension 2, while the Zika tree has 6 leaves and log-mapped trees of dimension 3. Since for a given sam-

Table 3.2: Estimated coverage of the confidence set procedure when sequence data are generated by an HKY process. Two different trees from Section 3.5 were used to generate base pair alignments, and then estimates of the true trees were calculated based on the alignments. These estimates were grouped together into 1000 samples of size n, and the described procedure was used to construct the confidence set. The proportion of confidence sets containing the Fréchet mean of the data generating process is reported. Exact coverage would be signified by $(90, 95, 99)\%$.

τ: True tree	n: Sample size	Coverage (%): $\alpha = (0.10, 0.05, 0.01)$
HIV Fréchet mean tree	20	$(86.1, 92.9, 97.5)$
HIV Fréchet mean tree	50	$(84.4, 91.4, 98.0)$
HIV Fréchet mean tree	100	$(82.8, 89.3, 96.8)$
Zika Fréchet mean tree	20	$(81.1, 86.7, 93.9)$
Zika Fréchet mean tree	50	$(81.0, 86.3, 92.6)$
Zika Fréchet mean tree	100	$(81.0, 86.7, 92.6)$

ple size statistical power generally decreases with an increase in dimension, the distribution of the log maps of the Zika trees is harder to distinguish from a normal distribution compared to the lower dimensional HIV trees. This hypothesis would explain the decrease in coverage for the HIV trees, and the relative stability of coverage for the Zika trees. However, it does not explain the lower absolute coverage in the Zika tree case, which is due to the structural differences between the trees. The HIV tree has a much greater gradient in the branch lengths compared to the Zika tree, with the longer branch 10 times longer than the shorter branch. The longer branch on the Zika tree is only 4 times the length of the middle branch, which is 5 times the length of the shortest branch.

3.5 Case studies

Having explored the performance of the coverage procedure, I turn to two case studies to demonstrate the potential applicability of the method in phylogenetic

studies.

3.5.1 Case study: Zika biogeography

The implications of the Zika virus' spread has caught worldwide attention. The virus is known to have originated in Africa, with media releases in South America purporting that the virus arrived across the Atlantic ocean [6, 38], while the academic literature agrees that the virus arrived from the Asia-Pacific [70, 69, 54]. I investigate this by tracing the biogeography of the current Zika outbreak in South America.

All available complete Zika genome sequences with complete location and year information were obtained from GenBank on June 7, 2016. I categorized the sequences by location and year (see the leaf labels on Figure 3.1 for the categories), and considered different samples within the same category as block replicates. I then draw one sample from each category, align the sequences using Clustal [33], and fit a simple HKY model to the phylogeny using PhyML [21]. I repeat this 108 times to have 108 evolutionary histories reflecting the within-virus variability. The choice of 108 trees was based on computational constraints. The Zika sequences are approximately 11,000 base pairs, while the simulated sequences were only 350, which makes this analysis more computationally intensive than the coverage simulations.

Figure 3.1 shows the sample Fréchet mean of the 108 trees. A branch separating recent South American strains and recent Pacific strains is present on the

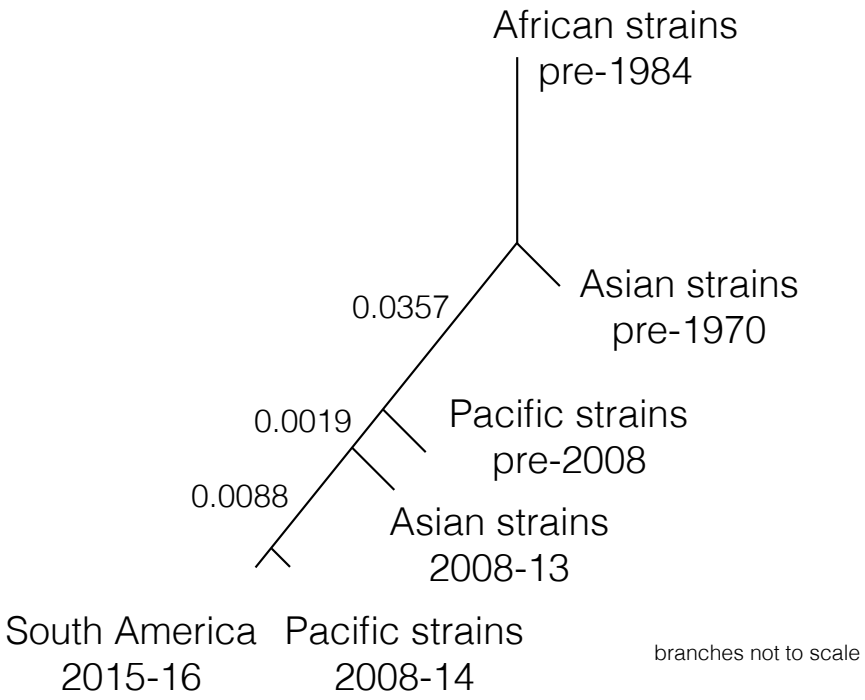

Figure 3.1: The Fréchet mean of 108 Zika phylogenies obtained by permuting the representative of each strain and estimating the phylogeny under a HKY model.

sample mean tree. However, the sample mean tree alone is insufficient to assess if this branch is present on the true mean tree, however, and for this I employ the proposed confidence procedure.

The log maps of the 108 trees (relative to the sample Fréchet mean tree) are shown in Figure 3.2, along with the 99.9% confidence set for the log map of the true Fréchet mean tree. The confidence set for the log-mapped tree does not contain any vectors with negative coordinates, equivalently, the confidence set for the tree only contains trees with the same topology shown in Figure 3.1.

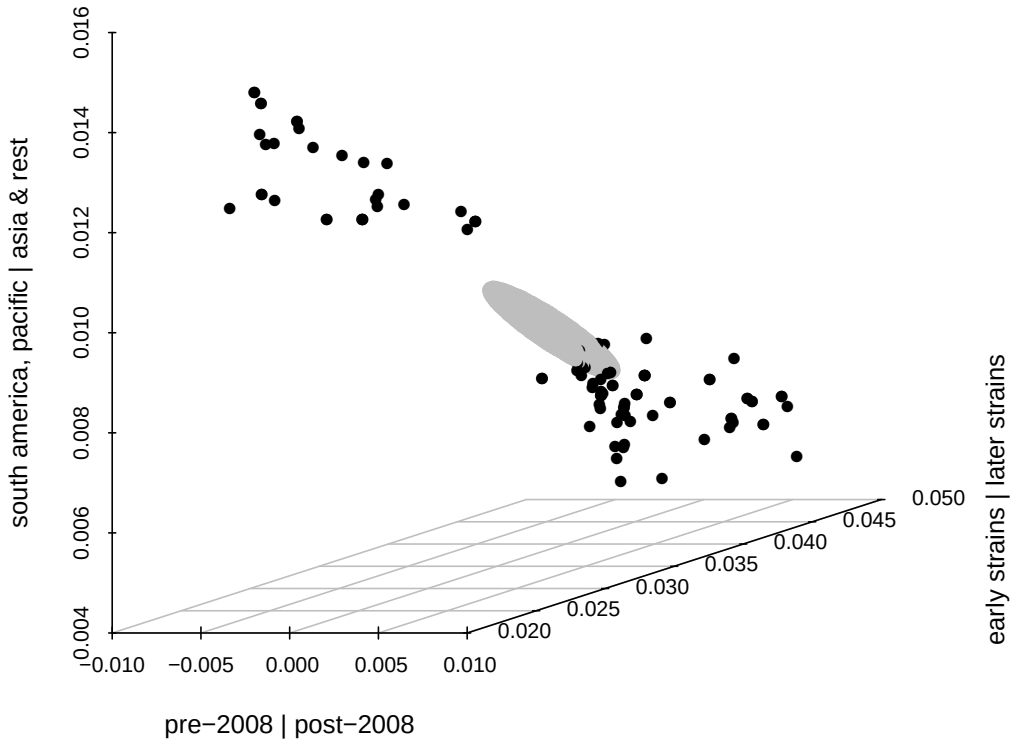

Figure 3.2: The log maps of 108 Zika phylogenies with respect to their Fréchet mean (black points), and the 99.9% confidence set for the log map of the true Fréchet mean of the tree generating process (gray ellipsoid). The log-mapped confidence set does not contain any vectors with negative coordinates. Equivalently, the confidence set is wholly contained in a single orthant of tree space.

In particular, all trees in the confidence set contain a branch that separates the South American and recent Pacific strains of the Zika virus. I therefore conclude that the virus travelled to South America via the Pacific, rather than descending from an African strain.

3.5.2 Case study: HIV forensics

In this example I investigate the hypothesis that two HIV-positive patients of a Floridian dentist with AIDS contracted HIV from the dentist. This question was formally investigated by the National Centre for Infectious Diseases in 1992, culminating in a report concluding transmission to the patients from the dentist [44]. The report considered several different analyses, including the within-patient HIV variation (HIV is known to mutate rapidly), and a phylogenetic analysis. The proposed confidence procedure permits accounting for both within- and across-patient variation.

Amino acid sequences from the V3 region of the HIV virus of the dentist (D, 8 replicates), patient A (A, 6 replicates), patient B (B, 14), a local control (LC, 2), and a non-local control (NLC, 2) were obtained from GenBank. The total number of sequence combinations is $8 \times 6 \times 14 \times 2 \times 2 = 2688$, and I selected 100 of these combinations to avoid overpowering the hypothesis test. For each of these combinations I aligned the sequences and estimated the underlying tree using a HKY model. I then calculated the mean tree, log maps and the confidence set for the mean tree. The Fréchet mean is shown in Figure 3.3.

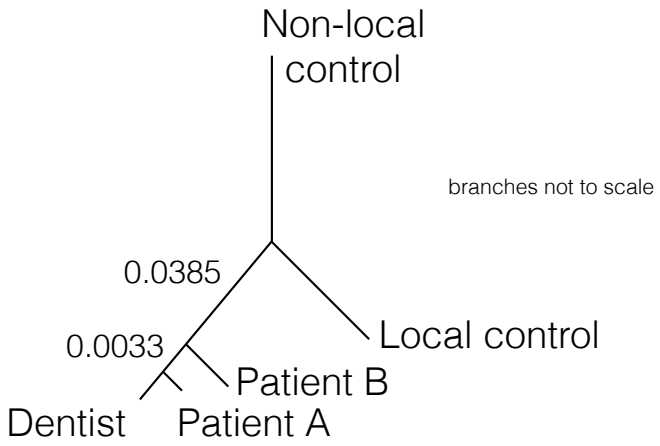

Figure 3.3: The Fréchet mean of 100 estimated phylogenies of the HIV viruses of a dentist, two patients of the dentist , a control from the local population, and a control from a distinct population. The different phylogenies were obtained by permuting the representative sequences of each individual.

The projected trees under the sample log map are shown in Figure 3.4. Approximately equal variance is observed in both branches, however the mean length of the branch separating the dentist and patients from the controls is large relative to its variability. The null hypothesis that this edge is not present on the true tree is rejected with $p < 10^{-13}$, and thus I conclude with high confidence that the dentist infected the two patients. The remaining branch indicates the relative similarity of the dentist's sequences to those of patient A and patient B (which patient was infected closer to the date of blood sample collection), and I do not reject the null hypothesis that there is a leaf more closely related to the dentist than patient A ($p = 0.075$). The coverage simulations suggest caution when interpreting p-values, but neither of the magnitudes of the two tests conducted here ($p < 10^{-13}$ and $p = 0.075$) are marginal.

By performing the analysis over trees estimated based on different within-patient sequences, I simultaneously utilize both the intra- and interperson se-

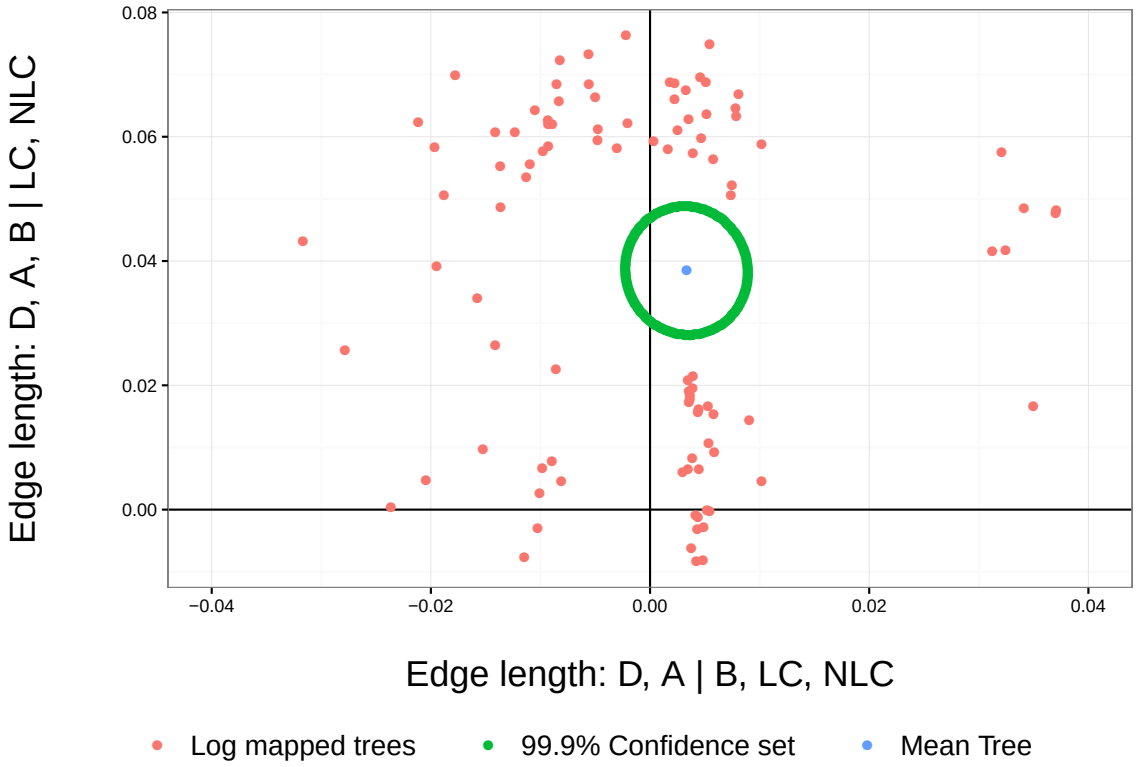

Figure 3.4: The log maps of 100 HIV phylogenies. A confidence interval for the true Fréchet mean tree suggests infection of the patients by the dentist, given small variability of this edge length relative to its mean (vertical direction). The horizontal direction indicates the relative similarity of the dentist clade; a positive coordinate indicates that patient A's virus is more similar to the dentist's virus than other groupings in this clade.

quence variability. The original investigation considered phylogenetic analysis separately to the within-patient analysis (the latter was investigated using a modified ANOVA).

3.6 Discussion

I now outline some limitations of the confidence set procedure and describe some directions of future research that may address them.

3.6.1 Degeneracy

The asymptotic behavior of the sample Fréchet mean changes when the true Fréchet mean falls on a stratum of co-dimension 1 or higher. In the co-dimension 1 case, the log map remains multivariate normal on the branches whose means do not correspond to co-faces, with the co-facing branches converging to either a degenerate distribution or a multivariate normal distribution ([5, Theorem 3], [30]). Unfortunately, no statistical tests currently exist for assessing whether a Fréchet mean lies on a co-dimensional face, and so it is not yet possible to assess whether top-dimensional stratum or co-dimensional stratum asymptotics might apply. However, the case of convergence to a degenerate distribution suggests that a sample mean on a co-dimensional stratum is a good indicator that the true mean falls on a co-dimensional stratum. I did not observe any sample means on strata of co-dimension 1 in the examples that I discussed above, and so did not investigate confidence set construction in this case. In particular, construction of a statistical test for the true Fréchet mean falling on an orthant boundary is an open problem with potential use cases in the biological sciences.

3.6.2 Data compression

A common criticism of the log map is that it is a compression of tree space, and both topological and branch length information is lost when trees are assigned to vectors. No compression occurs for trees with the same topology as the base tree of the log map, but the information loss increases as the argument to the log map becomes an increasing number of NNI moves from the base tree. The confidence set procedure described in this chapter utilizes the log map with base tree set to the sample Fréchet mean, and so the local structure of tree space around the sample Fréchet mean is well preserved. While information about trees in the sample that are far from the sample mean is lost, collections of trees that arise in practice are generally highly structured around a central tree or cluster of trees, as shown for two datasets in Section 3.5. These datasets and their high degree of structure are typical of tree-valued datasets, and it is this structure that results in the utility of the log map with base tree as the sample mean.

3.6.3 Dependence structures

When using the proposed confidence set, ideally, tree–building information for n different individuals from each of the taxa on the tree would be obtained, and each individual from each taxon would be used only once to build n independent trees. However, when differing numbers of individuals are obtained from each taxon, a choice must be made between discarding information (to equalize the number individuals from each group) or inducing dependence by repeating some individuals when building the trees. In Section 3.5, I chose the latter op-

tion. Unfortunately, modeling this source of dependence and incorporating it into the covariance estimate is extremely challenging, because the extent of dependence between two trees $T_i, T_j \in \mathcal{T}_{m+3}$ depends not only on the number of shared individuals used to build the trees, but also on how closely related these individuals are on the tree (a function of the unknown true tree). I conjecture that ignoring this dependence is a second–order issue compared to violations of identicality and ignoring uncertainty in estimating the trees. Investigation of this conjecture is an ongoing project.

3.6.4 Sources of tree-valued observations

The case studies in this dissertation have exclusively focused on using within–species variability to more accurately reflect variability in genetic data. However, many different processes give rise to phylogenetic tree–valued observations that could be used as inputs to this method. Gene trees, where each tree represents the phylogeny of a different location on the genome, provide another natural source of variability. However, there are two key considerations when using gene trees in the above method. Firstly, it may not be a plausible assumption that gene trees would be observed from the same distribution, because unusual biological processes (e.g. horizontal gene transfer) may give rise to outlying gene trees. Furthermore, since genes have a spatial structure, gene trees of genes spaced more closely together may not be independent. This relates to the dependence issue discussed in Section 3.6.3.

3.6.5 Extension to incorporate tree uncertainty

In the examples that I investigated, I treated each tree as an observation that was observed exactly. In fact, all of these trees were estimated based on a model that describes the observed sequence data as a function of the underlying tree. This estimation step induces another layer of uncertainty, because each tree's estimate has an uncertainty which I ignored in this chapter. I discuss and address this issue further in Chapters 4 and 5.

3.7 Concluding remarks

The framework discussed in this chapter for representing collections of trees as points in Euclidean space via the log map opens phylogenetic tree analysis to many multivariate analysis methods. This paper considers only confidence set construction for means, but testing two-sample hypotheses, discriminant analysis, multidimensional scaling and factor analysis could all be applied given an appropriate scientific question.

I believe that the most important contribution made here is the application of the statistical framework of variance modeling to tree space. Furthermore, the proposal for using species replicates to generate collections of trees for summary and analysis may prove fruitful by providing realistic measures of tree uncertainty. This is a known issue in phylogenetics and I hope that the sampling method and the confidence set construction procedure described here contributes to understanding of both of these issues.

The confidence set procedure is an inferential method, and may be useful in

situations where a formal hypothesis test for the value of the true Fréchet mean of a tree-generating-process is necessary. However, it is not always the case that the experimenter has well-defined hypotheses that they wish to test. In the following chapter, I propose an exploratory tool for visualizing collections of phylogenetic trees that may be useful when formal inference is not appropriate or necessary.

CHAPTER 4

VIZUALISATION USING THE LOG MAP

Visualizing high dimensional objects, especially high dimensional objects with with complex structure, is difficult without dimension reduction. In this chapter I argue that the log map, discussed in Section 2.3, is a useful tool for visualizing phylogenetic trees and their uncertainties. Before proposing a new visualization procedure, I briefly review existing tree visualization methods.

4.1 Literature review

4.1.1 Multidimensional scaling

A common method of visualizing collections of trees is multidimensional scaling [25, 13, 31]. Multidimensional scaling (MDS), first proposed by Torgerson [66], is a technique for mapping a collection of n objects to vectors in $\mathbb{R}^k$, where k is usually chosen to be 2 or 3. The only requirement is a distance (or dissimilarity measure) between the objects. MDS involves finding a matrix that minimizes a stress function, which encodes the difference between the distances under the map and the true distances. The minimizing matrix, in $\mathbb{R}^{n\times k}$, can then be visualized as n points in $\mathbb{R}^k$. A common choice of stress function is the Kruskal-1 function [32, 25], with the resulting map given by

$$\underset{x\in\mathbb{R}^{n\times k}}{\arg\min}\left(\sum_{i\neq j} D_{ij} - |p_i - p_j|^2\right)^{1/2},$$

where $\{D_{ij}\}_{(i,j)\in\{1,\dots,n\}^2} = d(\hat{T}_i, \hat{T}_j)$ is the $n \times n$ matrix of distances between the tree-valued estimates $\hat{T}_i, \hat{T}_j$.

An advantage of MDS is that the distance can be chosen to best highlight the differences of importance between the trees in the collection. Hillis *et al.* proposed using the Robinson-Foulds (RF) distance to view the topological differences between trees [25, 50]. Chakerian and Holmes discussed the advantages of using the BHV distance to incorporate both topological and branch length features [13], and Gori *et al.* used this approach to cluster genes by phylogeny [19]. Kendall and Colijn recently proposed a method for comparing differences between the most recent common ancestors of the tips [31]. In Section 4.4 I will contrast the advantages of MDS with the advantages of the procedure that I propose here.

4.1.2 DensiTree

The program DensiTree [11, 10] is a popular program for viewing collections of trees, and integrates with most Bayesian tree estimation programs. DensiTree overlays the trees transparently, thus darker regions of the image imply greater confidence. Furthermore, small numbers of alternative topologies with comparable levels of support are easily observed. This tool can also show other parameters that are used in coalescent-based phylogenetic tree estimation, such as population size, by indexing the widths of the branches to these parameters. DensiTree has the advantage that the graphical representation shows the collection of trees as trees, rather than mapping the trees to Euclidean space. It effectively illustrates both topological and branch length disagreements, but performs most effectively when only a small number of topologies conflict and the tree has a relatively small number of leaves. Large leaf sets, or many conflicting topologies, are difficult to distinguish by eye. The procedure that I propose

in this chapter can clearly show conflicting clusters of trees, and works well when there are many such clusters. It is at the expense, however, of visualizing the tree directly.

4.1.3 Related methods

Other methods for comparing phylogenetic trees have been proposed, and I briefly mention some that are designed for comparing more than two trees. Sundberg *et al.* [64] argue that the space of resolved trees can be mapped to an n-torus, and then uses cartographic projections to reduce dimension. However, this loses key branch length information and non-resolved trees are not permitted. A dimension reduction tool that removes selected branches to aid longitudinal analysis has recently been proposed [75], though the authors' goal was not visualization. Heatmaps can be used to summarize the dissimilarity matrix of MDS [47]. Treemaps, which partition a rectangle into panels representing each tree and then display the nodes in a space-filling manner, may be used to see hierarchical dissimilarities [67], though its scalability with the size of the collection is limited. Trees of trees, proposed by Nye [39], assigns the topologies of the trees in the collection to a *meta-tree's* leaf nodes, then uses neighbor-joining methods to cluster the nodes (trees) by topological similarity. Hess *et al.* [23] append histograms of related parameter estimates to leaves to show variation in these parameters (eg. population size) across the clades, while Bremm *et al.* [12] developed PhyloComp to enable targeted comparison of differences at recent or deep divergence times. Most of these methods are designed to highlight topological differences, and few scale well both visually and computationally.

Each of the methods described above have distinct advantages in different situations. However, none have the ability to illustrate uncertainties in a collection of tree estimates $\{\hat{T}_i\}_i$. Trees are usually estimated and not known exactly, and therefore visualization of tree uncertainties is of equal importance as visualization of the tree itself.

4.2 Visualizing tree uncertainty using extrinsic information

Consider trees $T_1, \ldots, T_n$ on the same m taxa. Suppose that it is not possible to observe these trees directly, but that noisy estimates are available. Specifically, for each tree T_i, suppose we observe $\hat{T}_i = B_{T_i,q_i}$, where $B_{t,q}$ is the tree Brownian motion defined in [41] with origin t and diffusion parameter q (see Chapter 5 for a complete description). Furthermore, suppose q_i is unknown but $q_i = \sigma^2 r_i$ and r_i is known. A detailed example where the r_i are gene mutation rates is explored in Section 4.2.1.

The contours of the Brownian motion distribution are spherical, therefore, if only a single orthant is considered, the projection of the contours onto a hyperplane will be also be spherical. However, the folded nature of tree space makes it difficult to construct global hyperplanes, and for this reason I propose to compress tree space to Euclidean space via the log map and project the contours onto a low dimensional subspace. The proposed procedure is as follows:

1. Calculate the weighted sample Fréchet mean

$$\overline{T} := \underset{t \in \mathcal{T}_m}{\arg\min} \sum_{i=1}^{n} \frac{1}{r_i} \gamma(\hat{T}_i, t)^2, \tag{4.1}$$

2. Assign the observed trees to vectors via the log map centered at $\overline{T}$. Each tree is now represented by $\Phi_{\overline{T}}(\hat{T}_1), \ldots, \Phi_{\overline{T}}(\hat{T}_n)$.

3. Project the vectors onto their first 2 principal components to obtain two-dimensional representations.

4. Define

$$\widehat{\Phi_T(T)} = \frac{\sum_i \Phi_{\overline{T}}(\hat{T}_i)/r_i}{\sum_i 1/r_i}, \tag{4.2}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i \frac{(\Phi_{\overline{T}}(\hat{T}_i) - \widehat{\Phi_T(T)})^2}{r_i}. \tag{4.3}$$

5. Represent the relative uncertainty of the trees as the minimum volume sets of measure $(1-\alpha)$ under a normal distribution with mean $\Phi_{\overline{T}}(\hat{T}_i)$ and variance $\hat{\sigma}^2 r_i I_m$. Project these sets onto the principal components of the log-mapped trees.

I now demonstrate this approach in a situation where the evolutionary rate of genes is available.

4.2.1 Evolutionary rate and phylogenetic uncertainty

Comparing sets of phylogenetic trees, each estimated using different regions of the genome (e.g. loci), is complicated by variation in the evolutionary rates of those regions. Loci that evolve slowly contain few informative sites for estimating recent divergence events, while loci that evolve more quickly often contain more natural variation, or mutational saturation, than phylogenetic signal for resolving older divergence events [68, 3]. The informativeness of a particular locus for resolving a particular tree (i.e. ancient versus recent divergence) is thus related to its evolutionary rate.

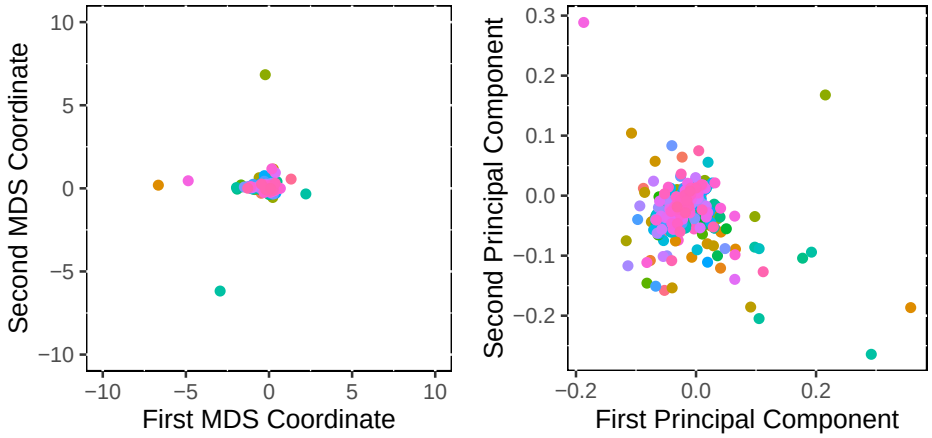

Figure 4.1: Euclidean representations of 574 gene trees shared by 42 mammals [15]. (left) Multidimensional scaling of the BHV distances between the trees. (right) The first two principal components of the log map of the trees with respect to their weighted Fréchet mean. Note that both representations suggest that the trees are known rather than estimated.

To illustrate how the visualization method described above can be used to incorporate uncertainty from covariate information, I consider the relative evolutionary rate of 574 different genes shared by 42 mammals. The OrthoMam database [49, 15] contains estimates of the gene trees of these genes, which I call $\hat{T}_1, \ldots, \hat{T}_{574}$ (estimation details available in [49]), along with the rates, which we call $r_1, \ldots, r_{574}$.

Multidimensional scaling of the trees with respect to the BHV distance [13] is shown in Figure 4.1 (left), along with the first two principal components of the log maps of the trees (right). Both representations suggest that there are 4 trees that differ from the rest. MDS emphasizes this more heavily, most likely because the objective of MDS is to best capture *all* of the pairwise distances in the mapping, while the log map instead preserves distances to the base tree $\overline{T}$. These 4 trees are topologically distinct from $\overline{T}$ by multiple nearest neighbor interchanges.

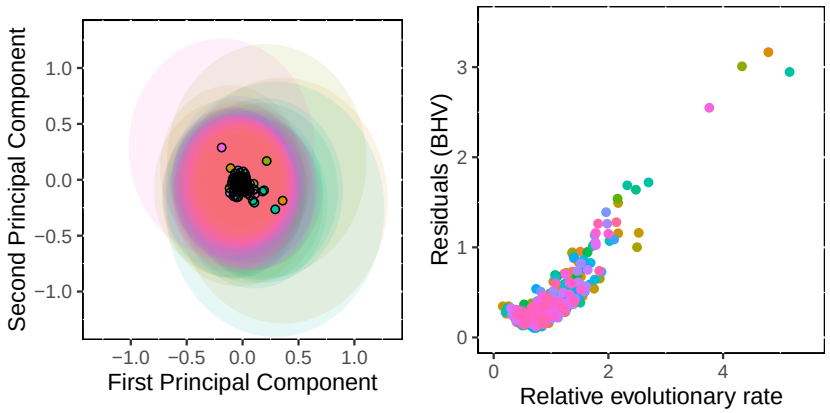

Figure 4.2: To account for estimation error, we construct a model for tree uncertainty (Section 4.2.1). (left) The 40-dimensional sets of volume 0.95 representing each tree, projected onto the first two principal components of the log maps of the tree estimates. The sets appear large because they were constructed in a much higher dimensional space, and not because of the 4 apparent outliers (see text). (right) The BHV distance from each gene tree to the weighted Fréchet mean shown against the relative evolutionary rates of the genes. Lower evolutionary rates, but not the lowest, correspond to the minimum distance trees.

I first calculate the weighted Fréchet mean of the gene trees $\overline{T}$ with weights $1/r_i$ (Algorithm 4.2 of [2] setting $\lambda_k = 1/k$), and found the log maps of the gene trees with respect to $\overline{T}$, $\Phi_{\overline{T}}(\hat{T}_i)$. I then calculate the maximum likelihood estimates of the model parameters (Equations (4.2) and (4.3)), and construct 0.95-measure sets with respect to a $\mathcal{N}(\Phi_{\overline{T}}(\hat{T}_i), \hat{\sigma}^2 r_i I_{40})$ distribution to illustrate the disagreement between the trees after accounting for the variance model. Because these sets are in $\mathbb{R}^{40}$, I project them onto the first two principal components of the $\{\Phi_{\overline{T}}(\hat{T}_i)\}_i$. These projected sets are shown in Figure 4.2 (left). Since constructing the sets in $\mathbb{R}^{40}$ and then projecting them is computationally wasteful, I use only the algebraic form of the $\mathbb{R}^{40}$ sets to determine the $\mathbb{R}^2$ sets, and do not construct the $\mathbb{R}^{40}$ set themselves.

The impression given by Figure 4.1 is that there are a small number of trees that are very different from the rest of the collection. However, Figure 4.2 (left)

makes it clear that relative to the uncertainty in the tree estimates, these trees are not especially outlying. Removing these 4 points reduces the variance by only 5.4%, suggesting that these points are not driving the size of the sets. However, both variance and dimension contribute to the size of the projected sets. These log maps have been estimated in a 40-dimensional space, and the radius of these sets grows with the square root of the dimension. It is the dimensionality of the tree estimation problem that leads to the uncertainty in estimation that we observe in Figure 4.2 (left). Note that while the sets in $\mathbb{R}^{40}$ sets are spherical, the rotation is not, and for this reason the sets do not appear isotropic.

While the form of Equation (4.1) places the greatest weight on gene trees with low evolutionary rate, Figure 4.2 (right) shows that these are not the closest gene trees to the weighted average tree: trees corresponding to genes with small but non-minimal evolutionary rates are the minimum distance trees. I conjecture that the weighted Fréchet mean may provide a good estimate of the overall evolutionary process by incorporating the evolutionary rate information into tree estimation.

4.3 Intrinsic uncertainty information

Having discussed the case where covariate information can inform relative tree uncertainty, I now consider the case where the trees themselves can be used for inferring the precision in estimation. I call this intrinsic information because it is tree-valued.

Suppose for each of k different phylogenies $T^{(1)}, \ldots, T^{(k)}$, I have n_i estimates of $T^{(i)}$, which I call $\hat{T}_1^{(i)}, \ldots, \hat{T}_{n_i}^{(i)}$. This collection contains multivariate informa-

tion about tree uncertainty: branches that can be estimated precisely would be present on all trees and with low variance in their lengths, while contentious branches may only appear on some trees or have large variation in their lengths. I wish to visualize these estimates and their multivariate uncertainty.

Define $\overline{T}^{(i)}$ to be the unweighted Fréchet mean of the $\{\hat{T}_j^{(i)}\}_j$'s, $\overline{T}$ to be the unweighted Fréchet mean of the $\overline{T}^{(i)}$, and T to be the population analogue of $\overline{T}$. The latter may not have any biological significance, but I construct it in order to have a common base for the log map. Consider the model

$$\Phi_{\overline{T}}(\hat{T}_j^{(i)}) = \Phi_T(T^{(i)}) + \mathcal{N}(0, \Sigma_i).$$

While I would prefer a model in tree space as in Section 4.2.1, no nonspherical analogue of Brownian motion in tree space has yet been proposed (see Chapter 5), and I wish to incorporate directional uncertainty.

The key difference between this model and that of Section 4.2 is that the uncertainty of the estimates is not constrained to be spherical: I permit Σ_i to be unstructured. However, I use the collection $\{\Phi_{\overline{T}}(\hat{T}_j^{(i)})\}_j$ to estimate it, using maximum likelihood if n_i is large relative to the dimension of tree space, or a structured estimator if not.

Similar to the proposal of Section 4.2.1, I construct the $(1-\alpha)$ minimum volume sets of the distribution of the $\{\Phi_{\overline{T}}(\hat{T}_j^{(i)})\}$. Again the sets will be m-dimensional, and so I project them to the subspace in $\mathbb{R}^2$ that spans the first two principal components of the $\{\Phi_{\overline{T}}(\hat{T}_j^{(i)})\}_{i,j}$.

A key element of this visualization strategy is that it maintains m-dimensional uncertainty of tree estimation. Constructing a set based on a model for the principal components gives the impression of substantially more preci-

sion than truly exists, and MDS prohibits meaningful model constructions because the coordinate system is sample-dependent (discussed in Section 4.4). Visualizing uncertainty in tree estimates is an extremely important issue because of documented overconfidence in phylogeny estimates. Strong support for conflicting topologies can arise among phylogenies constructed from different sets of loci (e.g. sequence capture versus restriction site associated DNA sequencing [34]) or from the same dataset analyzed under different models (e.g. concatenation versus multispecies coalescent [16]). I propose this procedure to assist with visualizing the uncertainty of phylogenetic estimates under these different scenarios.

4.3.1 Multivariate tree uncertainty

Here I use samples from a posterior distribution on tree space to generate collections of tree estimates. However, the procedure proposed is agnostic with respect to the origin of the estimates. Bootstrap resampling is another plausible method for generating collections of tree estimates [17, 26].

Bayesian methods for estimating phylogenies naturally give rise to collections of trees as samples from the posterior. Here I consider visualizing discordance between mitochondrial and nuclear gene phylogenies. In general, phylogenetic inferences based on different genes for a given set of taxa may differ with respect to topology and/or relative branch length due to poor gene tree reconstruction or because the gene trees differ from the underlying species tree [36, 51]. In particular, mitochondrial DNA can be especially problematic for resolving phylogenetic relationships at deeper evolutionary timescales due to

its higher evolutionary rate of change. A reasonable visualization procedure should showcase the relative uncertainties in inferring the phylogenies of different loci.

Wiens *et al.* [73] and Spinks *et al.* [60] investigated discordance between mitochondrial and nuclear loci in reconstructing evolutionary relationships among species of Emydid turtles. The large number of splits on the resulting tree estimates challenge fast identification of the differences between the mtDNA and nuDNA phylogeny estimates (see [73, Figures 1 and 2]). [74] combined data from [73], [60] and [1] to create a complete data matrix for 1 mitochondrial and 9 nuclear loci. Using this information, they used Bayesian methods to generate 10 gene trees $\times$ 100 tree estimates on the same 10 species: $\{\hat{T}_j^{(i)}\}_{\{i=1,\ldots,10, j=1,\ldots,100\}}$ where i indexes over the 10 genes and j indexes over the posterior trees. I apply the procedure of Section 4.3 to generate Figure 4.3 using these trees.

In Figure 4.3 there is obvious discordance between the 1 mitochondrial (*cytochrome b* gene) and the 9 nuclear genes, but no strong disagreement between the nuclear gene phylogenies. Most importantly, the discordance between mitochondrial and nuclear loci is present even after accounting for the uncertainty in estimating the gene trees. I believe that the proposed procedure is useful to distinguish situations where the uncertainty swamps the group differences (Figure 4.2) from situations where the differences exceed the uncertainty (Figure 4.3).

4.4 Contrasting MDS with the log map

MDS has its advantages over the procedure proposed here, especially when the collection of trees is unstructured (discussed in Section 4.5). However, it also has

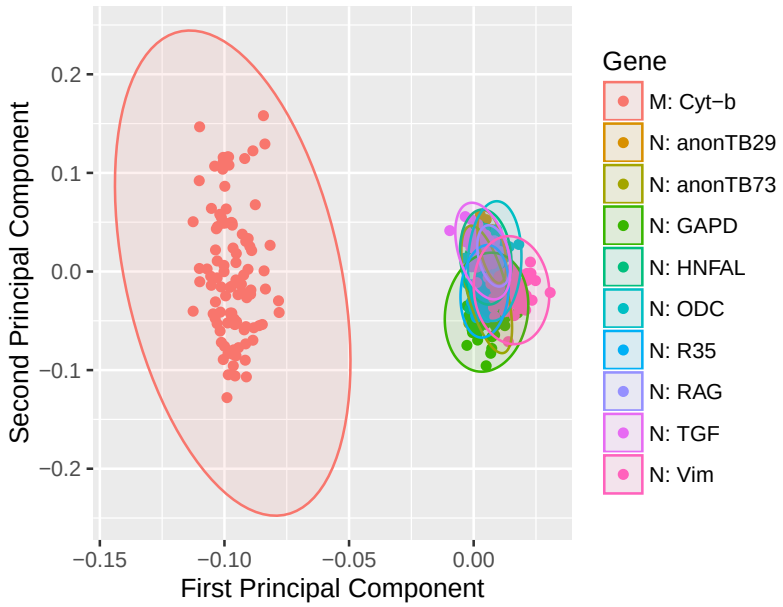

Figure 4.3: The 7-dimensional sets reflecting the variability in estimating the phylogenies of 9 nuclear (N) and 1 mitochondrial (M) gene in 10 species of Emydid turtles. The projection onto the first two principal components of the estimates is shown. We see that the difference between the nuclear and mitochondrial trees are large relative to the within-phylogeny estimation error.

serious disadvantages, some of which are not shared by the log map. I briefly describe five drawbacks of MDS that are improved upon by the procedures described here.

Visualizing uncertainty: MDS is a mapping rather than a rotation or projection, with the result that absolute measures of uncertainties in tree estimates cannot be preserved. In contrast, the same projection applied to the log maps can be applied to a set, and so sizes of uncertainties can be reflected (such as in Figures 4.2 and 4.3).

Distortions surrounding equidistance: An ideal visualization procedure would communicate clearly which trees are equally distant from a central tree. Unfortunately, MDS fails to do this in many situations. To illustrate, we uniformly at random select a fully resolved tree topology with 50 leaves, and as-

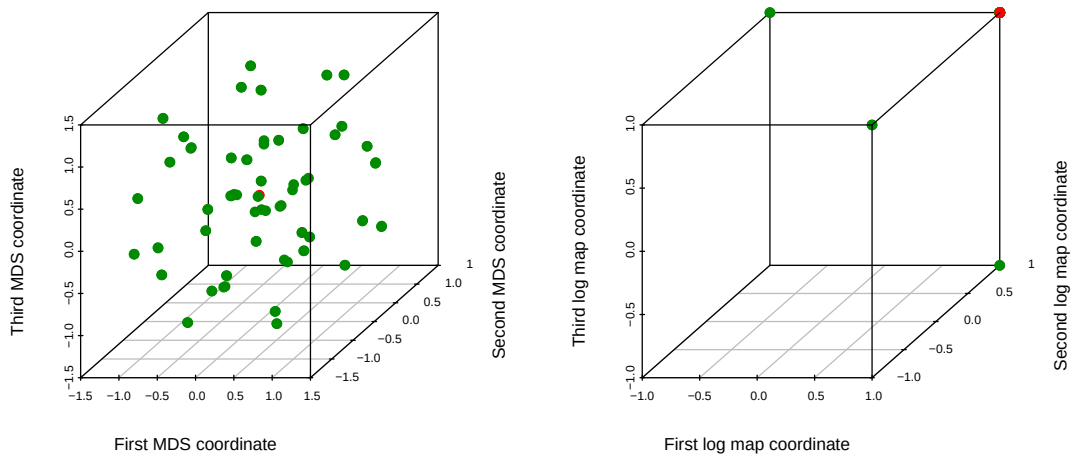

Figure 4.4: Multidimensional scaling may distort visualization of equidistant trees. Visualization of NNI trees (green) from a 50-taxon tree (red) under MDS (left) and the log map (right). All trees are equidistant from the base tree. MDS distorts this, but the log map compresses trees onto one another.

sign all branch lengths to unit length, and designate this tree as our base tree. We then consider the collection of all trees that are one NNI from this tree. According to both the RF and BHV distance, all trees are distance 2 from the base tree. However, the projection to 3 dimensions by MDS would suggest that some trees are substantially closer than others (Figure 4.4, left, modified from Hillis *et al.* [25, Figure 10]). This can be very misleading when trying to identify representative trees or outliers. The log map, by construction, preserves distances to the base tree. However, it compresses multiple topologies onto a single point (Figure 4.4, right).

Reproducibility: The stress-minimizing vector in MDS depends on every tree in the sample. As a result, a single new tree added to the sample may completely change the projection of all other points. Furthermore, visualizations

cannot be compared across different studies, making the method inherently unreproducible. As reproducibility becomes an increasing focus of genetic studies, the importance of reproducible figures and visualization-based results is no less than reproducible quantitative analyzes. As long as the base tree $\overline{T}$ and the PCA rotation matrices are maintained, the results of a new study can be compared alongside those of an original study *ex post facto*.

Topology versus branch length information: Negative coordinates in a log map indicate that the topology of the tree is different to that of the base tree. A single negative coordinate in a log map indicates that the target tree is a NNI move from the base tree. Thus the log map is capable of distinguishing topological versus branch length differences. This information is lost under MDS if the chosen metric is not RF distance (which counts the number of NNI moves between topologies). Furthermore, if there is particular interest in a certain branch of the Fréchet mean tree, this coordinate could be plotted in order to investigate if a particular set of models or genes characterize the presence of the branch. In Figure 4.5, I show two coordinates (branches) of the log map projection for the OrthoMam trees. The x-coordinate indicates the length of the branch separating the platypus from other marsupials (supported by 92% of trees), while the y-coordinate indicates the length of the branch separating the Human-Chimp-Gorilla clade from the remaining mammals (supported by 81% of trees). This information may be relevant to determining which clades on the mean tree are also supported by different genes (e.g. by color coding or labeling the genes).

Speed: Multidimensional scaling necessitates construction of the matrix of pairwise distances between trees, or calculation of $\frac{n\times(n-1)}{2}$ distances. This can be computationally prohibitive when distance calculations are expensive, as is

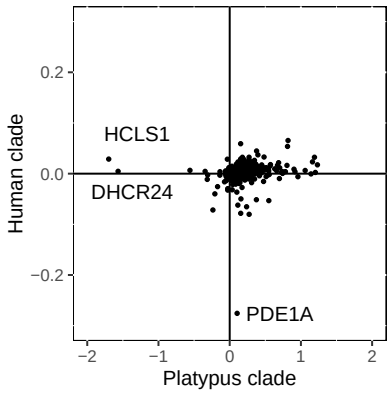

Figure 4.5: Two coordinates of the log-mapped OrthoMam gene trees. A negative coordinate in a log map indicates that the branch is absent on the tree. In this way the log map can distinguish trees that are topologically distinct from trees that have only different branch lengths. The x-coordinate indicates the length of the branch separating the platypus from other marsupials, and the y-coordinate indicates the length of the branch separating the Human-Chimp-Gorilla clade from the remaining mammals. Names of three genes that are highly discordant on these branches are shown.

the case for BHV distances. For example, in Section 4.2.1, constructing the BHV distance matrix for MDS required finding $\frac{574\times573}{2} = 164,451$ geodesics. The procedure proposed here required only 10,000 geodesics to find $\overline{T}$, followed by 574 geodesics to calculate the $\{\Phi_{\overline{T}}(\hat{T}_i)\}_i$. In practice, using the implementation of the geodesic calculation by [40] of the algorithm by [46], this amounted to 16 hours of computation for MDS compared to 5 minutes for our log map procedure, because the geodesics calculated using Algorithm 4.2 become progressively shorter, reducing the computational intensity of successive geodesic calculations.

4.5 Limitations and open problems

Because the log map assigns trees to vectors relative to a central tree, and is not a bijection between Euclidean space and tree space, it necessarily compresses information about its arguments. If a collection of trees was uniformly distributed through tree space, the log map could dissolve the most important information about the collection. However, in practice, collections of trees are rarely distributed uniformly throughout tree space, and a small number of topologies usually characterize most trees in the sample. For this reason, dimension reduction via the log map can preserve much of the structure present in the collection of trees. Nevertheless, unstructured tree datasets will be poorly reflected by the procedures proposed here.

The model of Section 4.3 could be greatly improved by using a non-spherical Brownian motion for tree space. However, no such analogue yet exists, and I believe this is a promising direction for future research.

A major limitation of almost all methods utilizing tree space is that missing leaves on some gene trees precludes them from the analysis. The procedure proposed here is no exception. While modern phylogenetic and phylogenomic data collection approaches (such as exon capture) are steadily reducing the amount of missing data, this remains a limitation.

4.6 Concluding remarks

In this chapter I have proposed a method for incorporating tree-valued and non-tree-valued information into visualizations of trees and their uncertainty.

Both methods use the log map centered at a sample Fréchet mean, an idea first proposed in Chapter 3. I have argued that these methods are faster, more reproducible, and better able to show relative uncertainties between trees than existing visualization methods. Most importantly, it provides biologists with a method for diagnosing whether differences between gene trees are biologically meaningful, or due to uncertainty in estimation.

The choice to centre the log map at a weighted Fréchet mean was not justified by theory in this chapter. I begin to address some of these issues and related problems in Chapter 5.

CHAPTER 5

INCORPORATING TREE UNCERTAINTY

Extensive theory has been developed that describes the asymptotic behavior of tree means based on observations from a tree-generating process [5, 14, 76]. However, it is not always possible to sample directly from this process. In many cases, the trees that are available for analysis are noisy estimates of the true trees of interest. For example, when we are interested in functions (such as means) of the distribution in tree space of gene trees, we usually can only estimate the gene trees. The consequences of analyzing trees that are imprecise using tools that were developed for precisely observed trees have not yet been studied.

In this chapter I describe some preliminary results that demonstrate the asymptotic unbiasedness and normality of Fréchet means calculated based on noisy realizations of the data generating process of interest. I begin by outlining the noise model for trees before stating some results.

5.1 Brownian motion on the space of trees

Nye [41] proposed a stochastic process on tree space that I use here to model the deviation of trees from their original source. The stochastic process is called tree space Brownian motion, B_{p_0,t_0}, so called because it behaves exactly as Euclidean Brownian motion in the top-dimensional strata. When the path of the process intersects a co-dimensional stratum, it uniformly at random selects an adjacent top-dimensional stratum (including the stratum from which it originated) and continues into this orthant.

Define a k-step random walk in tree space, $W^k_{t,r}$, with starting tree p_0 and

diffusion parameter t_0 by setting the k-th tree of the walk to be the $(k-1)$-th tree with a uniformly at random internal branch deviated by a realization of a $\mathcal{N}(0, t_0/k)$-distributed random variable. If the new branch length is of negative length then a randomly chosen NNI move is made at that branch. The key result of [41] was that as $m \to \infty$, this random walk weakly converges to the Brownian motion process on tree space. Consequently, tree space Brownian motion paths do not attract towards the origin.

5.2 Uncertainty model

I now use this process to model observed trees as noisy realizations of observations from the data-generating process of interest. Call the data generating process F, and let $T : \Omega \to (\mathcal{T}_m)^n$ be the random variable such that $T(\omega) = (T_1(\omega), \ldots, T_n(\omega))$ are drawn independently from F.

Now suppose that we do not observe T, but instead observe perturbations of these trees by Brownian motion deviations. That is, we observe a new random variable $Y : \Omega^* \to (\mathcal{T}_m)^n$, where

$$\begin{aligned} Y(\omega^*) &= Y(\omega, \omega^*) \\ &= (Y_1(\omega, \omega^*), \ldots, Y_n(\omega, \omega^*)) \\ &= (B_{T_1(\omega),r}(\omega^*), \ldots, B_{T_n(\omega),r}(\omega^*)). \end{aligned}$$

I now investigate if it is possible to analyze the perturbed trees Y using tools developed for analyzing the unperturbed trees T.

5.3 Analysis of perturbed tree means

5.3.1 Consistency

I first show that the sample mean of the perturbed trees is consistent for the true mean of the unperturbed trees, provided that both the diffusion parameter of the perturbation process and the variance of the unperturbed distribution are finite.

Theorem 5.3.1. *Suppose* $r < \infty$ *and that there exists some* $\alpha \in \mathcal{T}_m$ *such that* $\mathbb{E}(\gamma(T_1, \alpha)^2) < \infty$. *Then* $\hat{Y}_n \to \mu$ *as* $n \to \infty$.

Proof. Showing that $\mathbb{E}(\gamma(Y_1, \alpha)^2) < \infty$ will give that $\hat{Y}_n$ converges to its population mean [76]. I begin by showing this.

Consider the random walk $W_{t,r}^k$ defined in [41]. Let the state of a walk with k steps at step i be denoted $W_{t,r}^{(k,(i))}$. Then by the triangle inequality,

$$\begin{aligned}\gamma(t, W_{t,r}^k) &\leq \gamma(t, W_{t,r}^{k,(1)}) + \sum_{j=1}^{k-1} \gamma(W_{t,r}^{k(i)}, W_{t,r}^{k(i+1)}) \\ &:= A_1 + \ldots + A_m,\end{aligned}$$

where $A_i \overset{iid}{\sim} TN(0, r/k, 0, \infty)$, for $TN(\mu, \sigma^2, a, b)$ the truncated normal distribution. I wish to show that the variance of this sum is finite.

The moment generating function for the sum is

$$m_k(t) = m_{A_1+\ldots+A_k}(t) = 2^k \left(1 - \Phi\left(\frac{-rt}{k}\right)\right)^k e^{rt^2/2},$$

so

$$\begin{aligned}
\mathbb{E}[A_1 + \ldots + A_k] &= \frac{d}{dt} m_k(t)\Big|_{t=0} \\
&= 2^k \Bigg[e^{rt^2/2} \times \frac{d}{dt} \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^k \\
&\qquad + \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^k \times rte^{rt^2/2} \Bigg] \Bigg|_{t=0} \\
&= 2^k e^{rt^2/2} \Bigg[rt \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^k \\
&\qquad + m \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^{m-1} \times \frac{r}{k} \phi\left(-\frac{rt}{k}\right) \Bigg] \Big|_{t=0} \\
&:= 2^k r A \left[B + C\right] \Big|_{t=0} \\
&= 2^k r \left[0 + \frac{1}{2^{k-1}} \times \phi(0)\right] \\
&= 2r\phi(0)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(A_1 + \ldots + A_k)^2] &= \frac{d^2}{dt^2} m_k(t) \Big|_{t=0} \\
&= 2^k r \left(A'(B + C) + A(B' + C')\right) \Big|_{t=0} \\
&= 2^k r \left(0 \times (B + C) + 1 \times (B' + C')\right) \Big|_{t=0} \\
&= 2^k r \Bigg(\left(\left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^k + t \frac{d}{dt} \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^k \right) \Big|_{t=0} \\
&\qquad + \left(\left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^{k-1} \frac{d}{dt} \phi\left(-\frac{rt}{k}\right) \right) \Big|_{t=0} \\
&\qquad + \phi\left(-\frac{rt}{k}\right) \frac{d}{dt} \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^{k-1} \Big|_{t=0} \Bigg) \\
&= 2^k r \Bigg(\frac{1}{2^k} + 0 + 0 \\
&\qquad + \phi(0)(k-1) \left(1 - \Phi\left(-\frac{rt}{k}\right)\right)^{m-2} \times \frac{r}{k} \times \phi\left(-\frac{rt}{k}\right) \Bigg) \\
&= r + 4r^2 \left(\frac{k-1}{k}\right) \left(\phi(0)\right)^2 \\
&\to r + 4r^2 \left(\phi(0)\right)^2 < \infty
\end{aligned}$$

Because $\gamma(Y_1, T_1) \stackrel{D}{=} \gamma(t, B_{t,r})$ and $r < \infty$, $\mathbb{E}(\gamma(Y_1, T_1)^2) < \infty$. Then since $\gamma(Y_1, \alpha) \leq \gamma(Y_1, T_1) + \gamma(T_1, \alpha)$, showing that $\mathbb{E}[\gamma(Y_1, T_1)\gamma(T_1, \alpha)] < \infty$ proves the theorem. But

$$\begin{aligned}\mathbb{E}[\gamma(Y_1, T_1)\gamma(T_1, \alpha)] &= Cov(\gamma(Y_1, T_1), \gamma(T_1, \alpha)) + \mathbb{E}[\gamma(Y_1, T_1)]\mathbb{E}[\gamma(T_1, \alpha)] \\ &\leq \sqrt{(Var(A_1 + \ldots + A_m))(Var(\gamma(T_1, \alpha)))} \\ &\qquad + \mathbb{E}[A_1 + \ldots + A_m] \times \max(1, \mathbb{E}(\gamma(T_1, \alpha)^2) \\ &< \infty\end{aligned}$$

by Cauchy-Schwartz. So $\mathbb{E}[\gamma(Y_1, \alpha)^2] < \infty$.

It remains to prove that the Fréchet mean of Y_1 is equal to the Fréchet mean of T_1. By construction,

$$\arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, W^1_{t,r})^2 dF(t) = \arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, t)^2 dF(t), \tag{5.1}$$

since the expected perturbation is symmetric in all directions (this can be shown by conditioning on the edge selected to be perturbed and using the law of iterated expectation). But this implies

$$\arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, W^2_{t,r})^2 dF(t) = \arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, W^1_{t,r})^2 dF(t), \tag{5.2}$$

since $W^2_{t,r} \stackrel{\mathcal{D}}{=} W^1_{W^1_{(t,r)},r}$. So

$$\arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, t)^2 dF(t) = \arg\min_{u \in \mathcal{T}_m} \int_{\mathcal{T}_m} \gamma(u, W^k_{t,r})^2 dF(t), \qquad m = 1, 2, \ldots$$

but $W^k_{t,r} \stackrel{D}{\to} B_{t,r}$, and so we have the required result. □

5.3.2 Asymptotic normality

Having verified consistency, I now address asymptotic normality. Define

$$G(A) = \int_A B_{t,r} dF(t),$$

the distribution of Y_1.

Lemma 5.3.2. *In addition to the conditions of Theorem 5.3.1, suppose that the Fréchet function of F is everywhere finite, μ does not lie on an orthant boundary, and the measure with respect to F of the maximal cells at μ is zero. Then*

$$\sqrt{n}(\hat{Y}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

for Σ determined by F and r.

Proof. By the previous theorem, μ is the Fréchet mean of G, the distribution generating the Y's. Furthermore, since $\mathbb{E}_F[\gamma(Y_1, \alpha)^2] < \infty$, if the Fréchet function of F is everywhere finite then the Fréchet function of G is everywhere finite. The measure of the maximal cells under G is equal to the measure of the maximal cells under F. These conditions fall under the framework of [5, Theorem 2], and asymptotic normality follows. □

5.4 Open questions and conclusion

While some progress was made in this chapter regarding the applicability of existing theory to noisy tree observations, there are many remaining open questions. Extending the Brownian motion model to heteroscedastic tree noise, where each tree's perturbation has a potentially distinct diffusion parameter, is of practical significance, as discussed in Chapter 4. Unfortunately, generalizing the above theory to the non-identically distributed case is challenging, because even the most fundamental mathematical infrastructure (e.g. laws of large numbers) has not been developed in this case. I believe these extensions are important and I am continuing research in this area.

CHAPTER 6

CONCLUSION

Since the publication of tree space in 2001 [9], many advances have been made in computation [46, 45, 37, 42, 43, 2, 56] and probability [41, 4, 30, 5, 52] on this metric space. However, with some exceptions [28, 27, 72, 71, 58, 18, 7], statistics in this context has been untreated, especially relative to its potential to answer biological questions. This is especially true for inferential statistics. This thesis began to address this by describing a number of statistical methods for the analysis of tree-valued data (Chapters 3 and 4), and investigating the effect of noisy observations on tree parameter recovery (Chapter 5). A recurring theme of all work presented was incorporating uncertainty in tree estimation into phylogenetic estimates. I believe that tree uncertainty is commonly understated by biologists, and that this can explain many of the disagreements between biologists over phylogeny.

I hope that the tools presented here are of use to biologists in analyzing and visualizing phylogenetic estimates and errors. However, more importantly, I hope that the treatment of tree uncertainty as multivariate becomes incorporated into the phylogenetic literature as a tool that better reflects the actual structure of the metric space of phylogenetic trees.

CHAPTER 7

APPENDIX A: TREE SPACE AND THE HEINE-BOREL PROPERTY

I proved the following theorem to show consistency of a phylogenetic tree estimate [52]. Because the estimate was for both the internal and external branches of a tree, the parameter space is the product space of $\mathcal{T}_m$ and $\mathbb{R}^m$, and the distance metric is

$$\delta(\theta_1, \theta_2) = \gamma(T_1, T_2) + ||R_1 - R_2||,$$

where T_i refers to the internal branches, R_i refers to the external branches, and $||\cdot||$ is Euclidean L_2-distance.

Theorem 7.0.1. *The parameter space $\mathcal{T}_m \times \mathbb{R}^m$ is a metric space with metric $\delta(\cdot, \cdot)$ and has the property that every closed and bounded subset is compact.*

Proof. The distance $\delta(\cdot, \cdot)$ is a metric in the product space $\mathcal{T}_m \times \mathbb{R}^m$ [55, p. 203]. Because closed subsets of compact sets are compact [62, p. 102], I show that the set

$$T_0 = \{x \in \mathcal{T}_m : \gamma(0_{\mathcal{T}_m}, x) \leq a\} \times [-a, a]^m$$

is compact for all $a > 0$, where $0_{\mathcal{T}_m}$ denotes the origin in $\mathcal{T}_m$. Assume that T_0 is not compact, that is, there exists an open cover of T_0 that does not admit a finite subcover, which we call C. By the piecewise Euclidean structure of $\mathcal{T}_m$, I can decompose T_0 into $(2m-5)!! \times 2^m$ closed boxes in $\mathcal{T}_m \times \mathbb{R}^m$ with side lengths a, each entirely contained in a single Euclidean orthant. Since the union of these boxes gives T_0, at least one of these boxes must have a cover that requires an infinite subcover. Call one of these boxes that requires an infinite subcover T_1. I can bisect T_1, and again, and again, and contrive a decreasing sequence of boxes

$T_0 \supset T_1 \supset \ldots \supset T_k \supset \ldots$. By Cantor's intersection theorem, the intersection of these boxes contains some point $p_0 \in T_0$. There must be some open set $U \in C$ such that $p_0 \in U$. Since U is open, I can create a ball B around p_0 such that B is contained in U. But for large enough k, $T_k \subseteq B \subseteq U$. But then T_k is covered by U, a single open set, and does not need an infinite open cover. Hence every open cover of T_0 admits a finite subcover, that is, T_0 is compact. $\square$

BIBLIOGRAPHY

[1] Angielczyk, K. D., Feldman, C. R., and Miller, G. R. (2011). Adaptive evolution of plastron shape in Emydine turtles. *Evolution*, **65**(2), 377–394.

[2] Bačák, M. (2014). Computing medians and means in Hadamard spaces. *SIAM Journal on Optimization*, **24**(3), 1542–1566.

[3] Baeza, J. A. and Fuentes, M. S. (2013). Exploring phylogenetic informativeness and nuclear copies of mitochondrial DNA (numts) in three commonly used mitochondrial genes: mitochondrial phylogeny of peppermint, cleaner, and semi-terrestrial shrimps (Caridea: Lysmata, Exhippolysmata, and Merguia). *Zoological Journal of the Linnean Society*, **168**(4), 699–722.

[4] Barden, D., Le, H., and Owen, M. (2013). Central limit theorems for Fréchet means in the space of phylogenetic trees. *Electron. J. Probab*, **18**(25), 1–25.

[5] Barden, D., Le, H., and Owen, M. (2016). Limiting behaviour of Fréchet means in the space of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*.

[6] BBC Mundo (2016). Zika: el bosque de Uganda de donde salió el virus que afecta a América Latina . *www.bbc.com*.

[7] Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics*, **10**(1), 198–218.

[8] Bhattacharya, R. and Lin, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proceedings of the American Mathematical Society*, **145**, 413–428.

[9] Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, **27**(4), 733–767.

[10] Bouckaert, R. and Heled, J. (2014). DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*.

[11] Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**(10), 1372–1373.

[12] Bremm, S., von Landesberger, T., Hess, M., Schreck, T., Weil, P., and Hamacherk, K. (2011). Interactive visual comparison of multiple trees. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 31–40. IEEE.

[13] Chakerian, J. and Holmes, S. (2012). Computational Tools for Evaluating Phylogenetic and Hierarchical Clustering Trees. *Journal of Computational and Graphical Statistics*, **21**(3), 581–599.

[14] Dinh, V., Ho, L. S. T., Suchard, M. A., and Matsen, F. A. (2016). Consistency and convergence rate of phylogenetic inference via regularization. *arXiv:1606.03059*.

[15] Douzery, E. J., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., and Ranwez, V. (2014). OrthoMaM v8: A database of orthologous exons and coding sequences for comparative genomics in mammals. *Molecular Biology and Evolution*, **31**(7), 1923–1928.

[16] Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., and others (2016). Implementing and testing the multispecies coalescent model: a valuable

paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, **94**, 447–462.

[17] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, **93**(23), 13429–13429.

[18] Feragen, A., Petersen, J., Owen, M., Lo, P., Thomsen, L. H., Wille, M. M., Dirksen, A., and de Bruijne, M. (2012). A hierarchical scheme for geodesic anatomical labeling of airway trees. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–155. Springer.

[19] Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. (2016). Clustering Genes of Common Evolutionary History. *Molecular Biology and Evolution*, **33**(6), 1590–1605.

[20] Gromov, M. (1987). Hyperbolic groups. In *Essays in group theory*, pages 75–263.

[21] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.

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

[23] Hess, M., Bremm, S., Weissgraeber, S., Hamacher, K., Goesele, M., Wiemeyer, J., and von Landesberger, T. (2014). Visual Exploration of Parameter Influence on Phylogenetic Trees. *IEEE Computer Graphics and Applications*, **34**(2), 48–56.

[24] Hillis, D. M. and Huelsenbeck, J. P. (1994). Support for dental HIV transmission. *Nature*, **369**, 24–25.

[25] Hillis, D. M., Heath, T. A., and St John, K. (2005). Analysis and visualization of tree space. *Systematic Biology*, **54**(3), 471–482.

[26] Holmes, S. (2003a). Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, pages 241–255.

[27] Holmes, S. (2003b). Statistics for phylogenetic trees. *Theoretical Population Biology*, **63**(1), 17–32.

[28] Holmes, S. (2005). Statistical approach to tests involving phylogenies. In *Mathematics of Evolution and Phylogeny*, pages 91–120.

[29] Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, **2**(3), 360–378.

[30] Hotz, T., Huckemann, S., Le, H., Marron, J. S., Mattingly, J. C., Miller, E., Nolen, J., Owen, M., Patrangenaru, V., and Skwerer, S. (2013). Sticky central limit theorems on open books. *The Annals of Applied Probability*, **23**(6), 2238–2258.

[31] Kendall, M. and Colijn, C. (2016). Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution*, **33**(10), 2735–2743.

[32] Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**(1), 1–27.

[33] Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson,

J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–2948.

[34] Leaché, A. D., Chavez, A. S., Jones, L. N., Grummer, J. A., Gottscho, A. D., and Linkem, C. W. (2015). Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome biology and evolution*, **7**(3), 706–719.

[35] Loève, M. (1978). Sums of independent random variables. In *Probability theory*, pages 243–252. Springer.

[36] Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, **46**(3), 523–536.

[37] Miller, E., Owen, M., and Provan, J. S. (2015). Polyhedral computational geometry for averaging metric phylogenetic trees . *Advances in Applied Mathematics*, **68**, 51–91.

[38] Notimérica (2016). ¿De dónde procede el virus Zika? . *www.notimerica.com*.

[39] Nye, T. (2008). Trees of Trees: An Approach to Comparing Multiple Alternative Phylogenies. *Systematic Biology*, **57**(5), 785–794.

[40] Nye, T. M. (2011). Principal components analysis in the space of phylogenetic trees. *The Annals of Statistics*, **39**(5), 2716–2739.

[41] Nye, T. M. (2015). Convergence of random walks to Brownian motion in phylogenetic tree-space. *arXiv:1508.02906*.

[42] Nye, T. M. W. (2014). An Algorithm for Constructing Principal Geodesics in Phylogenetic Treespace. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **11**(2), 304–315.

[43] Nye, T. M. W. (2016). Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees . *arXiv.org:1609.03045*.

[44] Ou, C.-Y., Ciesielski, C. A., Myers, G., and others (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science*, **256**(5060), 1165–1171.

[45] Owen, M. (2011). Computing geodesic distances in tree space. *SIAM Journal on Discrete Mathematics*, **25**(4), 1506–1529.

[46] Owen, M. and Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**(1), 2–13.

[47] Puigbò, P., Garcia-Vallvé, S., and McInerney, J. O. (2007). TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, **23**(12), 1556–1558.

[48] Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences*, **13**(3), 235–238.

[49] Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., and Douzery, E. J. (2007). OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, **7**(1), 1.

[50] Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**(1), 131–147.

[51] Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**(5), 380–390.

[52] RoyChoudhury, A., Willis, A., and Bunge, J. (2015). Consistency of a phylogenetic tree maximum likelihood estimator. *Journal of Statistical Planning and Inference*, **161**, 73–80.

[53] Scaduto, D. I., Brown, J. M., Haaland, W. C., Zwickl, D. J., Hillis, D. M., and Metzker, M. L. (2010). Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proceedings of the National Academy of Sciences*, **107**(50), 21242–21247.

[54] Shen, S., Shi, J., Wang, J., Tang, S., Wang, H., Hu, Z., and Deng, F. (2016). Phylogenetic analysis revealed the central roles of two African countries in the evolution and worldwide spread of Zika virus. *Virologica Sinica*, **31**(2), 118–130.

[55] Shirali, S. and Vasudeva, H. L. (2006). *Metric spaces*. Springer.

[56] Skwerer, S. (2014). *Tree Oriented Data Analysis* . Ph.D. thesis, University of North Carolina at Chapel Hill, University of North Carolina at Chapel Hill.

[57] Skwerer, S. and Provan, S. (2015). Dynamic Geodesics in Treespace via Parametric Maximum Flow. *arXiv:1512.03115*.

[58] Skwerer, S., Bullitt, E., Huckemann, S., Miller, E., Oguz, I., Owen, M., Patrangenaru, V., Provan, S., and Marron, J. S. (2014). Tree-oriented analysis of brain artery structure. *Journal of Mathematical Imaging and Vision*, **50**(1-2), 126–143.

[59] Skwerer, S., Provan, S., and Marron, J. S. (2015). Relative Optimality Conditions and Algorithms for Treespace Fréchet Means . *arXiv:1605.02082*.

[60] Spinks, P. Q. and Shaffer, H. B. (2009). Conflicting mitochondrial and nuclear phylogenies for the widely disjunct Emys (Testudines: Emydidae)

species complex, and what they tell us about biogeography and hybridization. *Systematic Biology*, **58**, 1–20.

[61] Spinks, P. Q., Thomson, R. C., Pauly, G. B., Newman, C. E., Mount, G., and Shaffer, H. B. (2013). Misleading phylogenetic inferences based on single-exemplar sampling in the turtle genus Pseudemys. *Molecular Phylogenetics and Evolution*, **68**(2), 269–281.

[62] Stromberg, K. R. (1981). *An introduction to classical real analysis*. Wadsworth International.

[63] Sturm, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. *Contemporary mathematics*, **338**, 357–390.

[64] Sundberg, K., Clement, M., and Snell, Q. (2009). Visualizing Phylogenetic Treespace Using Cartographic Projections. In *Algorithms in Bioinformatics*, pages 321–332. Springer Berlin Heidelberg, Berlin, Heidelberg.

[65] Timm, N. H. (2002). Multivariate Distributions and the Linear Model. In *Applied Multivariate Analysis*, pages 79–184. Springer.

[66] Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**(4), 401–419.

[67] Tu, Y. and Shen, H.-W. (2007). Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics*, **13**(6), 1286–1293.

[68] Wägele, J. W. and Mayer, C. (2007). Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evolutionary Biology*, **7**(1), 1.

[69] Wang, L., Valderramos, S. G., Wu, A., Ouyang, S., and others (2016). From mosquitos to humans: genetic evolution of Zika virus. *Cell host & microbe*, **19**(5), 561–565.

[70] Weaver, S. C., Costa, F., Garcia-Blanco, M. A., Ko, A. I., Ribeiro, G. S., Saade, G., Shi, P.-Y., and Vasilakis, N. (2016). Zika virus: history, emergence, biology, and prospects for control. *Antiviral research*, **130**, 69–80.

[71] Weyenberg, G. (2016). Statistics in the Billera-Holmes-Vogtmann treespace.

[72] Weyenberg, G., Huggins, P. M., Schardl, C. L., Howe, D. K., and Yoshida, R. (2014). KDETREES: non-parametric estimation of phylogenetic tree distributions. *Bioinformatics*, **30**(16), 2280–2287.

[73] Wiens, J. J., Kuczynski, C. A., and Stephens, P. R. (2010). Discordant mitochondrial and nuclear gene phylogenies in Emydid turtles: implications for speciation and conservation. *Biological Journal of the Linnean Society*, **99**(2), 445–461.

[74] Willis, A. and Bell, R. C. (2016). Uncertainty in phylogenetic tree estimates. *arXiv:1611.03456*.

[75] Zairis, S., Khiabanian, H., Blumberg, A. J., and Rabadan, R. (2016). Genomic data analysis in tree spaces. *arXiv:1607.07503*.

[76] Ziezold, H. (1977). On Expected Figures and a Strong Law of Large Numbers for Random Elements in Quasi-Metric Spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, pages 591–602. Springer Netherlands, Dordrecht.