Information Geometry for Phylogenetic Trees

Maryam Kashia Garba

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

September, 2019

To my family. I couldn't have done it without you.

Acknowledgements

In the name of Allah, the Most Gracious, the Most Merciful. All praise and thanks is due to Allah. I send peace and blessings upon our beloved prophet Muhammad (peace be upon him). I am deeply grateful to Allah for the guidance and blessings He has bestowed upon me.

I am grateful to Newcastle University for the award of a PhD studentship that provided the necessary financial support for this research. I would like to thank my supervisors, Dr. Tom Nye and Prof. Richard Boys, for the guidance, support and encouragement they have provided throughout my study. This thesis would not have been possible without your valuable comments and feedback, not to mention your patience and generosity. I am thankful to Prof. Stephan Huckemann for the insightful discussions on Chapter 6.

My sincere appreciation goes to my husband, Abdul and my three children, Musa, Anisa and Abdurrahim, who kept our life going and endured this process with me, always offering their support and love. To my mum, thanks for always being there for me and my dad, who has always being proud of me, may your soul rest in peace. The support and comfort of my dear brother, Mutari and beloved sisters, Hadiza and Aisha, are much appreciated. Special thanks to my friends, especially Zuwaira and extended family for inspiring me at all times and for always being there for me.

Abstract

Phylogenetic trees represent evolutionary relationships between existing organisms, and are fundamental to many applications in molecular biology. These applications often require comparisons to be made between different phylogenetic trees, and this is generally achieved by using a metric or distance defined on pairs of trees. Distances between trees are used to perform hypothesis testing, cluster trees to identify differing patterns of evolution, averaging of trees, the postprocessing of results of phylogenetic analysis, among many applications. Most existing measures of distance between phylogenetic trees are based purely on the branching structure and edge lengths of the trees, and thus ignore the fact that phylogenetic trees represent probability models for gene sequence data. This project concerns the development of distance metrics and geodesics between trees based on the underlying probability distributions on genetic sequence data induced by trees. The field of information geometry offers specific methods for constructing distance metrics and geodesics on spaces of probability distributions, and hence on spaces of phylogenetic trees.

The opening chapters of the thesis give background information on phylogenetic models, inference of phylogenies from sequence data, various notions of tree space, and the fundamental ideas of information geometry. Two main areas are then developed in the rest of the thesis. First, we present methods for computing distances between trees based on the probability distributions on genetic sequence they induce. This enables metrics such as the Hellinger distance and Jenson-Shannon distance to be pulled back from the space of distributions on sequence data to tree space. Approximate calculation of these metrics on tree space involves Monte Carlo simulation methods. We compare these probabilistic metrics to existing metrics on trees, and describe various interesting properties, such as their behaviour when trees have some leaves which are not in common.

The second area concerns the construction of geodesics between trees using methods from information geometry. In the most widely studied tree space, Billera-Holmes-Vogtmann tree space, the local metric is taken to be Euclidean, and this metric extends to give a well-defined global geodesic geometry on the whole space. Existence of geodesics enables basic statistical procedures such as computation of means and variances, or principal component analysis, to be carried out in Billera-Holmes-Vogtmann tree space. This part of the thesis is motivated by the aim of reproducing such methods using an alternative and more meaningful geometry on the space of trees. As an alternative to the local

Euclidean metric, we consider the metric and corresponding geodesics on trees induced by embedding tree space in the space of $n \times n$ symmetric positive definite matrices where n is the number of leaves on each tree. Equivalently, this corresponds to the information geometry arising when a certain multivariate normal distribution is associated to each phylogenetic tree. Geodesics in the space of symmetric positive definite matrices can be computed via existing exact methods. We describe algorithms for constructing geodesics with respect to the metric on tree space induced by the embedding. These are based on projecting geodesics between symmetric positive definite matrices down into the embedded tree space. In addition to the change in local geometry relative to Billera-Holmes-Vogtmann tree space, it is necessary to change the underlying topology of tree space by gluing together parts of tree space corresponding to edges with infinite length. The resulting space is known as the phylogenetic orange space, or edge-product space, and the computational tools we have developed are used to explore our proposed geometry for this space. Many open questions remain about this geometry, and the thesis closes with a discussion of future work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Phylogenetic trees

Phylogenetic trees are a fundamental tool in biology for understanding the evolutionary history of organisms. A phylogenetic tree, also known as a phylogeny, is a graphical structure that depicts the evolutionary relationship among a group of organisms or species as descended from a common ancestor. The internal points of the phylogeny represent ancestral species while the tips or leaves represent present-day species. The internal vertices depict speciation events, where a single species diverges to become two or more distinct species and the edges on the phylogeny describe the evolutionary lineages of the species. Biologists often use genetic sequences (protein or DNA) from existing species to reconstruct phylogenetic trees (De Bruyn *et al.*, 2014). Recent advances in sequencing technologies has made available a large amount of sequence data for phylogenetic analysis, such as GenBank (Benson *et al.*, 2018) and EMBL-Ban (Cochrane *et al.*, 2009). The first step in any phylogenetic reconstruction is to identify homologous sequences of the species under consideration, that is, sequences that have shared ancestry (Pearson, 2013). These sequences are arranged to represent rows within a set of sequences known as an alignment (Phillips *et al.*, 2000). Many methods have been developed for inferring phylogenetic trees from alignment or sequence data. The methods are either character-based (which use directly the alignment) or distance-based (which use a matrix of pairwise distances between sequences to reconstruct phylogenetic trees). Character-based methods include maximum parsimony, maximum likelihood and Bayesian inference methods, and examples of distance-based methods are the neighbor-joining method of Saitou & Nei (1987) and the least-squares method of Cavalli-Sforza & Edwards (1967). Some methods require a probabilistic substitution model to describe the evolutionary process (Lió & Goldman,

1998), and statistical inference can be performed using Bayesian or maximum likelihood techniques.

The most basic use of phylogenetic trees is to discover the evolutionary history and relationships among species (Clucas *et al.*, 2010; Kellogg, 2001). However, this can lead to other practical applications especially in solving biological problems. For example, Siljic *et al.* (2018) use phylogenetic trees to make inference about infectious disease transmission. In fact, since a well-known case of HIV transmission in a dental practice in Florida (Ou *et al.*, 1992), phylogenetics has been used as a forensic tool for investigating HIV transmission among individuals (Abecasis *et al.*, 2018; Goujon *et al.*, 2000). In the use of phylogenetic trees to aid drug discovery, Saslis-Lagoudakis *et al.* (2012) reveal shared phylogenetic patterns across three medicinal plants from different regions, and this provides a measure of relatedness among the plants. Smith & Wheeler (2006) provide a predictive phylogenetic tree for understanding the evolution and diversity of venomous fishes and this can enhance bioprospecting in fish venoms. Phylogenetic trees are also applied to problems in biological diversity and conservation (Davenport *et al.*, 2006; Vézquez & Gittleman, 1998). Other applications include the use of phylogenetic trees to understand the origin of diseases (Kenah *et al.*, 2016; Lam *et al.*, 2010; Bush *et al.*, 1999), to predict gene function (Eisen & Wu, 2002), to understand human origin through the tree of life (Doolittle & Brunet, 2016), and to infer evolutionary process at the molecular level (Mooers & Heard, 1997), among others.

1.2 Distances between phylogenetic trees

There is usually more than one possible phylogenetic tree which is compatible with any given data set. It therefore becomes necessary to consider a number of different possible phylogenetic trees, and in fact this number can be quite large ($100 - 10^6$ phylogenetic trees). Many methods for post-processing phylogenetic trees rely on some measure of distance between pairs of phylogenetic trees. A variety of different distances are used, for example, the matching distance (Lin *et al.*, 2012), subtree prune and regraft distance (Hickey *et al.*, 2008), Billera Holmes Vogtmann (BHV) metric (Billera *et al.*, 2001; Owen & Provan, 2011), path-length-difference metric (Penny *et al.*, 1993), quartet distance (Estabrook *et al.*, 1985), partition metric (Penny & Hendy, 1985), Robinson-Foulds metric (Robinson & Foulds, 1979, 1981), nearest neighbor interchange distance (Waterman & Smith, 1978), among others. Distances between phylogenetic trees are used to perform hypothesis testing (Arnaoudova *et al.*, 2010), cluster phylogenetic trees to identify differing

patterns of evolution (Gori *et al.*, 2016; Whidden & Matsen, 2015; Kendall & Colijn, 2015; Stockham *et al.*, 2002), averaging of phylogenetic trees (Miller *et al.*, 2012), the postprocessing of results of phylogenetic analysis (Hillis *et al.*, 2005; Stockham *et al.*, 2002; Kuhner & Felsenstein, 1994), among many applications.

Billera *et al.* (2001) first introduced the phylogenetic tree space $\mathcal{T}_n$, which contains all possible phylogenetic trees on n leaves. Each bifurcating phylogenetic tree topology on n leaves is associated with an Euclidean region called an orthant, and $\mathcal{T}_n$ is formed by gluing together the different orthants along common boundaries. In other words, points within the same orthant have the same topology but different edge lengths. Orthant boundaries correspond to phylogenetic trees which contain trifurcations or similar "singular" phylogenetic trees. Billera *et al.* (2001) proved the existence of a unique path, called a geodesic, between any pair of points in the space, which minimizes a certain distance between the points. The distance is locally Euclidean in each ortant. The space provides an excellent platform for analysing data sets of trees and performing several statistical procedures (Nye *et al.*, 2017; Chakerian & Holmes, 2012; Miller *et al.*, 2012; Nye, 2011). However, analysis in this space relies on certain geometrical assumptions. The main aim of this thesis is to construct an alternative geometry for statistical analysis of data sets of phylogenetic trees.

1.3 Thesis outline

Distances between phylogenetic trees are generally defined by directly comparing the branching pattern and/or edge lengths in a given pair of phylogenetic trees. However, phylogenetic trees also represent probability models for genetic sequence data, and for some applications it might be more appropriate to use a distance measure which compares the probability distributions on characters induced by phylogenetic trees, rather than comparing the phylogenetic trees as geometric objects. By adopting suitable techniques from the field of information geometry, we aim to develop distance metrics and geodesics between phylogenetic trees based on the underlying probability distributions on genetic sequence data induced by phylogenetic trees. Chapters 2 and 3 provide background material on the thesis. While Chapter 2 contains an overview of the idea of representing phylogenetic trees as probability models for gene sequence data and the notion of tree space, Chapter 3 introduces the relevant concepts of information geometry: a means of doing geometry on statistical models in a principled way.

Chapter 4 describes methodology for calculating distances and metrics between phy-

logenetic trees when they are regarded as probability models for gene sequence data. The chapter begins with a review of existing distances and metrics between phylogenetic trees before describing simulation methods that approximately calculate Hellinger distance, total variation distance, Jensen-Shannon distance and Kullback-Liebler divergence between phylogenetic trees. These methods extend to phylogenetic trees with missing taxa. Using a simulation approach, we show how to estimate an adequate sample size for the simulation methods. The chapter ends with applications of the distance measures in various scenarios which clearly demonstrate their desirable properties over existing measures. This chapter has been published in Garba *et al.* (2018).

Chapter 5 explores information geometry in the context of tree space, a first step towards the construction of information geometry geodesics in tree space. For computational speed, we consider 5-taxon unrooted phylogenetic trees, and use numerical integration of ODEs to construct information geodesics in tree space. We show the geodesics obtained are very different from the BHV geodesics, but computational cost limits this approach.

Chapter 6 focuses on the construction of approximate geodesics in the edge-product space, which is a compactified version of BHV tree space with a different parametrization. Motivated by the findings of Chapter 5, the chapter defines formally the edge-product space and a natural way to embed this space into the space of covariance matrices. The space of covariance matrices offers an analytical way of computing geodesics and hence through embedding in this space, we formulate algorithms for constructing approximate geodesics with respect to the induced geometry in edge-product space. In contrast with the results of Chapter 5, we show that geodesic firing with respect to the induced covariance geometry in tree space produces similar geodesics as information geometry geodesics.

The main results in this thesis are

1. New information-based distance metrics on tree space with interesting properties that are very different from existing metrics, published in Garba *et al.* (2018).

2. Investigation of information geometry geodesics on "orange space" or edge-product space. Geodesics behave differently from existing geodesics in tree space.

3. A new geometry for phylogenetic trees. The embedding of edge-product space into the space of covariance matrices opens a new geometry for the space, and we establish some fundamental computational methods.

Chapter 2

Background

2.1 Evolution

Over many generations, changes in the characteristics of living organisms are the main driving force of evolution. Deoxyribonucleic acid (DNA) contained in every cell of all living organisms is the genetic material responsible for these characteristics. DNA is a polymeric molecule in which each piece of the polymer consists of one of four different building blocks called *nucleotides* or *bases* - adenine (A), cytosine (C), guanine (G) and thymine (T). It can therefore be represented as a sequence of letters, with the letter at each site or position representing the corresponding nucleotides along the polymer chain. DNA molecules usually comprise of two complementary sequences with nucleotide A in one sequence complementing nucleotide T on the other, and nucleotide G complementing nucleotide C. This joint structure is called the *double helix*. Certain regions in DNA are called *genes*, and these encode various characteristics of the organism. Many genes contain the instructions for making proteins. Proteins are another type of molecule found in cells that carry out various functions within the body, such as transporting oxygen around the body and catalysing chemical reactions. They are built up from 20 different building blocks called *amino acids*. In a specific DNA sequence, every group of three nucleotides codes for a different amino acid and each group is called a *codon*. Genes are spread along the DNA and direct the production of proteins through a process called *expression*. Another level of genetic structure is the *genome*, the set of DNA in an organism's cell. Different species have different genome structures, for example, the human genome consists of 23 pairs of chromosomes with over 3 billion DNA base pairs, while the bacterium E. Coli has a single circle of DNA carrying 5 million base pairs. However, there are differences in the DNA sequences among individuals within each species, called *genetic variation*.

During reproduction, organisms pass a copy of their DNA to their offspring through DNA *replication*. Errors or *mutations* can arise in the process, for instance, one nucleotide can be replaced by another in the new DNA copy. This is called *point mutation*. Other kinds of mutations apart from point mutation can occur - insertion, deletion or rearrangements but we will be concerned only with point mutations in DNA. Mutations that occur in coding regions might alter the corresponding protein leading to abnormal functions. However, beneficial mutations improve the ability of the organism to survive and reproduce, while other mutations are neutral. A beneficial or neutral mutation tends to be carried through generations until it eventually becomes *fixed* in the population, or in other words, the mutation is found in a large proportion of the population. A fixed mutation involving a change of a single nucleotide is called *substitution*.

Together, DNA replication and random mutation provide the source of evolutionary change. Evolution also involves *natural selection*. This is the process whereby within a population, organisms with certain genes better adapted to the environment are more likely to survive and reproduce than others. Therefore, their genes are more likely to be passed from one generation to another. Mutation ensures genetic variation, without which natural selection would bring about a population in a genetic steady state. Under the action of both mutation and selection, mutations gradually accumulate over time and the corresponding characteristics of organisms *evolve*. A population of organisms that are capable of reproducing offspring is called a *species*. When different sub-populations of fixed species diverge to become distinct species, *speciation events* are said to occur.

2.2 Phylogenetic trees

A tree is a non-empty set of vertices V and a set of edges E connecting them, such that a unique path of edges connects any two vertices. The *degree* of the vertex is the number of edges that join to the vertex. Degree 1 vertices are called *leaves* and vertices other than the leaves are called *internal vertices*.

In the last section, we saw how evolutionary processes give rise to new species. The branching pattern of speciation can be represented in the form of a *phylogenetic tree* (or simply a *phylogeny*). A phylogenetic tree is a tree for which the tips or leaves represent extant species, while the internal vertices represent speciation events (the ancestral species) that occurred in the past. The leaf vertices are labelled with the species they represent, which gives a bijection from the set of leaves to a set of labels $\mathcal{L}$. Edges on phylogenies are weighted. The weight or length of each edge represents the degree of divergence between

the species at its ends. On certain types of phylogeny, the edge weights are proportional to the time between speciation events. The *topology* is a phylogeny with edge lengths ignored.

Phylogenies are either rooted or unrooted. Unrooted phylogenies have no degree 2 vertices. A rooted phylogeny has a unique degree 2 vertex identified as the root and therefore has a direction associated to each edge. The root represents the common ancestor of the set of species $\mathcal{L}$, usually referred to as descendants.

2.3 Markov models of nucleotide substitution

2.3.1 Characters

A character is an assignment of letter or state to each taxon (species) on a phylogeny. Thus, a DNA character is an assignment of {A, C, G, T} to each taxon or equivalently a map $\mathcal{L} \to \{\text{A, C, G, T}\}$, so there are a total of 4^n characters for a phylogeny with $|\mathcal{L}| = n$ leaves. We let Ω denote the set of states and use Ω^n to denote the set of all characters.

If we consider the letter at a single genomic site in the most recent common ancestor (MRCA) of the species $\mathcal{L}$, then it can potentially evolve into a different letter at a corresponding site in each extant descendant species $\mathcal{L}$. Each genomic site in the MRCA therefore determines a character.

We will consider the binary alphabets $\Omega = \{0,1\}$ in addition to the DNA alphabets. This could be used to model presence or absence; for example, it can be used to model whether a particular gene is present in a species or not. In fact, any DNA sequence can be rewritten as a binary sequence, but the resulting sequence will be twice as long with no independence between columns.

2.3.2 Markov process

We focus on how to model the evolution of a single genomic site. Consider the nucleotide (letter) at this site to be a random variable $X(t)$, defined at time t, that assumes values in a discrete finite space $\Omega = \{\text{A, C, G, T}\}$. The process of change from one letter to another over a certain period of time is described by the continuous-time Markov process, $X(t)$. In other words, the process of substitution satisfies the *Markov property*

$$\Pr(X(t_{n+1}) = i_{n+1} | X(t_1) = i_1, \ldots, X(t_n) = i_n) = \Pr(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n)\,,$$

for any times $t_1 < t_2 < \cdots < t_{n+1}$ and states $i_1, \ldots, i_{n+1} \in \Omega$. In other words, given previous states $i_1, \ldots, i_n$ of the process, the distribution of the future state i_{n+1} at time t_{n+1} depends only on the present state i_n at time t_n.

We assume that the process $X(t)$ is homogeneous, that is

$$\Pr(X(t+h) = j | X(t) = i) = \Pr(X(s+h) = j | X(s) = i)\,,$$

for any time interval $h > 0$ and for all times s, t, and $i, j \in \Omega$. This allows us to describe the process by a transition probability matrix $P(t) = (p_{ij}(t))$:

$$p_{ij}(t) = \Pr(X(t) = j | X(0) = i)\,.$$

The element $p_{ij}(t)$ represents the probability that a nucleotide in its initial state i will be in state j after time t has elapsed. The row sums of $P(t)$ are equal to 1 for all t and the homogeneous assumption implies that $P(t)$ satisfies the Chapman-Kolmogorov equations:

$$P(t+h) = P(t)P(h), \tag{2.1}$$

for any times t, h. In addition $P(0) = I$, where I is the identity matrix.

Suppose $P(t)$ is differentiable. A Taylor series expansion of $P(t)$ about $t = 0$ is

$$P(t) = P(0) + tQ + \mathcal{O}(t^2),$$

where

$$Q = \frac{dP}{dt}\Big|_{t=0} = \lim_{t \to 0} \frac{P(t) - I}{t}. \tag{2.2}$$

The matrix $Q = (q_{ij})$ is called the *instantaneous rate matrix*. The non-diagonal entries of Q represent the instantaneous rate of change from nucleotide state i to nucleotide state j, that is, for $i \neq j$

$$q_{ij} = \lim_{t \to 0} \frac{p_{ij}(t)}{t}$$

and by definition, $q_{ij} \geq 0$. Also every row of Q sums to zero, which can be seen by

summing (2.2) over j:

$$\sum_j q_{ij} = \sum_j \lim_{t\to 0} \frac{p_{ij} - \delta_{ij}}{t} = \lim_{t\to 0} \frac{\sum_j (p_{ij} - \delta_{ij})}{t} = \lim_{t\to 0} \frac{1-1}{t} = 0,$$

where δ_{ij} is the Kronecker delta. Hence $q_{ii} = -\sum_{j\neq i} q_{ij}$.

The transition matrix P and the rate matrix Q are related by the forward and backward Kolmogorov equations. Using (2.1)

$$P(t+h) = P(t)P(h) = P(t)\left\{I + hQ + O(h^2)\right\}$$

and so, as $h \to 0$

$$\frac{P(t+h) - P(t)}{h} = P(t)Q.$$

Hence $P(t)$ satisfies the forward Kolmogorov equation

$$\frac{dP(t)}{dt} = P(t)Q,$$

and similarly the backward Kolmogorov equation

$$\frac{dP(t)}{dt} = QP(t).$$

The solution to these equations is

$$\begin{aligned} P(t) &= I + tQ + \frac{1}{2!}t^2Q^2 + \frac{1}{3!}t^3Q^3 + \dots \\ &= \exp(tQ). \end{aligned}$$

If the rate matrix Q can be diagonalised into

$$Q = U\mathrm{diag}(\lambda_1, \dots, \lambda_n)U^{-1},$$

where the λ_i are eigenvalues of Q and the eigenvectors of Q form the columns of U, then P can be easily computed using

$$P(t) = \exp(tQ) = U \times \mathrm{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}) \times U^{-1}, \quad |\Omega| = n. \tag{2.3}$$

2.3.3 Stationary distributions

The vector $\boldsymbol{\pi}$ is called the *stationary distribution* for the states if

$$\boldsymbol{\pi} = \boldsymbol{\pi} P(t),$$

for all times t. Therefore, if the Markov process $X(t)$ has distribution $\boldsymbol{\pi}$, then $X(s)$ will have the same distribution $\boldsymbol{\pi}$, for all $s \geq t$. Suppose $X(t)$ is irreducible, that is, there is a positive probability of changing from any state i to any other state j at some future time. It can be shown that, after sufficient time, the process is distributed according to the state vector $\boldsymbol{\pi}$ independent of the starting state, that is, for all $i, j \in \Omega$

$$\lim_{t \to \infty} p_{ij}(t) = \pi_j.$$

Given the rate matrix (q_{ij}), $-\sum_i \pi_i q_{ii}$ is the overall substitution rate. Typically, the rate matrix Q is normalised so that

$$\sum_i \pi_i q_{ii} = -1. \tag{2.4}$$

2.3.4 Reversibility

A Markov process $X(t)$ is reversible if it satisfies the detailed balance equation

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \quad \text{for all } i, j \in \Omega \text{ and time } t. \tag{2.5}$$

This means that the probability of sampling nucleotide state i from the stationary distribution and changing to nucleotide state j over a time t is the same as the probability of sampling state j from the stationary distribution and changing to state i over t (Gascuel, 2005).

With this assumption, the rate matrix Q can be decomposed as $Q = \mathcal{R}\Delta$, where $\mathcal{R} = (\rho_{ij})$ is a symmetric matrix whose entries ρ_{ij} are referred to as the exchangeability parameters, and $\Delta = \text{diag}(\boldsymbol{\pi})$ is a diagonal matrix of the stationary distribution $\boldsymbol{\pi}$. Every row of $\mathcal{R}$ sums to zero, that is, $\rho_{ii} = -\sum_{j \neq i} \rho_{ij}$.

2.4 Distribution of characters on phylogenies

We will model substitutions via a continuous-time Markov process defined along each edge of the phylogeny. The edge length ℓ on the phylogeny is a measure of the expected number of substitutions that occured during a time t. Therefore, if substitutions occur at a rate μ, we expect $\ell = \mu t$ substitutions to have occured in time t. The transition probability matrix defined over an edge length ℓ of a phylogeny is given by

$$P(\ell) = \exp(\ell Q),$$

where the instantaneous rate matrix has been normalised as in (2.4).

In phylogenetic analysis, the evolutionary process at each site is commonly assumed to be homogeneous, stationary and reversible. However, these assumptions are only for computational efficiency rather than biological reality. In fact Squartini & Arndt (2008) found stationarity and reversibility to be violated in real data sets. Despite this, we work with a standard class of models which are homogeneous, stationary and reversible.

2.4.1 Simulating characters under the model

A character is an assignment of letters to the leaves, and the Markov process induces a distribution on characters. Characters can be sampled from this distribution in the following way. We start by generating a starting letter $\hat{s}_\rho$ for the root ρ of the phylogeny by sampling a nucleotide state independently according to the stationary distribution $(\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$ of the model. If the phylogeny is unrooted, we select a root vertex ρ randomly from the set of internal vertices on the phylogeny. The starting letter is then allowed to evolve so as to generate letters at descendant vertices of the root. To evolve a letter along an edge of length ℓ, the transition matrix is used to sample a new letter. Conditional on the letter at the ancestral vertex, this sampling is independent between edges. In other words, given letter $\hat{s}_\rho$ at the root, the corresponding letter in any vertex just below ρ is randomly drawn from {A, C, G, T} with probabilities $\big(p_{\hat{s}_\rho \mathrm{A}}(\ell), p_{\hat{s}_\rho \mathrm{C}}(\ell), p_{\hat{s}_\rho \mathrm{G}}(\ell), p_{\hat{s}_\rho \mathrm{T}}(\ell)\big)$ respectively. This is repeated for every edge of the phylogeny, simulating descendant letters once the ancestral letter has been simulated. This process determines the letters at the leaves, i.e. a character, and the entire procedure can be repeated to obtain different characters. The algorithm is described as follows:

1. Select a root vertex ρ.
2. Sample a starting letter $\hat{s}_\rho$ at ρ according to the stationary distribution $\boldsymbol{\pi}$.
3. Let W be the set of vertices already assigned a letter. For each immediate descendant vertex $v \notin W$ of W, suppose ℓ is the length of the edge from v to its ancestor and x is the letter assigned to its ancestor. Then the letter at v is sampled using probabilities $(p_{x\mathrm{A}}(\ell), p_{x\mathrm{C}}(\ell), p_{x\mathrm{G}}(\ell), p_{x\mathrm{T}}(\ell))$.
4. Repeat Step 3 until all letters at the leaves are sampled, i.e. no more descendant vertices.
5. Repeat steps 1-4 to obtain different characters.

2.4.2 Probability of characters under the model

We now consider how to calculate the probability of a particular character having evolved from a phylogeny with n species. Suppose the phylogeny has topology τ and each edge is of the form $e = (u, v)$, $u, v \in V$, the set of vertices in the phylogeny. Consider a character as a function $s : \mathcal{L} \to \Omega$ with $s(i) = s_i$ and define an extension of s by $\hat{s} : V \to \Omega$, which assigns letters to all the vertices in the phylogeny that are consistent with a character s on the set of leaves $\mathcal{L}$. In other words, the function $\hat{s}$ extends the letters observed at the leaves to the interior vertices of the phylogeny. The probability of an extension is the product of the probability of the letter at the root ρ and the transition probabilities along every edge in the phylogeny, that is

$$\Pr(\hat{s}|\boldsymbol{\vartheta}) = \pi_{\hat{s}_\rho} \prod_{\text{edges } e=(u,v)} p_{\hat{s}_u \hat{s}_v}(\ell_e),$$

where $\boldsymbol{\vartheta} = (\tau, \boldsymbol{\ell}, \boldsymbol{\theta})$. The vectors $\boldsymbol{\theta}$ and $\boldsymbol{\ell}$ represent the substitution model parameters and the set of edge lengths on the phylogeny. Since the letters at the internal vertices are unknown, the probability of a character s is obtained by summing over all possibilities for the letters at the internal vertices:

$$\Pr(s|\boldsymbol{\vartheta}) = \sum_{\hat{s} \in \Omega} \Pr(\hat{s}|\boldsymbol{\vartheta}), \tag{2.6}$$

where the sum is taken over $\hat{s}$ that extends s.

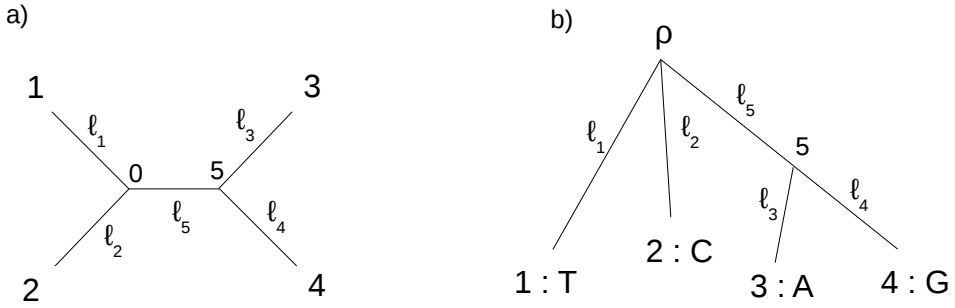


Figure 2.1: a) Unrooted 4-species phylogeny showing all the vertices and the length of the edges. b) The phylogeny in (a) rooted at vertex 0, used to demonstrate the probability calculation. The observed letters at the leaves are shown.

As an example, the probability of the character given the phylogeny in Figure 2.1, where the observed letters at the leaves are shown, is

$$p(s) = \Pr(s|\boldsymbol{\vartheta}) = \sum_{\hat{s}_\rho \in \Omega} \sum_{\hat{s}_5 \in \Omega} \pi_{\hat{s}_\rho} p_{\hat{s}_\rho \mathrm{T}}(\ell_1) p_{\hat{s}_\rho \mathrm{C}}(\ell_2) p_{\hat{s}_\rho \hat{s}_5}(\ell_5) p_{\hat{s}_5 \mathrm{A}}(\ell_3) p_{\hat{s}_5 \mathrm{G}}(\ell_4). \tag{2.7}$$

Suppose instead the position of the root is changed and the phylogeny is rooted at vertex 5. The probability now becomes

$$p(s) = \sum_{\hat{s}_5 \in \Omega} \sum_{\hat{s}_\rho \in \Omega} \pi_{\hat{s}_5} p_{\hat{s}_5 \mathrm{A}}(\ell_3) p_{\hat{s}_5 \mathrm{G}}(\ell_4) p_{\hat{s}_5 \hat{s}_\rho}(\ell_5) p_{\hat{s}_\rho \mathrm{T}}(\ell_1) p_{\hat{s}_\rho \mathrm{C}}(\ell_2),$$

which in fact equals (2.7) provided the detailed balance equation (2.5) is satisfied. This is an important outcome of the reversibility assumption, which implies that the probability of a character is independent of the root position used for the calculation.

Computing the probability (2.6) is computationally expensive for large phylogenies. The number of terms in this equation rises exponentially with the number of species, as the sum involves $|\Omega|^{n-2}$ terms for the $n-2$ interior vertices. Felsenstein (1973, 1981) introduced the Felsenstein pruning algorithm which makes the computation practicable. The basic idea of the algorithm is to use the conditional probability of subtrees given the nucleotide at their root vertex. The conditional probability for the different subtrees of the phylogeny can be computed recursively, starting from subtrees whose immediate descendants are leaves. Let $I_i(\hat{s}_i)$ be the probability of observing data at the tips that

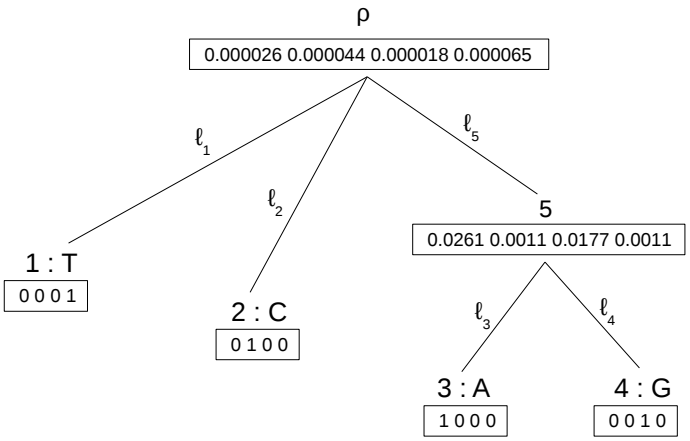


Figure 2.2: Illustration of probability calculation using the pruning algorithm for the phylogeny in Fig. 2.1. All edge lengths on the phylogeny and the model are fixed. At each vertex is the vector of conditional probabilities of observing letters at descendant vertices, given that the vertex has A, C, G or T respectively.

are descendants of vertex i, given that the nucleotide at vertex i is $\hat{s}_i$. For example, tips 3,4 are descendants of vertex 5, so $I_5(A)$ is the probability of observing $\hat{s}_3\hat{s}_4 = AG$ given that vertex 5 has nucleotide state $\hat{s}_5 = A$.

1. Initialise at the leaves: for each leaf vertex i, $I_i(\hat{s}_i) = 1$ if $\hat{s}_i$ is the observed nucleotide and 0 otherwise.

2. Follow the ancestry from the leaves to the root, applying the recursion

$$I_i(\hat{s}_i) = \sum_{\hat{s}_j} p_{\hat{s}_i\hat{s}_j}(\ell_j) I_j(\hat{s}_j) \times \sum_{\hat{s}_k} p_{\hat{s}_i\hat{s}_k}(\ell_k) I_k(\hat{s}_k)$$

   for each internal vertex i with descendant vertices j and k. If vertex i has more than two descendant vertices, $I_i(\hat{s}_i)$ will be a product of as many terms.

3. Suppose vertex ρ is the root. Compute the probability of a character s at the set of leaf vertices through

$$p(s) = \sum_{\hat{s}_\rho} \pi_{\hat{s}_\rho} I_\rho(\hat{s}_\rho).$$

We illustrate the pruning algorithm by computing the probability of the observed nucleotides T,C,A,G at the leaves of the phylogeny in Figure 2.1. Suppose the probability transition matrix is given as

$$P(0.1) = \begin{pmatrix} 0.8954 & 0.0561 & 0.0291 & 0.0194 \\ 0.0374 & 0.9141 & 0.0291 & 0.0194 \\ 0.0194 & 0.0291 & 0.9141 & 0.0374 \\ 0.0194 & 0.0291 & 0.0561 & 0.8954 \end{pmatrix},$$

where all the edge lengths of the phylogeny are fixed at 0.1. Starting from leaf vertex 1, $I_1(T) = 1$ and 0 otherwise, since T is the observed nucleotide at vertex 1. Similarly, this is computed for all other leaf vertices as shown in Figure 2.2. The algorithm then proceeds to the vertex whose all descendant vertices have been visited, in this case vertex 5. We compute the conditional probability of observing A,G at vertices $3, 4$ given that the nucleotide at vertex 5 is $\hat{s}_5$. But since 5 is an interior vertex, $I_5(\hat{s}_5)$ is computed for all possible nucleotides $\hat{s}_5 \in \Omega$ as

$$\begin{aligned} I_5(\mathrm{A}) =& \{p_{\mathrm{AA}} I_3(\mathrm{A}) + p_{\mathrm{AC}} I_3(\mathrm{C}) + p_{\mathrm{AG}} I_3(\mathrm{G}) + p_{\mathrm{AT}} I_3(\mathrm{T})\} \\ & \times \{p_{\mathrm{AA}} I_4(\mathrm{A}) + p_{\mathrm{AC}} I_4(\mathrm{C}) + p_{\mathrm{AG}} I_4(\mathrm{G}) + p_{\mathrm{AT}} I_4(\mathrm{T})\} \\ =& 0.8954 \times 0.0291 \\ =& 0.0261. \end{aligned}$$

Similarly, $I_5(\mathrm{C}) = 0.0011$, $I_5(\mathrm{G}) = 0.0177$ and $I_5(\mathrm{T}) = 0.0011$. Then the conditional probabilities at the root ρ are computed (see Figure 2.2), for example

$$\begin{aligned} I_\rho(\mathrm{A}) =& \{p_{\mathrm{AA}} \times I_1(\mathrm{A}) + p_{\mathrm{AC}} \times I_1(\mathrm{C}) + p_{\mathrm{AG}} \times I_1(\mathrm{G}) + p_{\mathrm{AT}} \times I_1(\mathrm{T})\} \\ & \times \{p_{\mathrm{AA}} \times I_2(\mathrm{A}) + p_{\mathrm{AC}} \times I_2(\mathrm{C}) + p_{\mathrm{AG}} \times I_2(\mathrm{G}) + p_{\mathrm{AT}} \times I_2(\mathrm{T})\} \\ & \times \{p_{\mathrm{AA}} \times I_5(\mathrm{A}) + p_{\mathrm{AC}} \times I_5(\mathrm{C}) + p_{\mathrm{AG}} \times I_5(\mathrm{G}) + p_{\mathrm{AT}} \times I_5(\mathrm{T})\} \\ =& \{0.0194 \times 1\} \times \{0.0561 \times 1\} \\ & \times \{0.8954 \times 0.0261 + 0.0561 \times 0.0011 + 0.0291 \times 0.0177 + 0.0194 \times 0.0011\} \\ =& 0.000026. \end{aligned}$$

2.5 Some models of nucleotide substitution

2.5.1 The Jukes-Cantor (JC69) model

The simplest model of DNA substitution is the JC69 model (Jukes & Cantor, 1969) which assumes that substitution from one nucleotide to the other occurs at the same rate λ. Thus the instantaneous rate matrix is given by

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix},$$

where the letters (rows and columns) are ordered A, C, G and T. This model assumes, under stationarity, that all nucleotide frequencies are the same i.e. $\pi_i = 1/4$, for all i. The transition probability matrix $P(t)$ over time t obtained using (2.3) has elements

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, & i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}, & i \neq j \quad \text{for } i, j \in \Omega. \end{cases}$$

Assuming that the overall substitution rate is 1, λ is fixed at 1/3 to allow for parameter identification.

2.5.2 The Hasegawa, Kishino and Yano (HKY85) model

The HKY85 model (Hasegawa *et al.*, 1985) allows different probabilities for the nucleotides in the stationary distribution so that $\boldsymbol{\pi} = (\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$. In addition, the model distinguishes two types of substitution - transition and transversion. Transitions are substitutions between two purines A $\leftrightarrow$ G or between two pyrimidines C $\leftrightarrow$ T and transversions are all other substitutions between purines and pyrimidines which are known to be less likely than transitions (Fitch, 1967). Let the rate of transition and transversion be α and β respectively. The model can be parametrized by a transition-tranversion ratio $\kappa = \alpha/\beta$.

Hence, the rate matrix is given as

$$Q = \begin{pmatrix} a_1 & \pi_{\mathrm{C}} & \kappa\pi_{\mathrm{G}} & \pi_{\mathrm{T}} \\ \pi_{\mathrm{A}} & a_2 & \pi_{\mathrm{G}} & \kappa\pi_{\mathrm{T}} \\ \kappa\pi_{\mathrm{A}} & \pi_{\mathrm{C}} & a_3 & \pi_{\mathrm{T}} \\ \pi_{\mathrm{A}} & \kappa\pi_{\mathrm{C}} & \pi_{\mathrm{G}} & a_4 \end{pmatrix},$$

where

$$\begin{aligned} a_1 &= -(\pi_{\mathrm{C}} + \kappa\pi_{\mathrm{G}} + \pi_{\mathrm{T}}), \\ a_2 &= -(\pi_{\mathrm{A}} + \pi_{\mathrm{G}} + \kappa\pi_{\mathrm{T}}), \\ a_3 &= -(\kappa\pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{T}}), \\ a_4 &= -(\pi_{\mathrm{A}} + \kappa\pi_{\mathrm{C}} + \pi_{\mathrm{G}}). \end{aligned}$$

2.5.3 The general time-reversible (GTR) model

The GTR model (Tavaré, 1986) is the most general form of reversible model from which all other models are derived. The model assumes different instantaneous rate of substitution between each of the six pairs of nucleotides. The six rate parameters $\rho_{ij}, i = 1, 2, 3$, $j = i + 1, \ldots, 4$ sometimes refered to as exchangeability parameters are expressed in the rate matrix as

$$Q = \begin{pmatrix} c_1 & \rho_{12}\pi_{\mathrm{C}} & \rho_{13}\pi_{\mathrm{G}} & \rho_{14}\pi_{\mathrm{T}} \\ \rho_{12}\pi_{\mathrm{A}} & c_2 & \rho_{23}\pi_{\mathrm{G}} & \rho_{24}\pi_{\mathrm{T}} \\ \rho_{13}\pi_{\mathrm{A}} & \rho_{23}\pi_{\mathrm{C}} & c_3 & \rho_{34}\pi_{\mathrm{T}} \\ \rho_{14}\pi_{\mathrm{A}} & \rho_{24}\pi_{\mathrm{C}} & \rho_{34}\pi_{\mathrm{G}} & c_4 \end{pmatrix},$$

where

$$\begin{aligned} c_1 &= -(\rho_{12}\pi_{\mathrm{C}} + \rho_{13}\pi_{\mathrm{G}} + \rho_{14}\pi_{\mathrm{T}}), \\ c_2 &= -(\rho_{12}\pi_{\mathrm{A}} + \rho_{23}\pi_{\mathrm{G}} + \rho_{24}\pi_{\mathrm{T}}), \\ c_3 &= -(\rho_{13}\pi_{\mathrm{A}} + \rho_{23}\pi_{\mathrm{C}} + \rho_{34}\pi_{\mathrm{T}}), \\ c_4 &= -(\rho_{14}\pi_{\mathrm{A}} + \rho_{24}\pi_{\mathrm{C}} + \rho_{34}\pi_{\mathrm{G}}). \end{aligned}$$

It is necessary to impose a constraint on the exchangeability parameters, typically $\rho_{34} = 1$, to ensure parameter identifiability. This prevents arbitrary rescaling of the

edge lengths and the exchangeability parameters, and the overall substitution rate can be expressed relative to the constrained parameter.

2.6 Rate variation

All the models introduced in Section 2.5 assume that the rate of substitution is a constant for every site in a genome. This is not true in general. Rate variation across sites has long been detected in DNA sequences (Fitch & Margoliash, 1967; Uzzell & Corbin, 1971; Wakeley, 1993; Excoffier & Yang, 1999). Typically, the substitution rate may vary at different sites due to different evolutionary pressures acting on nucleotides as a result of the functional structure of the gene (Bofkin & Goldman, 2006). While other sites may change freely, many functionally useful sites are maintained in the process of evolution.

We previously specified the distribution of characters under the assumption of a shared fixed rate. Here, we generalise this distribution by introducing the concept of rate variation. One way to account for variable substitution rate is to assume that the rate of substitution r is a random variable modelled by a statistical distribution. Yang (1993) suggested the gamma distribution, with probability density function

$$f(r|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1} \quad \alpha > 0,\ \beta > 0,\ r > 0,$$

where α and β are the shape and scale parameters. It is often assumed that $\alpha = \beta$, so that the distribution has mean 1 and variance $1/\alpha$. Thus a single parameter α governs the distribution by determining the level of rate variation: a small α implies significant variation in rates across sites while a very large α indicates nearly the same rate for all sites.

Due to computational cost of tree reconstruction under this model, the continuous gamma rate distribution is usually approximated with a discrete-gamma model (Yang, 1994). In this model, several rate categories $r_1, \ldots, r_k$ are used to approximate the continuous gamma distribution, where the average rate of the section of gamma distribution lying in each category represents the rate for that category (Yang, 1994). The rate is assumed to fall into each of these categories with equal probability $p_c = 1/k$, $c = 1, \ldots, k$. Also, each character with a rate of substitution r has a specific rate matrix rQ, where the instantaneous rate matrix Q is the same over all characters. Thus, the likelihood of a

character becomes

$$\Pr(s|\boldsymbol{\vartheta}) = \sum_{c=1}^{k} p_c \times \Pr(s|r_c, \boldsymbol{\vartheta}) \,,$$

where r_c is the substitution rate of the c-th category and $\boldsymbol{\vartheta} = (\tau, \boldsymbol{\ell}, \boldsymbol{\theta})$. The conditional probability $\Pr(s|r_c, \boldsymbol{\vartheta})$ is the same as (2.6) but with all edge lengths scaled by the rate r_c depending on the character category. Yang (1994) recommends $k = 4$ as a plausible value for the number of categories.

2.7 Tree reconstruction

One of the aims of phylogenetic analysis is to establish relationships between species by inferring the common evolutionary history of the species. Given DNA sequences from species under investigation, a biologist first has to identify sites related by common descent. Such sites, called homologous sites, yield a collection of different characters. These are represented by an *alignment*, typically $D = (D_{ij})$, where D_{ij} is the letter observed in character j for species i. Each column of D is a character, and characters are typically modelled as evolving independently from site to site. Tree inference is based on D. Tree reconstruction methods include distance-based methods which use a matrix of pairwise distance between sequences to reconstruct the tree. Several methods have been developed. For example, the widely used neighbor-joining (Saitou & Nei, 1987) uses a clustering algorithm on the distance matrix to come up with a tree, and the least-squares (LS) method (Cavalli-Sforza & Edwards, 1967) minimizes the sum of squared differences between the distance matrix and the path length difference matrix between any pair of species. Another approach to tree reconstruction is maximum parsimony which minimizes the total number of letter changes over the tree. By assigning states to the interior vertices of a tree, a minimum number of substitutions is calculated for each site. A score for each tree is then obtained as the sum over all sites in the alignment. The maximum parsimony tree is the tree with the smallest score and provides an estimate of the true tree. However, inference under the probabilistic models described above is currently regarded as the best way to reconstruct trees. Inference can be performed in a maximum likelihood or Bayesian framework.

2.7.1 Maximum likelihood

Maximum likelihood (ML) is a statistical method for estimating unknown parameters in the evolutionary model. The likelihood is defined as the probability of the data given a tree (with topology τ and edge lengths $\boldsymbol{\ell}$) and a substitution model of evolution (with parameters $\boldsymbol{\theta}$). Assuming that all sites evolve independently, the likelihood is the product of the site likelihoods:

$$L(\boldsymbol{\vartheta}) = \Pr(D|\boldsymbol{\vartheta}) = \prod_{j=1}^{m} \Pr(s_j|\boldsymbol{\vartheta}),$$

where m is the number of sites and the likelihood L is considered as a function of the parameters $\boldsymbol{\vartheta} = (\tau, \boldsymbol{\ell}, \boldsymbol{\theta})$. The maximum likelihood method aims to find the parameter values that maximize the likelihood function and the corresponding values are called the maximum likelihood estimates (MLEs). Felsenstein (1973, 1981) developed the first algorithm for computing the maximum likelihood parameters from DNA sequence data (see Section 2.4.2). The basic idea is to compute the likelihood of a given data set and maximize it over all possible trees. For a given tree topology, edge lengths are altered in order to maximize the likelihood, and the tree is assigned a score given by the maximum likelihood value. Then, a new topology is tried usually obtained by some rearrangement of the previous topology. The tree with the best score is identified as the maximum likelihood tree. Several maximum likelihood programs are available, PAML (Yang, 1997), PhyML (Guindon & Gascuel, 2003) and RAxML (Stamatakis, 2006), among others.

2.7.2 Bayesian inference

Bayesian inference is a way of modelling uncertainty about the unknown parameters using prior knowledge (about the parameters) in addition to the data. Therefore the parameters are associated with a random variable following a statistical distribution, as opposed to being fixed constants in the maximum likelihood method. The prior distribution of the parameters before the analysis, together with the likelihood of the data is used to obtain the posterior distribution of the parameters. The way these are combined is known as Bayes theorem: the posterior density of the parameters $\boldsymbol{\vartheta} = (\tau, \boldsymbol{\ell}, \boldsymbol{\theta})$ given the data D is

$$p(\boldsymbol{\vartheta}|D) = \frac{p(\boldsymbol{\vartheta})p(D|\boldsymbol{\vartheta})}{p(D)},$$

where $p(\boldsymbol{\vartheta})$ is the prior probability density function of the parameters $\boldsymbol{\vartheta}$, $p(D|\boldsymbol{\vartheta})$ is the likelihood or probability of the data given $\boldsymbol{\vartheta}$ values and $p(D) = \int_{\boldsymbol{\vartheta}} p(D|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}$ is the marginal probability of the data. In general, calculating $p(D)$ is difficult analytically, as it involves integration over all possible parameter values: all edge lengths $\boldsymbol{\ell}$ and all parameters $\boldsymbol{\theta}$ of the substition model for each tree topology τ. Instead, Markov Chain Monte Carlo (MCMC) algorithms are used to generate a sample from the posterior distribution of trees. The development of efficient MCMC algorithms through programs such as MrBayes (Huelsenbeck & Ronquist, 2001) and BEAST (Drummond & Rambaut, 2007) has made Bayesian inference methods very popular in the systematics community.

Bayesian inference is typically used to infer phylogenies which are not time-like. For a non time-like phylogeny, the likelihood only does not depend on the root position if the continuous time Markov process is assumed to be reversible and in its stationary distribution. Sometimes, molecular clock phylogenies are used, in which every leaf has the same distance from the root. Relaxed molecular clock models allow rates to vary between edges, and under these models both a molecular clock phylogeny and unconstrained phylogeny are inferred.

2.8 Tree space

We use two related notions of the space of all possible trees for a fixed set $\mathcal{L}$ of labelled leaves.

2.8.1 Billera, Holmes and Vogtmann (BHV) space

Consider unrooted phylogenies on $|\mathcal{L}| = n$ leaves. A phylogeny is said to be *resolved* if each interior vertex has degree 3, while unresolved if at least one interior vertex is of degree greater than 3. We denote the set of all such phylogenies as $\mathcal{U}_n$. Any phylogeny in $\mathcal{U}_n$ has n pendant edges, so we represent their lengths as a point in $\mathbb{R}^n_{\geq 0}$, and so we can write $\mathcal{U}_n = \mathbb{R}^n_{\geq 0} \times \mathrm{BHV}_n$. BHV_n is a space that parametrizes internal edge lengths and topology, which we describe in the following way.

A *split* $X|X'$ is a bipartition of the set of leaves $\mathcal{L}$ into two non-overlapping non-empty sets X and X'. Cutting any edge in a phylogeny results in a split which divides the phylogeny into two subtrees, with subtrees containing exactly the leaves in X and X' respectively. Two splits $X|X'$ and $Y|Y'$ are *compatible* if one of the subsets $X \cap Y$, $X \cap Y'$, $X' \cap Y$ or $X' \cap Y'$ is empty. In other words, two splits are compatible if and

only if they can both be displayed in the same phylogenetic tree. Each fully resolved topology τ corresponds to a set of $n-3$ internal splits. Lengths of these splits are parametrized by $\mathcal{O}_\tau = \mathbb{R}^{n-3}_{\geq 0}$. This is called a maximal orthant. There are $(2n-5)!! = 1 \times 3 \times 5 \times \ldots \times (2n-5)$ different fully resolved topologies, so BHV_n consists of $(2n-5)!!$ maximal orthants. BHV_n is constructed by gluing the orthants together along their boundaries via an equivalence relation. Under this relation, two phylogenies are equivalent if they are the same modulo removal of zero-length internal edges (see Figure 2.3 for example). A pictorial representation of BHV_4 space is shown in Figure 2.4, where three orthants corresponding to the different fully resolved topologies are glued together at the origin (boundary). Phylogenies on the same orthant, e.g. x and u, have the same topology. One moves around the space by varying edge lengths. When one edge is contracted to length zero, we end up at the boundary of the orthant which is shared by two other orthants. Hence, there are two possible ways of expanding out a replacement edge each resulting in a different topology. These moves are based on nearest neighbor interchange (NNI) between splits. Thus, phylogeny z can be obtained from x by continously shrinking split $12|34$ to the origin, the replacing $12|34$ with $13|24$ and expanding this split out.

Any rooted phylogeny is equivalent to an unrooted phylogeny with an additional special vertex labelled "root" attached to the phylogeny via a zero length pendant. Therefore, we use $\mathcal{T}_n = \mathbb{R}^n_{\geq 0} \times \mathrm{BHV}_{n+1}$ to denote the space of rooted phylogenies on n leaves. Billera *et al.* (2001) explored the geometry of BHV_n (and hence $\mathcal{U}_n$ and $\mathcal{T}_n$): it is equipped with a natural metric, which is locally the Euclidean metric on each orthant. Two points in different orthants can be joined by series of line sections, with each section lying in a single orthant. The length of the path can then be measured by summing up the length of the sections. The distance between two points is defined as the minimum of the lengths of paths joining the two points. A minimum length path is called *geodesic*. Billera *et al.* (2001) proved there exists a unique geodesic between every pair of points. They also showed that this natural metric has a non-positive curvature, which guarantees the uniqueness of geodesic paths. Owen & Provan (2011) provided a polynomial time algorithm for computing geodesics in this space.

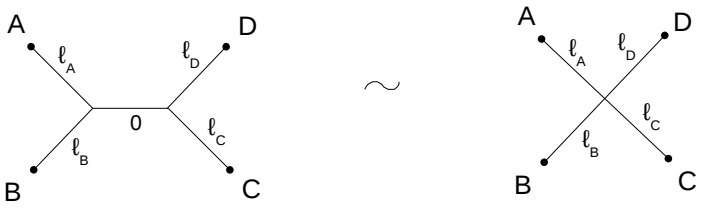


Figure 2.3: Two equivalent phylogenies showing removal of zero length internal edge. Here A, B, C, D are subtrees.

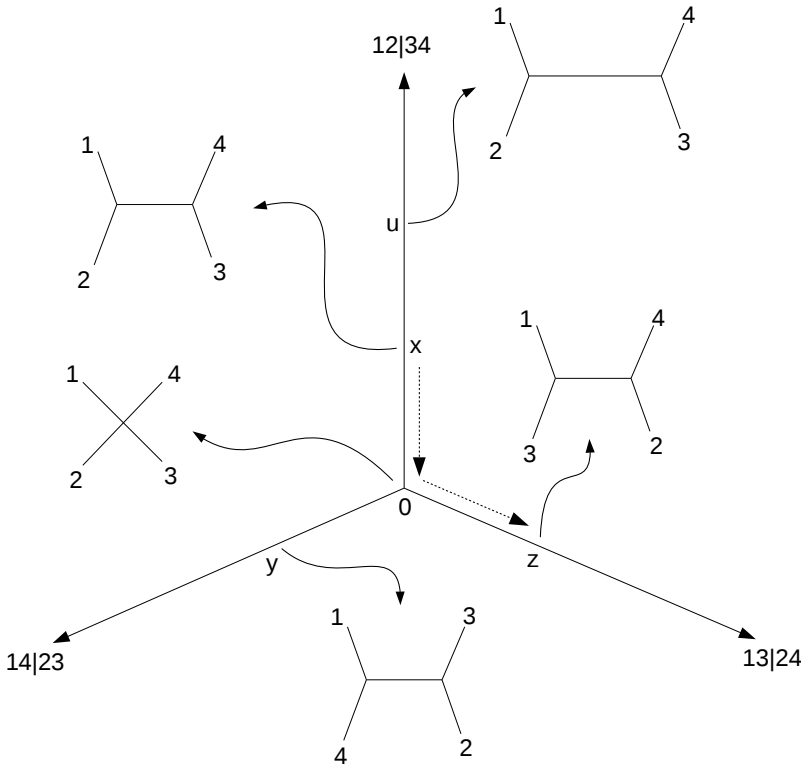


Figure 2.4: A pictorial representation of BHV_4 space. The distinct fully resolved topologies on four taxa correspond to three copies of $\mathbb{R}_{\geq 0}$ joined together at the origin. Phylogenies of the same topology (e.g. x and u) belong to the same orthant. Movement across boundary of orthants is through nearest neighbor interchange (NNI): phylogeny z can be obtained from phylogeny x through a nearest neighbor interchange of split 12|34 into split 13|24.

2.8.2 Edge-product space

In the edge-product space, phylogenetic trees are identified with points in a space of distributions on characters. This idea was first considered by Kim (2000). The space is often also referred to as the space of "hyperdimensional oranges" or "phylogenetic oranges". Moulton & Steel (2004) studied the topological and combinatorial aspects of this space.

The space can be constructed by first re-parametrizing $\mathcal{U}_n$. If ℓ is the length of an edge in a phylogeny $(\tau, \boldsymbol{\ell}) \in \mathcal{U}_n$, we re-parametrize by using $\lambda^* = e^{-\ell}$. Hence, maximal orthants in $\mathcal{U}_n$ are replaced by $(n-3)$-dimensional unit cubes. The length of the path on a phylogeny between two leaves i, j is given as

$$\sum_{e \text{ on path } i,j} \ell_e,$$

which transforms to

$$\prod_{e \text{ on path } i,j} \lambda_e^*.$$

It is this re-parametrization which gives rise to the name edge-product space. The edge-product space includes phylogenetic trees with $\lambda^* = 0$, which corresponds to the boundary at infinity of $\mathcal{U}_n$. Hence, as a set, the BHV tree space is a subset of the edge-product space, but with a different parametrization. In fact, the edge-product space parametrizes certain Markov models defined on phylogenetic trees (Moulton & Steel, 2004). No geometry exists on this space to date, and we give a more rigorous definition of the space later in Chapter 6.

Let Y_i denote the character at vertex i and $d_j(Y_i)$ denote the character at the vertex which is left ($j = 0$) or right ($j = 1$) immediate descendant of vertex i. Let $l(Y_i) = \{l_0(Y_i), l_1(Y_i)\}$ denote the characters at the set of leaf vertices which are descendants of vertex i, partitioned according to whether they result from the left ($j = 0$) or right ($j = 1$) branch at i.

Chapter 3

Information geometry

In this chapter, we consider spaces of probability distributions as (i) metric spaces, and (ii) Riemannian manifolds. We start by defining some notation. Let X denote a discrete set. A probability distribution on X is described by a function $p : X \to \mathbb{R}$ defined by $p(x) = \Pr(X = x)$ which satisfies

$$p(x) \geq 0, \forall x \in X \qquad \text{and} \qquad \sum_{x \in X} p(x) = 1. \tag{3.1}$$

The function p is referred to as a probability function or a probability mass function (pmf for short). If instead X is a continuous set, then the summation symbol in (3.1) becomes an integral over the domain of X, in which case p is called a probability density function (pdf). Define $\mathcal{D}(X)$ to be the set of probability distributions on X when X is a finite discrete set.

3.1 Distances between discrete probability distributions

Although we will consider $\mathcal{D}(X)$ as a Riemannian manifold later, we start by considering it simply as a metric space. It is often necessary in applications to compare two statistical models which are specified by some distributions in $\mathcal{D}(X)$. Many measures of distance between probability distributions exist, which include Hellinger distance, Wasserstein metric, Kullback-Leibler divergence, total variation distance, Jensen-Shannon distance, Kolmogorov-Smirnov distance, etc. Some of these distances are not metrics, since they are not symmetric and do not satisfy the triangle inequality, but they nonetheless

have useful properties.

For any arbitrary set Z, a function $d : Z \times Z \to \mathbb{R}$ is a *metric* if for all $x, y, z \in Z$, it satisfies the following properties:

(i) Nonnegativity: $d(x, y) \geq 0$,

(ii) Identity: $d(x, y) = 0 \iff x = y$,

(iii) Symmetry: $d(x, y) = d(y, x)$,

(iv) Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Our concern is on distances between probability distributions i.e. metrics and other measures of distance d whose domain is $Z = \mathcal{D}(X)$. Suppose $p(x)$ and $q(x)$, with $x \in X$, are pmfs in $\mathcal{D}(X)$. We choose to study some important and commonly used distances between p and q, stated in Table 3.1. Note that versions of these distances exist when X is a continuous set, but we are only concerned with the discrete case.

Abbreviation	Distance measures
H	Hellinger distance
TV	Total variation distance
KL	Kullback-Leibler divergence
JS	Jensen-Shannon distance

Table 3.1: Distances between probability distributions and their abbreviations.

3.1.1 Hellinger distance

The Hellinger distance between $p(x)$ and $q(x)$ is defined by

$$d_H(p, q)^2 = \frac{1}{2} \sum_{x \in X} \left\{ \sqrt{p(x)} - \sqrt{q(x)} \right\}^2 .$$

Lemma 3.1. *The Hellinger distance d_H is a metric bounded above by 1.*

Proof. (i) Obviously $d_H(p, q) \geq 0$ since $\left\{ \sqrt{p(x)} - \sqrt{q(x)} \right\}^2 \geq 0, \forall x \in X$.

(ii) $d_H(p,q) = 0 \iff \frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{q(x)}\right\}^2 = 0$

$$\iff \left\{\sqrt{p(x)} - \sqrt{q(x)}\right\}^2 = 0, \ \forall x \in X$$
$$\iff \sqrt{p(x)} = \sqrt{q(x)}, \ \forall x \in X$$
$$\iff p(x) = q(x), \ \forall x \in X$$
$$\iff p = q.$$

(iii) $d_H(p,q)^2 = \frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{q(x)}\right\}^2 = \frac{1}{2}\sum_x \left\{\sqrt{q(x)} - \sqrt{p(x)}\right\}^2 = d_H(q,p)^2.$

Therefore $d_H(p,q) = d_H(q,p)$.

(iv) $$d_H(p,q) = \left(\frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{q(x)}\right\}^2\right)^{1/2}$$
$$= \left(\frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{r(x)} + \sqrt{r(x)} - \sqrt{q(x)}\right\}^2\right)^{1/2}$$
$$\leq \left(\frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{r(x)}\right\}^2\right)^{1/2} + \left(\frac{1}{2}\sum_x \left\{\sqrt{r(x)} - \sqrt{q(x)}\right\}^2\right)^{1/2}$$
$$= d_H(p,r) + d_H(r,q),$$

by Minkowski's inequality for sums. Hence, the triangle inequality is satisfied.

We now establish an upper bound for d_H:

$$\begin{aligned} d_H(p,q)^2 &= \frac{1}{2}\sum_x \left\{\sqrt{p(x)} - \sqrt{q(x)}\right\}^2 \\ &= \frac{1}{2}\left\{\sum_x p(x) - 2\sum_x \sqrt{p(x)q(x)} + \sum_x q(x)\right\} \\ &\leq \frac{1}{2}\left\{\sum_x p(x) + \sum_x q(x)\right\} \\ &= \frac{1}{2}\{1+1\} \\ &= 1. \end{aligned}$$

Therefore, we conclude that $0 \leq d_H(p,q) \leq 1$. This occurs when $p(x) = 0 \iff q(x) \neq 0$, for all $x \in X$ or vice versa. □

3.1.2 Total variation distance

The total variation distance between $p(x)$ and $q(x)$ is defined as

$$d_{TV}(p,q) = \frac{1}{2}\sum_{x\in X}|p(x)-q(x)|.$$

Lemma 3.2. *d_{TV} is a metric and is bounded above by 1.*

Proof. (i) By definition, $d_{TV}(p,q) \geq 0$.

(ii) $$\begin{aligned} d_{TV}(p,q) = 0 &\iff \frac{1}{2}\sum_{x}|p(x)-q(x)| = 0 \\ &\iff |p(x)-q(x)| = 0,\ \forall x \in X \\ &\iff p(x) = q(x),\ \forall x \in X \\ &\iff p = q. \end{aligned}$$

Therefore $d_{TV}(p,q) = 0 \iff p = q$.

(iii) $$\begin{aligned} d_{TV}(p,q) = \frac{1}{2}\sum_{x}|p(x)-q(x)| &= \frac{1}{2}\sum_{x}|-\{q(x)-p(x)\}| \\ &= \frac{1}{2}\sum_{x}|q(x)-p(x)| \\ &= d_{TV}(q,p). \end{aligned}$$

This implies that $d_{TV}(p,q)$ is symmetric.

(iv) $$\begin{aligned} d_{TV}(p,q) &= \frac{1}{2}\sum_{x}|p(x)-q(x)| \\ &= \frac{1}{2}\sum_{x}|p(x)-r(x)+r(x)-q(x)| \\ &\leq \frac{1}{2}\sum_{x}\{|p(x)-r(x)|+|r(x)-q(x)|\} \\ &= \frac{1}{2}\sum_{x}|p(x)-r(x)| + \frac{1}{2}\sum_{x}|r(x)-q(x)| \\ &= d_{TV}(p,r) + d_{TV}(r,q), \end{aligned}$$

adopting the triangle inequality for absolute value.

To find an upper bound for d_{TV}, we have

$$\begin{aligned} d_{TV}(p,q) &= \frac{1}{2}\sum_x |p(x)-q(x)| \\ &\leq \frac{1}{2}\sum_x \{|p(x)|+|q(x)|\} \\ &= \frac{1}{2}\left\{\sum_x p(x)+\sum_x q(x)\right\} \\ &= \frac{1}{2}\{1+1\} \\ &= 1. \end{aligned}$$

As a result, $0 \leq d_{TV}(p,q) \leq 1$. □

3.1.3 Kullback-Leibler divergence

The Kullback-Leibler divergence from $q(x)$ to $p(x)$ is defined as

$$d_{KL}(p;q) = \sum_{x\in X} p(x)\log\left\{\frac{p(x)}{q(x)}\right\}.$$

The KL divergence is not a metric because it is not symmetric and does not satisfy the triangle inequality but it has some nice properties:

(i) $$\begin{aligned} d_{KL}(p;q) &= \sum_x p(x)\log\left\{\frac{p(x)}{q(x)}\right\} \\ &= -\sum_x p(x)\log\left\{\frac{q(x)}{p(x)}\right\} \\ &= E_p\left[-\log\left\{\frac{q(x)}{p(x)}\right\}\right] \\ &\geq -\log\left\{E_p\left[\frac{q(x)}{p(x)}\right]\right\}, \quad \text{by } \textit{Jensen's inequality}^{1} \text{ since } -\log x \text{ is } \textit{convex} \\ &= -\log\left\{\sum_x p(x)\frac{q(x)}{p(x)}\right\} \\ &= -\log 1 \\ &= 0. \end{aligned}$$

Hence the KL divergence is always non-negative.

(ii) $d_{KL}(p;q) = 0 \iff \sum_x p(x) \log\left\{\frac{p(x)}{q(x)}\right\} = 0$

$$\iff E_p\left[-\log\left\{\frac{q(x)}{p(x)}\right\}\right] = 0,$$

which is possible if $q/p = c$ (through the *Jensen's inequality*[1]). However, the only way the distance can be zero is when $c = 1$, that is $p = q$.

Therefore $d_{KL}(p,q) = 0 \iff p = q$.

3.1.4 Jensen-Shannon distance

The Jensen-Shannon distance is defined using the Kullback-Leibler divergence, with some useful differences: it is a metric with an upper bound of $\sqrt{\log 2}$. It is defined by

$$d^2_{JS}(p,q) \quad = \tfrac{1}{2} d_{KL}\left(p; \tfrac{p+q}{2}\right) + \tfrac{1}{2} d_{KL}\left(q; \tfrac{p+q}{2}\right).$$

The squared Jensen-Shannon distance, known as the Jensen-Shannon divergence, is not a metric.

Lemma 3.3. *d_{JS} is a metric bounded above by $\sqrt{\log 2}$.*

Proof. (i) $d_{JS}(p,q) \geq 0$ holds trivially since $d_{KL}(p;q) \geq 0$.

(ii) $d_{JS}(p,q) = 0 \iff p = q$. This follows from the identity property of KL divergence.

(iii) $d_{JS}(p,q)$ is symmetric by definition.

(iv) The triangle inequality has been proven rigorously in Endres & Schindelin (2003). We present a sketch of this proof:

Define a function $L(p,q) = p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}$.

Since $\sqrt{L(p,q)} \leq \sqrt{L(p,r)} + \sqrt{L(r,q)}$ (see Endres & Schindelin (2003) for a proof),

where r is a pmf defined over X, it follows that

$$L(p,q) \leq \left\{\sqrt{L(p,r)} + \sqrt{L(r,q)}\right\}^2.$$

[1] *Jensen's inequality* If f is a *convex* function and X is a random variable, then $E[f(X)] \geq f(E[X])$, with equality when f is a straight line or X is any constant.

Taking sums over x gives

$$\sum_x L\big(p(x), q(x)\big) \leq \sum_x \left\{ \sqrt{L\big(p(x), r(x)\big)} + \sqrt{L\big(r(x), q(x)\big)} \right\}^2 .$$

Therefore

$$\begin{aligned} d_{JS}(p,q) &= \left[\frac{1}{2} \sum_x L\big(p(x), q(x)\big) \right]^{1/2} \\ &\leq \left[\frac{1}{2} \sum_x \left(\sqrt{L\big(p(x), r(x)\big)} + \sqrt{L\big(r(x), q(x)\big)} \right)^2 \right]^{1/2} \\ &\leq \left[\frac{1}{2} \sum_x L\big(p(x), r(x)\big) \right]^{1/2} + \left[\frac{1}{2} \sum_x L\big(r(x), q(x)\big) \right]^{1/2} \\ &= d_{JS}(p,r) + d_{JS}(r,q), \end{aligned}$$

applying the Minkowski's inequality.

An upper bound for $d_{JS}(p,q)$ (Endres & Schindelin, 2003) is given by

$$\begin{aligned} d_{JS}^2(p,q) &= \frac{1}{2} \sum_{x \in X} \left(p(x) \log \left\{ \frac{2p(x)}{p(x)+q(x)} \right\} + q(x) \log \left\{ \frac{2q(x)}{p(x)+q(x)} \right\} \right) \\ &= \frac{1}{2} \sum_x \left[p(x) \left(\log 2p(x) - \log \left\{ p(x)+q(x) \right\} \right) + q(x) \left(\log 2q(x) - \log \left\{ p(x)+q(x) \right\} \right) \right] \\ &= \frac{1}{2} \sum_x \left(\left\{ p(x)+q(x) \right\} \log 2 + p(x) \log \left\{ \frac{p(x)}{p(x)+q(x)} \right\} + q(x) \log \left\{ \frac{q(x)}{p(x)+q(x)} \right\} \right) \\ &\leq \frac{1}{2} \sum_x \left(\left\{ p(x)+q(x) \right\} \log 2 + p(x) \left\{ 0 \right\} + q(x) \left\{ 0 \right\} \right) \\ &= \frac{1}{2} \left(\sum_x p(x) + \sum_x q(x) \right) \log 2 \\ &= \frac{1}{2} \left(1 + 1 \right) \log 2 \\ &= \log 2, \end{aligned}$$

so $d_{JS} \leq \sqrt{\log 2}$. □

All the results presented on metric properties of probability distances were stated, sometimes without proofs in Gibbs & Su (2002).

3.1.5 f-divergence

The f-divergence of $p(x)$ from $q(x)$ is defined as

$$d_f(p,q) \;=\; \sum_{x\in X} q(x) f\left\{\frac{p(x)}{q(x)}\right\},$$

where f is any convex function such that $f(1) = 0$. First introduced by Csiszár (1963) and Ali & Silvey (1966), the f-divergences are also known as Csiszár's divergences. Several distances between probability distributions are instances of f-divergence, depending on a particular choice of f. Table 3.2 shows some divergences and their corresponding functions $f(y)$ (Sason & Verdú, 2016).

Divergence	$f(y)$
Squared Hellinger distance d_H^2	$(\sqrt{y}-1)^2/2$
Total variation distance d_{TV}	$\lvert y-1\rvert/2$
Kullback-Leibler divergence d_{KL}	$y\log y$
Jensen-Shannon divergence d_{JS}^2	$\left(y\log y-(1+y)\log\left\{\frac{1+y}{2}\right\}\right)/2$

Table 3.2: Some f-divergences and their corresponding functions.

3.2 Riemannian geometry

We now consider subsets $S \subset \mathcal{D}(X)$ which are Riemannian manifolds. We are concerned with a family S of probability distributions on X such that

$$S = \left\{p\left(x|\boldsymbol{\theta}\right) : \boldsymbol{\theta} = \left(\theta^1,\ldots,\theta^m\right) \in \Theta \subset \mathbb{R}^m\right\} \tag{3.2}$$

and the function $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}$ is injective, that is, the parameters are identifiable. Hence, Θ is called a parameter space and the set S, an m-dimensional *statistical* or *parametric model*.

The parameter space Θ is assumed to be an open subset of $\mathbb{R}^m$ and for each $x \in X$, the mapping $p(x|\cdot) : \Theta \to \mathbb{R}$ is C^∞, that is, it has derivatives of all orders everywhere in its domain. Therefore, differentiation is carried out freely with respect to the parameters and it makes sense to define expressions such as $\partial_i p(x|\boldsymbol{\theta})$ and $\partial_i\partial_j p(x|\boldsymbol{\theta})$. Furthermore,

we assume summation and differentiation can be interchanged freely, for example,

$$\sum_x \partial_i p(x|\boldsymbol{\theta})dx = \partial_i \sum_x p(x|\boldsymbol{\theta})dx = 0.$$

Consider the binomial distribution as an example of a statistical model, with density

$$p(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $X = \{0, 1, 2, \ldots, n\}$ and $\Theta = \{(n,p) : n \in \mathbb{N}, p \in [0,1]\}$, with $m = 2$ and $(\theta^1, \theta^2) = (n,p)$.

3.2.1 Fundamentals of information geometry

In this section, we introduce some fundamental concepts of information geometry adopting some notation and terminologies from Amari & Nagaoka (2000). Most of the techniques in information geometry involve characterising the properties of manifolds in statistical settings. A manifold S can be thought of as a set with local coordinates in $\mathbb{R}^m$: every point p in S is contained in some open subset $U \subset S$, and there is a homeormorphism $\varphi : U \to \mathbb{R}^m$ usually called a *chart*. We write

$$\varphi(q) = (\theta^1(q), \ldots, \theta^m(q)) = (\theta^i(q))$$

for $q \in U$ so that $(\theta^1, \ldots, \theta^m)$ are local coordinates on S near p. On a differentiable manifold where two charts φ and ψ overlap, the chart transformation $\psi \circ \varphi^{-1}$ is a diffeomorphism between subsets of $\mathbb{R}^m$. We refer the reader to Bröcker & Jänich (1982) for a detailed discussion on manifolds.

We introduce some common notation in Table 3.3 to be used through out this chapter. Also we shall adopt the *Einstein summation convention* where an index appearing twice in a product is summed, for example, $u^i \partial_i$ is an abbreviation for $\sum_{i=1}^m u^i \partial_i$.

Differentiable functions

A function $f : S \to \mathbb{R}$ defined on a manifold S can be rewritten in terms of the local coordinate system φ for S: $\hat{f} = f \circ \varphi^{-1}$ is a real-valued function defined on the domain $\varphi(U) \subseteq \mathbb{R}^m$. Suppose $\hat{f}$ is partially differentiable at each point of its domain, thus $\partial_i \hat{f}(\theta^1, \ldots, \theta^m)$ is also a function on $\varphi(U)$. Then by back transforming the domain, the

Notation	Definition
$\gamma^i(t)$	$\theta^i(\gamma(t))$
∂_i	$\frac{\partial}{\partial\theta^i}$
$g_{ij,k}$	$\frac{\partial g_{ij}}{\partial\theta^k}$

Table 3.3: Some common adopted notations

partial derivative of f is defined by

$$\partial_i f \stackrel{\text{def}}{=} (\partial_i \hat{f}) \circ \varphi : U \to \mathbb{R}.$$

If $f \circ \varphi^{-1}$ is C^∞ for all charts, then f is called a C^∞ *function* on S. The partial derivatives of f including higher-order are also C^∞ functions. Such class of functions are closed under addition, multiplication as well as scalar multiplication.

Tangent space

Let $\gamma : I \to S$ be an injective function from some interval $I \subset \mathbb{R}$ to S. For $t \in I$, a point $\gamma(t)$ can be written using coordinates $\hat{\gamma}(t) = (\gamma^1(t), \ldots, \gamma^m(t)) \in \mathbb{R}^m$. When $\hat{\gamma}(t)$ is C^∞, γ is called a C^∞ *curve* on S independent of the choice of coordinate system. Consider a real-valued C^∞ function $f : S \to \mathbb{R}$ and define $\frac{d}{dt} f(\gamma(t))$ in terms of coordinates:

$$\frac{d}{dt} f(\gamma(t)) = (\partial_i f)_{\gamma(t)} \frac{d\gamma^i(t)}{dt}.$$

This is called the directional derivative of f along the curve γ. Define the tangent vector of γ at $p = \gamma(a)$ to be the linear operator that maps f to $\frac{d}{dt} f(\gamma(a))$, that is

$$\frac{d\gamma(t)}{dt}\Big|_{t=a} = \left(\frac{d\gamma}{dt}\right)_p = \frac{d\gamma^i(t)}{dt}\Big|_{t=a} (\partial_i)_p \,, \tag{3.3}$$

where $(\partial_i)_p$ is the operator that maps $f \mapsto (\partial_i f)_p$. The operator $(\partial_i)_p$ is the tangent vector at point p of the i-th coordinate curve: a curve obtained by changing θ^i alone and fixing values of all θ^j $(j \neq i)$. By considering all the curves that pass through p, the set of their

corresponding tangent vectors denoted as $T_p(S)$ (or simply T_p) is given from (3.3) as

$$T_p(S) = \left\{ c^i \left(\partial_i\right)_p : (c^1, \ldots, c^m) \in \mathbb{R}^m \right\}.$$

The linear space T_p is called the *tangent space* of S at point p and its elements are called the *tangent vectors*. The dimension of T_p is the same as the dimension of S, since $\left\{ \left(\partial_i\right)_p : i = 1, \ldots, m \right\}$ are linearly independent operators.

A mapping $V : p \mapsto V_p$ from each point p in S to a tangent vector $V_p \in T_p$ is called a *vector field*. Given a vector field V, the tangent vector $V_p = V_p^i(\partial_i)_p$ is determined uniquely by m real numbers $(V_p^1, \ldots, V_p^m)$ for each point p. Hence, we can write $V = V^i\partial_i$ and the m functions $V^i : p \mapsto V_p^i$ on S are refered to as the *components* of V with respect to (θ^i). We consider vector fields that are C^∞, in the sense that the components of the vector field are C^∞ with respect to local coordinates, and this definition invariant under coordinate change. We denote by $\mathcal{T}(S)$ the set of all C^∞ vector fields.

Affine connections

For any two different points p and q in a manifold S, the corresponding tangent spaces $T_p(S)$ and $T_q(S)$ are not directly related in anyway. To establish a relationship between the two spaces, the structure of the manifold S can be augmented through an affine connection. Suppose a point p' in S is very close to p such that the difference in coordinates $\mathrm{d}\theta^i = \theta^i(p') - \theta^i(p)$ between p and p' is infinitesimally small. An affine connection is introduced on the manifold S through a linear mapping $\mathit{\Pi}_{p,p'} : T_p \to T_{p'}$ defined as

$$\mathit{\Pi}_{p,p'}\big((\partial_j)_p\big) = (\partial_j)_{p'} - \mathrm{d}\theta^i \big(\mathit{\Gamma}_{ij}^k\big)_p (\partial_k)_{p'}, \tag{3.4}$$

and the m^3 functions $\mathit{\Gamma}_{ij}^k : p \to \big(\mathit{\Gamma}_{ij}^k\big)_p$ are all C^∞. The $\mathit{\Gamma}_{ij}^k$ functions are called the *connection coefficients* of the affine connection with respect to the coordinate system (θ^i). They must satisfy some conditions, for example, the symmetry relation $\mathit{\Gamma}_{ij}^k = \mathit{\Gamma}_{ji}^k$. The concept of affine connection is invariant under a change of coordinate system.

For close points p and p', the affine connection establishes a relationship between T_p and $T_{p'}$. For arbitrary distant points p and q, a relationship can still be established between T_p and T_q by connecting a sequence of relationships between intermediate close points. However, this depends on the curve γ connecting p and q.

Suppose $\gamma : [a, b] \to S$ is a curve connecting points p and q in S, such that $\gamma(a) = p$ and $\gamma(b) = q$. Given a vector field along γ, that is $V : t \mapsto V(t)$ that maps each point

$\gamma(t)$ to a tangent vector $V(t)$, V is *parallel* along γ if

$$V(t+\mathrm{d}t) = \Pi_{\gamma(t),\gamma(t+\mathrm{d}t)}\big(V(t)\big)$$

for all $t \in [a,b]$ with infinitesimal $\mathrm{d}t$. This can be written with respect to a coordinate system (θ^i), hence $V(t) = V^i(t)(\partial_i)_{\gamma(t)}$ and $V(t+\mathrm{d}t) = V^i(t+\mathrm{d}t)(\partial_i)_{\gamma(t+\mathrm{d}t)}$. Using (3.4), we get

$$\frac{\mathrm{d}V^k(t)}{\mathrm{d}t} + \frac{d\gamma^i(t)}{dt} V^j(t)\big(\Gamma^k_{ij}\big)_{\gamma(t)} = 0, \tag{3.5}$$

where $\gamma^i \overset{\text{def}}{=} \theta^i \circ \gamma$ and $\frac{\mathrm{d}V^k(t)}{\mathrm{d}t} = \frac{V^k(t+\mathrm{d}t)-V^k(t)}{\mathrm{d}t}$. Given an initial condition, this differential equation has a unique solution which is a parallel vector field V along γ. Therefore, given $\boldsymbol{u} \in T_{\gamma(a)} = T_p$, there exist a unique V such that $V(a) = \boldsymbol{u}$. If $\Pi_\gamma(\boldsymbol{u})$ denote the vector $V(b) \in T_{\gamma(b)} = T_q$, then Π_γ is a linear isomorphism between T_p and T_q, called the *parallel translation along* γ.

Moreover, both $V(t)$ and $V(t+\mathrm{d}t)$ lie in different tangent spaces, thus the derivative $\frac{\mathrm{d}V(t)}{\mathrm{d}t} = \lim_{\mathrm{d}t\to 0} \frac{V(t+\mathrm{d}t)-V(t)}{\mathrm{d}t}$ doesn't make sense. However, an affine connection on S allows the derivative $\lim_{\mathrm{d}t\to 0} \frac{V_t(t+\mathrm{d}t)-V(t)}{\mathrm{d}t}$ to be considered within $T_{\gamma(t)}$, where $V_t(t+\mathrm{d}t) = \Pi_{\gamma(t+\mathrm{d}t),\gamma(t)}\big(V(t+\mathrm{d}t)\big)$ is the parallel translation of $V(t+\mathrm{d}t) \in T_{\gamma(t+\mathrm{d}t)}$ to the space $T_{\gamma(t)}$ along γ. This is refered to as the *covariant derivative* of $V(t)$ and denoted as $\frac{\delta V}{\mathrm{d}t}$:

$$\frac{\delta V}{\mathrm{d}t} = \frac{\Pi_{\gamma(t+\mathrm{d}t),\gamma(t)}\big(V(t+\mathrm{d}t)\big) - V(t)}{\mathrm{d}t} = \left\{ \frac{\mathrm{d}V^k(t)}{\mathrm{d}t} + \frac{d\gamma^i(t)}{dt} V^j(t)\big(\Gamma^k_{ij}\big)_{\gamma(t)} \right\} (\partial_k)_{\gamma(t)}. \tag{3.6}$$

This implies that (3.5) can be written as $\frac{\delta V}{\mathrm{d}t} = 0$. The covariant derivative of a vector field $V = V^i\partial_i \in \mathcal{T}(S)$ along a curve whose tangent vector at a point p is $\boldsymbol{u} = u^i(\partial_i)_p \in T_p$ is given as

$$\nabla_u V = u^i \left\{ \big(\partial_i V^k\big)_p + V^j_p \big(\Gamma^k_{ij}\big)_p \right\} (\partial_k)_p \in T_p. \tag{3.7}$$

Therefore, if $V_\gamma : t \mapsto V_{\gamma(t)}$ for a curve γ, we see from (3.6) and (3.7) that

$$\frac{\delta V_\gamma}{\mathrm{d}t} = \nabla_{\frac{d\gamma}{dt}} V,$$

Given $U = U^i\partial_i$ and $V = V^i\partial_i$ both in $\mathcal{T}(S)$, the covariant derivative of V with respect

to U is defined as

$$\nabla_U V = U^i \left\{ \partial_i V^k + V^j \Gamma_{ij}^k \right\} \partial_k. \tag{3.8}$$

In particular, if $U = \partial_i$ and $V = \partial_j$, we get the component expression of the covariant derivative, that is

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k. \tag{3.9}$$

The operator $\nabla : \mathcal{T}(S) \times \mathcal{T}(S) \to \mathcal{T}(S)$ maps (U, V) to $\nabla_U V$ and for $U, V, W \in \mathcal{T}(S)$ and a C^∞ function f, it satisfies the following:

(i) $\nabla_{U+V} W = \nabla_U W + \nabla_V W$,

(ii) $\nabla_U (V + W) = \nabla_U V + \nabla_U W$,

(iii) $\nabla_{fU} V = f \nabla_U V$,

(iv) $\nabla_U (fV) = f \nabla_U V + (Uf) V$,

where Uf denotes the C^∞ function $p \mapsto U_p f$. An *affine connection* on S is defined to be the mapping ∇ which satisfies (i)-(iv) (Amari & Nagaoka, 2000). In fact, both (3.8) and (3.9) can be derived from conditions (i)-(iv), which determines the connection coefficients Γ_{ij}^k of the affine connection ∇.

3.2.2 Riemannian metrics

For each point p in a manifold S, assume an inner product $\langle \ , \ \rangle_p : T_p(S) \times T_p(S) \to \mathbb{R}$ is defined on the tangent space $T_p(S)$, such that for any tangent vectors $\boldsymbol{u}, \boldsymbol{v} \in T_p(S)$, the following are satisfied:

(i) Linearity: $\langle a\boldsymbol{u} + b\boldsymbol{v}, \boldsymbol{w} \rangle_p = a\langle \boldsymbol{u}, \boldsymbol{w} \rangle_p + b\langle \boldsymbol{v}, \boldsymbol{w} \rangle_p$ $(\forall a, b \in \mathbb{R})$,

(ii) Symmetry: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_p = \langle \boldsymbol{v}, \boldsymbol{u} \rangle_p$,

(iii) Positive-definiteness: If $\boldsymbol{u} \neq \boldsymbol{0}$ then $\langle \boldsymbol{u}, \boldsymbol{u} \rangle_p > 0$.

The function $g : p \to \langle \ , \ \rangle_p$ that maps every point p in S to its inner product is called a *Riemannian metric* on S. This metric is not defined by the manifold structure and thus, infinitely many Riemannian metrics can be defined on S and (S, g) is called a *Riemannian*

manifold. Furthermore, if the manifold S is C^∞ and the function $p \mapsto \langle U, V \rangle_p$ is C^∞ for all C^∞ vector fields $U, V \in \mathcal{T}(S)$, then g is called a C^∞ Riemannian metric.

With respect to a coordinate system (θ^i) for S, the Riemannian metric g determines the components $\{g_{ij}; i, j = 1, \ldots, m\}$ such that for each point p in S, $g_{ij}(p) = \langle (\partial_i)_p, (\partial_j)_p \rangle_p$. The inner product of two tangent vectors $\boldsymbol{u} = u^i(\partial_i)_p$ and $\boldsymbol{v} = v^i(\partial_i)_p$ can be written as

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_p = g_{ij}(p) u^i v^j,$$

and the length $\|\boldsymbol{u}\|$ of the tangent vector $\boldsymbol{u}$ is given by

$$\|\boldsymbol{u}\|^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_p = g_{ij}(p) u^i u^j. \tag{3.10}$$

The $m \times m$ matrix $G(p) = (g_{ij}(p))$ formed by the components $g_{ij}(p)$ is symmetric positive definite by definition. However, given a coordinate system (θ^i) for S and m^2 C^∞ functions g_{ij}, if $G(p) = (g_{ij}(p))$ is symmetric positive definite for every p in S, then the corresponding Riemannian metric g with components g_{ij} is determined in a unique way. The relationship between g_{ij} and another components $\hat{g}_{kl}$ with respect to a different coordinate system (ρ^k) is given by the coordinate transformations

$$\hat{g}_{kl} = g_{ij} \left(\frac{\partial \theta^i}{\partial \rho^k} \right) \left(\frac{\partial \theta^j}{\partial \rho^l} \right) \quad \text{and} \quad g_{ij} = \hat{g}_{kl} \left(\frac{\partial \rho^k}{\partial \theta^i} \right) \left(\frac{\partial \rho^l}{\partial \theta^j} \right). \tag{3.11}$$

3.2.3 Riemannian connection

An affine connection ∇ on a Riemannian manifold (S, g) is a *metric connection* with respect to g if for all vector fields $U, V, W \in \mathcal{T}(S)$

$$W\langle U, V \rangle = \langle \nabla_W U, V \rangle + \langle U, \nabla_W V \rangle.$$

This may be rewritten using the coordinate expressions of g and ∇ as

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i},$$

for all i, j, k where $\Gamma_{ij,k} = \langle \nabla_{\partial_i} \partial_j, \partial_k \rangle = \Gamma^h_{ij} g_{hk}$.

A connection ∇ is symmetric if $\nabla_U V - \nabla_V U = [U, V]$, where $[U, V]$ is the Lie bracket of vector fields U and V. For a given g, there exists a unique connection that is both metric and symmetric, called the *Levi-Civita connection* or the *Riemannian connection*

with respect to g. Thus, using the symmetry $\Gamma_{ij,k} = \Gamma_{ji,k}$, we see that

$$\Gamma_{ij,k} = \frac{1}{2}\big(g_{jk,i} + g_{ki,j} - g_{ij,k}\big).$$

3.2.4 Fisher information metric

For a discrete set X, the Fisher information matrix of S at a given point $\boldsymbol{\theta} \in \Theta$ is the $m \times m$ matrix $G(\boldsymbol{\theta}) = (g_{ij}(\boldsymbol{\theta}))$, $i, j = 1, \ldots, m$ such that

$$g_{ij}(\boldsymbol{\theta}) = \sum_{x \in X} p(x|\boldsymbol{\theta}) \partial_i \log p(x|\boldsymbol{\theta}) \partial_j \log p(x|\boldsymbol{\theta}). \tag{3.12}$$

The fuction $g_{ij} : \Theta \to \mathbb{R}$ is assumed to be C^∞ and for all $\boldsymbol{\theta}$ and for all i, j, g_{ij} is finite. Since

$$\begin{aligned} E[\partial_i \partial_j l_{\boldsymbol{\theta}}] &= \sum_x p \partial_i \left(\frac{1}{p}\partial_j p\right) = \sum_x \left(\partial_i \partial_j p - \frac{1}{p}\partial_i p \partial_j p\right) \\ &= 0 - \sum_x p \left(\frac{1}{p}\partial_i p\right)\left(\frac{1}{p}\partial_j p\right) = -E[\partial_i l_{\boldsymbol{\theta}} \partial_j l_{\boldsymbol{\theta}}], \end{aligned}$$

where $l_{\boldsymbol{\theta}} = \log p(x|\boldsymbol{\theta})$, the elements g_{ij} may be written as

$$g_{ij}(\boldsymbol{\theta}) = E\left[\partial_i l_{\boldsymbol{\theta}} \partial_j l_{\boldsymbol{\theta}}\right] = -E\left[\partial_i \partial_j l_{\boldsymbol{\theta}}\right].$$

For any vector $\boldsymbol{z} = (z^1, \ldots, z^m)^t$,

$$\begin{aligned} \boldsymbol{z}^t G(\boldsymbol{\theta}) \boldsymbol{z} = z^i z^j g_{ij}(\boldsymbol{\theta}) &= z^i z^j E\left[\partial_i l_{\boldsymbol{\theta}} \partial_j l_{\boldsymbol{\theta}}\right] \\ &= E\left[z^i z^j \partial_i l_{\boldsymbol{\theta}} \partial_j l_{\boldsymbol{\theta}}\right] \\ &= E\left[z^i \partial_i l_{\boldsymbol{\theta}} z^j \partial_j l_{\boldsymbol{\theta}}\right] \\ &= E\left[(z^i \partial_i l_{\boldsymbol{\theta}})^2\right] \geq 0 \end{aligned}$$

which implies that G is positive semidefinite. It is further assumed that G is positive definite which from the above equation requires the elements of $\{\partial_1 l_{\boldsymbol{\theta}}, \ldots, \partial_m l_{\boldsymbol{\theta}}\}$ to be linearly independent as functions. Also, the matrix $G(\boldsymbol{\theta})$ is symmetric since $g_{ij}(\boldsymbol{\theta}) = g_{ji}(\boldsymbol{\theta})$. Therefore, defining the inner product of the natural basis of (θ^i) as $\langle \partial_i, \partial_j \rangle = g_{ij}$ determines uniquely a Riemannian metric $g = \langle\ ,\ \rangle$ on S. This metric is called the *Fisher information metric* or the *Fisher information* (when $m = 1$). It can be seen that the Fisher information metric (3.12) satisfies the coordinate transformations (3.11), hence it

is invariant under a change of coordinate system.

3.2.5 Geodesics

Let $\gamma : [a, b] \to S$ be a smooth curve in a Riemannian manifold S. The *length* of γ is defined using (3.10) as

$$L(\gamma) = \int_a^b \left\| \frac{d\gamma(t)}{dt} \right\| dt,$$

and the energy of γ is given as

$$E(\gamma) = \frac{1}{2} \int_a^b \left\| \frac{d\gamma(t)}{dt} \right\|^2 dt.$$

The energy functional $E(\gamma)$ measures the total kinetic energy of an object traveling along γ with speed stated by $\frac{d\gamma}{dt}$. Both L and E can be written in local coordinates $(\theta^i\{\gamma(t)\}, \ldots, \theta^m\{\gamma(t)\})$ such that

$$L(\gamma) = \int_a^b \sqrt{g_{ij}(\theta\{\gamma(t)\}) \frac{d\theta^i}{dt} \frac{d\theta^j}{dt}} dt \quad \text{and} \quad E(\gamma) = \frac{1}{2} \int_a^b g_{ij}(\theta\{\gamma(t)\}) \frac{d\theta^i}{dt} \frac{d\theta^j}{dt} dt. \quad (3.13)$$

Let γ_p^q be the set of piecewise smooth paths joining arbitrary points p and q in a manifold S. The distance between p and q is the minimum of the lengths of all $\gamma \in \gamma_p^q$ such that $\gamma(a) = p$ and $\gamma(b) = q$, that is

$$d(p, q) = \min_{\gamma_p^q} \{L(\gamma)\}. \quad (3.14)$$

The length $L(\gamma)$ is invariant under parameter changes. If $\phi : [\alpha, \beta] \to [a, b]$ is a parameter change, then $L(\gamma \circ \phi) = L(\gamma)$. Also, using the Hölder's inequality, we see that

$$\begin{aligned} L(\gamma) &= \int_a^b 1 \left\| \frac{d\gamma}{dt} \right\| dt \\ &\leq \sqrt{\int_a^b 1^2 dt} \sqrt{\int_a^b \left\| \frac{d\gamma}{dt} \right\|^2 dt} \\ &= \sqrt{(b-a) \int_a^b \left\| \frac{d\gamma}{dt} \right\|^2 dt} \\ &= \sqrt{2(b-a)E(\gamma)}, \end{aligned}$$

with equality iff $\left\|\frac{d\gamma}{dt}\right\|$ is a constant. This relationship between $L(\gamma)$ and $E(\gamma)$ establishes the following result.

Proposition 3.1. *A constant speed curve $\sigma \in \gamma_p^q$ (i.e $\left\|\frac{d\sigma}{dt}\right\|$ is a constant) that minimizes $L(\gamma)$, over all $\gamma \in \gamma_p^q$ also minimizes $E(\gamma)$, over all $\gamma \in \gamma_p^q$. The converse is also true* (Peterson, 2006).

Proof. Let $\sigma \in \gamma_p^q$ be a constant speed curve that minimizes L and $\gamma \in \gamma_p^q$. Thus

$$\begin{aligned} E(\sigma) &= \frac{1}{2(b-a)}\{L(\sigma)\}^2 \\ &\leq \frac{1}{2(b-a)}\{L(\gamma)\}^2 \\ &\leq E(\gamma), \end{aligned}$$

so σ also minimizes $E(\gamma)$.

Conversely, let $\sigma \in \gamma_p^q$ be a minimizer of E and $\gamma \in \gamma_p^q$. If the speed of γ is not constant, γ can be reparametrised using arc length to a smooth curve $\bar{\gamma}$ that has constant speed almost everywhere without changing $L(\gamma)$. Then

$$\begin{aligned} L(\sigma) &\leq \sqrt{2(b-a)E(\sigma)} \\ &\leq \sqrt{2(b-a)E(\bar{\gamma})} \\ &= L(\bar{\gamma}) \\ &= L(\gamma). \end{aligned}$$

□

We have seen that to find the minimising curve of the length functional $L(\gamma)$, it suffices to minimise the energy functional. In this case, it is enough to find the Euler-Lagrange equations for the energy functional E, whose solution is a critical point of E. This argument is presented in the following result.

Proposition 3.2. *The Euler-Lagrange equations for the energy E are*

$$\frac{d^2\theta^k(t)}{dt^2} + \Gamma_{ij}^k\big(\theta(t)\big)\frac{d\theta^i(t)}{dt}\frac{d\theta^j(t)}{dt} = 0, \quad k = 1, \ldots, m \tag{3.15}$$

with

$$\Gamma_{ij}^k = \frac{1}{2}g^{kl}\big(g_{il,j} + g_{jl,i} - g_{ij,l}\big), \tag{3.16}$$

where g^{ij} is the inverse of g_{ij} i.e. $g^{il}g_{lj} = \delta_{ij}$. The coefficients Γ_{ij}^{k} are called Christoffel symbols (Jost, 2017).

Proof. A function

$$I(x) = \int_a^b f\big(t, x(t), x'(t)\big)dt,$$

where $x'(t) = \frac{dx(t)}{dt}$ has the following Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial f}{\partial x'^k} - \frac{\partial f}{\partial x^k} = 0, \quad k = 1, \dots, m.$$

In this case, since

$$E(\gamma) = \frac{1}{2}\int_a^b g_{ij}\big(\theta(t)\big)\frac{d\theta^i(t)}{dt}\frac{d\theta^j(t)}{dt}dt,$$

we get

$$\frac{d}{dt}\left(g_{kj}\frac{d\theta^j(t)}{dt} + g_{ik}\frac{d\theta^i(t)}{dt}\right) - g_{ij,k}\frac{d\theta^i(t)}{dt}\frac{d\theta^j(t)}{dt} = 0,$$

and hence

$$g_{kj}\frac{d^2\theta^j(t)}{dt^2} + g_{ik}\frac{d^2\theta^i(t)}{dt^2} + g_{kj,l}\frac{d\theta^l(t)}{dt}\frac{d\theta^j(t)}{dt} + g_{ik,l}\frac{d\theta^l(t)}{dt}\frac{d\theta^i(t)}{dt} - g_{ij,k}\frac{d\theta^i(t)}{dt}\frac{d\theta^j(t)}{dt} = 0.$$

Using the symmetry of g and renaming some indices, we have

$$\begin{aligned}
&2g_{ln}\frac{d^2\theta^n(t)}{dt^2} + \big(g_{jl,i} + g_{il,j} - g_{ij,l}\big)\frac{d\theta^i(t)}{dt}\frac{d\theta^j(t)}{dt} \\
&= g^{kl}g_{ln}\frac{d^2\theta^n}{dt^2} + \frac{1}{2}g^{kl}\big(g_{jl,i} + g_{il,j} - g_{ij,l}\big)\frac{d\theta^i}{dt}\frac{d\theta^j}{dt} \\
&= 0.
\end{aligned}$$

Since $g^{kl}g_{ln} = \delta_{kn}$, we have $g^{kl}g_{ln}\frac{d^2\theta^n}{dt^2} = \frac{d^2\theta^k}{dt^2}$ and hence

$$\frac{d^2\theta^k}{dt^2} + \frac{1}{2}g^{kl}\big(g_{jl,i} + g_{il,j} - g_{ij,l}\big)\frac{d\theta^i}{dt}\frac{d\theta^j}{dt} = 0,$$

the required result. □

A curve γ that satisfies (3.15) is called a *geodesic*, which is known to correspond locally with the curve of minimum length joining two points (3.14). Equation (3.15) is refered to as the *geodesic equation*.

Let us consider the following simple example of a univariate normal distribution with pdf given by

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2},$$

as adapted from Costa *et al.* (2015). A natural parameter space for this family of probability distributions is given as

$$\Theta = \left\{(\mu,\sigma) \in \mathbb{R}^2 : \sigma > 0\right\},$$

so that each point in the space represents a univariate normal pdf. Given any two points $P = (\mu_1, \sigma_1)$ and $Q = (\mu_2, \sigma_2)$ in the half-plane Θ, a measure of distance between them is the Riemannian metric distance. The Fisher information matrix $G(\boldsymbol{\theta}) = \big(g_{ij}(\boldsymbol{\theta})\big)$ provides a proper distance measure of the amount of information of the unknown parameter $\boldsymbol{\theta}$. In this case, $\boldsymbol{\theta} = (\theta^1, \theta^2) = (\mu, \sigma)$, and the Fisher information matrix is given by

$$G = \big(g_{ij}\,(\mu,\sigma)\,\big) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

(Costa *et al.*, 2015). To obtain the geodesic equation (3.15), we first deduce the coefficients using (3.16):

$$\Gamma^1_{ij} = \frac{1}{2} g^{11} \left(\partial_i g_{j1} + \partial_j g_{i1} - \partial_1 g_{ij}\right) + \frac{1}{2} g^{12} \left(\partial_i g_{j2} + \partial_j g_{i2} - \partial_2 g_{ij}\right),$$

hence $\Gamma^1_{11} = \Gamma^1_{22} = 0$ and $\Gamma^1_{12} = \Gamma^1_{21} = -\frac{1}{\sigma}$. Similarly $\Gamma^2_{11} = \frac{1}{2\sigma}$, $\Gamma^2_{12} = \Gamma^2_{21} = 0$ and $\Gamma^2_{22} = -\frac{1}{\sigma}$. By substituting in (3.15), the corresponding geodesic equations are given as

$$\begin{aligned} \frac{d^2\mu}{dt^2} - \frac{2}{\sigma}\frac{d\mu}{dt}\frac{d\sigma}{dt} &= 0, \\ \frac{d^2\sigma}{dt^2} + \frac{1}{2\sigma}\left(\frac{d\mu}{dt}\right)^2 - \frac{1}{\sigma}\left(\frac{d\sigma}{dt}\right)^2 &= 0. \end{aligned}$$

The solution to these equations is a geodesic curve that encompases the shortest path

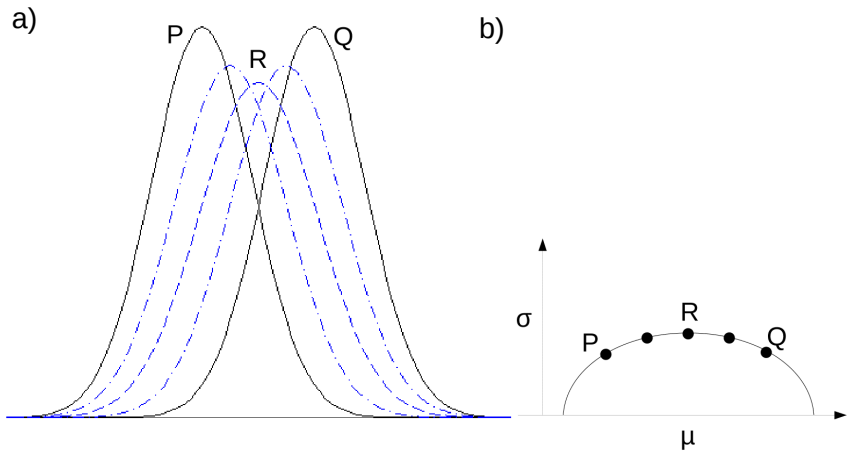


Figure 3.1: a) Univariate normal ditributions P, Q and R. b) Shortest path between P and Q passing through R in the (μ, σ) half-plane. It is a segment of an ellipse in $\mathbb{R}^2$ with focal points on the μ axis. The path joining P and Q is not the Euclidean line segment because σ changes (increases and decreases) along the path.

between P and Q as shown in Figure 3.1b, and is a segment of an ellipse in $\mathbb{R}^2$. Geodesics are ellipses with focal points on the μ axis. Figure 3.1a shows the univariate normal distributions P, Q and R, whose corresponding pdfs are represented by points in the (μ, σ) half plane in Figure 3.1b. The distance between points in the (μ, σ) half-plane representing normal distributions cannot be the usual Euclidean distance. This is evident from the path joining P and Q where σ changes (increases and decreases) along the path (Figure 3.1b).

3.3 Relationships between Fisher information metric and probability metrics

The KL divergence is an essential concept of information theory (Cover & Thomas, 2006). For instance, the KL divergence between two infinitesimally close ditributions is proportional to the Fisher information metric. Suppose $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}$ are two parametrized distributions defined over a discrete set X, with $\delta\boldsymbol{\theta}$ a small change in $\boldsymbol{\theta}$. Since

$$\sum_{x\in X} p_{\boldsymbol{\theta}} \partial_i \log p_{\boldsymbol{\theta}} = \sum_x p_{\boldsymbol{\theta}} \frac{1}{p_{\boldsymbol{\theta}}} \partial_i p_{\boldsymbol{\theta}} = \partial_i \sum_x p_{\boldsymbol{\theta}} = \partial_i(1) = 0 \quad \text{and}$$

$$\begin{aligned}
\sum_{x\in X} p_{\boldsymbol{\theta}}\partial_i\partial_j \log p_{\boldsymbol{\theta}} &= \sum_x p_{\boldsymbol{\theta}}\partial_i\left(\partial_i \log p_{\boldsymbol{\theta}}\right) \\
&= \sum_x p_{\boldsymbol{\theta}}\partial_i\left(\frac{1}{p_{\boldsymbol{\theta}}}\partial_j p_{\boldsymbol{\theta}}\right) \\
&= \sum_x p_{\boldsymbol{\theta}}\left(\frac{1}{p_{\boldsymbol{\theta}}}\partial_i\partial_j p_{\boldsymbol{\theta}} - \frac{1}{p_{\boldsymbol{\theta}}^2}\partial_i p_{\boldsymbol{\theta}}\partial_j p_{\boldsymbol{\theta}}\right) \\
&= \partial_i\partial_j \sum_x p_{\boldsymbol{\theta}} - \sum_x p_{\boldsymbol{\theta}}\partial_i \log p_{\boldsymbol{\theta}}\partial_j \log p_{\boldsymbol{\theta}} \\
&= \partial_i\partial_j(1) - \sum_x p_{\boldsymbol{\theta}}\partial_i \log p_{\boldsymbol{\theta}}\partial_j \log p_{\boldsymbol{\theta}} \\
&= -\sum_x p_{\boldsymbol{\theta}}\partial_i \log p_{\boldsymbol{\theta}}\partial_j \log p_{\boldsymbol{\theta}},
\end{aligned}$$

we see that

$$\begin{aligned}
d_{KL}(p_{\boldsymbol{\theta}}; p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) &= \sum_{x\in X} p_{\boldsymbol{\theta}} \log\left\{\frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}}\right\} \\
&= \sum_x p_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}} - \log p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) \\
&= \sum_x p_{\boldsymbol{\theta}}\left(-\delta\boldsymbol{\theta}\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}} - \frac{1}{2}\delta\boldsymbol{\theta}^T H \delta\boldsymbol{\theta} + \mathcal{O}(\|\delta\boldsymbol{\theta}\|^3)\right) \qquad (3.17) \\
&= \frac{1}{2}\delta\boldsymbol{\theta}^T R \delta\boldsymbol{\theta} + \mathcal{O}(\|\delta\boldsymbol{\theta}\|^3),
\end{aligned}$$

where $H = (h_{ij}) = (\partial_i\partial_j \log p_{\boldsymbol{\theta}})$ is the Hessian matrix and R denotes the Fisher information matrix. Equation (3.17) is as a result of the Taylor series expansion of $\log p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}$ about $\boldsymbol{\theta}$.

The Fisher information metric is also related to the Jensen-Shannon divergence (squared Jensen-Shannon distance) through the length of a curve $\gamma : [a, b] \to S$ on a Riemannian

metric S (3.13). Now consider $d^2_{JS}(p,q)$ where $q(x) = p(x) + \delta p(x)$, for all x:

$$\begin{aligned}
d^2_{JS}(p,q) &= \frac{1}{2}\sum_x p(x)\log\left\{\frac{2p(x)}{2p(x)+\delta p(x)}\right\} + \frac{1}{2}\sum_x \{p(x)+\delta p(x)\}\log\left\{\frac{2p(x)+2\delta p(x)}{2p(x)+\delta p(x)}\right\} \\
&= \frac{1}{2}\sum_x p(x)\log\left\{\left(1+\frac{\delta p(x)}{2p(x)}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\sum_x \{p(x)+\delta p(x)\}\log\left\{\frac{2p(x)+2\delta p(x)}{2p(x)+2\delta p(x)-\delta p(x)}\right\} \\
&= -\frac{1}{2}\sum_x p(x)\log\left\{1+\frac{\delta p(x)}{2p(x)}\right\} - \frac{1}{2}\sum_x \{p(x)+\delta p(x)\}\log\left\{1-\frac{\delta p(x)}{2p(x)+2\delta p(x)}\right\} \\
&= -\frac{1}{2}\sum_x p\left[\frac{\delta p}{2p} - \frac{\delta p^2}{8p^2}\right] - \frac{1}{2}\sum_x p\left[-\frac{\delta p}{2p+2\delta p} - \frac{\delta p^2}{8(p+\delta p)^2}\right] \\
&\quad -\frac{1}{2}\sum_x \delta p\left[-\frac{\delta p}{2p+2\delta p} - \frac{\delta p^2}{8(p+\delta p)^2}\right] \\
&= \sum_x p\frac{\delta p^2}{16p^2} + \sum_x p\frac{\delta p^2}{16p^2} \\
&= \frac{1}{8}\sum_x p\left(\frac{\delta p}{p}\right)^2,
\end{aligned}$$

by applying the Taylor expansion of $f(y) = \log(1+y)$ about $y = 0$, where higher order terms in δp are neglected. Therefore

$$\begin{aligned}
L(\gamma) &= \int_a^b \left(g_{ij}\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\right)^{1/2} dt \\
&= \int_a^b \left(\sum_x p\frac{d\log p}{d\gamma^i}\frac{d\gamma^i}{dt}\frac{d\log p}{d\gamma^j}\frac{d\gamma^j}{dt}\right)^{1/2} dt \\
&= \int_a^b \left(\sum_x p\left(\frac{1}{p}\frac{dp}{dt}\right)^2\right)^{1/2} dt \\
&= \int_a^b \left(\sum_x \frac{1}{p}\left(\frac{dp}{dt}\right)^2\right)^{1/2} dt \\
&= \sqrt{8}\int_a^b d_{JS}dt,
\end{aligned}$$

so $L(\gamma)$ is the length of curve if arc length is d_{JS}.

3.4 Some applications of probability distances

Distances between probability distributions have several statistical applications. For example, the Hellinger distance has been used as a test statistic in hypothesis testing for two populations (Basu *et al.*, 2010). Let $X_1, \ldots, X_{m_1}$ and $Y_1, \ldots, Y_{m_2}$ be two random independent samples from two discrete populations X and Y with common support $\mathcal{X} = \{x_0, x_1, \ldots\}$ and corresponding pmfs p_{θ_1} and p_{θ_2}. A test statistic $d_H(p_{\hat{\theta}_1}, p_{\hat{\theta}_2})^2$ is used in testing the null hypothesis $\theta_1 = \theta_2$. The unknown parameters are calculated via

$$\hat{\theta}_i = \arg\min_{\theta_i} d_H(\boldsymbol{d}_i, p_{\theta_i})^2,$$

with $\boldsymbol{d}_i = \left(\frac{n_i(x_0)}{m_i}, \ldots, \frac{n_i(x_j)}{m_i}, \ldots\right), i = 1, 2$, where $n_i(x)$ denote the number of elements in the ith sample that coincide with $x \in \mathcal{X}$. Thus, $\hat{\theta}_i \in \Theta$ is called the minimum Hellinger distance estimator of θ_i. The Hellinger distance is just one of several probability distances used as test statistics. Salicru *et al.* (1994) consider test statistics from a family of f-divergence. Similarly, Sarkar & Basu (1995) suggest a test statistic based on the Kullback-Leibler (KL) divergence, where the minimum parameter is the maximum likelihood estimator (MLE). This points to an important application of KL divergence to maximum likelihood estimation.

The process of maximizing the likelihood function $p(x|\theta)$ is similar to minimizing the KL divergence between $p(x|\theta)$ and the empirical distribution $\hat{p}(x)$. The empirical distribution function of a sample $(x_1, \ldots, x_n)$ of n independent and identically distributed (i.i.d) observations is a function $\hat{p} : \mathbb{R} \to [0, 1]$ given as $\hat{p}(x) = 1/n \sum_{i=1}^{n} \delta(x - x_i)$, where δ is the delta function. Therefore

$$\begin{aligned} d_{KL}(\hat{p}; p_\theta) &= \sum_x \hat{p}(x) \log \left\{ \frac{\hat{p}(x)}{p(x|\theta)} \right\} \\ &= \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \log p(x|\theta) \qquad (3.18) \\ &\propto -\sum_x \hat{p}(x) \log p(x|\theta) \\ &= -\sum_x \left[\frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \right] \log p(x|\theta) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta), \end{aligned}$$

which is indeed the negative of the log-likelihood function. The first term in (3.18) is dropped because it is independent of the parameter θ. Maximum likelihood estimation seek to identify parameter models or distributions that are close to our observation. In this case, it is reasonable to think of the process as minimizing some measure of distance between the empirical distribution and our observation model.

The total variation distance is a widely used measure of distance between probability distributions. It has been used for bounding rates of convergence for Markov chains (Rosenthal, 1995; Tierney, 1994; Gilks *et al.*, 1996). An important application of this is in Markov Chain Monte Carlo algorithms, where the chain aims to converge to a target distribution (Rosenthal, 1995). The convergence time is vital for the proper implementation of the MCMC algorithm, but it has always been difficult to ascertain. However, the coupling characterisation of total variation has made it possible to bound convergence rates in terms of coupling times. Two Markov processes X_k and Y_k, with initial distributions μ_0 and π, changing according to the same transition matrix become equal in distribution after some time T (known as the coupling time), that is,

$$|\mu_k - \pi| = d_{TV}(p, q) \leq \Pr(X_k \neq Y_k) \leq \Pr(T > k),$$

where p and q are the corresponding pdfs of the two processes. In other words, there is a random time T such that $X_k = Y_k$ for all $k \geq T$. Here

$$d_{TV}(p, q) = \sup_{A \subset X} |p(A) - q(A)|.$$

Chapter 4

Probabilistic distances between trees

4.1 Introduction

In the previous chapter, we saw some measures of distance between probability distributions. In this chapter, we describe methodology and software for calculating distances between phylogenetic trees based on the underlying probability distributions on genetic sequence data induced by the phylogenies. The results discussed in this chapter have been published in Garba *et al.* (2018); this chapter is an expanded version of that article. The software is available from `www.mas.ncl.ac.uk/~ntmwn/probdist`.

We begin by reviewing some existing distances between phylogenetic trees in Section 4.2. In Section 4.3, we introduce our idea of probabilistic distances between phylogenetic trees when they are regarded as probability models for gene sequence data. In this regard, we describe simulation methods which calculate (approximately) Hellinger distance, total variation distance, Jensen-Shannon distance and Kullback-Liebler divergence between phylogenetic trees (Section 4.4). These distance measures also have a natural extension to situations when phylogenies do not share the same set of taxa, which is described in Section 4.6. Unlike existing distance measures, the distance measures we propose can be defined between pairs $(T_i, \boldsymbol{\theta}_i)$, $i = 1, 2$, where T_i is a phylogeny and $\boldsymbol{\theta}_i$ is a vector of DNA substitution model parameters, rather than between phylogenies alone. Phylogenetic trees are usually inferred with an associated substitution model, and so information is lost if comparison is only carried out on inferred phylogenetic trees without the associated substitution models. These models were studied in Section 2.5. When obtaining the distance measures, the parameters $\boldsymbol{\theta}_i$ are fixed which means that the instantaneous rate matrix Q of the associated substitution model is fixed.

Determining sample size is an important procedure for any statistical method. Through

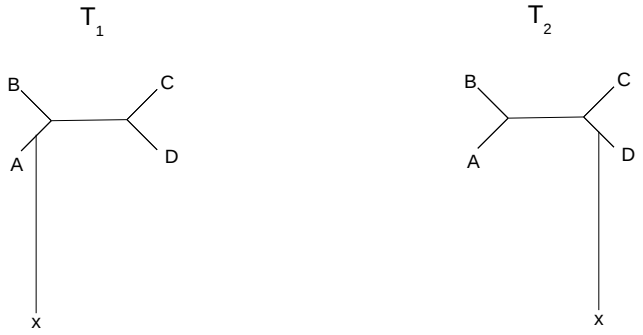


Figure 4.1: Two phylogenetic trees T_1 and T_2 that differ only with respect to the position of taxon x.

a simulation approach, we devised an appropriate method to estimate an adequate sample size for our simulation methods. This is presented in Section 4.5. The sample size was improved using a variance reduction technique. The chapter ends with possible applications of our probabilistic distance measures in several scenarios, in Section 4.7.

Existing distances between phylogenetic trees are purely based on topology and geometry of the phylogenies. However, we propose probabilistic distances between phylogenetic trees when they are regarded as sequence models. The following simple example illustrates the difference between the two approaches. Suppose we have two phylogenetic trees T_1 and T_2 with a common leaf-set so that the two phylogenies differ only with respect to the position of a single taxon x, as shown in Figure 4.1. In other words, the subtrees of T_1 and T_2 obtained by removing x are identical. Then, in the limit that the edge leading to x gets increasingly long, distance measures which compare the topology or geometry of the phylogenies will generally view T_1 and T_2 as being bounded away from each other (or getting further apart). However, under the same limit, the genetic sequence of x effectively becomes independent of the other taxa. Since the relationships between the other taxa are fixed, the probability distributions on characters induced by T_1 and T_2 become identical in the limit and so the distance tends to zero.

The idea of identifying phylogenetic trees with points in a space of distributions on characters was first considered by Kim (2000). The space is usually referred to as the space of "hyperdimensional oranges" or "phylogenetic oranges" as described in Section 2.8.2. Topological and combinatorial aspects of the space were studied by Moulton & Steel (2004). The methods developed in this chapter enable the computation of metrics on this space, a first step towards more involved geometrical methods such as computation of sample means and variances.

4.2 Existing distances between trees

These distances are used extensively, but they are all based on topology and geometry.

4.2.1 Robinson-Foulds distance

Robinson-Foulds (RF) distance (Robinson & Foulds, 1981) is the most widely used measure of dissimilarity between trees. A tree T can be uniquely represented by $S(T)$, its set of splits. Given two trees T_1 and T_2 on the same leaf set, the Robinson-Foulds distance between them is the number of splits that differ between the trees, that is,

$$d_{RF}(T_1, T_2) = \frac{1}{2}\Big(|S(T_1) \setminus S(T_2)| + |S(T_2) \setminus S(T_1)|\Big).$$

The Robinson-Foulds distance is in fact a metric and it can be computed in linear time (Felsenstein, 2004). The same approach has been used to measure distance between edge-weighted trees. In this case, instead of assigning weight 1 to the splits, the splits are weighted by their lengths. This is called the weighted Robinson-Foulds distance (Robinson & Foulds, 1979).

4.2.2 The quartets distance

Estabrook *et al.* (1985) suggest a metric between trees based on subtrees induced by four leaves (quartets). The quartet distance between two trees on the same set of species is the number of quartets that differ in topology between the two trees scaled by the total number of quartets. The number of quartets in a tree with n species is proportional to n^4, thus the quartet distance could potentially be $\mathcal{O}(n^4)$ to compute. However, Brodal *et al.* (2004) discovered an algorithm for computing it that is almost linear in n. A similar distance was developed for rooted trees by Critchlow *et al.* (1996), which instead uses subtrees of three leaves (triplets).

4.2.3 The nearest-neighbor interchange distance

Another well known distance between trees is the nearest-neighbor interchange (NNI) distance of Waterman & Smith (1978). It is a metric and it is defined on the set of unrooted binary trees on some fixed set of taxa. The NNI distance between two trees is the minimum number of nearest-neighbor interchange operations needed to change one tree to the other. NNI operation has been discussed in Section 2.8.1. Computing the NNI

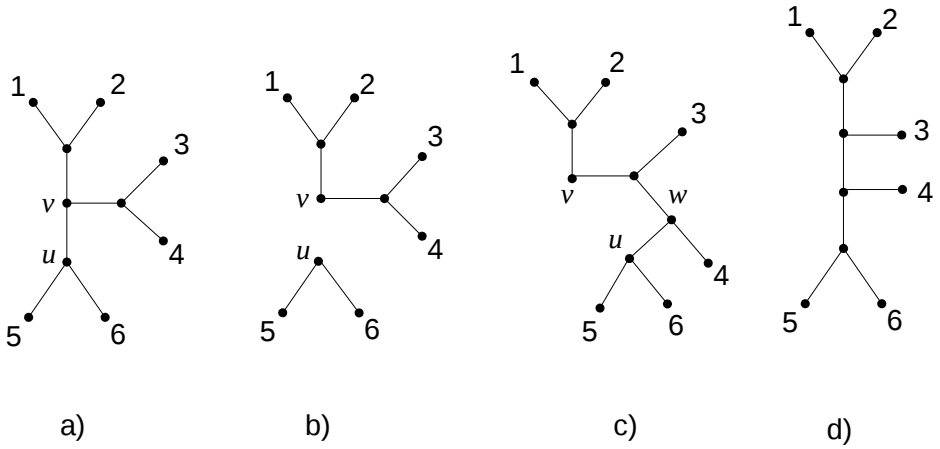


Figure 4.2: Example of subtree-prune-and-regraft (SPR) operation. a) The original tree. b) Prunning subtree rooted at u by removing edge (u, v). c) Regrafting the subtree by subdividing an edge, forming a new vertex w. d) Degree 2 vertex v is removed.

distance for very different large trees is infeasible: Li & Zhang (1999) proved that the computation is NP-complete.

4.2.4 The subtree-prune-and-regraft distance

The subtree-prune-and-regraft (SPR) distance defines a metric between two trees as the minimum number of SPR operations needed to change one tree to the other (Hickey *et al.*, 2008). SPR operation on a tree is illustrated in Figure 4.2. Suppose E is the set of edges and V is the set of vertices in the tree such that each $e \in E$ is written as $e = (u, v)$, where $u, v \in V$. The SPR operation is defined as follows: an edge $(u, v) \in E$ is removed at vertex u thereby prunning a subtree that is rooted at u from the tree. Then a new vertex w is created by subdividing an edge in the tree and attaching (regrafting) the subtree, thus creating a new edge (u, w). Finally, the degree 2 vertex v is suppressed. It has been shown that computing the SPR distance between trees is NP-hard (Bordewich & Semple, 2005; Hickey *et al.*, 2008).

4.2.5 The path-length-difference metric

Penny *et al.* (1993) suggested a metric between trees based on the length of the path between pairs of species on the tree. For each pair of species on a tree, the number of edges that separate them is counted. The path-length-difference distance between two

trees is the square root of the sum of squared differences between the two lists of numbers correspondingly, that is $(\sum(a_{ij} - b_{ij})^2)^{1/2}$, where a_{ij} is the number of edges between leaves i and j and same for b_{ij}.

4.2.6 Billera, Holmes and Vogtmann (BHV) metric

This is a natural metric in the continuous space of trees introduced by Billera *et al.* (2001) (refer to Section 2.8.1 for a discussion of the space). In the BHV space $\mathcal{T}_n$, any two trees T_1 and T_2 can be connected by a unique minimum length path called a *geodesic*. The distance between the two trees is the length of the geodesic between them. In other words, there exist a parametrized set $\Gamma = \{\gamma(\lambda) : 0 \leq \lambda \leq 1\}$ of trees $\gamma(\lambda) \in \mathcal{T}_n$ connecting any two trees T_1 and T_2 in $\mathcal{T}_n$. In a simple case, this path is the cone path which consists of a straight line segment from T_1 to the origin and then from the origin to T_2. However, we can think of any path Γ between the trees as a sequence of connected line segments with each segment lying in a single orthant. If $L(\Gamma)$ denotes the length of a path Γ, then $L(\Gamma)$ is defined as the sum of the Euclidean lengths of the line segments. Hence the distance between T_1 and T_2 in $\mathcal{T}_n$ is the length of the shortest path (geodesic) in $\mathcal{T}_n$ between T_1 and T_2. Owen & Provan (2011) provided an algorithm for computing the BHV metric which is $\mathcal{O}(n^4)$ to compute.

4.3 Probabilistic distances

Given a phylogenetic tree $T \in \mathcal{T}_n$ and a Markov process model with parameter $\boldsymbol{\theta}$, we saw in Section 2.4 that the Markov process determines a distribution on characters at the leaves, that is, a distribution on Ω^n. Let $\mathcal{D}(\Omega^n)$ be the set of distributions on Ω^n. We look at distances between pairs $(T_i, \boldsymbol{\theta}_i)$, $i = 1, 2$ of phylogenies and model parameters to be the distance between the induced probability distributions on Ω^n, that is

$$d\big((T_1, \boldsymbol{\theta}_1), (T_2, \boldsymbol{\theta}_2)\big) = d\big(p_{(T_1, \boldsymbol{\theta}_1)}, q_{(T_2, \boldsymbol{\theta}_2)}\big),$$

where $p_{(T_1, \boldsymbol{\theta}_1)}$ is the pmf associated with $(T_1, \boldsymbol{\theta}_1)$ and likewise $q_{(T_2, \boldsymbol{\theta}_2)}$. This defines a metric provided the map from $\mathcal{T}_n \times \Theta \to \mathcal{D}(\Omega^n)$ is injective.

When $\Omega = \{0, 1\}$, we assume the Markov substitution process is the unique symmetric process on two states. This Markov process has no parameters. When $\Omega = \{\text{A,C,G,T}\}$, we assume a general time-reversible (GTR) model with across-site Gamma rate heterogeneity. The parameters $\boldsymbol{\theta}$ for this model determine the rates of character substitution and the

stationary distribution of the process. Since the GTR model is identifiable (Allman *et al.*, 2008), the map from pairs $(T, \boldsymbol{\theta})$ to probability distributions on Ω^n is injective, so that distinct phylogenies always induce distinct distributions. This is also the case for the two-state symmetric model and GTR with across-site Gamma rate heterogeneity. However, the methodology we develop below can be applied to arbitrary alphabets and substitution models, in particular, to amino acid models.

For $p(\boldsymbol{s})$, $q(\boldsymbol{s}) \in \mathcal{D}(\Omega^n)$, with $\boldsymbol{s} \in \Omega^n$, we consider (i) Hellinger (H) distance, (ii) total variation (TV) distance, (iii) Kullback-Leibler (KL) divergence, and (iv) Jensen-Shannon (JS) distance (as discussed in Section 3.1).

4.4 Simulation strategy and expectation

Computing the probabilistic distance measures exactly for large phylogenetic trees is computationally expensive as they are expressed as sums over Ω^n. However, we can estimate these distances via simulation since each can be expressed in terms of expectations with respect to the distributions $p, q \in \mathcal{D}(\Omega^n)$. Suppose that $\boldsymbol{s}_{p,i}$, $i = 1, \ldots, m$ are a set of m characters simulated on phylogeny T_1, and $\boldsymbol{s}_{q,i}$, $i = 1, \ldots, m$ are a set of m characters simulated on phylogeny T_2. In other words, the characters $\boldsymbol{s}_{p,i}$ are independent samples from distribution p and similarly for the characters $\boldsymbol{s}_{q,i}$. We can think these samples as each being equivalent to a simulated alignment with m independent columns from the phylogenies T_1, T_2 respectively. Then it can be shown that the Hellinger distance can be estimated via

$$\begin{aligned}
d_H(p,q)^2 &= \frac{1}{2}\sum_{\boldsymbol{s}\in\Omega^n}\left\{\sqrt{p(\boldsymbol{s})}-\sqrt{q(\boldsymbol{s})}\right\}^2 \\
&= \frac{1}{2}\sum_{\boldsymbol{s}}\left\{p(\boldsymbol{s})-2\sqrt{p(\boldsymbol{s})q(\boldsymbol{s})}+q(\boldsymbol{s})\right\} \\
&= 1-\sum_{\boldsymbol{s}}\sqrt{p(\boldsymbol{s})q(\boldsymbol{s})} \\
&= 1-\sum_{\boldsymbol{s}}p(\boldsymbol{s})\sqrt{\frac{q(\boldsymbol{s})}{p(\boldsymbol{s})}} \\
&= 1-E_p\left[\sqrt{\frac{q(\boldsymbol{s})}{p(\boldsymbol{s})}}\right] \\
&\simeq 1-\frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{q(\boldsymbol{s}_{p,i})}{p(\boldsymbol{s}_{p,i})}},
\end{aligned}$$

where the $\boldsymbol{s}_{p,i}$ are a random sample drawn from distribution p. By symmetry, an equivalent approximation which uses samples from both p and q is

$$d_H(p,q)^2 \simeq 1 - \frac{1}{2m}\sum_{i=1}^{m}\left\{\sqrt{\frac{q(\boldsymbol{s}_{p,i})}{p(\boldsymbol{s}_{p,i})}} + \sqrt{\frac{p(\boldsymbol{s}_{q,i})}{q(\boldsymbol{s}_{q,i})}}\right\},$$

where the $\boldsymbol{s}_{q,i}$ are a random sample drawn from distribution q. A similar estimator can be derived for the total variation distance:

$$\begin{aligned} d_{TV}(p,q) &= \frac{1}{2}\sum_{\boldsymbol{s}\in\Omega^n}|p(\boldsymbol{s}) - q(\boldsymbol{s})| \\ &= \frac{1}{2}\sum_{\boldsymbol{s}}\left|q(\boldsymbol{s})\frac{p(\boldsymbol{s})}{q(\boldsymbol{s})} - q(\boldsymbol{s})\right| \\ &= \frac{1}{2}\sum_{\boldsymbol{s}} q(\boldsymbol{s})\left|\frac{p(\boldsymbol{s})}{q(\boldsymbol{s})} - 1\right| \\ &= \frac{1}{2}E_q\left[\left|\frac{p(\boldsymbol{s})}{q(\boldsymbol{s})} - 1\right|\right] \\ &\simeq \frac{1}{2m}\sum_{i=1}^{m}\left|\frac{p(\boldsymbol{s}_{q,i})}{q(\boldsymbol{s}_{q,i})} - 1\right|. \end{aligned}$$

Similarly, the Kullback-Leibler divergence can be estimated using

$$\begin{aligned} d_{KL}(p;q) &= \sum_{\boldsymbol{s}\in\Omega^n} p(\boldsymbol{s})\log\left\{\frac{p(\boldsymbol{s})}{q(\boldsymbol{s})}\right\} \\ &= E_p\left[\log\left\{\frac{p(\boldsymbol{s})}{q(\boldsymbol{s})}\right\}\right] \\ &\simeq \frac{1}{m}\sum_{i=1}^{m}\log\left\{\frac{p(\boldsymbol{s}_{p,i})}{q(\boldsymbol{s}_{p,i})}\right\} \\ &= \frac{1}{m}\sum_{i=1}^{m}\left\{\log p(\boldsymbol{s}_{p,i}) - \log q(\boldsymbol{s}_{p,i})\right\}. \end{aligned}$$

The Jensen-Shannon distance can be estimated in a similar way, as it is a sum of two Kullback-Leibler divergences:

$$\begin{aligned} d_{JS}^2(p,q) &= \frac{1}{2}d_{KL}(p;r) + \frac{1}{2}d_{KL}(q;r) \\ &= \frac{1}{2}\sum_{\boldsymbol{s}\in\Omega^n} p(\boldsymbol{s}) \log\left\{\frac{p(\boldsymbol{s})}{r(\boldsymbol{s})}\right\} + \frac{1}{2}\sum_{\boldsymbol{s}\in\Omega^n} q(\boldsymbol{s}) \log\left\{\frac{q(\boldsymbol{s})}{r(\boldsymbol{s})}\right\} \\ &= \frac{1}{2}E_p\left[\log\left\{\frac{p(\boldsymbol{s})}{r(\boldsymbol{s})}\right\}\right] + \frac{1}{2}E_q\left[\log\left\{\frac{q(\boldsymbol{s})}{r(\boldsymbol{s})}\right\}\right] \\ &\simeq \frac{1}{2m}\sum_{i=1}^{m}\left(\log\left\{\frac{p(\boldsymbol{s}_{p,i})}{r(\boldsymbol{s}_{p,i})}\right\} + \log\left\{\frac{q(\boldsymbol{s}_{q,i})}{r(\boldsymbol{s}_{q,i})}\right\}\right), \end{aligned}$$

where $r(\cdot) = \{p(\cdot) + q(\cdot)\}/2$.

We can compare the exact distance between a pair of phylogenetic trees with the simulated distance between the same pair for different sample sizes m for phylogenies with relatively small n, so that the sum over Ω^n can be evaluated computationally. We consider two fixed random 6-taxon phylogenies each with topology sampled from a Yule distribution and the edge lengths sampled from a Gamma distribution with mean 0.1 and variance 0.005. Figure 4.3 shows the sampling distribution of simulated Hellinger distance, total variation distance, Kullback-Leibler divergence and Jensen-Shannon distance between the two phylogenies for different sample sizes. The dashed horizontal line is the exact distance between the pair of phylogenies. While the graphs show, as expected, that estimates improve as sample size m increases, we ideally require some means of determining m automatically in order to achieve a given level of accuracy.

4.5 Sample size calculation

When estimating the probabilistic distances, it is helpful to determine the smallest size sample needed to be reliably within a given tolerance of the true distance. This can be achieved using the central limit theorem to obtain normal approximations (Kwak & Kim, 2017). The estimate

$$R_m = 1 - \frac{1}{2m}\sum_{i=1}^{m}\left(\sqrt{\frac{q(\boldsymbol{s}_{p,i})}{p(\boldsymbol{s}_{p,i})}} + \sqrt{\frac{p(\boldsymbol{s}_{q,i})}{q(\boldsymbol{s}_{q,i})}}\right)$$

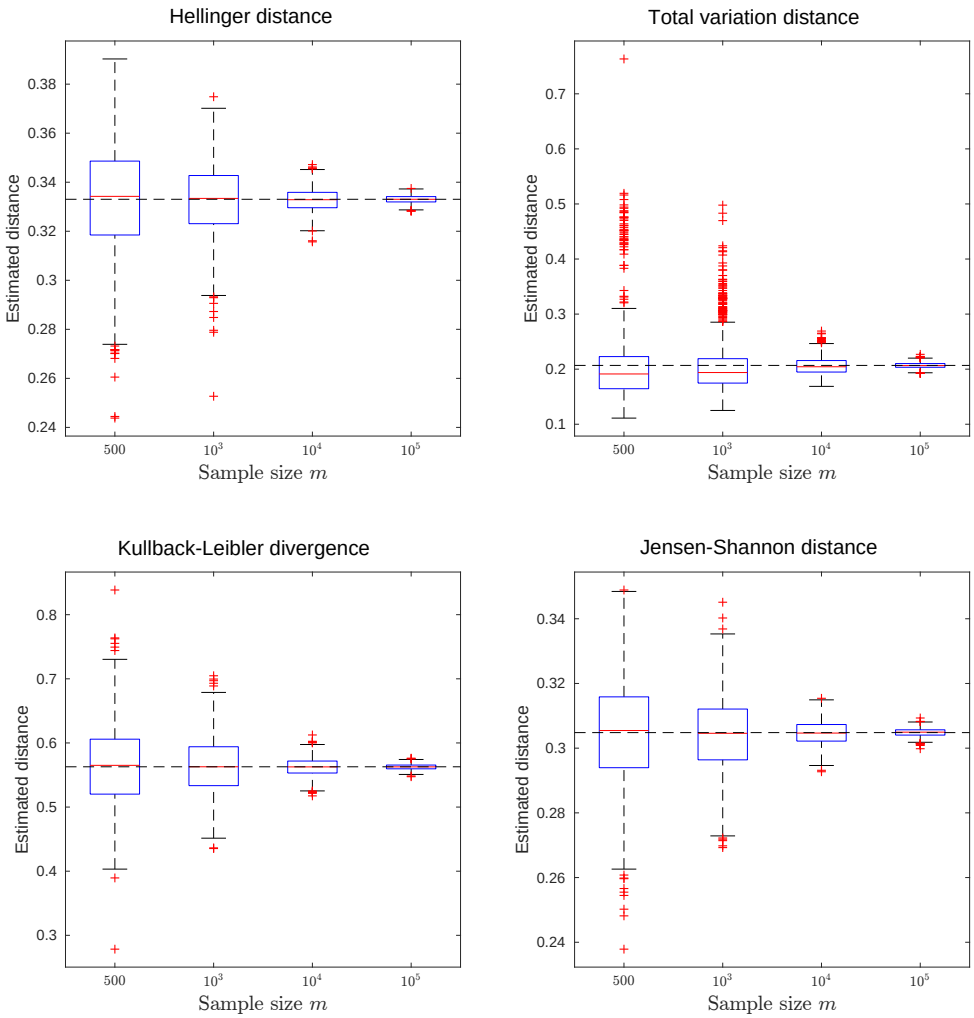


Figure 4.3: Sampling distribution of simulated Hellinger distance, total variation distance, Kullback-Leibler divergence and Jensen-Shannon distance between two random 6-taxon phylogenetic trees for different sample sizes m. The dashed horizontal line (on each figure) is the exact distance between the pair of phylogenies.

is unbiased for the squared Hellinger distance $\mu_0 = d_H(p,q)^2$. Also, for large m, R_m is approximately normally distributed with variance σ_0^2/m, where σ_0^2 is the variance of $R_{m=1}$. We can obtain provisional estimates μ_0 and σ_0^2 from a pilot run of size m_0. Each of the m_0 realisations produces an estimate of R_1 and so their mean and variance are unbiased estimates for μ_0 and σ_0^2. Absolute or relative error are standard criteria for determining an appropriate sample size m in this situation. For example, to require the estimate R_m to have an absolute error of ω with probability $1-\beta$ requires

$$\Pr\left(\left|\sqrt{\mu_0} - \sqrt{R_m}\right| < \omega\right) = 1 - \beta.$$

Consequently

$$\begin{aligned}\left|\sqrt{\mu_0}-\sqrt{R_m}\right|<\omega \quad &\Leftrightarrow \quad \sqrt{\mu_0}-\omega\le\sqrt{R_m}\le\sqrt{\mu_0}+\omega\\ &\Leftrightarrow \quad (\sqrt{\mu_0}-\omega)^2\le R_m\le(\sqrt{\mu_0}+\omega)^2\\ &\Leftrightarrow \quad \mu_0-\omega\,(2\sqrt{\mu_0}-\omega)\le R_m\le\mu_0+\omega\,(2\sqrt{\mu_0}+\omega)\,,\end{aligned}$$

provided $\sqrt{\mu_0}-\omega$ is non-negative. Therefore

$$\Pr\Big(\mu_0-\omega\,(2\sqrt{\mu_0}-\omega)\le R_m\le\mu_0+\omega\,(2\sqrt{\mu_0}+\omega)\Big)=1-\beta.$$

Since R_m is normal

$$\frac{z_{\beta/2}\sigma_0}{\sqrt{m}}\le\omega(2\sqrt{\mu_0}\pm\omega)\quad\Leftrightarrow\quad m\ge\frac{z^2_{\beta/2}\sigma_0^2}{\omega^2(2\sqrt{\mu_0}\pm\omega)^2}=\frac{z^2_{\beta/2}\sigma_0^2}{\omega^2(2\sqrt{\mu_0}-\omega)^2},\tag{4.1}$$

where z_β is the upper β point of the standard normal distribution (e.g. $z_{0.025}=1.96$). If instead we require a relative error of α, this is equivalent to using absolute error $\omega=\alpha\sqrt{\mu_0}$. In this case, we require

$$\begin{aligned}(1-\alpha)\sqrt{\mu_0}<\sqrt{R_m}<(1+\alpha)\sqrt{\mu_0}\quad &\Leftrightarrow\quad (1-\alpha)^2\mu_0<R_m<(1+\alpha)^2\mu_0\\ &\Leftrightarrow\quad \mu_0-\alpha(2-\alpha)\mu_0<R_m<\mu_0+\alpha(2+\alpha)\mu_0.\end{aligned}$$

This happens when

$$\frac{z_{\beta/2}\sigma_0}{\sqrt{m}}\le\alpha(2\pm\alpha)\mu_0\quad\Leftrightarrow\quad m\ge\frac{z^2_{\beta/2}\sigma_0^2}{\alpha^2(2\pm\alpha)^2\mu_0^2}=\frac{z^2_{\beta/2}\sigma_0^2}{\alpha^2(2-\alpha)^2\mu_0^2}.\tag{4.2}$$

Estimated values of μ_0 and σ_0^2 from the pilot run are used in (4.1) or (4.2) to estimate m. Note that it is possible to have $\sqrt{\mu_0}-\omega<0$ for two trees that are very similar in which case it is advisable to use the relative error in order to estimate m.

Furthermore, the estimate

$$R_m=\frac{1}{2m}\sum_{i=1}^{m}\left|\frac{p(\boldsymbol{s}_{q,i})}{q(\boldsymbol{s}_{q,i})}-1\right|$$

is unbiased for the total variation distance $\mu_0=d_{TV}(p,q)$. Also, for large m, R_m is approximately normally distributed with variance σ_0^2/m, where σ_0^2 is the variance of $R_{m=1}$.

We require R_m to be within $\alpha\%$ of μ_0 with probability $1-\beta$. This happens when

$$\frac{z_{\beta/2}\sigma_0}{\sqrt{m}} \leq \alpha\mu_0 \Leftrightarrow m \geq \frac{z_{\beta/2}^2\sigma_0^2}{(\alpha\mu_0)^2}.$$

Similarly, the estimate

$$R_m = \frac{1}{m}\sum_{i=1}^{m} \log\left\{\frac{p(\boldsymbol{s}_{p,i})}{q(\boldsymbol{s}_{p,i})}\right\}$$

is unbiased for the Kullback-Leibler divergence $\mu_0 = d_{KL}(p;q)$. Also, for large m, R_m is approximately normally distributed with variance σ_0^2/m, where σ_0^2 is the variance of $R_{m=1}$. Thus, for R_m to be within $\alpha\%$ of μ_0 with probability $1-\beta$, we require

$$\frac{z_{\beta/2}\sigma_0}{\sqrt{m}} \leq \alpha\mu_0 \Leftrightarrow m \geq \frac{z_{\beta/2}^2\sigma_0^2}{(\alpha\mu_0)^2}.$$

Finally, an unbiased estimate for the squared Jensen-Shannon distance $\mu_0 = d_{JS}(p,q)^2$ is

$$R_m = \frac{1}{2m}\sum_{i=1}^{m}\left(\log\left\{\frac{2p(\boldsymbol{s}_{p,i})}{p(\boldsymbol{s}_{p,i})+q(\boldsymbol{s}_{p,i})}\right\} + \log\left\{\frac{2q(\boldsymbol{s}_{q,i})}{p(\boldsymbol{s}_{q,i})+q(\boldsymbol{s}_{q,i})}\right\}\right).$$

For large m, R_m is approximately normally distributed with variance σ_0^2/m, where σ_0^2 is the variance of $R_{m=1}$. To require the estimate R_m to have an absolute error of ω with probability $1-\beta$ requires

$$m \geq \frac{z_{\beta/2}^2\sigma_0^2}{\omega^2(2\sqrt{\mu_0}-\omega)^2}.$$

If instead we require a relative error of α, we have

$$m \geq \frac{z_{\beta/2}^2\sigma_0^2}{\alpha^2(2-\alpha)^2\mu_0^2}.$$

Both the Hellinger and Jensen-Shannon distance can be estimated using sample size obtained from either absolute or relative error. Based on our simulation study, none of the two errors outperforms the other in terms of requiring smaller sample size.

The estimators described above can be improved using the control variate method (Morgan, 1984, Chapter 7, p. 171) to give unbiased estimators which achieve the same

precision with fewer samples i.e. smaller values of m. The control variate method is a variance reduction technique which works by adding in a scaled version of a zero-mean estimator which is negatively correlated with the original estimator (R_m). For example, it is easily shown that $E_p[q/p] = 1$ and so we can construct a new unbiased estimator for the Hellinger distance as $\theta_c = \sqrt{q/p} - c(q/p - 1)$ and choose c to minimise its variance:

$$\mathrm{Var}(\theta_c) = \mathrm{Var}\left(\sqrt{q/p}\right) + c^2\mathrm{Var}\left(q/p\right) - 2c\,\mathrm{Cov}\left(\sqrt{q/p}, q/p\right).$$

By differentiating with respect to c and equating to zero, an optimal value of c is obtained as

$$c^* = \frac{\mathrm{Cov}\left(\sqrt{q/p}, q/p\right)}{\mathrm{Var}\left(q/p\right)},$$

where the covariance and variance are estimated from the pilot run. In the software, a pilot run is used to estimate c^*, and then in the main sampling run, the modified estimator θ_{c^*} is used to estimate the Hellinger distance. The control variate technique has also been implemented for the total variation, Kullback-Leibler and Jensen-Shannon distances. We found the reduction in m varied very substantially with each data set.

In order to explore the possible values for m that might be required for experimental data sets, we estimated m for the distance (H, TV, KL and JS) between every pair of gene trees contained in the data set of 106 yeast gene trees on 8 yeast species (Rokas *et al.*, 2003), using the two-state symmetric model. Figure 4.4 show histograms of estimated values for m. Estimation was performed in order to achieve a relative error of $\alpha = 5\%$ with probability $1 - \beta = 80\%$. The figures show that fairly accurate distances can be obtained using reasonably small sample sizes.

4.6 Missing taxa

Suppose T_A and T_B are phylogenetic trees with taxon sets A and B. Let p and q denote probability mass functions on characters induced by T_A and T_B respectively. Very commonly $A \neq B$ and $A \cap B \neq \emptyset$ but many tree-metrics cannot be computed under these assumptions. We consider two approaches for computing distances when taxon sets differ between phylogenies.

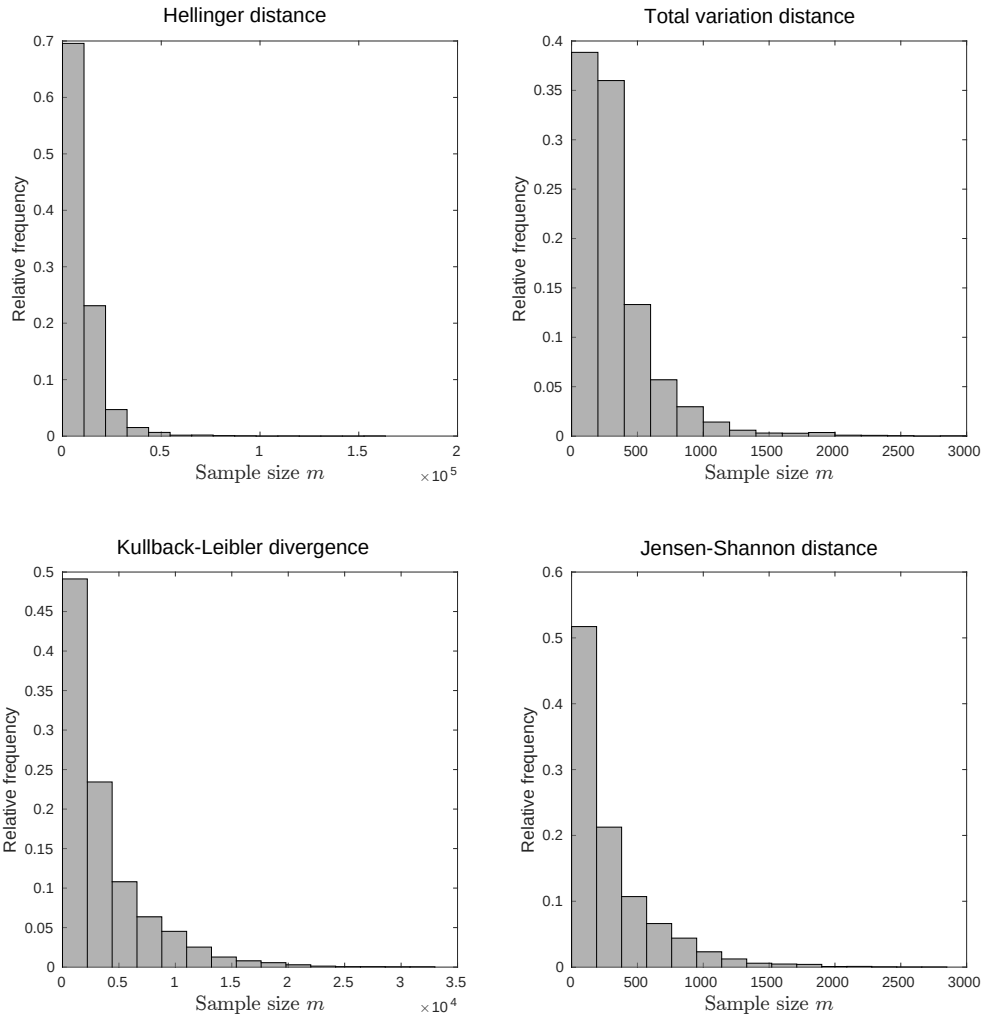


Figure 4.4: Histograms of estimated values of m for the comparison of every pair of gene trees in the data set of 106 gene trees due to Rokas *et al.* (2003), using the two-state symmetric model and probabilistic distances. It can be seen that the TV and JS distance between most pairs of trees in the data set can be estimated accurately with fewer than 2000 samples but the Hellinger and KL require more. In terms of computational cost, this is similar to the cost of computing the likelihood for an alignment of length 2000 for each pair of trees.

4.6.1 Common taxa method

The analysis is restricted to $A \cap B$ by cropping the phylogenies down to the common taxon set. This is achieved by deleting from T_A all subtrees whose leaves lie in $A \setminus B$, and similarly for T_B. The resulting reduced phylogenies can then used to obtain the distance measures considered in this chapter.

4.6.2 Augmentation method

This method can only be used for our probabilistic distance measures and does not apply to the BHV metric. The strategy here is that we extend the definition of p from $\Omega^{|A|}$ to $\Omega^{|A\cup B|}$ in such a way that the extended distribution is uniform on $\Omega^{|B\setminus A|}$. This is done as follows. Any element $\boldsymbol{s} \in \Omega^{|A\cup B|}$ can be decomposed into three parts corresponding to the taxa in each of the sets $A \setminus B$, $A \cap B$ and $B \setminus A$ and these parts are denoted $\boldsymbol{s}_{A\setminus B}$, $\boldsymbol{s}_{A\cap B}$ and $\boldsymbol{s}_{B\setminus A}$ respectively. Since there are $|\Omega|^{|B\setminus A|}$ possibilities for $\boldsymbol{s}_{B\setminus A}$, the uniform assumption implies that the probability of each possibility is $|\Omega|^{-|B\setminus A|}$. If the extension of p to $\Omega^{|A\cup B|}$ is denoted $p_{A\cup B}$, then we define

$$p_{A\cup B}(\boldsymbol{s}) = p_{A\cup B}(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}, \boldsymbol{s}_{B\setminus A}) = \frac{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})}{|\Omega|^{|B\setminus A|}}$$

for all $\boldsymbol{s} \in \Omega^{|A\cup B|}$ where $p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})$ denotes the original distribution on $\Omega^{|A|}$, based on the assumption that the distribution of $\boldsymbol{s}_{B\setminus A}$ is independent of that of $(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})$. The subscript $A \cup B$ in $p_{A\cup B}$ is simply a label rather than a set, so $p_{A\cup B}$ does not mean p depends on $A \cup B$ as a set. Probabilistic distances between T_A and T_B can be computed by extending p and q to $A \cup B$ and basing the distance on these extended distributions. The uniform distribution is used as it represents a condition of maximal indifference of the position on the phylogenies of the missing taxa.

Proposition 4.1. *If d is Hellinger, Kullback-Leibler or Jensen-Shannon, then $d(p_{A\cup B}, q_{A\cup B}) \neq d(p_{A\cap B}, q_{A\cap B})$ in general.*

Proof. Suppose d is the Hellinger distance, it is enough to show that

$$\sum_{\boldsymbol{s}\in\Omega^{|A\cup B|}} p^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s})\, q^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s}) \neq \sum_{\boldsymbol{s}_{A\cap B}\in\Omega^{|A\cap B|}} p^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B})\, q^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B}).$$

If $\boldsymbol{s}_{A\cap B}\in\Omega^{|A\cap B|}$, $\boldsymbol{s}_{A\setminus B}\in\Omega^{|A\setminus B|}$ and $\boldsymbol{s}_{B\setminus A}\in\Omega^{|B\setminus A|}$,

$$\begin{aligned}
\sum_{\boldsymbol{s}\in\Omega^{|A\cup B|}} p^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s})\, q^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s}) &= \sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}}\sum_{\boldsymbol{s}_{B\setminus A}} \frac{p^{\frac{1}{2}}\left(\boldsymbol{s}_{A\setminus B},\boldsymbol{s}_{A\cap B}\right)}{|\Omega|^{|B\setminus A|/2}} \frac{q^{\frac{1}{2}}\left(\boldsymbol{s}_{B\setminus A},\boldsymbol{s}_{A\cap B}\right)}{|\Omega|^{|A\setminus B|/2}} \\
&= \sum_{\boldsymbol{s}_{A\cap B}} \left(\frac{1}{|\Omega|^{|B\setminus A|/2}} \sum_{\boldsymbol{s}_{A\setminus B}} p^{\frac{1}{2}}\left(\boldsymbol{s}_{A\setminus B},\boldsymbol{s}_{A\cap B}\right)\right) \\
&\quad \times \left(\frac{1}{|\Omega|^{|A\setminus B|/2}} \sum_{\boldsymbol{s}_{B\setminus A}} q^{\frac{1}{2}}\left(\boldsymbol{s}_{B\setminus A},\boldsymbol{s}_{A\cap B}\right)\right) \\
&= \sum_{\boldsymbol{s}_{A\cap B}} \left(|\Omega|^{-|B\setminus A|/2} \sum_{\boldsymbol{s}_{A\setminus B}} p^{\frac{1}{2}}\left(\boldsymbol{s}_{A\setminus B},\boldsymbol{s}_{A\cap B}\right)\right) \\
&\quad \times \left(|\Omega|^{-|A\setminus B|/2} \sum_{\boldsymbol{s}_{B\setminus A}} q^{\frac{1}{2}}\left(\boldsymbol{s}_{B\setminus A},\boldsymbol{s}_{A\cap B}\right)\right).
\end{aligned}$$

But

$$|\Omega|^{-|B\setminus A|/2} \sum_{\boldsymbol{s}_{A\setminus B}} p^{\frac{1}{2}}\left(\boldsymbol{s}_{A\setminus B},\boldsymbol{s}_{A\cap B}\right) \neq \left(\sum_{\boldsymbol{s}_{A\setminus B}} p\left(\boldsymbol{s}_{A\setminus B},\boldsymbol{s}_{A\cap B}\right)\right)^{\frac{1}{2}} = p^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B}),$$

and similarly

$$|\Omega|^{-|A\setminus B|/2} \sum_{\boldsymbol{s}_{B\setminus A}} q^{\frac{1}{2}}\left(\boldsymbol{s}_{B\setminus A},\boldsymbol{s}_{A\cap B}\right) \neq \left(\sum_{\boldsymbol{s}_{B\setminus A}} q\left(\boldsymbol{s}_{B\setminus A},\boldsymbol{s}_{A\cap B}\right)\right)^{\frac{1}{2}} = q^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B}).$$

Therefore

$$\sum_{\boldsymbol{s}\in\Omega^{|A\cup B|}} p^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s})\, q^{\frac{1}{2}}_{A\cup B}(\boldsymbol{s}) \neq \sum_{\boldsymbol{s}_{A\cap B}} p^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B})\, q^{\frac{1}{2}}_{A\cap B}(\boldsymbol{s}_{A\cap B}).$$

Suppose now d is the Kullback-Leibler divergence, we have

$$\begin{aligned}
d\left(p_{A\cup B}, q_{A\cup B}\right) &= \sum_{\boldsymbol{s}_{A\cup B}} p_{A\cup B}(\boldsymbol{s}) \log\left\{\frac{p_{A\cup B}(\boldsymbol{s})}{q_{A\cup B}(\boldsymbol{s})}\right\} \\
&= \sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}}\sum_{\boldsymbol{s}_{B\setminus A}} \frac{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})}{|\Omega|^{|B\setminus A|}} \log\left\{\frac{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})}{q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})}|\Omega|^{|A\setminus B|-|B\setminus A|}\right\} \\
&= \sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}) \sum_{\boldsymbol{s}_{B\setminus A}} \frac{1}{|\Omega|^{|B\setminus A|}} \log\left\{\frac{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})}{q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})}\right\} \\
&\quad +(|A\setminus B| - |B\setminus A|)\log|\Omega| \\
&= \sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}) \sum_{\boldsymbol{s}_{B\setminus A}} \frac{1}{|\Omega|^{|B\setminus A|}} \log\{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})\} \\
&\quad -\sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}) \sum_{\boldsymbol{s}_{B\setminus A}} \frac{\log\{q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})\}}{|\Omega|^{|B\setminus A|}} \\
&\quad +(|A\setminus B| - |B\setminus A|)\log|\Omega| \\
&= \sum_{\boldsymbol{s}_{A\cap B}}\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}) \log\{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})\} \\
&\quad -\sum_{\boldsymbol{s}_{A\cap B}} p_{A\cap B}(\boldsymbol{s}_{A\cap B}) \sum_{\boldsymbol{s}_{B\setminus A}} \frac{\log\{q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})\}}{|\Omega|^{|B\setminus A|}} \\
&\quad +(|A\setminus B| - |B\setminus A|)\log|\Omega|.
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B}) \log\{p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})\} &\neq \left(\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})\right) \log\left\{\sum_{\boldsymbol{s}_{A\setminus B}} p(\boldsymbol{s}_{A\setminus B}, \boldsymbol{s}_{A\cap B})\right\} \\
&= p_{A\cap B}(\boldsymbol{s}_{A\cap B}) \log\{p_{A\cap B}(\boldsymbol{s}_{A\cap B})\}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{\boldsymbol{s}_{B\setminus A}} \frac{\log\{q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})\}}{|\Omega|^{|B\setminus A|}} + (|A\setminus B| - |B\setminus A|)\log|\Omega| &\neq \log\left\{\sum_{\boldsymbol{s}_{B\setminus A}} q(\boldsymbol{s}_{B\setminus A}, \boldsymbol{s}_{A\cap B})\right\} \\
&= \log\{q_{A\cap B}(\boldsymbol{s}_{A\cap B})\},
\end{aligned}$$

the result follows.

The result of Jensen-Shannon distance follows easily from that of KL divergence since JS distance is defined in terms of KL divergence. However, this result has not been

proved for the total variation distance. Therefore, we conclude that $d(p_{A\cup B}, q_{A\cup B}) \neq d(p_{A\cap B}, q_{A\cap B})$ for Hellinger, KL and JS distance measures. In computing distance between pairs of phylogenetic trees with different number of taxa, one can choose between two methods, i.e. $d(p_{A\cup B}, q_{A\cup B})$ or $d(p_{A\cap B}, q_{A\cap B})$ referred to as augmentation method or common taxa method respectively. □

4.7 Results

We now look at properties of the probabilistic distance measures in several scenarios. The aim is to illustrate possible advantages and disadvantages in comparison to existing metrics, especially the BHV metric.

4.7.1 Scaling edges

We consider two phylogenies $T_1 = (\tau_1, \boldsymbol{\ell}_1)$ and $T_2 = (\tau_2, \boldsymbol{\ell}_2)$ on a shared set of n taxa, where τ_i and $\boldsymbol{\ell}_i$ are the topology and vector of edge lengths on the phylogenies respectively. Here the two topologies were sampled from a Yule distribution and the edge lengths were sampled from a Gamma distribution with mean 0.1 and variance 0.005. The edge lengths on both phylogenies were then scaled by a factor s and the quantity

$$d\big((\tau_1, s\boldsymbol{\ell}_1), (\tau_2, s\boldsymbol{\ell}_2)\big)$$

was computed for different values of s using the probabilistic distances and the BHV metric: here $s\boldsymbol{\ell}_i$ is the vector of edge lengths $\boldsymbol{\ell}_i$ multiplied by s. Probabilistic distances were calculated exactly using the two-state symmetric substitution model. Figure 4.5 shows how the distances do not behave like the BHV metric under this scaling. Note that the absolute values of probabilistic distances and BHV metric cannot be compared directly (in fact the BHV axis has been rescaled). As we make the edge lengths on both phylogenies increasingly long, the BHV metric increases linearly. However, under the same limit, the distribution of sequence data at each leaf becomes independent of the distribution at any other leaf, and so both phylogenies give rise to the same distribution on characters under this limit. Therefore, the probabilistic distance between the phylogenies reduces to zero as s increases whereas BHV increases linearly. For values of s which are more meaningful from a biological perspective (for example, s between zero and five) our distances also depend non-linearly on s. This clearly demonstrates the fundamental difference between probabilistic distances and existing distances which use edge-length information.

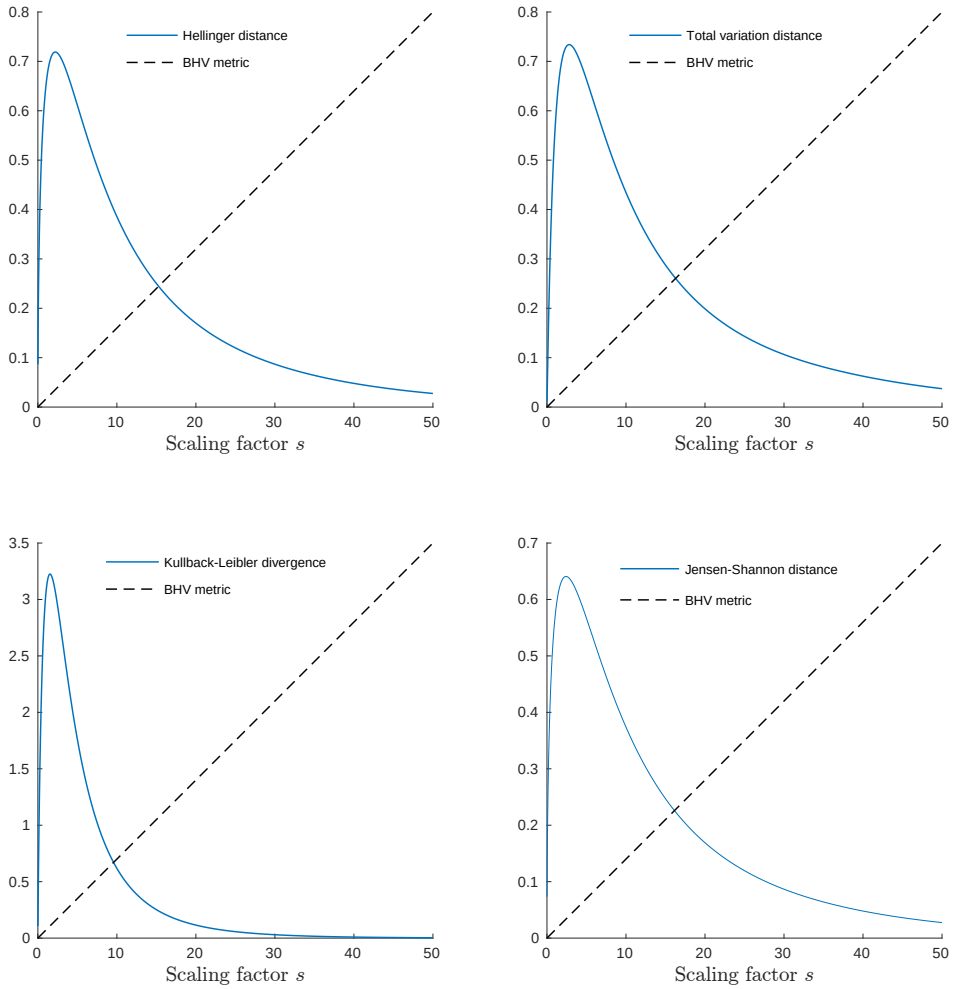


Figure 4.5: Probabilistic distances and Billera-Holmes-Vogtmann (BHV) metric between two random 16-taxon phylogenetic trees T_1 and T_2 with branch lengths scaled by a factor s.

Probabilistic distances behave differently from existing metrics when the phylogenies contain long edges. This is particularly relevant to situations when phylogenies might be subject to long branch attraction artefacts (Bergsten, 2005). To illustrate this, we consider phylogenetic trees in the so-called Felsenstein zone and Faris zone (Felsenstein, 2004). The phylogenies are shown in Figure 4.6. Each represents an alternative hypothesis in the presence of two long edges. As the edge length s increases, the probabilistic distances between the phylogenies decrease to a constant, as shown in Figure 4.7. In contrast, the BHV distance does not vary with s. The probabilistic distances correctly capture the fact that distinguishing one phylogeny from the other as s increases is difficult since they induce similar distributions on sequence data.

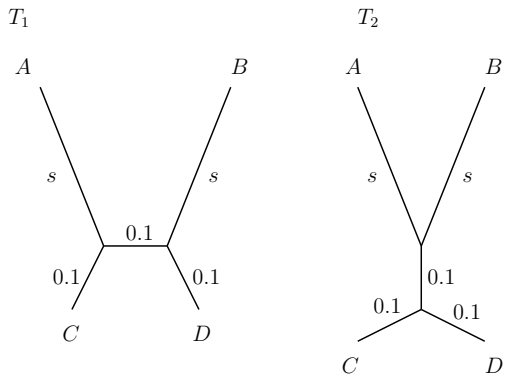


Figure 4.6: Two phylogenetic trees in the Felsenstein zone and Faris zone representing an alternative hypothesis in the presence of two long edges.

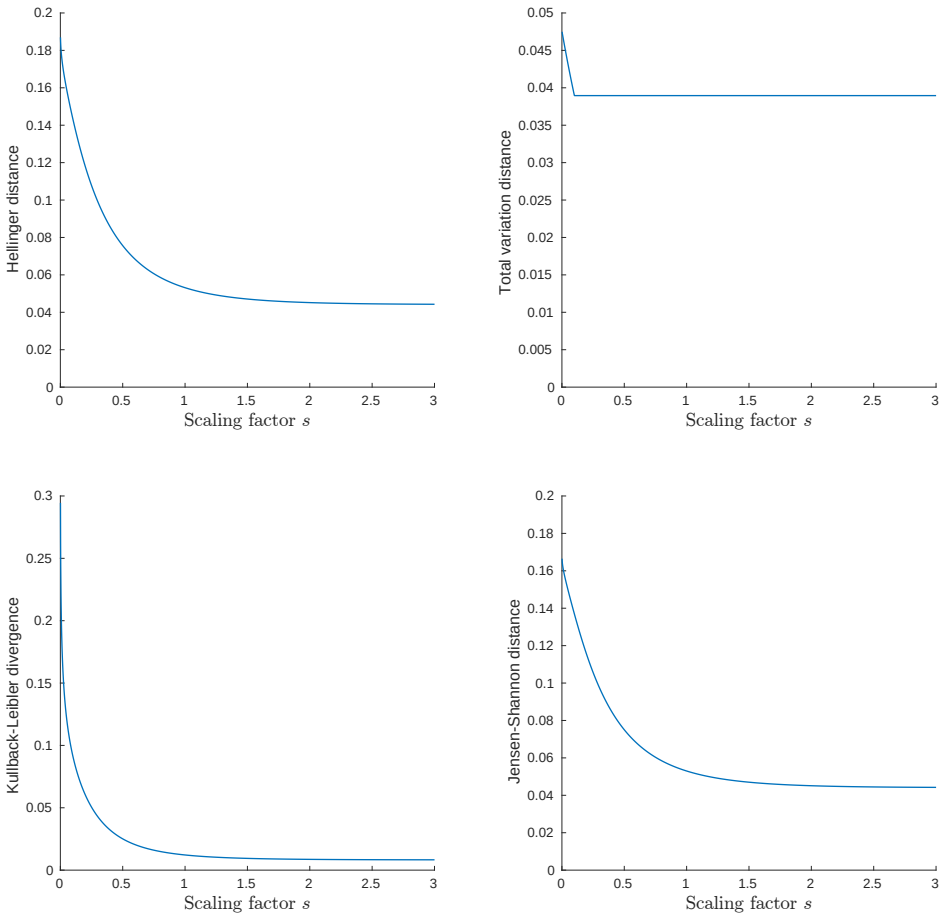


Figure 4.7: Probablistic distances between the two phylogenetic trees shown in Figure 4.6, as a function of s, the length of the long pendant edges. The BHV distance does not vary with s.

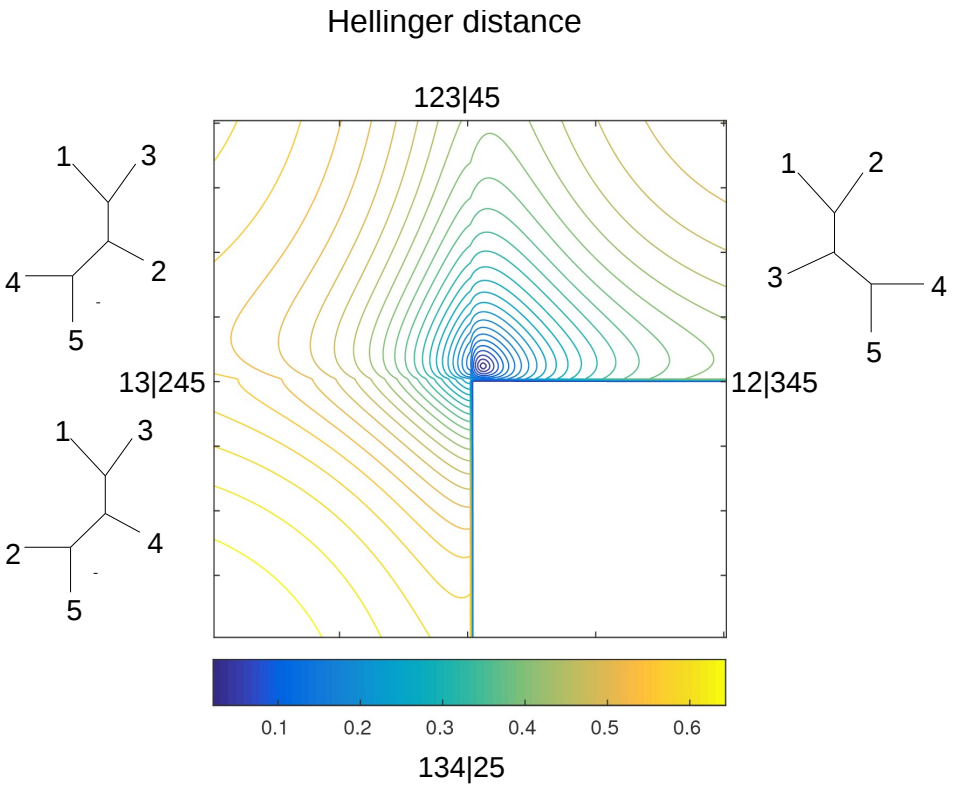


Figure 4.8: Contours of Hellinger distance between T_0 and T_1 each with the HKY85 model in $\mathcal{U}_5$ space, where T_0 is a fixed phylogeny (at the centre of the small blue circle) and T_1 varies over the space. The HKY85 parameter values used are $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ and $\kappa = 1$.

4.7.2 Probabilistic distances on $\mathcal{U}_5$

We compute the probabilistic distances between pairs of phylogenetic trees on $\mathcal{U}_5$, the space of five taxa unrooted phylogenies. Figure 4.8 shows contours of Hellinger distance between T_0 and T_1 each with the HKY85 model, where T_0 is a fixed phylogeny (at the centre of the small blue circle) and T_1 varies over the space. The HKY85 parameter values used are $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ and $\kappa = 1$. The figure consists of three orthants stuck together along common faces, with each orthant representing a valid phylogeny topology as shown. Each point in this space is a five taxa phylogeny with one of the three topologies and the position of the point determines the two internal edge lengths on the phylogeny, whereas all pendant edge lengths are fixed at 0.1. Similar results obtained using TV, KL and JS distances are shown in Figure 4.9. Also, we compute contours of probabilistic distances using the two-state symmetric model and similar trends were seen.

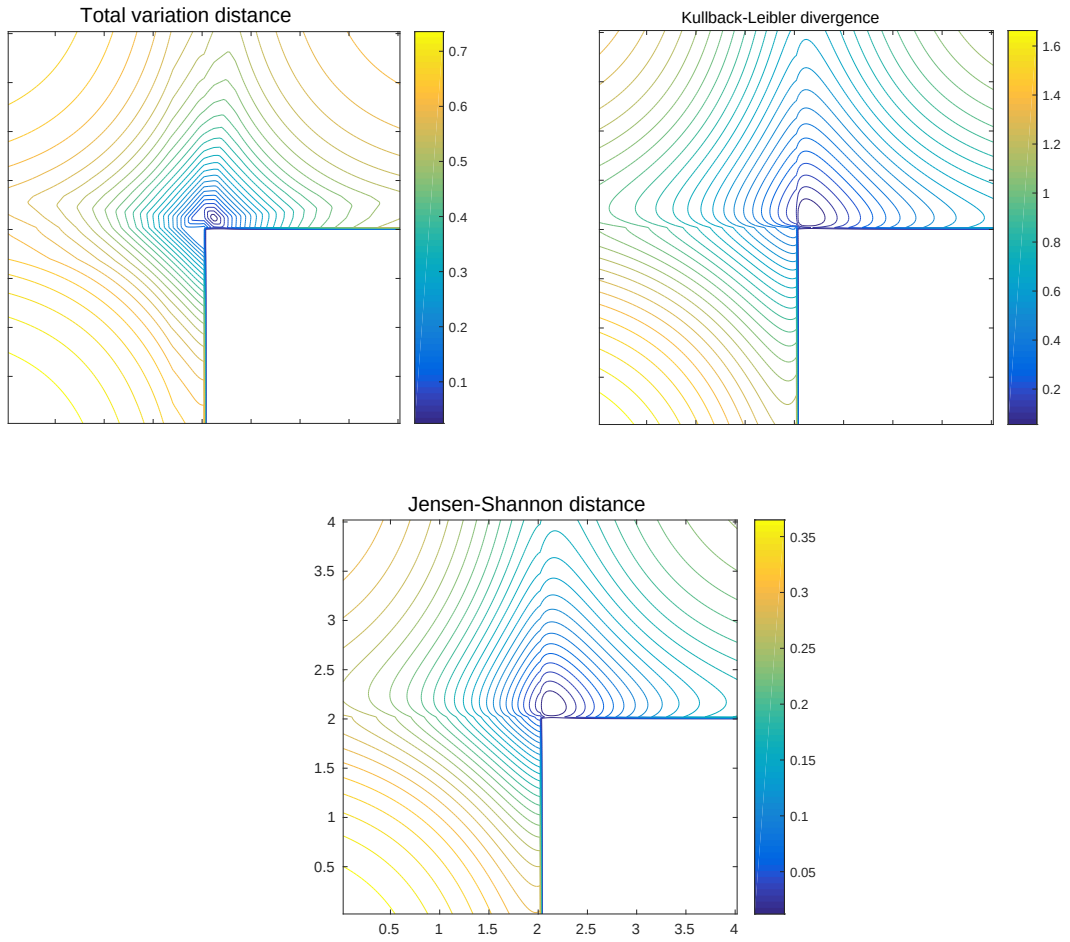


Figure 4.9: Contours of distance (total variation, Kullback-Leibler and Jensen-Shannon) between T_0 and T_1 each with the HKY85 model in $\mathcal{U}_5$ space, where T_0 is a fixed phylogeny (at the centre of the small blue circle) and T_1 varies over the space. The HKY85 parameter values used are $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ and $\kappa = 1$.

4.7.3 Missing Taxa

We investigate the effect of missing taxa on phylogenetic tree distances by first considering a phylogeny with 100 taxa, again with topology sampled from a Yule distribution. Two copies of this topology were made and then phylogenies constructed by assigning edge lengths independently to each topology from a Gamma distribution with mean 0.1 and variance 0.005. Each phylogeny was then subjected to random deletions of the same number of taxa. By repeating the random deletion many times on the fixed pair of phylogenies, and computing the distance each time, we obtain a distribution of distances between the two phylogenies for a given proportion of deletion on each phylogeny; see

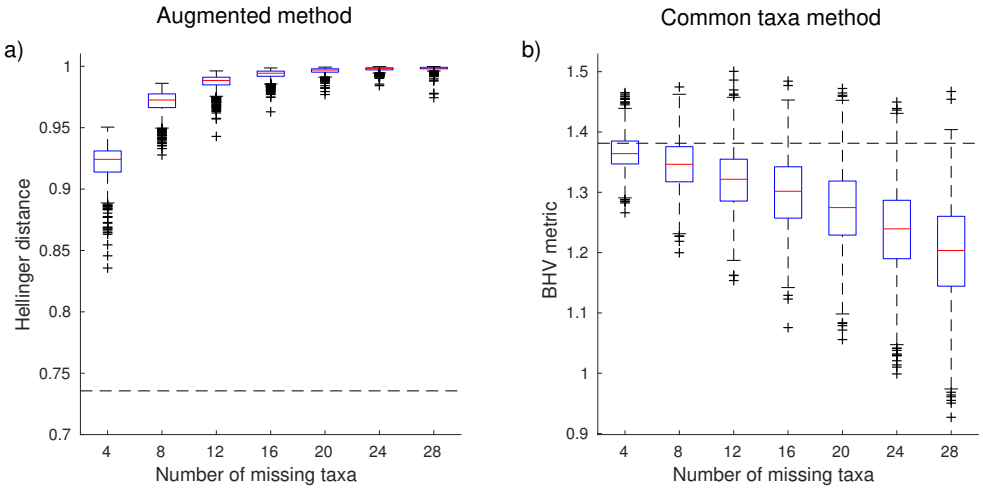


Figure 4.10: Sampling distribution of distance between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method with Hellinger distance, and b) common taxa method with the BHV metric. The initial pair of phylogenies have the same 100 taxa with the same topology but different (random) edge lengths. The dashed horizontal line is the distance/metric between the initial pair of phylogenies (before any deletions).

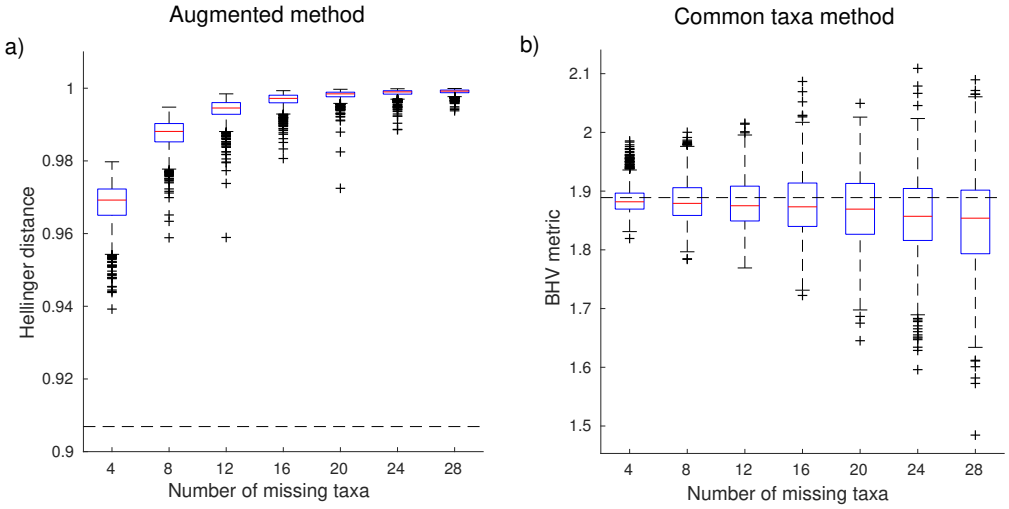


Figure 4.11: Sampling distribution of distance between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method with Hellinger distance, and b) common taxa method with the BHV metric. Both phylogenies have 100 taxa with different (random) edge lengths, with one phylogeny generated at random and the other phylogeny determined using ten subsequent subtree pruning and regraft (SPR) operations. The dashed horizontal line is the distance/metric between the initial pair of phylogenies (before any deletions).

Figure 4.10. We use the two-state symmetric model to compute the probabilistic distances throughout this section.

We also looked at the effect of random deletions of the same number of taxa from two phylogenies with different topologies: the first being another phylogeny with 100 taxa and the second being determined by applying ten random subtree pruning and regraft (SPR) operations to the first; see Figure 4.11. In both figures, we compare the augmentation method using Hellinger distance with the common taxa method using the BHV metric. The augmentation method cannot be applied to the BHV metric, since it is intrinsically probabilistic.

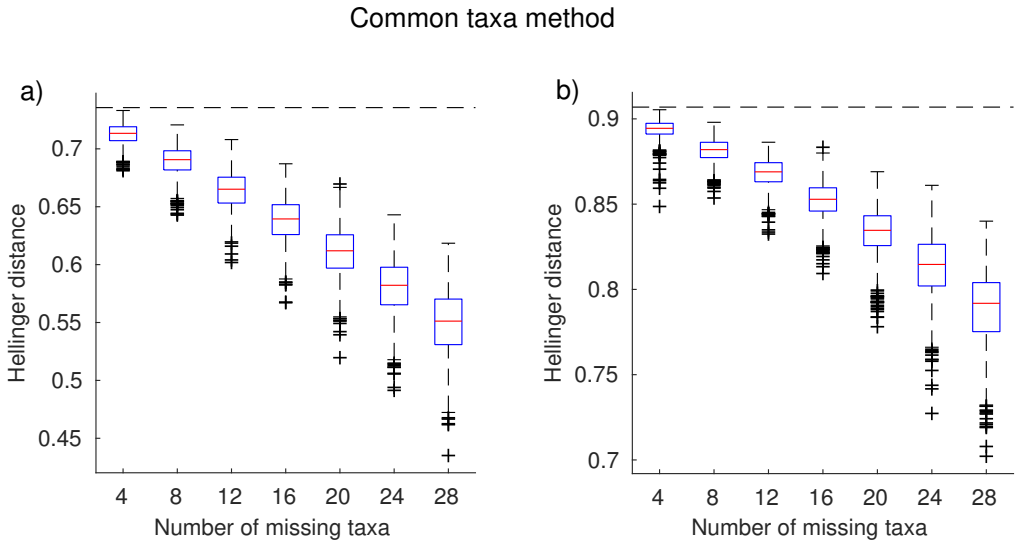


Figure 4.12: Sampling distribution of Hellinger distance between two phylogenetic trees for different levels of random deletions of taxa using common taxa method where a) The initial pair of phylogenies have the same 100 taxa with the same topology but different (random) edge lengths and b) Both phylogenies have 100 taxa with different (random) edge lengths, with one phylogeny generated at random and the other phylogeny determined using ten subsequent subtree pruning and regraft (SPR) operations. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

The figures show that as the number of missing taxa increases, the Hellinger distance increases towards its upper bound of 1 and the BHV metric decreases. The decrease in the BHV metric is more substantial in Figure 4.10 where both initial phylogenies have the same topology (before deletions). These trends were observed in several replicate experiments and in different sizes of phylogenetic trees; results are given in Figures 4.17 and 4.18. Overall these figures show a desirable property of using the augmentation method over the common taxa method, namely that distances between phylogenetic trees

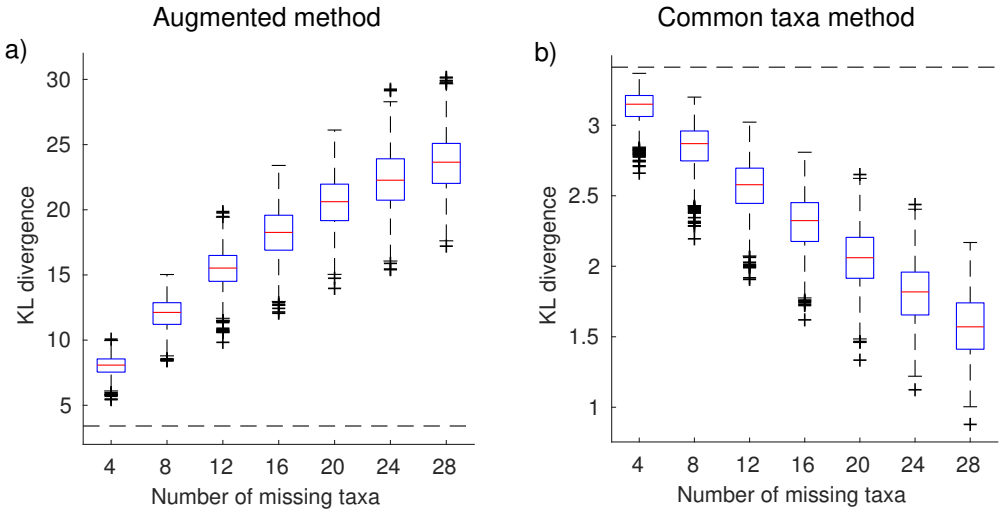


Figure 4.13: Sampling distribution of Kullback-Leibler divergence between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method, and b) common taxa method. The initial pair of phylogenies have the same 100 taxa with the same topology but different (random) edge lengths. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

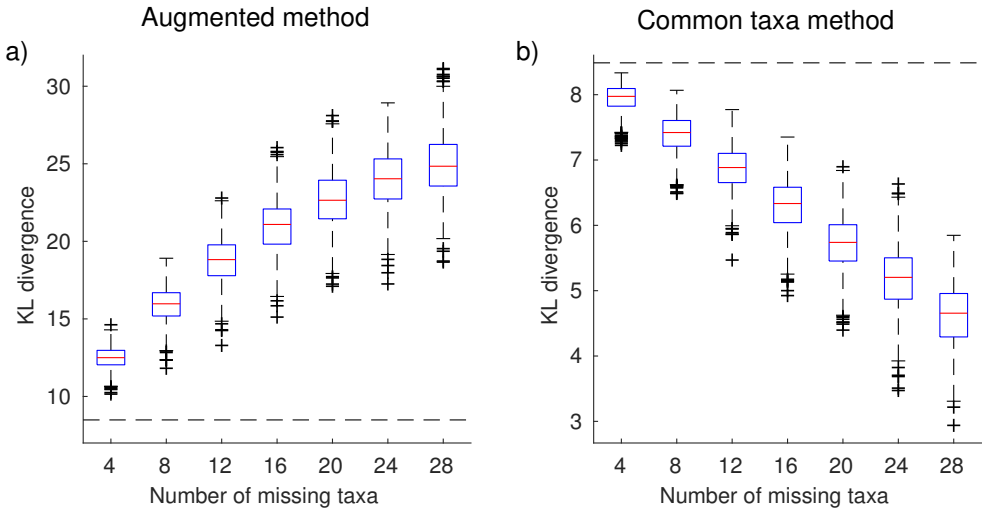


Figure 4.14: Sampling distribution of Kullback-Leibler divergence between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method, and b) common taxa method. Both phylogenies have 100 taxa with different (random) edge lengths, with one phylogeny generated at random and the other phylogeny determined using ten subsequent subtree pruning and regraft (SPR) operations. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

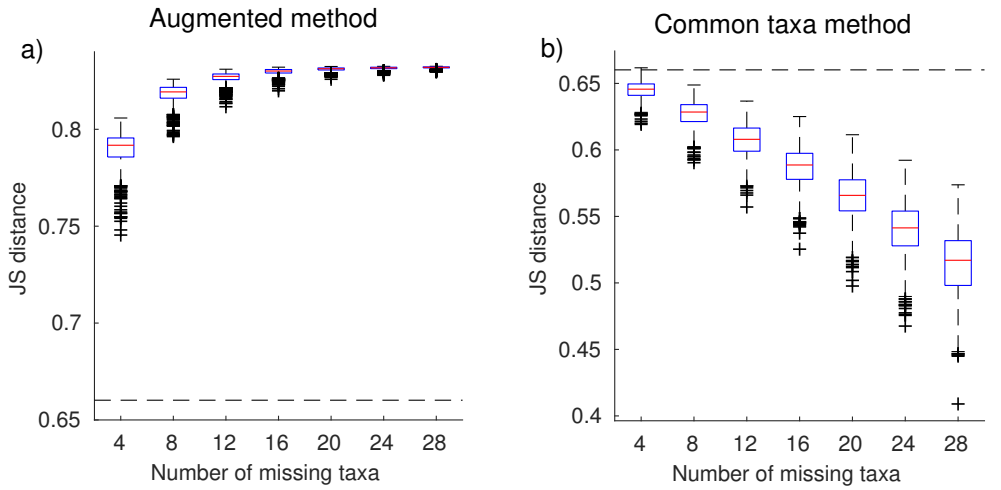


Figure 4.15: Sampling distribution of Jensen-Shannon distance between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method, and b) common taxa method. The initial pair of phylogenies have the same 100 taxa with the same topology but different (random) edge lengths. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

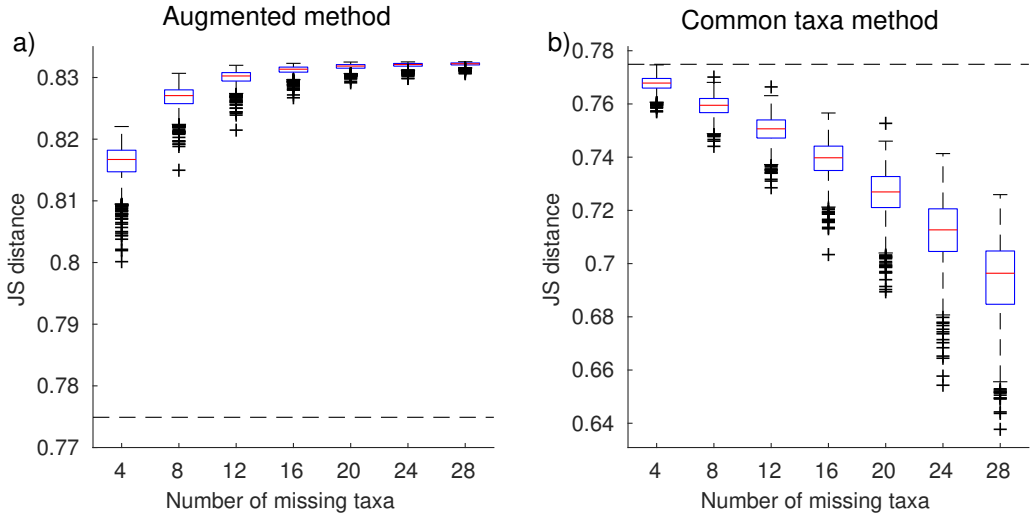


Figure 4.16: Sampling distribution of Jensen-Shannon distance between two phylogenetic trees for different levels of random deletions of taxa using a) augmented method, and b) common taxa method. Both phylogenies have 100 taxa with different (random) edge lengths, with one phylogeny generated at random and the other phylogeny determined using ten subsequent subtree pruning and regraft (SPR) operations. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

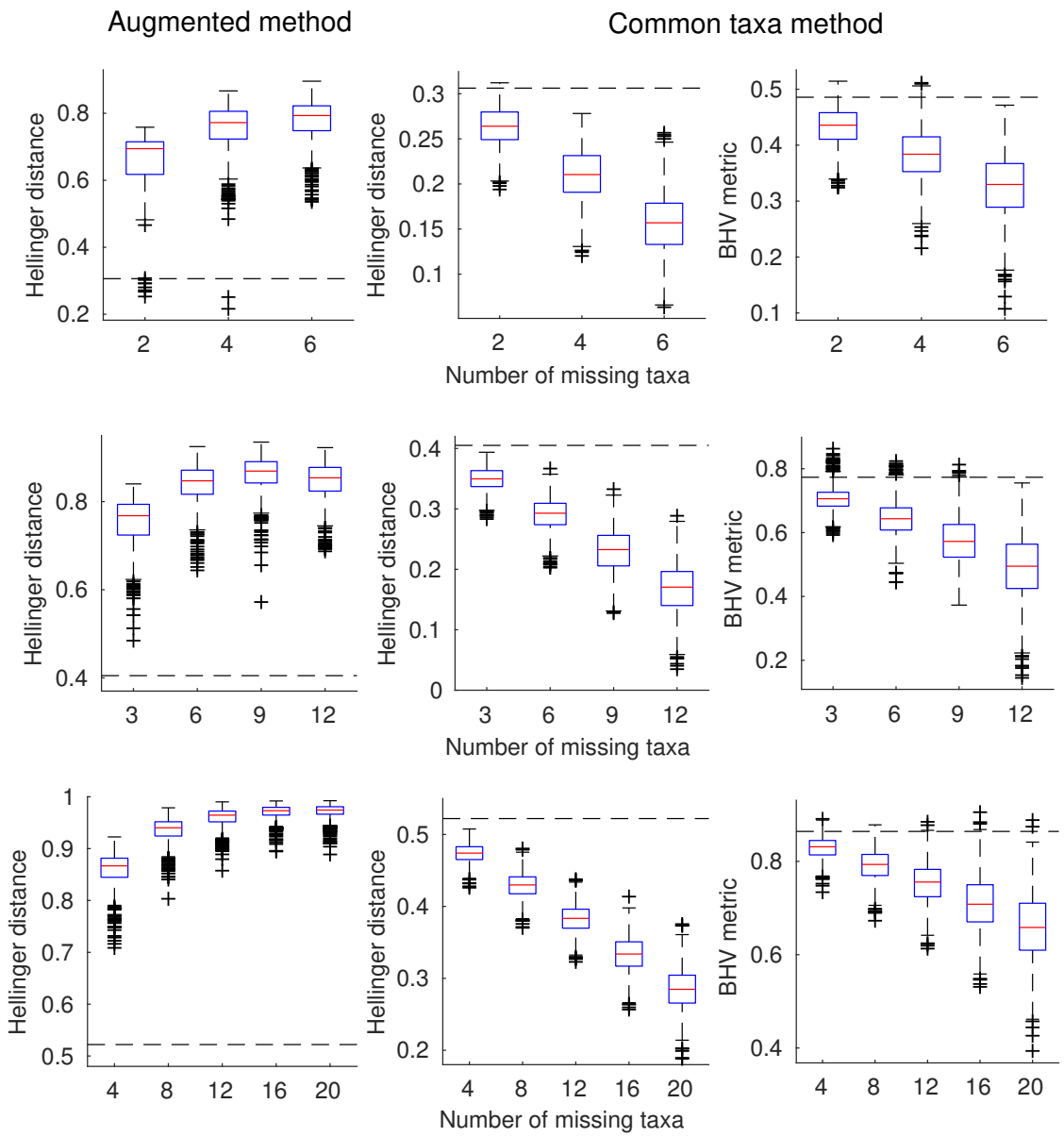


Figure 4.17: Sampling distribution of Hellinger distance between two phylogenetic trees for different levels of random deletions of taxa where the initial pair of phylogenies, 16, 25 and 50 taxa (row-wise), have the same topology but different (random) edge lengths. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

increase as we have more uncertainty about the phylogenies due to missing taxa. The probabilistic distance measures with common taxa method behave similarly to the BHV metric with the same method: results obtained with the Hellinger distance and common taxa method are given in Figure 4.12. Similar analysis was repeated using Kullback-Leibler divergence (see Figures 4.13 and 4.14) and Jensen-Shannon distance (see Figures

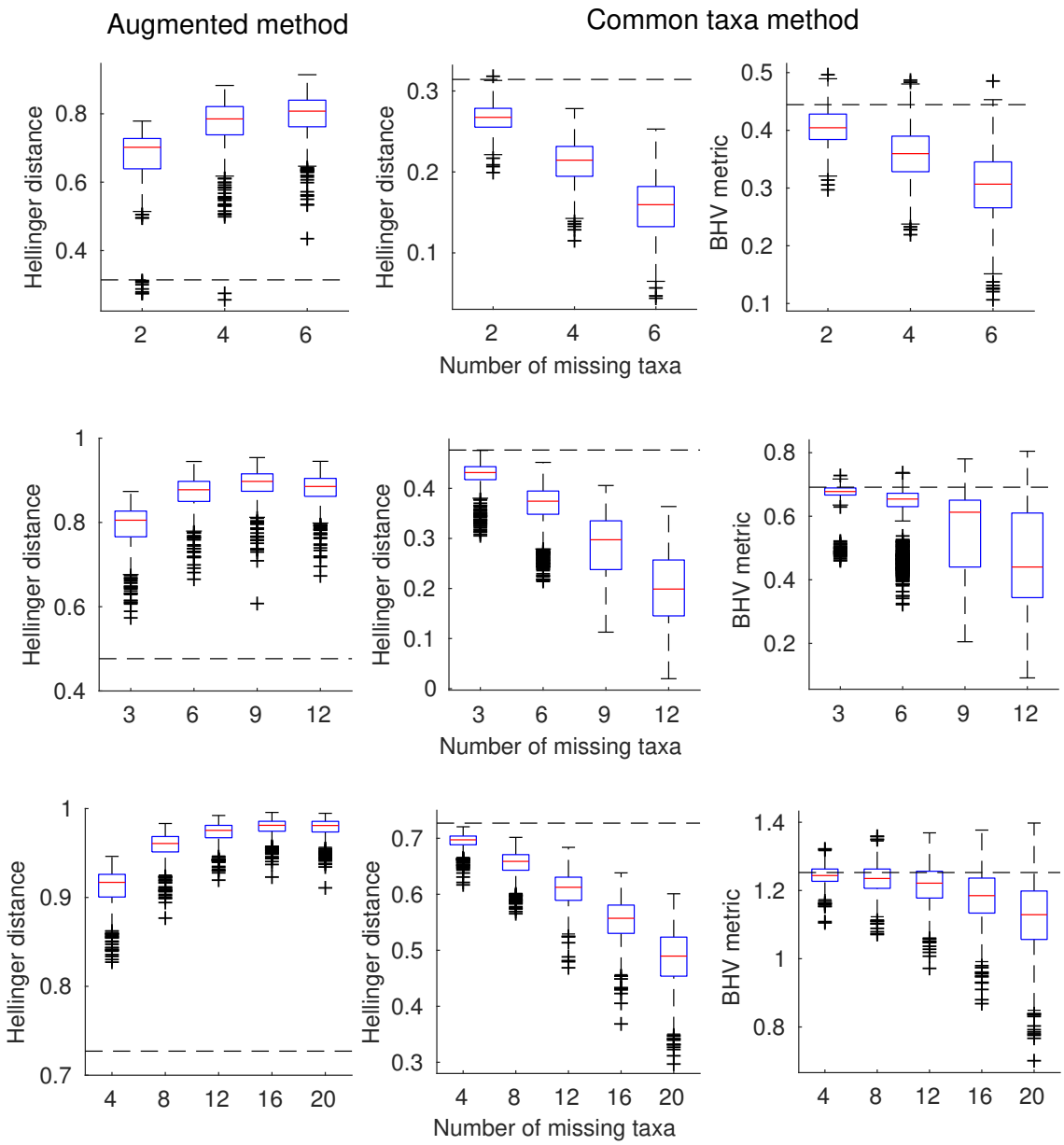


Figure 4.18: Sampling distribution of Hellinger distance between two phylogenetic trees for different levels of random deletions of taxa. The initial pair of phylogenies are: row 1 is 16 taxa and phylogeny topologies differ by randomly selected 2 SPRs, row 2 is 25 taxa and phylogeny topologies differ by randomly selected 3 SPRs and row 3 is 50 taxa and phylogeny topologies differ by randomly selected 5 SPRs. The dashed horizontal line is the distance between the initial pair of phylogenies (before any deletions).

4.15 and 4.16). We see that as the number of missing taxa increases, the Jensen-Shannon distance increases towards its upper bound of $\sqrt{\log 2}$.

4.7.4 Incorporating substitution model parameters

We investigated the distribution of distances calculated over biologically plausible phylogenetic trees and their substitution parameters for an experimental data set of primate DNA data (Huelsenbeck & Ronquist, 2001). We analysed the data set using the PHYML program (Guindon & Gascuel, 2003) assuming the GTR model with Gamma rate heterogeneity. This gave us the maximum likelihood (ML) phylogeny and its model parameters $\boldsymbol{\theta}_{ML}$ together with a set of 100 bootstrap replicates of phylogenetic trees T_i, each with their (ML) model parameters $\boldsymbol{\theta}_i$. For each pair of phylogenies in the bootstrap sample, we calculated the distance (Hellinger, TV, KL and JS) between phylogenies using the same (ML) model parameters, $d\{(T_i, \boldsymbol{\theta}_{ML}), (T_j, \boldsymbol{\theta}_{ML})\}$ and between phylogenies using the tree-parameter pairs, $d\{(T_i, \boldsymbol{\theta}_i), (T_j, \boldsymbol{\theta}_j)\}$. We will see how changes in the parameters $\boldsymbol{\theta}_i$ can change the distance between pairs of trees. Pairwise plots of these measures are given in Figure 4.19. It is clear that the distances are nearly always increased when taking proper account of the substitution parameter values.

4.7.5 Clustering

Gori *et al.* (2016) applied different combinations of metrics and clustering algorithms to cluster a data set of gene trees from 344 loci on 18 yeast species. We consider the same data set, but we remove all loci with at least one edge length greater than 1 leaving a total of 229 loci. This is due to the presence of gene trees with very long edge lengths which are potential outliers. In contrast to Gori's approach, we use Hellinger and BHV distances both with spectral clustering algorithm for our cluster analysis. First, we construct a symmetric non-negative matrix $D = (d_{ij})$ describing the distance between every pair of trees in this data set, such that d_{ij} is the distance between tree i and tree j. The data set is partitioned into k clusters by applying the spectral clustering algorithm (Ng *et al.*, 2001) to the distance matrix D. In this algorithm, the relationship between data points is represented as a graph which is described by a Laplacian matrix. The algorithm works by embedding the data points into a subspace of k largest eigenvectors of the Laplacian matrix, then applying k-means clustering algorithm on the embedded points to form clusters. Figure 4.20 is a visualisation of the distribution of the 229 loci using multidimensional scaling on the matrix $D = (d_{ij})$, where d_{ij} is the Hellinger distance in Figure 4.20a, and the BHV metric in Figure 4.20b. In each figure, three clusters obtained by spectral clustering are shown, with each cluster indicated by a different colour: cluster 1 (black) is the largest cluster with 137 loci and 99 loci from Figure 4.20a and 4.20b

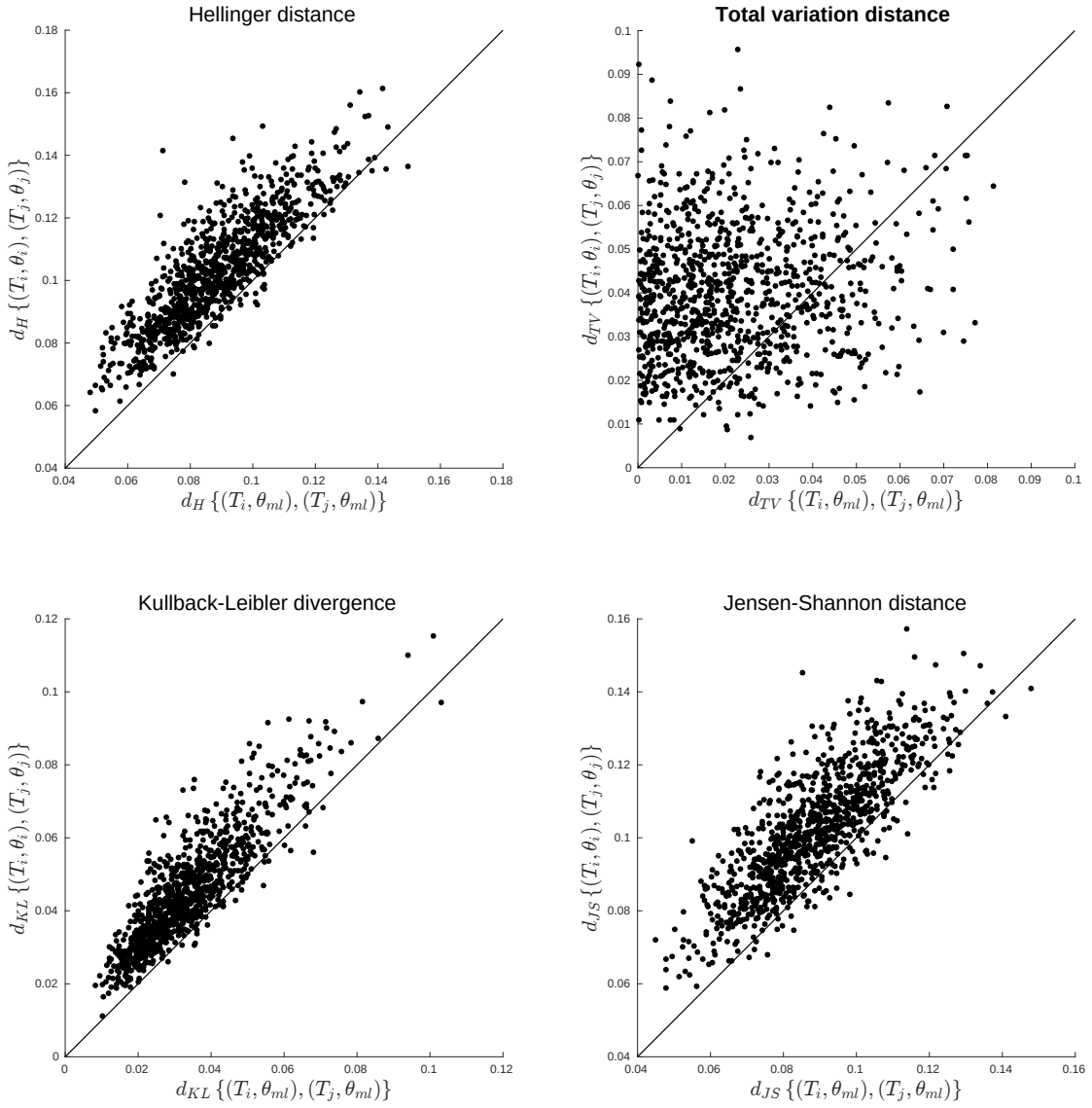


Figure 4.19: Comparison of distance (Hellinger, TV, KL and JS) between pairs of phylogenetic trees T_i and T_j using overall ML substitution parameters $\boldsymbol{\theta}_{ML}$ with that using individual ML substitution parameter $\boldsymbol{\theta}_i$. The phylogenetic trees used are maximum likelihood phylogenies $T_1, \ldots, T_{100}$ obtained from 100 bootstrap replicates of the primate data set.

respectively, cluster 2 (red) consists of 81 loci and 65 loci, and the remaining 11 loci and 65 loci belong to cluster 3 (green). All loci in the smallest cluster of 4.20a are members of cluster 3 of 4.20b and also, members of the first cluster in 4.20b also belong to cluster 1 of 4.20a, with additional loci, 40 out of 65 from cluster 2 of 4.20b. The rest of the 25 loci in this cluster are found in cluster 2 of 4.20a. Despite similarities in cluster memberships between the two groups, we see differences in cluster sizes due to the characteristics of

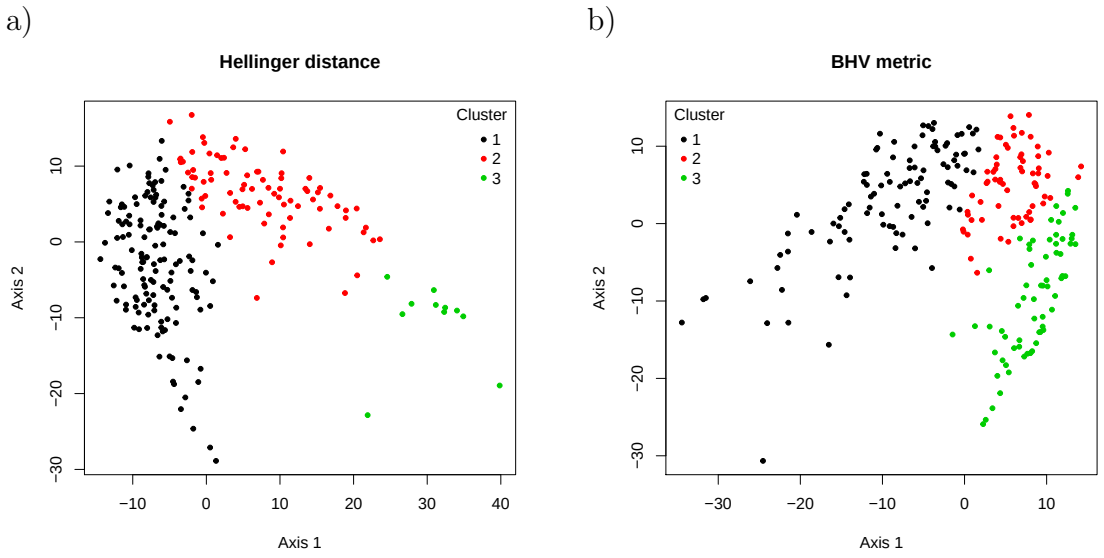


Figure 4.20: Visualisation of 229 loci on 18 yeast species using multidimensional scaling of a) Hellinger distance, and b) BHV metric. Three clusters were obtained by spectral clustering: cluster 1 (black) is the largest cluster with 137 loci and 99 loci from a) and b) respectively, cluster 2 (red) consists of 81 loci and 65 loci, and the remaining 11 loci and 65 loci belong to cluster 3 (green).

the metrics involved. While the 11 loci in cluster 3 of Figure 4.20a are separated from other loci in terms of Hellinger distance, they remain close to other members in cluster 3 of Figure 4.20b using the BHV metric, hence belong to the same cluster. These results are inconclusive in the sense that it is not clear what clusters mean with either metric. This is likely due to the main conclusion following Gori's analysis: trees with long edge lengths arise as a result of "data collection errors" and correspond to errors labelling genes or species in the original data set. These trees are easy to identify with almost any metric. Beyond that, clustering on this data set does not produce biologically meaningful clusters. As a consequence, we were motivated to look at data sets with clearer biological interpretations and so considered phylogenetic islands, as described next.

4.7.6 Phylogenetic islands

The term *phylogenetic island* has been used to refer to modes in multimodal posterior distributions, especially when these modes correspond to distinct phylogeny topologies. In this section we study two data sets for which posterior samples have previously been found to contain distinct clusters of phylogenies when the samples are analysed with metrics

based on topological differences between phylogenies. We compute probabilistic distances between phylogenetic trees in posterior samples and perform multidimensional scaling (MDS) using these distances (Hillis *et al.*, 2005). This leads to contrasting probabilistic interpretations of the results for the two data sets.

The first data set consists of 1949 nucleotides from 27 tetrapod species (Hedges *et al.*, 1990). The alignment was analysed in MrBayes (Huelsenbeck & Ronquist, 2001) using the GTR model with Gamma rate heterogeneity. The analysis used a burn-in of 1 million iterations followed by another 1 million iterations, sampled every 1000 iterations, in order to obtain a posterior sample of 1000 phylogenetic trees T_i and their associated model parameters $\boldsymbol{\theta}_i$, $i = 1, \ldots, 1000$. The Hellinger distance was estimated for each pair $(T_i, \boldsymbol{\theta}_i), (T_j, \boldsymbol{\theta}_j)$, $i \neq j$, and these distances were analysed with MDS. The results are shown in Figure 4.21a. The second data set consisted of 1485 nucleotides from 17 dengue virus serotype 4 sequences (Drummond & Rambaut, 2007). This alignment was analysed using a GTR model with Gamma rate heterogeneity and invariant sites using an uncorrelated lognormal distributed relaxed molecular clock. The BEAST software was used to perform the analysis (Drummond & Rambaut, 2007), using an xml file provided with the sofware, and 500 pairs $(T_i, \boldsymbol{\theta}_i)$ were sampled from the posterior. Figure 4.21b shows the results of applying MDS to the Hellinger distances between these pairs.

Previous analyses of these data sets have revealed distinct clusters in posterior samples when distances are measured using topological information alone. Whidden & Matsen (2015) found clusters in the tetrapod phylogenies using the subtree prune-and-regraft (SPR) metric. For the dengue fever sequences, Kendall & Colijn (2015) considered a family of metrics parametrized by $\lambda \in [0, 1]$. In the case $\lambda = 0$ the metric retains only topological information, and using this metric to perform MDS reveals several distinct clusters in the posterior sample (see Figure 4.21c), described by Colijn and Kendall as phylogenetic islands. MDS with the unweighted Robinson-Foulds metric gives similar results for this data set.

The MDS results obtained using the Hellinger distance differ for the two data sets. For the tetrapod data set, the MDS plot shows distinct clusters of phylogenies. The three clusters correspond to distinct topological regions in tree space. On the other hand, MDS for the probabilistic distances between dengue fever phylogenies did not reveal any clusters in the posterior sample, as shown in Figure 4.21b. The clusters obtained with the Kendall-Colijn metric as shown in Figure 4.21c do not correspond to separate regions in this plot. As seen in previous examples in this chapter, two phylogenies with different topologies can induce similar distributions on sequence data for particular choices of

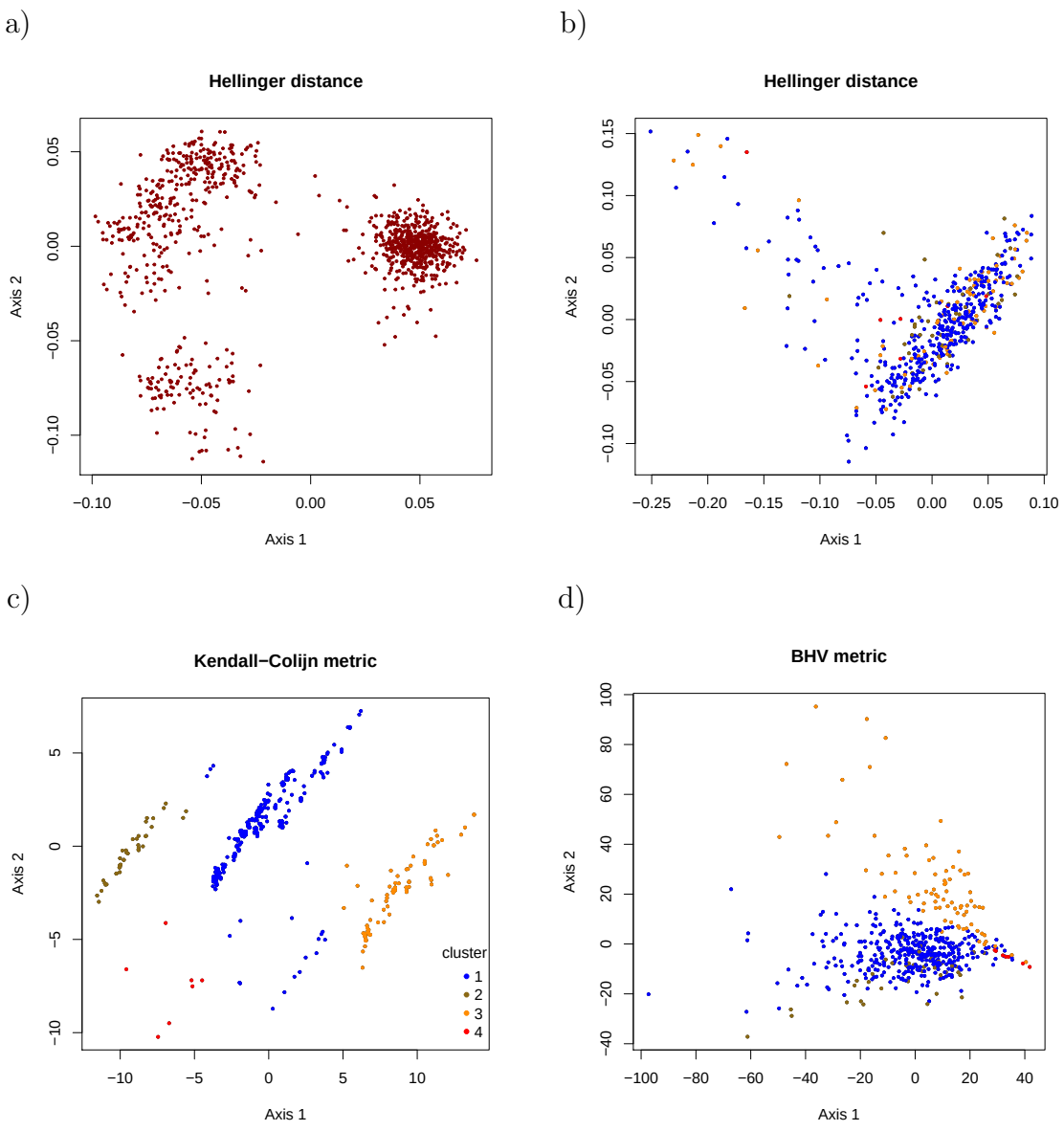


Figure 4.21: Multidimensional scaling of the pairwise a) Hellinger distance between posterior sample of 1000 phylogenetic trees from the tetrapod data set under GTR+Γ model, b) Hellinger distance between posterior sample of 500 phylogenetic trees from dengue fever data set under GTR+Γ+I substitution model with uncorrelated lognormal-distributed relaxed molecular clock, c) Kendall Colijn metric (with $\lambda = 0$) between posterior sample of 500 phylogenetic trees from dengue fever data set, and d) BHV metric between posterior sample of 500 phylogenetic trees from dengue fever data set. Clusters obtained in b) were indicated by the same color in c) and d).

edge lengths and substitution model parameters. The same phenomenon is at play for

the dengue fever phylogenies: although the posterior sample contains distinct clusters of topologies, phylogenies in different clusters are in fact giving rise to similar distributions of nucleotides. The interpretation of the results is therefore different in the two cases. First, for the tetrapod data, it appears that a single phylogeny together with the GTR model and Gamma rate heterogeneity is not able to explain the information in the sequence alignment. One possibility is that the substitution model is mis-specified and a more sophisticated model is required; a second is that the data have arisen from a non-tree-like process, such as a mixture of phylogenies. Secondly, for the dengue fever data, it appears that several distinct groupings of topologies are consistent with the data, but regarding the phylogenies as probability models, these groupings lack meaning as phylogenies in different clusters represent similar distributions on characters. If more sequence data were available, and under the assumption that these sequences were generated by the same evolutionary process, we would expect the single cluster in Figure 4.21b to become tighter, and correspondingly, for the variability in topology in the posterior sample to be reduced.

4.7.7 Computing times

The time taken to estimate distances depends on the sample size and hence on the degree of accuracy required by the user. The time taken to compute all 5565 distances for the yeast data in Figure 4.1 was 3 minutes. Similarly, the time taken to compute all 4950 distances between phylogenetic trees in the primate bootstrap sample using the GTR model with Gamma rate heterogeneity was 151 minutes. In both cases, the sample size was estimated to achieve a relative error of $\alpha = 5\%$ with probability $1 - \beta = 80\%$. Calculations were performed using a desktop computer with an Intel(R) Core i7-4790S processor running at 3.20HGz.

Chapter 5

Information geometry in tree space

In this chapter, the idea is to explore information geometry as defined and described in Chapter 3, but in the context of tree space rather than on a manifold. While Chapter 4 gave some novel information-based distance metrics on tree space, no construction of geodesics was given. We therefore aim to explore information geometry on tree space by constructing geodesics. This is achieved by solving the geodesic differential equation (3.15) numerically within orthants. We work largely with 5-taxon unrooted phylogenetic trees. We consider three motivating ideas:

1. Do information geodesics resemble BHV geodesics?
2. Do information geodesics ever pass through infinity? (i.e. through phylogenetic trees with infinitely long edges).
3. Can computations be performed sufficiently quickly that the geometry can be used in practical applications?

The idea of geodesics passing through infinity or more precisely, which contain phylogenetic trees with infinitely long edges, arises from the example seen in Section 4.7.1: as the edge lengths on two phylogenetic trees become increasingly long, the BHV metric between them increases linearly. However, the probabilistic distance between them reduces to zero because both phylogenetic trees induce similar distribution of characters. It is therefore possible that the information geodesic between two phylogenetic trees might pass through phylogenetic trees with infinitely long edge lengths rather than through BHV tree space (in which all edges have finite length). We deal with this "boundary at infinity" more formally in the next chapter, with a less formal approach here, looking for evidence for the existence of geodesics through infinity.

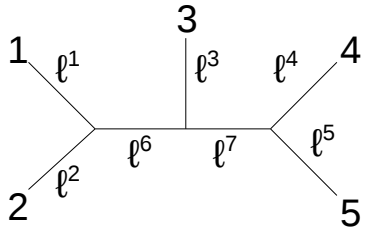


Figure 5.1: Unrooted phylogeny with 5 leaves labelled $1, \ldots, 5$ and edge lengths $\ell^1, \ldots, \ell^7$.

5.1 Fisher information metric on an orthant

Consider an unrooted phylogenetic tree on n leaves labelled $1, \ldots, n$ and some fixed fully resolved topology. Throughout this chapter we work entirely with the symmetric two-state Markov model. The $2n-3$ edge lengths are denoted $\boldsymbol{\ell} = (\ell^1, \ldots, \ell^{2n-3})$. In the notation of Chapter 3, the orthant corresponding to the fixed topology forms a parameter space

$$\mathcal{S} = \{\boldsymbol{\ell} = (\ell^1, \ldots, \ell^{2n-3}) \in \mathbb{R}^{2n-3} : \ell^i \geq 0\}.$$

The distribution $p(\cdot|\boldsymbol{\ell})$ on Ω^n ($\Omega = \{0, 1\}$) induced by the phylogeny is given by Equation (2.6). (Note that the vector of model parameters $\boldsymbol{\theta}$ is empty for the two-state symmetric model). If $\boldsymbol{u}$, $\boldsymbol{v}$ are tangents vectors at the point $\boldsymbol{\ell}$, then the inner product between $\boldsymbol{u}$ and $\boldsymbol{v}$ in the tangent space at $\boldsymbol{\ell}$ is given by

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{\ell}} = g_{ij}(\boldsymbol{\ell}) u^i v^j, \quad \text{and} \quad \|\boldsymbol{u}\|^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\boldsymbol{\ell}},$$

where

$$g_{ij}(\boldsymbol{\ell}) \;=\; \sum_{\boldsymbol{s} \in \Omega^n} p(\boldsymbol{s}|\boldsymbol{\ell}) \frac{\partial \log p(\boldsymbol{s}|\boldsymbol{\ell})}{\partial \ell^i} \frac{\partial \log p(\boldsymbol{s}|\boldsymbol{\ell})}{\partial \ell^j},$$

for $i,\ j = 1, \ldots, 2n-3$. Defining $\langle \partial_i, \partial_j \rangle = g_{ij}$ with $\partial_i \stackrel{\text{def}}{=} \frac{\partial}{\partial \ell^i}$ determines the Fisher information metric $g = \langle\ ,\ \rangle$ on the family of probability distributions on Ω^n.

For speed of calculation and for ease of visualization we fix $n = 5$ at this point, so that there are $2n - 3 = 7$ edges in total. We work with the fixed topology shown in Figure 5.1: ℓ^6 and ℓ^7 are the internal edge lengths while $\ell^1, \ldots, \ell^5$ are the pendants.

5.2 Geodesics

We view $\mathcal{S}$ as a Riemannian manifold and compute geodesics in it in the standard way. The geodesic equation (3.15) becomes

$$\frac{d^2\ell^k}{dt^2} + \Gamma^k_{ij}\frac{d\ell^i}{dt}\frac{d\ell^j}{dt} = 0, \tag{5.1}$$

for each $k = 1, \ldots, 7$ and $i, j = 1, \ldots, 7$. We reduce the equation to first order ordinary differential equations

$$\begin{aligned} \frac{d\ell^k}{dt} &= f^k(t, \ell^1, \ldots, \ell^7, v^1, \ldots, v^7) = v^k \\ \frac{dv^k}{dt} &= h^k(t, \ell^1, \ldots, \ell^7, v^1, \ldots, v^7) = -\Gamma^k_{ij} v^i v^j. \end{aligned}$$

Given initial conditions, namely a vector of edge lengths $(\ell^1_0, \ldots, \ell^7_0)$ and a velocity vector $(v^1_0, \ldots, v^7_0)$, the system of ordinary differential equations were solved numerically using the Runge-Kutta fourth order (rk4) method of the form

$$\begin{aligned} c^k_0 &= \delta t\, f^k(t_n, \ell^1_n, \ldots, \ell^7_n, v^1_n, \ldots, v^7_n), \\ d^k_0 &= \delta t\, h^k(t_n, \ell^1_n, \ldots, \ell^7_n, v^1_n, \ldots, v^7_n), \\ c^k_i &= \delta t\, f^k\left(t_n + \frac{1}{2}\delta t, \ell^1_n + \frac{1}{2}c^k_{i-1}, \ldots, \ell^7_n + \frac{1}{2}c^k_{i-1}, v^1_n + \frac{1}{2}d^k_{i-1}, \ldots, v^7_n + \frac{1}{2}d^k_{i-1}\right), \\ d^k_i &= \delta t\, h^k\left(t_n + \frac{1}{2}\delta t, \ell^1_n + \frac{1}{2}c^k_{i-1}, \ldots, \ell^7_n + \frac{1}{2}c^k_{i-1}, v^1_n + \frac{1}{2}d^k_{i-1}, \ldots, v^7_n + \frac{1}{2}d^k_{i-1}\right), \quad i = 1, 2, 3. \end{aligned}$$

$$\begin{aligned} \ell^k_{n+1} &= \ell^k_n + \frac{1}{6}\left(c^k_0 + 2c^k_1 + 2c^k_2 + c^k_3\right), \\ v^k_{n+1} &= v^k_n + \frac{1}{6}\left(d^k_0 + 2d^k_1 + 2d^k_2 + d^k_3\right), \end{aligned} \tag{5.2}$$

where $\delta t = (t_N - t_0)/N$ is the step-size which depends on initial time t_0 and final time t_N, with N as the required number of steps. The algorithm for computing the geodesic within a single orthant is as follows:

Set $t_0 = 0$ and $t_N = 1$. Choose a required number of steps N, initial vector of edge lengths $(\ell^1_0, \ldots, \ell^7_0)$ and initial vector of directions $(v^1_0, \ldots, v^7_0)$. For $n = 0, \ldots, N-1$:

1. Compute ℓ_{n+1}^k and v_{n+1}^k $(k = 1, \ldots, 7)$ using (5.2) given ℓ_n^k and v_n^k.

2. If $\ell_{n+1}^k \leq 0$, for some k:

   (a) if it is a pendant edge, reset $\ell_{n+1}^k = 0$ and if $v_{n+1}^k < 0$ correspondingly, reset $v_{n+1}^k = 0$;

   (b) if it is an internal edge, terminate the algorithm, since this is the orthant boundary.

The output of the algorithm is a set of vectors of edge lengths and directions for each time step $t_{n+1} = t_n + \delta t$.

We consider examples of geodesics in a single orthant corresponding to a fixed topology (Figure 5.1) for different initial conditions. Figure 5.2 shows geodesics (blue lines) in several orthants fired in different directions from a fixed initial tree (at the centre). Pendant edge lengths on the initial tree are all the same and fixed at $0.1, 0.25, 0.5, 1.0$ respectively (row-wise), and all their initial directions are zero. The red lines are contours of distance travelled. The geodesics are not the same as BHV geodesics which are straight lines radiating out at even-angular interval with regularly spaced contours of distance. Here we have some curved geodesics with irregularly spaced non-circular contours of distance. Contours stack up towards the origin while the spacing between them increase as geodesics shoot out to infinity. This is more obvious when (initial) internal edges are long (see first column of Figure 5.2). Overall, this suggests that shortest paths between different orthants might sometimes correspond to moving through the boundary at infinity rather than contracting edges to length zero like BHV geodesics. However, geodesics tend to resemble BHV geodesics when pendant edge lengths are large; see last row of Figure 5.2.

The pendant edges do not behave as they do in BHV space. First, they can change value even when the initial velocity is zero for each pendant edge. In BHV tree space, the pendants do not change length if the phylogenies at either end of the geodesic have the same pendant edge lengths. In the information geometry, pendants have the potential to become zero along the geodesic, and in fact can adopt small negative values as the differential equations are integrated numerically. Small black circles along some geodesics in Figure 5.2 indicate that at least one pendant edge has gone negative or zero. In Figure 5.3, we show how edge lengths change along certain geodesics by plotting edge length versus time for some geodesics in Figure 5.2. Here time is proportional to distance along each geodesic. Figures a) and c) provide evidence that a point on the boundary

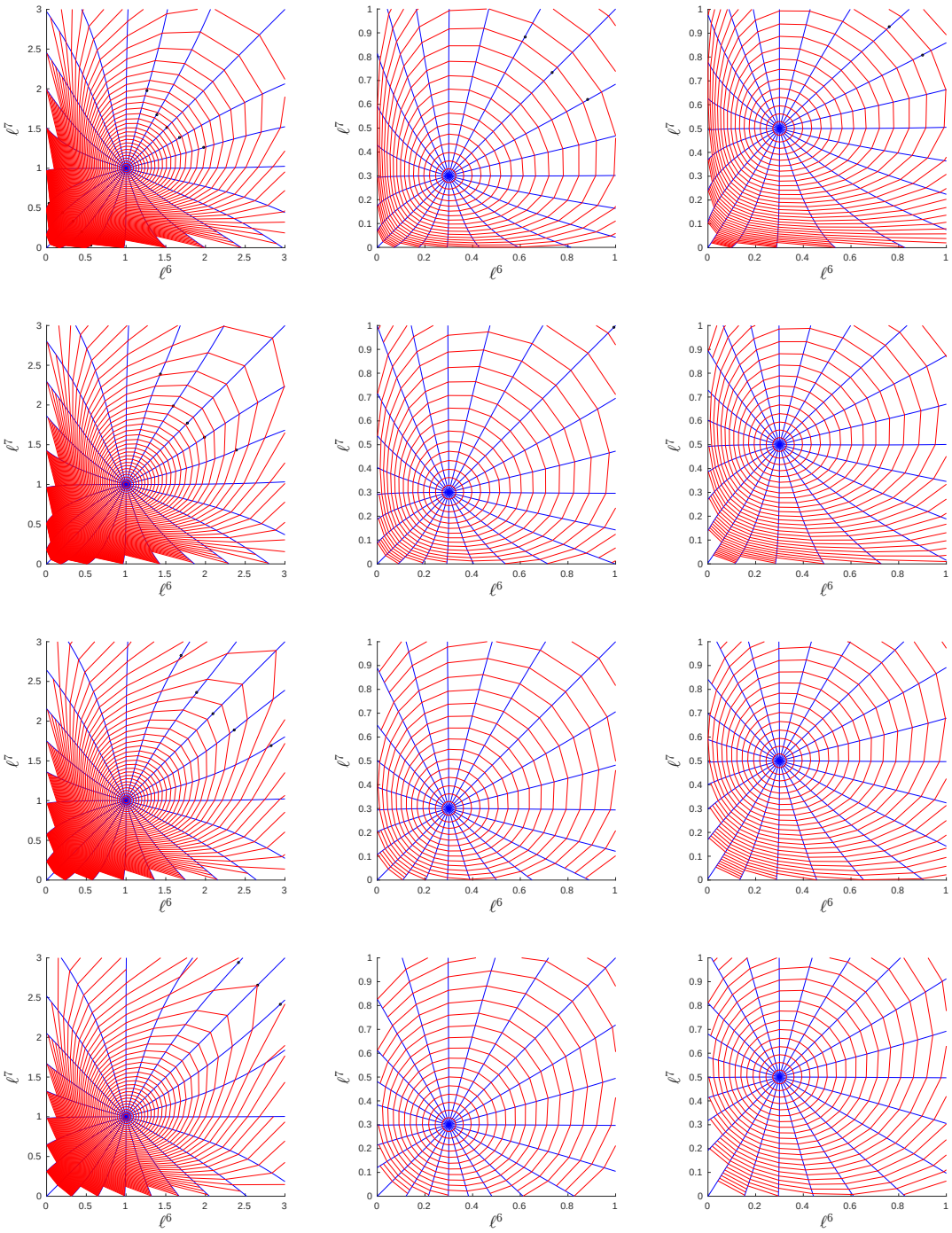


Figure 5.2: Geodesics (blue) and contours of distance (red) within a single orthant. Geodesics were fired from the central tree in each case. Rows correspond to different initial pendant edge lengths: $\ell^i = 0.1, 0.25, 0.5, 1.0$ (each fixed for all pendants $i = 1, \ldots, 5$) respectively. Columns correspond to different initial values for the internal edge lengths ℓ^6 and ℓ^7. The blue lines show the initial velocity vector at the starting point, which was in each case zero for the pendant edges. In the text, plot (i, j) refers to the ith row and jth column of this figure, counting down from the top.

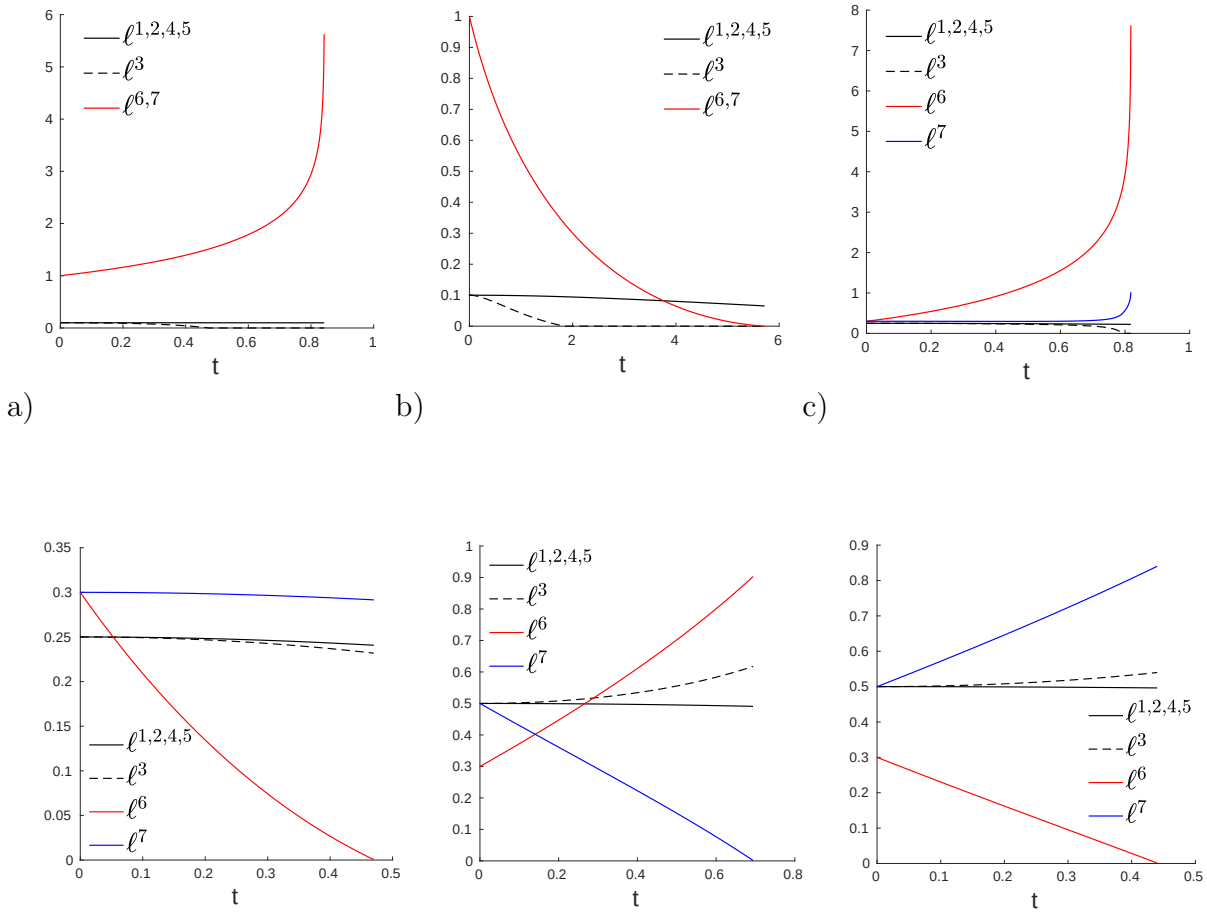
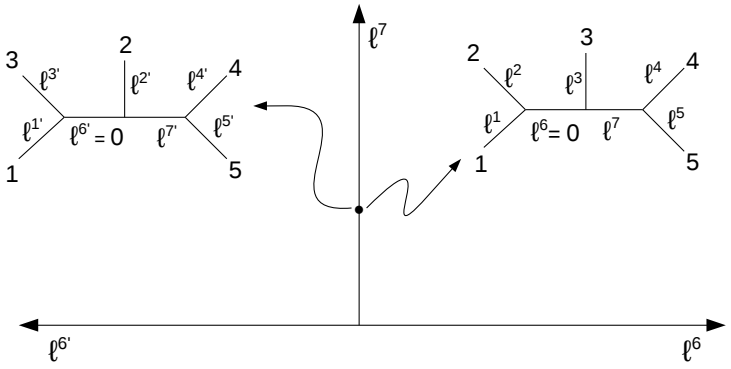


Figure 5.3: Edge lengths versus time along some geodesics in Figure 5.2. a) Geodesic heading in NE compass direction in plot(1,1). b) Geodesic heading SW in plot(1,1). c) Geodesic heading East in plot(2,2). d) Geodesic heading West in plot(2,2). e) Geodesic heading SE in plot(3,3). f) Geodesic heading NW in plot(3,3).

at infinity can be reached in finite time, or equivalently, is a finite distance away from the initial tree. This suggests that information geodesics can indeed pass through the boundary at infinity, which is a very significant difference from the BHV geodesics. The plots also show how the BHV boundaries (i.e. $\ell^6 = 0$ or $\ell^7 = 0$) are approached relatively slowly; see figures b), d), e) and f), corresponding to contours stacking up in Figure 5.2. The BHV boundaries are a finite distance from the starting tree in each case, and are eventually crossed. The cone point (origin) does not appear to be "attractive": there is no evidence of geodesics with different initial directions being pulled to pass through the origin.

We want to extend calculation of geodesics from the initial orthant into the neighbouring orthants. To do this, we extend the algorithm given above. The idea is that when we hit a codimension-1 boundary, we continue into either of the two possible neighbouring orthants arbitrarily, maintaining the direction vector across the boundary. In other words, a particular choice of orthant must be made when a boundary is crossed. Symmetry considerations show that the vector of edge lengths traced out by the two choices are exactly the same; it is simply that the tree topology is different in the two orthants. The extended algorithm does not terminate when an internal edge length becomes negative or zero, thus Step 2(b) of the algorithm becomes: if $\ell^k_{n+1} \leq 0$ for some k for an internal edge, (i) choose a neighboring orthant to expand out a replacement edge, (ii) perform a nearest neighbor interchange (NNI) operation on ℓ^k_{n+1} in order to cross to the chosen orthant. Reset $\ell^k_{n+1} = 0$ and direction $v^k_{n+1} = -v^k_{n+1}$. Reset other edge lengths and directions so that they are consistent with the new orthant; see Figure 5.4 for an example. Figure 5.5 shows geodesics starting from a single orthant and crossing over the BHV boundary to a neighboring othant. We can continue into other orthants. However, joining two arbitrary points in the space is very hard as we don't know which orthants to go through, and also the direction to the end point.



$$\boldsymbol{\ell} = (\ell^1, \ell^2, \ell^3, \ell^4, \ell^5, 0, \ell^7) \quad \boldsymbol{v} = (v^1, \ldots, v^7)$$

$$\boldsymbol{\ell}' = (\ell^1, \ell^3, \ell^2, \ell^4, \ell^5, 0, \ell^7) \quad \boldsymbol{v}' = (v^1, v^3, v^2, v^4, v^5, -v^6, v^7)$$

Figure 5.4: Illustration of nearest neighbor interchange operation on ℓ^6. When $\ell^6 = 0$, we cross a BHV boundary by resetting certain edge lengths and directions based on the new orthant. The new set of edge length and direction is given by $\boldsymbol{\ell}'$ and $\boldsymbol{v}'$.

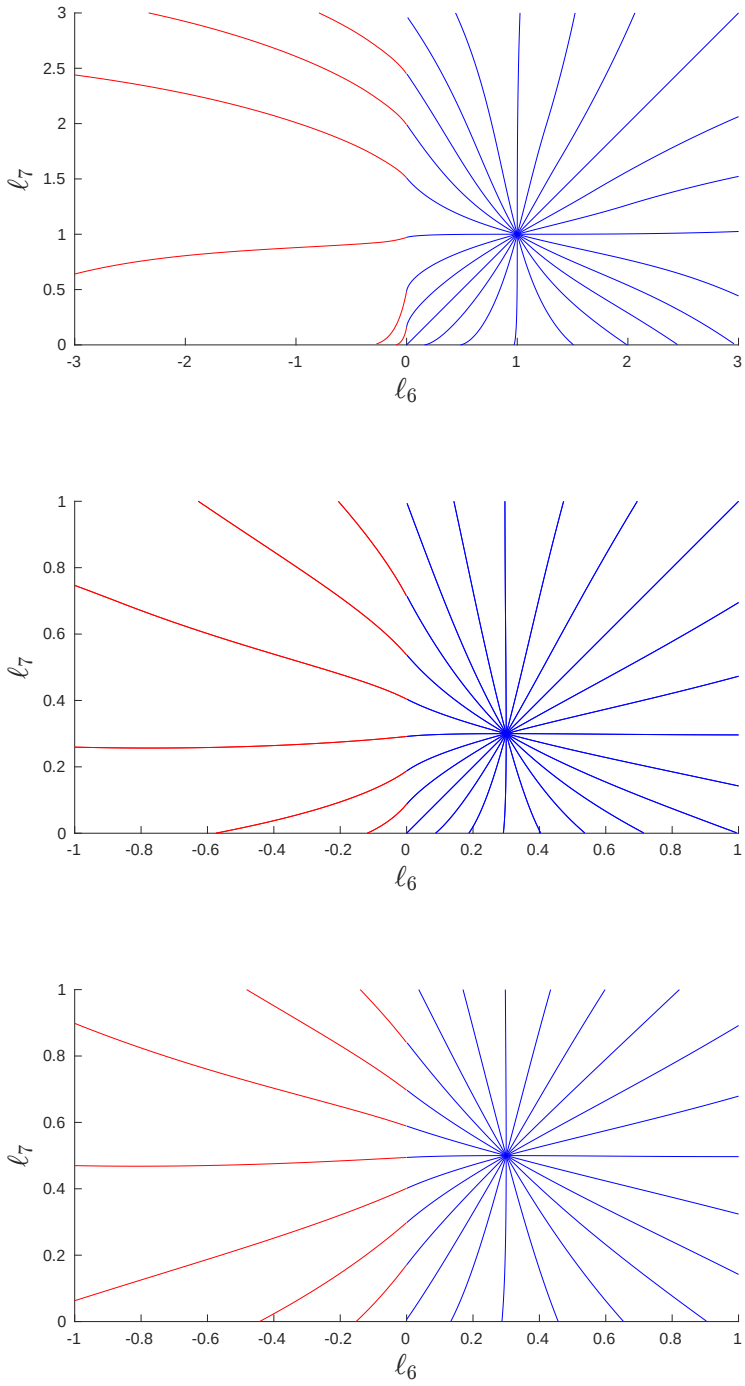


Figure 5.5: Geodesics beyond the BHV boundary within two neighboring orthants. Geodesics were fired from the central tree in each case represented with initial values for the internal edge lengths ℓ^6 and ℓ^7. Pendant edge lengths on the initial tree are all the same and fixed at 0.1, 0.25, 0.5 respectively (row-wise). The initial velocity vector was zero in each case for the pendant edges. Blue curves correspond to geodesics in the initial orthant; red curves are where the geodesics extend into neighboring orthant. Negative coordinates are due to graphical representation but they actually refer to positive edge lengths.

5.3 Conclusion

We return to the three motivational ideas behind this chapter. Firstly, we have seen that, in general, information geodesics are different from BHV geodesics within any single orthant: geodesics can curve and the distance contours can either stack up or spread out relative to BHV. The stack up corresponds to edge lengths shrinking to length zero while the spread out is where edge lengths become infinitely long. Secondly, it appears that some geodesics achieve an infinite edge length but cover a finite distance. We have seen numerical evidence from our plots. This is a significant difference with BHV where geodesics want to go through the origin whereas here is the opposite. Finally, the computation time suggests that the computation is infeasible for large trees and thus the geometry is not suitable for practical application.

Chapter 6

Geometry on the edge-product space of phylogenetic trees via embedding in the space of covariance matrices

In the previous chapter we explored information geometry for 5-taxon unrooted phylogenies. This was achieved by numerically integrating the geodesic equation given initial conditions (a phylogeny and a velocity vector for the edge lengths). Calculating geodesics in this way is not only very slow computationally, but also based on "firing" geodesics in different directions. However, in practice, given any two points in tree space, we want to compute a geodesic between them, or rather "join" the points. This is what the Owen-Provan algorithm achieves in BHV tree space (Owen & Provan, 2011). In this Chapter, we show that there is a natural way to embed tree space in the space of positive semi-definite matrices, or covariance matrices. Geodesics in the space of covariance matrices can be computed analytically, hence through embedding in this space, our aim is to obtain approximate geodesics with respect to the induced geometry in tree space. We show the geodesics in the induced geometry are closely related to the information geodesics. We also consider the boundary at infinity more formally and provide a clear definition of edge-product space.

6.1 Formal definition of the edge-product space

The edge-product space was originally defined by Moulton & Steel (2004). The definition we give is rather different, and is structured so as to show how it generalizes the definition

of BHV tree space.

A forest is a disjoint union of trees. Let $\mathcal{L} = \{1, \ldots, n\}$ be a set of labels and E_n be the collection of forests satisfying the following conditions:

(i) Each forest contains exactly n leaves, bijectively labelled with $\mathcal{L}$. A leaf vertex is a vertex of degree 0 or 1.

(ii) Edges are weighted, taking values in $[0, 1]$. Edge lengths on elements of E_n are denoted with the symbol λ. They are related to BHV edge lengths via a transformation which we describe below.

(iii) There are no degree 2 vertices.

We impose an equivalence relation ($\sim$) on E_n. This is defined by the following two rules.

BHV boundary rule. Under this rule, a tree which contains an edge with length zero is equivalent to a tree in which the edge is removed, and the two vertices at either end are merged. The rule applies only to internal edges and not to pendants. The rule is defined in Figure 6.1a, and applies for all subtrees A, B, C, D and forests F. The F term in the figure indicates that the rule applies to every tree in a forest, not only when the forest is a single tree.

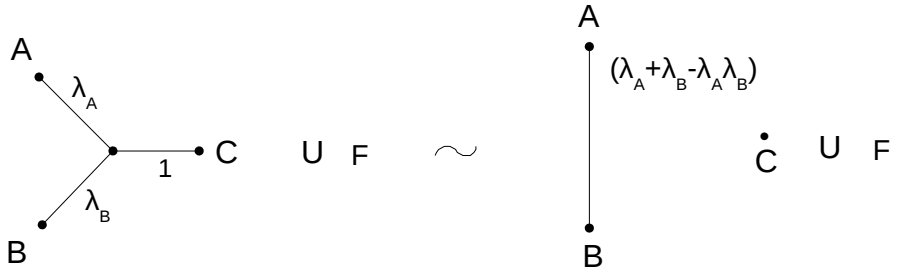
Infinity boundary rule. This states that having edge length 1 is equivalent to snapping the edge, then removing the stumps and any remaining degree 2 vertices. Figure 6.1b illustrates the infinity boundary rule. Unlike the BHV rule, which identifies a finite collection into a single equivalence class, this rule identifies infinitely many elements of E_n. This is because infinitely many combinations of edge lengths λ_A, λ_B give rise to the same edge length $\lambda_A + \lambda_B - \lambda_A\lambda_B$.

We denote the quotient space $E_n/\sim = \mathcal{E}_n$.

The formula for edge lengths in the rule for the boundary at infinity requires some justification. Given a BHV tree on $\mathcal{L}$, the corresponding element of $\mathcal{E}_n$ has the same topology, and edge weights

$$\lambda_e = 1 - \exp(-\ell_e) \tag{6.1}$$

a) BHV boundary



b) Infinity boundary

Figure 6.1: a) BHV boundary rule: zero length edge is equivalent to removing the edge but merging the two vertices at either end. b) Infinity boundary rule: edge length 1 is equivalent to removing the edge, then removing any degree 2 vertices. This is true for all subtrees A, B, C, D, and for all disconnected forests F.

for each edge e where ℓ_e is the BHV edge length. The boundary at infinity rule defined in Figure 6.1b maintains the BHV distance between vertices A and B under the equivalence. This distance is $\ell_A + \ell_B$ where ℓ_A and ℓ_B are the BHV lengths of the respective edges. Then the edge length in $\mathcal{E}_n$ is

$$\begin{aligned}\lambda_{AB} &= 1 - \exp(-\{\ell_A + \ell_B\}) \\ &= 1 - \exp(-\ell_A)\exp(-\ell_B) \\ &= 1 - (1-\lambda_A)(1-\lambda_B) \\ &= \lambda_A + \lambda_B - \lambda_A\lambda_B. \end{aligned} \tag{6.2}$$

Moulton & Steel (2004) worked with a similar parametrization of $\mathcal{E}_n$, in which edge lengths are defined by $\lambda^* = \exp(-\ell)$, that is, $\lambda^* = 1 - \lambda$, where ℓ is the BHV edge length. This has the disadvantage that the BHV boundary is at $\lambda^* = 1$ and the boundary at

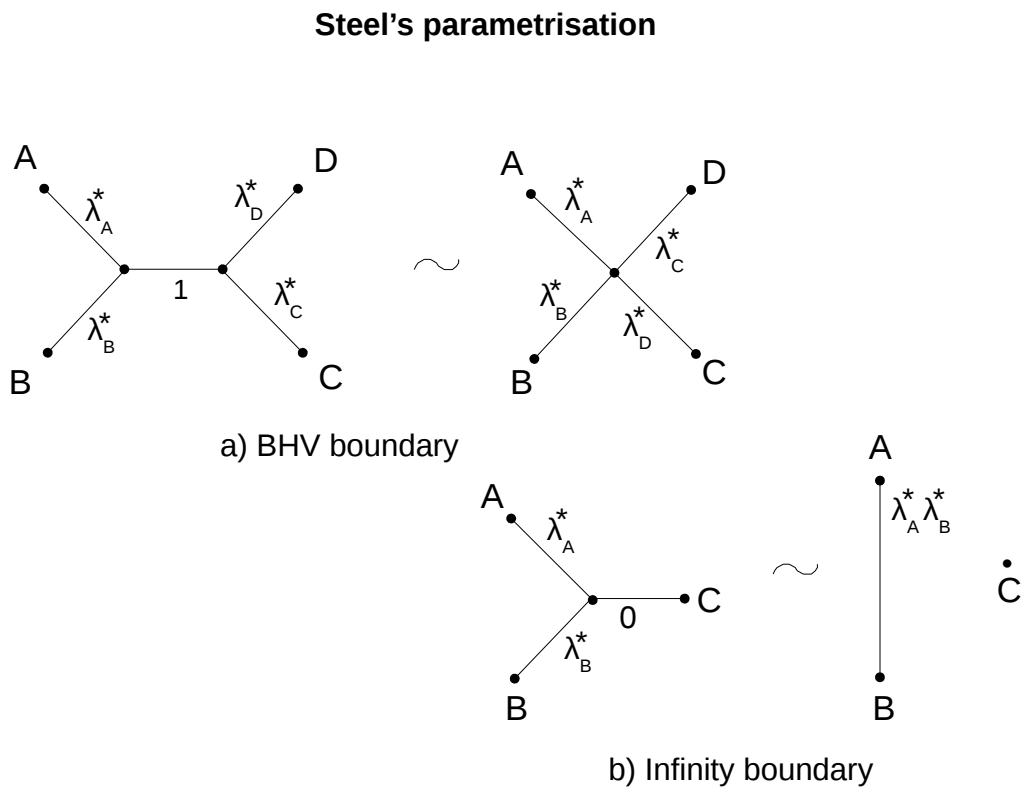


Figure 6.2: Steel's parametrization. a) BHV boundary at $\lambda^* = 1$ is equivalent to removing the edge but merging the two vertices at either end. b) Infinity boundary at $\lambda^* = 0$ is equivalent to snapping the edge then removing any degree two vertices. This happens for all subtrees A, B, C, D.

infinity is at $\lambda^* = 0$ as shown in Figure 6.2. However, the analog of (6.2) is $\lambda^* = \lambda_A^* \lambda_B^*$. This formula gives the space its name as the "edge-product" space. We prefer to work with the parametrization given first, despite the more complicated formula. This is because it is intuitive to think of BHV tree space as being formed from a set of unit cubes embedded in $\mathcal{E}_n$. The boundary at infinity corresponds to additional gluing rules on the faces of these cubes, in particular those faces with at least one unit length edge. It is easy to see how $\mathcal{E}_n$ is a compactification of BHV tree space in this way.

6.2 Embedding in the space of covariance matrices

Elements of $\mathcal{E}_n$ are naturally identified with distributions on $\{0,1\}^n$ induced by the two-state symmetric model on forests, in the following way. Consider an element of E_n which consists of a single tree. The Markov process on this tree and corresponding distribution on

characters have already been defined using the BHV edge lengths. When a forest consists of more than one component, the Markov process on each component is independent of the process on the other components. The distribution on characters is the same for any two points in E_n identified under the relation $\sim$. This is trivial to see for the BHV rule, and is true for the boundary at infinity because an edge with $\lambda = 1$ ($\ell = \infty$) results in independence between leaves on either side of the edge. We saw in Chapter 4 that the map from trees to distributions of characters under the two-state symetric model is injective, and it follows that this is the case for $\mathcal{E}_n$. It should be noted that Moulton & Steel (2004) first described the edge product space in terms of a parametrization of Markov models.

Suppose $T \in E_n$. The transition probability matrix over a certain edge with BHV length ℓ has entries given by

$$p_{ij}(\ell) = \begin{cases} \frac{1}{2}\{1 + \exp(-\ell)\}, & i = j \\ \frac{1}{2}\{1 - \exp(-\ell)\}, & i \neq j. \end{cases}$$

This becomes

$$p_{ij}(\lambda) = \begin{cases} 1 - \frac{1}{2}\lambda, & i = j \\ \frac{1}{2}\lambda, & i \neq j, \end{cases}$$

in our parametrization.

Let $i, j \in \mathcal{L}$ be any two leaf labels and let X_i, X_j be the random variables representing the letters at leaf i and j. If a path between i and j exists in T, we deduce the joint distribution for the associated random variables $X_i, X_j \in \{0, 1\}$ is

		X_i	
		0	1
X_j	0	$\frac{1}{4}\{1 + \exp(-\ell_{ij})\}$	$\frac{1}{4}\{1 - \exp(-\ell_{ij})\}$
	1	$\frac{1}{4}\{1 - \exp(-\ell_{ij})\}$	$\frac{1}{4}\{1 + \exp(-\ell_{ij})\}$

where ℓ_{ij} is the BHV length of the path between leaf i and j. Hence, since

$$E[X_i X_j] = \frac{1}{4}\{1 + \exp(-\ell_{ij})\} \quad \text{and} \quad E[X_i] = E[X_j] = \frac{1}{2},$$

we have

$$\mathrm{Cov}(X_i, X_j) = \frac{1}{4}\exp(-\ell_{ij}).$$

Under our parametrization (6.1), this becomes

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \frac{1}{4}\exp\left(-\ell_{ij}\right) \\ &= \frac{1}{4}\exp\Big(-\sum_{e\in(i,j)}\ell_e\Big) \\ &= \frac{1}{4}\prod_{e\in(i,j)}\exp\left(-\ell_e\right) \\ &= \frac{1}{4}\prod_{e\in(i,j)}(1-\lambda_e),\end{aligned}$$

where (i, j) is the path from leaf i to j. This is simply

$$\mathrm{Cov}(X_i, X_j) = \frac{1}{4}\prod_{e\in(i,j)}\lambda_e^*$$

under Steel's parametrization. However, if there is no path between i and j, then

$$\mathrm{Cov}(X_i, X_j) = 0. \tag{6.3}$$

Therefore, the covariance matrix associated with T is defined as

$$\Sigma(T) = (\Sigma_{ij}) \tag{6.4}$$

where $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Suppose $S^+(n, \mathbb{R})$ denotes the set of $n \times n$ real symmetric positive-definite matrices. In order to simplify notation, we ignore references to n and $\mathbb{R}$. The map $\Sigma : \mathcal{E}_n \to S^+$ is injective since two elements of E_n have the same covariance matrix if and only if they are related by $\sim$, and so we can think of $\mathcal{E}_n \subset S^+$ under this embedding. As we describe below, the space of covariance matrices S^+ is a non-positively curved Riemannian manifold. As a consequence, there exists a unique geodesic between every pair of points in S^+. We next describe the geometry of the space S^+, and then explore the geometry on $\mathcal{E}_n$ induced by the embedding.

6.3 Geometry of the space of covariance matrices

In this section, we describe the geometry in S^+. While these results have been established by other authors, we bring them together from a number of different sources in

the literature.

6.3.1 The multivariate normal model

Consider a statistical model S of multivariate normal distributions $N_n(\mathbf{0}, \Sigma)$ of fixed dimension n, with zero mean vector and covariance matrix Σ such that

$$S = \left\{ p\big(\boldsymbol{x}|\Sigma\big) : \Sigma \in S^+(n, \mathbb{R}) \right\},$$

where

$$p\big(\boldsymbol{x}|\Sigma\big) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} \right\}. \tag{6.5}$$

For each matrix $\Sigma \in S^+$, its elements $\{\sigma_{rs} : r \leq s; r, s = 1, \ldots, n\}$ are identified by parameters $\{\theta^i : i = 1, \ldots, m\}$ with $\theta^i = \sigma_{rs}$. In this case, S^+ is isomorphic to an open subset Θ of $\mathbb{R}^m$ with $m = \frac{1}{2}n(n+1)$. We consider S^+ as a manifold with a coordinate system (θ^i).

Let $TS^+(n, \mathbb{R})$ and $T^*S^+(n, \mathbb{R})$ be the tangent and cotangent space of S^+ respectively. With respect to the coordinate system for S^+, let $E_i, i = 1, \ldots, m$ denote the canonical basis of the tangent space TS^+ and $E_i^*, i = 1, \ldots, m$ the dual basis of the cotangent space T^*S^+. These are given by

$$E_i = \begin{cases} 1_{rr}, & r = s \\ (1_{rs} + 1_{sr}), & r \neq s \end{cases}$$

$$E_i^* = \begin{cases} 1_{rr}, & r = s \\ \frac{1}{2}(1_{rs} + 1_{sr}), & r \neq s \end{cases}$$

where 1_{rs} denotes $n \times n$ matrix with zero everywhere except 1 at row r and column s. The tangent space at any point of S^+ corresponds to the space of $n \times n$ symmetric matrices. The duality between the tangent space TS^+ and the cotangent space T^*S^+ is defined as (e.g. Skovgaard (1984))

$$\langle A, A^* \rangle_\Sigma = \mathrm{tr}(AA^*); \quad A \in TS^+, A^* \in T^*S^+.$$

In line with several studies (Skovgaard, 1984; Burbea, 1984; Calvo & Oller, 1991; Förstner & Moonen, 2003), we characterised S^+ as a Riemannian manifold for which the Riemannian metric, the Riemannian connection, the solution of the geodesic equation as well as the geodesic distance have closed form expressions.

6.3.2 Riemannian metric and Riemannian connection

The Riemannian metric for S^+ is given for all $\Sigma \in S^+$ as

$$g_{ij}(\boldsymbol{\theta}) = \langle E_i, E_j \rangle_\Sigma = -\mathrm{E}_p \left[\partial_i \partial_j l_\Sigma(\boldsymbol{x}) \right], \quad i,j = 1, \ldots, m$$

where $l_\Sigma(\boldsymbol{x}) = \log p(\boldsymbol{x}|\Sigma)$ and $\partial_i \stackrel{\text{def}}{=} \frac{\partial}{\partial \theta^i}$. From (6.5),

$$l_\Sigma = \log p(\boldsymbol{x}|\Sigma) = -\frac{1}{2}\left[n \log 2\pi + \log|\Sigma| + \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} \right],$$

so that

$$\begin{aligned}
\partial_j l_\Sigma &= -\frac{1}{2}\left[\partial_j \left(n \log 2\pi \right) + \partial_j \left(\log|\Sigma| \right) + x^T \left(\partial_j \Sigma^{-1} \right) x \right] \\
&= -\frac{1}{2}\left[0 + \mathrm{tr}\left(\Sigma^{-1} (\partial_j \Sigma) \right) - x^T \left(\Sigma^{-1} \left(\partial_j \Sigma \right) \Sigma^{-1} \right) x \right] \\
&= -\frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1} E_j \right) - x^T \Sigma^{-1} E_j \Sigma^{-1} x \right]
\end{aligned}$$

and

$$\begin{aligned}
\partial_i \partial_j l_\Sigma &= -\frac{1}{2}\left[\mathrm{tr}\left(\partial_i (\Sigma^{-1}) E_j \right) - x^T \left\{ \partial_i (\Sigma^{-1}) E_j \Sigma^{-1} + \Sigma^{-1} E_j \partial_i (\Sigma^{-1}) \right\} x \right] \\
&= \frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1} (\partial_i \Sigma) \Sigma^{-1} E_j \right) - x^T \left\{ \Sigma^{-1} (\partial_i \Sigma) \Sigma^{-1} E_j \Sigma^{-1} + \Sigma^{-1} E_j \Sigma^{-1} (\partial_i \Sigma) \Sigma^{-1} \right\} x \right] \\
&= \frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1} E_i \Sigma^{-1} E_j \right) - x^T \left\{ \Sigma^{-1} E_i \Sigma^{-1} E_j \Sigma^{-1} + \Sigma^{-1} E_j \Sigma^{-1} E_i \Sigma^{-1} \right\} x \right].
\end{aligned}$$

Thus

$$\begin{aligned}
g_{ij}(\boldsymbol{\theta}) = -\mathrm{E}_p \left[\partial_i \partial_j l_\Sigma \right] &= -\frac{1}{2}\mathrm{E}_p \left[\mathrm{tr}\left(\Sigma^{-1} E_i \Sigma^{-1} E_j \right) \right] + \frac{1}{2}\mathrm{E}_p \left[x^T \left(\Sigma^{-1} E_i \Sigma^{-1} E_j \Sigma^{-1} \right) x \right] \\
&\quad + \frac{1}{2}\mathrm{E}_p \left[x^T \left(\Sigma^{-1} E_j \Sigma^{-1} E_i \Sigma^{-1} \right) x \right] \\
&= -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1} E_i \Sigma^{-1} E_j \right) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1} E_i \Sigma^{-1} E_j \right) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1} E_j \Sigma^{-1} E_i \right) \\
&= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1} E_i \Sigma^{-1} E_j \right)
\end{aligned}$$

since $\mathrm{tr}\,(A\Sigma) = \mathrm{E}_p\left[x^T Ax\right]$ for all A. Hence

$$g_{ij}(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}E_i\Sigma^{-1}E_j\right). \tag{6.6}$$

In practice, the inner product with respect to Σ of any two vectors $A, B \in TS^+$ is $\langle A, B\rangle_\Sigma = \frac{1}{2}\mathrm{tr}\,(\Sigma^{-1}A\Sigma^{-1}B)$. Specifically, two infinitesimal close points Σ and $\Sigma + \delta\Sigma$ of S^+ have distance (with respect to Σ) as

$$\|\delta\Sigma\|_\Sigma = \sqrt{\frac{1}{2}\mathrm{tr}\,((\Sigma^{-1}\delta\Sigma)^2)}.$$

Equipped with the Riemannian metric g, S^+ is now a Riemannian manifold. The metric induces a unique affine connection called the Riemannian or Levi-Civita connection. The connection coefficients are given by the Christoffel symbols, which are defined in terms of the canonical and dual basis as

$$\begin{aligned}\Gamma_{ij}^k &= \Gamma(E_i, E_j; E_k^*) = E_k^*(\nabla_{E_i}E_j)\\ &= \frac{1}{2}g^{kl}(\partial_j g_{il} + \partial_i g_{jl} - \partial_l g_{ij}).\end{aligned} \tag{6.7}$$

Since

$$\partial_k g_{ij}(\boldsymbol{\theta}) = -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}E_k\Sigma^{-1}E_i\Sigma^{-1}E_j\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}E_i\Sigma^{-1}E_k\Sigma^{-1}E_j\right),$$

it follows that

$$\Gamma_{ij}^k = -\frac{1}{2}\mathrm{tr}\left(E_i\Sigma^{-1}E_jE_k^*\right) - \frac{1}{2}\mathrm{tr}\left(E_j\Sigma^{-1}E_iE_k^*\right) \tag{6.8}$$

(see Lenglet *et al.* (2006) for details).

The manifold S^+ can be formulated as both a homogeneous and a symmetric space which arises naturally by the action of Lie group on S^+ (see Fletcher & Joshi (2007)). This is essential for computing geodesics and distance expressions in S^+.

6.3.3 $S^+(n, \mathbb{R})$ as a homogeneous space

We begin with some basic terminologies on Lie groups. Suppose G is an algebraic group. If G forms a differentiable manifold, with the property that the two group operations, multiplication and inversion, are smooth, then G is said to be a Lie group.

Let M be a manifold. A group action of G on M is a smooth mapping $f : G \times M \to M$ such that for all $g, h \in G$ and all $x \in M$,

$$f(g, f(h, x)) = f(gh, x) \quad \text{and} \quad f(e, x) = x,$$

where e is the identity element of G.

The orbit of a point $x \in M$ is defined as $G(x) = \{f(g, x) : g \in G\}$. The group action f is transitive if it has only a single orbit and in this case the manifold M is called a homogeneous space.

The set $G_x = \{g \in G : f(g, x) = x\}$ is a subgroup of G that fixes the point $x \in M$, and is called the isotropy subgroup of x. If H is a closed Lie subgroup of G, the subsets of G defined (for any $g \in G$) as

$$\{gh : h \in H\} \quad \text{and} \quad \{hg : h \in H\}$$

are respectively the left/right coset of H. The space of left/right cosets of H, denoted as G/H is a differentiable manifold. The mapping $gx \mapsto gG_x$ defines a natural bijection between $G(x)$ and the quotient group (coset space) G/G_x, i.e. $G(x) \approxeq G/G_x$.

Now consider the Lie group of all $n \times n$ real non-singular matrices with positive determinant, denoted $GL^+(n, \mathbb{R})$ and its subset S^+ of symmetric positive-definite matrices. A group action on S^+ is defined through

$$f : GL^+ \times S^+ \to S^+, \quad f(X, \Sigma) = X \Sigma X^T.$$

Since any $\Sigma \in S^+$ can be written as $\Sigma = XX^T = f(X, I)$, where I is the $n \times n$ identity matrix in S^+, we see that f is transitive and S^+ is a homogeneous space. The rotation group

$$SO(n, \mathbb{R}) = \left\{X \in GL^+ : XX^T = I\right\}$$

is the isotropy subgroup of the identity matrix. We consider the quotient space GL^+/SO. This can be thought in terms of polar decomposition of a matrix which seperates $X \in GL^+$ into $X = \Sigma A$, where $\Sigma \in S^+$ and $A \in SO$ (Fletcher & Joshi, 2007). Therefore, the manifold $S^+ \approxeq GL^+/SO$, which results from crossing out the rotational element in the polar decomposition of GL^+.

6.3.4 $S^+(n, \mathbb{R})$ as a symmetric space

A Riemannian metric defines an inner product $\langle \cdot, \cdot \rangle_x$ on the tangent space $T_x M$ at each point x of a manifold M. A transitive group action $f : x \mapsto f(g, x)$ on M is said to be an isometry if it is a distance preserving homeomorphism on M with respect to the Riemannian metric. If for any $x \in M$, there exist some isometry f_x on M such that

$$f_x(x) = x \quad \text{and} \quad (df_x)_x = -I,$$

where df_x is the derivative map of f_x, then M is said to be a symmetric space. In other words, a symmetric space is exactly a homogeneous space M with a symmetry f_x at some point $x \in M$ (Eschenburg, 2018).

The Riemannian metric on the manifold S^+ is defined for any $U, V \in T_\Sigma S^+$ as

$$\langle U, V \rangle_\Sigma = \frac{1}{2} \text{tr}\left(\Sigma^{-1} U \Sigma^{-1} V\right).$$

The group GL^+ acts transitively on S^+ through $f_X(\Sigma) = X \Sigma X^T$, and due to linearity of this action, the derivative map is given by $df_X U = X U X^T$, where $X \in GL^+$. We see that the group action f_X is an isometry with respect to our Riemannian metric, that is

$$\begin{aligned}
\langle df_X U, df_X V \rangle_{f_X(\Sigma)} &= \frac{1}{2} \text{tr}\left(X U X^T (X \Sigma X^T)^{-1} X V X^T (X \Sigma X^T)^{-1}\right) \\
&= \frac{1}{2} \text{tr}\left(X U X^T (X^T)^{-1} \Sigma^{-1} X^{-1} X V X^T (X^T)^{-1} \Sigma^{-1} X^{-1}\right) \\
&= \frac{1}{2} \text{tr}\left(X U \Sigma^{-1} V \Sigma^{-1} X^{-1}\right) \\
&= \langle U, V \rangle_\Sigma.
\end{aligned}$$

Since any $\Sigma \in S^+$ can be written as $\Sigma = X X^T = f_X(I)$, where f_X is a transitive group action, a group action $f_{X^{-1}}$ maps Σ to I. The reverse mapping $f_I(\Sigma) = \Sigma^{-1}$ with $(df_I)_\Sigma U = -\Sigma^{-1} U \Sigma^{-1}$ is also an isometry of S^+:

$$\begin{aligned}
\langle (df_I)_\Sigma U, (df_I)_\Sigma V \rangle_{\Sigma^{-1}} &= \frac{1}{2} \text{tr}\left(\Sigma^{-1} U \Sigma^{-1} \Sigma \Sigma^{-1} V \Sigma^{-1} \Sigma\right) \\
&= \frac{1}{2} \text{tr}\left(\Sigma^{-1} U \Sigma^{-1} V\right) \\
&= \langle U, V \rangle_\Sigma.
\end{aligned}$$

Hence, since $f_I(I) = I$ and $(df_I)_I = -I$, the isometry f_I is a symmetry at I which makes S^+ a symmetric space. The symmetry at an arbitrary $X \in GL^+$ is $f_X(\Sigma) = X \Sigma^{-1} X$. The

symmetry at an arbitrary $X \in GL^+$ is defined by $f_X(\Sigma) = X\Sigma^{-1}X$. This construction appears in Eschenburg (2018).

6.3.5 Exponential and logarithm maps

As a consequence of the space S^+ being symmetric, it is geodesically complete, that is, given any two points in S^+, there exist a length minimizing (geodesic) curve connecting them (Eschenburg, 2018). In general, geodesics on Riemannian manifolds observe a local diffeomorphism, called an exponential map, from the tangent space at a given point of the manifold to the manifold. However, this mapping is global in the case of geodesically complete manifolds (see Hopf-Rinow theorem in Jost (2017)). Suppose $\gamma : [0, 1] \to S^+$ is a geodesic curve starting at a point $\gamma(0) = \Sigma_1$, with initial tangent vector $\gamma'(0) = V$. If $T_{\Sigma_1}S^+$ is the tangent space at Σ_1, the exponential map is given by

$$\exp_{\Sigma_1} : T_{\Sigma_1}S^+ \to S^+, \quad \exp_{\Sigma_1}(V) = \gamma(1),$$

for all $V \in T_{\Sigma_1}S^+$ and for all $\Sigma_1 \in S^+$. We can then define an inverse map, the logarithm map, which is a diffeomorphism from the manifold S^+ to the tangent space $T_{\Sigma_1}S^+$, by

$$\log_{\Sigma_1} : S^+ \to T_{\Sigma_1}S^+, \quad \log_{\Sigma_1}(\Sigma_2) = V,$$

where γ is a geodesic curve between any two points Σ_1, $\Sigma_2 \in S^+$ such that $\gamma(0) = \Sigma_1$ and $\gamma(1) = \Sigma_2$. This geodesic is guaranteed to be a unique and its length is the Riemannian distance between Σ_1 and Σ_2. If $V \in T_{\Sigma_1}S^+$ is the unique tangent vector in $T_{\Sigma_1}S^+$ such that $\Sigma_2 = \exp_{\Sigma_1}(V)$, then the geodesic curve is of the form

$$\gamma(t) = \exp_{\Sigma_1}(tV), \quad V = \log_{\Sigma_1}(\Sigma_2).$$

6.3.6 Computing geodesics

In the Euclidean geometry, a straight line is the shortest path connecting any given two points, and any object moving along this path has a constant velocity. However, a geodesic generalises the concept of straight line in the Riemannian geometry. Recall that a curve γ is a geodesic if and only if it satisfies the Euler-Lagrange equation (adopting Einstein's summation notation), that is

$$\frac{d^2\theta^k}{dt^2} + \Gamma^k_{ij}\frac{d\theta^i}{dt}\frac{d\theta^j}{dt} = 0, \quad k = 1, \ldots, m,$$

where the n^3 functions Γ_{ij}^k are the coefficients of the Riemannian connection. The Riemannian connection allows the tangent space at Σ_1 to be mapped to the tangent space at Σ_2, and this depends on the curve γ joining the two points. Using (6.8), the Euler-Lagrange equation reduces to (Lenglet *et al.*, 2006)

$$\frac{d^2\gamma(t)}{dt^2} - \frac{d\gamma(t)}{dt}\gamma(t)^{-1}\frac{d\gamma(t)}{dt} = 0. \tag{6.9}$$

Calvo & Oller (1991) obtain an explicit solution of the geodesic curve $\gamma(t)$, $t \in [a, b] \subset \mathbb{R}$ by solving a general system of differential equations for which (6.9) is a special case. Another approach to finding the geodesic equation is through a group action (Fletcher & Joshi, 2007). Since S^+ is isomorphic to the quotient space GL^+/SO, the geodesic on S^+ is invariant under the action of the group GL^+. The special geodesic starting from $\gamma(0) = I$ with initial tangent vector $\gamma'(0) = V$ is given by $\gamma(t) = \exp(tV)$. However, in general, for arbitrary initial point Σ_1 and tangent vector V, we use the group action to map this setting into the special case of the identity. Let $\Sigma_1 = \Sigma_1^{1/2}\Sigma_1^{1/2}$, where the matrix square root $\Sigma_1^{1/2}$ is well-defined since Σ_1 is positive definite. Then the group action $f_{\Sigma_1^{-1/2}}$ maps Σ_1 to the identity I and through the corresponding tangent map, the tangent vector V is mapped to a diagonal tangent vector $W = \Sigma_1^{-1/2}V\Sigma_1^{-1/2}$. Thus, we can compute a geodesic $\tilde{\gamma}$ with initial point $\tilde{\gamma}(0) = I$ and tangent vector $\tilde{\gamma}'(0) = W$ by $\tilde{\gamma}(t) = \exp(tW)$. Finally, $\tilde{\gamma}$ is mapped back to the original setting through the inverse group action $f_{\Sigma_1^{1/2}}$. Therefore, the geodesic starting from $\Sigma_1 \in S^+$ in the direction $\Sigma_1' = \Sigma_1^{1/2}W\Sigma_1^{1/2} \in T_{\Sigma_1}S^+$ is given by

$$\gamma(t) = \Sigma_1^{1/2}\exp(tW)\Sigma_1^{1/2}, \tag{6.10}$$

where $W = \log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) \in T_{\Sigma_1}S^+$.

6.3.7 Geodesic distance

The length of γ is defined by

$$L(\gamma) = \int_0^1 \sqrt{\langle\gamma'(t), \gamma'(t)\rangle_{\gamma(t)}}dt = \int_0^1 \sqrt{\frac{1}{2}\mathrm{tr}\left((\gamma^{-1}(t)\gamma'(t))^2\right)}dt,$$

where $\gamma'(t) = d\gamma/dt$. The distance between two matrices Σ_1 and Σ_2 in S^+ is defined as the infimum of the lengths of curves joining them, that is

$$D\left(\Sigma_1, \Sigma_2\right) = \inf\left\{L(\gamma) : \gamma : [0,1] \to S^+ \text{ with } \gamma(0) = \Sigma_1,\ \gamma(1) = \Sigma_2\right\}.$$

This is given by

$$\begin{aligned} D\left(\Sigma_1, \Sigma_2\right) &= \sqrt{\frac{1}{2}\mathrm{tr}\left\{\log^2\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)\right\}} \\ &= \sqrt{\frac{1}{2}\sum_{i=1}^{n}\log^2\left(\eta_i\right)}, \end{aligned} \tag{6.11}$$

where η_i denote the n eigenvalues of the matrix $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} \in S^+$ (Lenglet *et al.*, 2006). Indeed, D defines a metric on S^+ with nice properties and we refer to it as the covariance distance. We present and prove these properties as pointed in Förstner & Moonen (2003). However, no complete proof of the triangle inequality was given in this reference.

Lemma 6.1. *D has the following properties*

(i) Positivity: $D\left(\Sigma_1, \Sigma_2\right) \geq 0$, $D\left(\Sigma_1, \Sigma_2\right) = 0 \iff \Sigma_1 = \Sigma_2$.

(ii) Invariance under congruence transformations:

$D\left(\Sigma_1, \Sigma_2\right) = D\left(X\Sigma_1X^T, X\Sigma_2X^T\right), \quad \forall X \in GL(n, \mathbb{R})$.

(iii) Invariance under inversion: $D\left(\Sigma_1, \Sigma_2\right) = D\left(\Sigma_1^{-1}, \Sigma_2^{-1}\right)$.

(iv) Symmetry: $D\left(\Sigma_1, \Sigma_2\right) = D\left(\Sigma_2, \Sigma_1\right)$.

(v) Triangle inequality: $D\left(\Sigma_1, \Sigma_2\right) \leq D\left(\Sigma_1, \Sigma_3\right) + D\left(\Sigma_3, \Sigma_2\right)$.

Proof. Let η be a function defined on arbitrary $A \in S^+$ such that $\eta(A)$ gives the eigenvalues of A. We can think of η as a vector of functions $\eta_i, i = 1, \ldots, n$ corresponding to the n eigenvalues of A. It can be seen that $\eta(AB) = \eta(BA)$, since

$$\begin{aligned} \det(AB - \lambda I) &= \frac{1}{\det(A)}\det(AB - \lambda I)\det(A) \\ &= \det(A^{-1})\det(AB - \lambda I)\det(A) \\ &= \det(A^{-1}(AB - \lambda I)A) \\ &= \det(BA - \lambda I), \end{aligned}$$

for some scalar λ and identity matrix I.

(i) By definition, $D \geq 0$.

$$\begin{aligned} D = 0 &\iff \eta_i = 1, \text{ for all } i \\ &\iff \eta_i \left\{ \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} \right\} = \eta_i \{I\}, \text{ for all } i \\ &\iff \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} = I \\ &\iff \Sigma_1 = \Sigma_2. \end{aligned}$$

(ii) Let $X \in GL(n, \mathbb{R})$,

$$\begin{aligned} \eta \left\{ \left(X\Sigma_1 X^T\right)^{-1/2} X\Sigma_2 X^T \left(X\Sigma_1 X^T\right)^{-1/2} \right\} &= \eta \left\{ X\Sigma_2 X^T \left(X\Sigma_1 X^T\right)^{-1} \right\} \\ &= \eta \left\{ X\Sigma_2 X^T \left(X^T\right)^{-1} \Sigma_1^{-1} X^{-1} \right\} \\ &= \eta \left\{ X\Sigma_2 \Sigma_1^{-1} X^{-1} \right\} \\ &= \eta \left\{ \Sigma_2 \Sigma_1^{-1} \right\} \\ &= \eta \left\{ \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} \right\}. \end{aligned}$$

This implies that $D(\Sigma_1, \Sigma_2) = D\left(X\Sigma_1 X^T, X\Sigma_2 X^T\right)$.

(iii)
$$\begin{aligned} D^2\left(\Sigma_1^{-1}, \Sigma_2^{-1}\right) &= \frac{1}{2} \sum_{i=1}^{n} \log^2 \left(\eta_i \left\{ \Sigma_2^{-1} \Sigma_1 \right\}\right) \\ &= \frac{1}{2} \sum_{i=1}^{n} \left[\log \left(\eta_i \left\{ \Sigma_2^{-1} \Sigma_1 \right\}\right)\right]^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \left[\log \left(\left(\eta_i \left\{ \Sigma_1^{-1} \Sigma_2 \right\}\right)^{-1}\right)\right]^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \left[-\log \left(\eta_i \left\{ \Sigma_1^{-1} \Sigma_2 \right\}\right)\right]^2 \\ &= \frac{1}{2} \sum_{i=1}^{n} \left[\log \left(\eta_i \left\{ \Sigma_1^{-1} \Sigma_2 \right\}\right)\right]^2 \\ &= D^2(\Sigma_1, \Sigma_2). \end{aligned}$$

Therefore, $D\left(\Sigma_1^{-1}, \Sigma_2^{-1}\right) = D(\Sigma_1, \Sigma_2)$.

(iv) Since

$$\begin{aligned}\eta\left\{\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right\} &= \eta\left\{\Sigma_2\Sigma_1^{-1}\right\} \\ &= \left(\eta\left\{\Sigma_1\Sigma_2^{-1}\right\}\right)^{-1} \\ &= \left(\eta\left\{\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}\right\}\right)^{-1},\end{aligned}$$

we see that $D(\Sigma_1, \Sigma_2) = D(\Sigma_2, \Sigma_1)$.

(v) We refer the reader to Förstner & Moonen (2003) for a sketch of the proof of the triangle inequality.

□

6.4 The extrinsic geometry and projection

6.4.1 Comparing the induced metric on the embedded space with other tree space distances

The metric D on S^+ also defines a metric on $\mathcal{E}_n$ induced by the embedding: for any pair of trees T_1, $T_2 \in \mathcal{E}_n$, the induced distance between them is given by $D\big(\Sigma(T_1), \Sigma(T_2)\big)$ as defined in (6.11). We have previously studied properties of probabilistic distances as well as the BHV metric between trees from several data sets, and now we want to compare these distances with the covariance distance. We consider the data set consisting of 100 bootstrap replicates of trees obtained from primate DNA data (Huelsenbeck & Ronquist, 2001). Analysis of this data set has already been discussed in Section 4.7.4. We compare distance between every pair of trees in this data set using our probabilistic distances (Hellinger (H), total variation (TV), Kullback-Leibler (KL) and Jensen-Shannon(JS)) with the two-state model, BHV metric and the covariance (Cov) distance. Figure 6.3 shows the results obtained: the covariance distance, while correlated with the other distance measures, appears to differ considerably from the other distances especially BHV.

6.4.2 Projection from S^+ onto the embedded space $\mathcal{E}_n$

The projection of a point $\Sigma \in S^+$ onto tree space $\mathcal{E}_n$ is defined as the point $T \in \mathcal{E}_n$ such that $\Sigma(T)$ is closest to Σ in covariance distance. Hence, we define the projection mapping $P : S^+ \to \mathcal{E}_n$ as

$$P(\Sigma) = \underset{T \in \mathcal{E}_n}{\operatorname{argmin}}\ D^2(\Sigma, \Sigma(T))$$

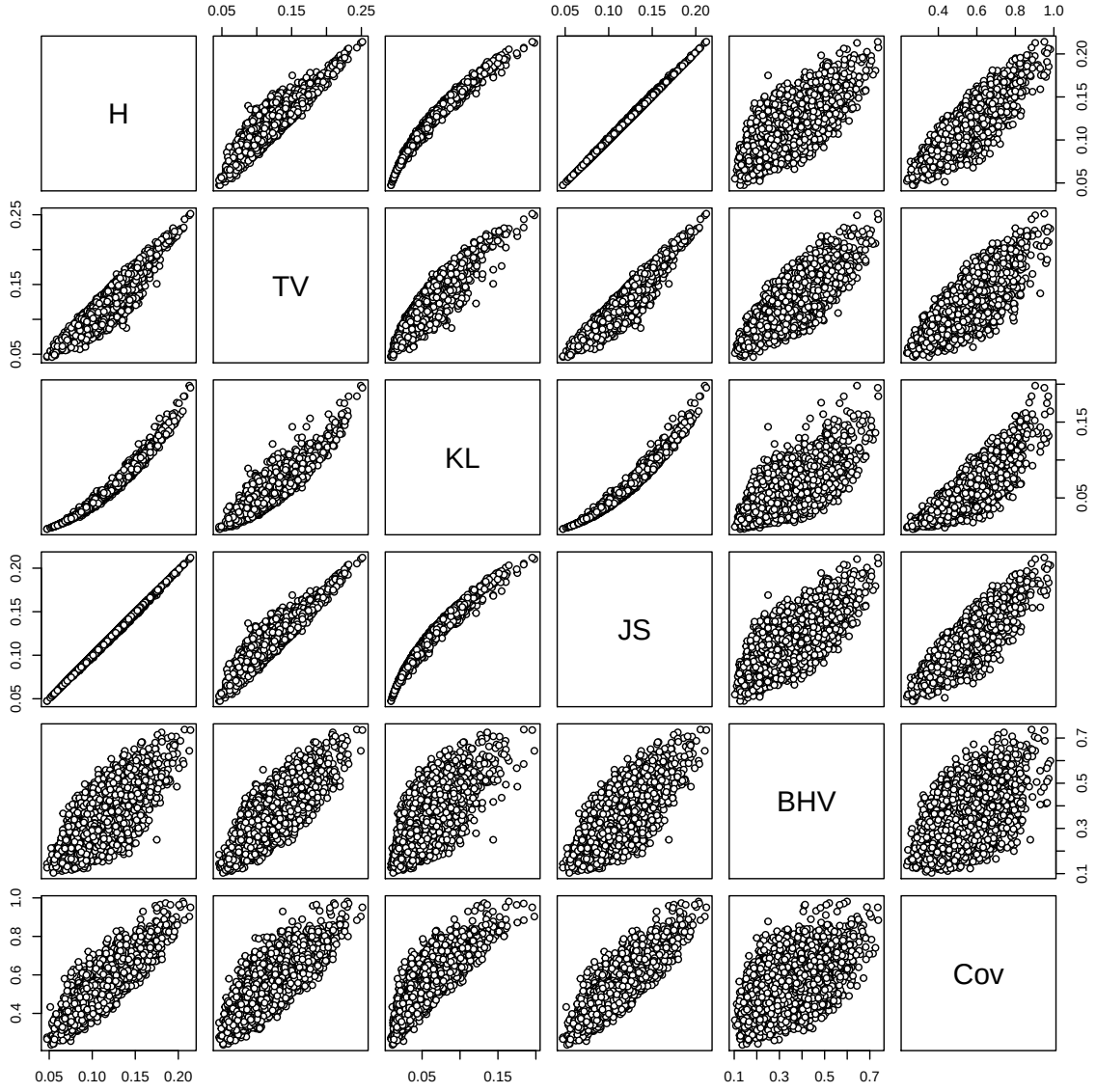


Figure 6.3: Comparison of probabilistic distances (Hellinger (H), total variation (TV), Kullback-Leibler (KL) and Jensen-Shannon(JS)) with two-state model, BHV metric and covariance (Cov) distance between every pair of trees in 100 bootstrap replicates of trees obtained from primate DNA data.

with D as defined in (6.11). The point $P(\Sigma)$ is not necessarily unique. The gradient of $D^2\big(\Sigma, \Sigma(T)\big)$, written as $\nabla_{\boldsymbol{\ell}} D^2\big(\Sigma, \Sigma_{\boldsymbol{\ell}}\big)$ where $\nabla_{\boldsymbol{\ell}}$ is the derivative with respect to edge lengths of the tree $T = (\tau, \boldsymbol{\ell})$, can be computed analytically as a vector in the orthant

corresponding to τ. Components of the gradient are

$$\partial_i D^2\left(\Sigma, \Sigma_{\boldsymbol{\ell}}\right) = \partial_i \left(\frac{1}{2}\mathrm{tr}\left\{\log^2(X)\right\}\right) = \mathrm{tr}\left\{\log(X)X^{-1}(\partial_i X)\right\} \tag{6.12}$$

where $X = \Sigma^{-1/2}\Sigma_{\boldsymbol{\ell}}\Sigma^{-1/2}$ and

$$\begin{aligned}
\partial_i X &= \partial_i\left(\Sigma^{-1/2}\Sigma_{\boldsymbol{\ell}}\Sigma^{-1/2}\right) \\
&= \left\{\partial_i\Sigma^{-1/2}\right\}\Sigma_{\boldsymbol{\ell}}\Sigma^{-1/2} + \Sigma^{-1/2}\left\{\partial_i\Sigma_{\boldsymbol{\ell}}\right\}\Sigma^{-1/2} + \Sigma^{-1/2}\Sigma_{\boldsymbol{\ell}}\left\{\partial_i\Sigma^{-1/2}\right\} \\
&= \Sigma^{-1/2}\left\{\partial_i\Sigma_{\boldsymbol{\ell}}\right\}\Sigma^{-1/2}.
\end{aligned}$$

Given initial set of edge lengths from a starting tree, we use the gradient descent algorithm (Chong & Zak, 2013, Chapter 8, p. 113). This gives an iterative scheme:

$$\boldsymbol{\ell}_{k+1} = \boldsymbol{\ell}_k - \alpha_k \nabla D^2(\Sigma, \Sigma_{\boldsymbol{\ell}}), \quad k \geq 0, \tag{6.13}$$

where α_k is the value of the step size which is allowed to change at every iteration. We adopt the Barzilai-Borwein method (Barzilai & Borwein, 1988) with

$$\alpha_k = \frac{(\boldsymbol{\ell}_k - \boldsymbol{\ell}_{k-1})^T[\nabla D^2(\Sigma, \Sigma_{\boldsymbol{\ell}_k}) - \nabla D^2(\Sigma, \Sigma_{\boldsymbol{\ell}_{k-1}})]}{\|\nabla D^2(\Sigma, \Sigma_{\boldsymbol{\ell}_k}) - \nabla D^2(\Sigma, \Sigma_{\boldsymbol{\ell}_{k-1}})\|^2}. \tag{6.14}$$

We implemented the projection procedure in Java, given as follows.

Projection algorithm

The algorithm proceeds just as in the Euclidean case within each orthant such that given initial values for α and edge lengths, we compute the components of the gradient using (6.12) and compute a new set of edge lengths (6.13). The parameter α is updated using (6.14). Then, at each step of the algorithm, a check is performed to see if, under that step, any edge lengths go negative. If one or more pendant edge lengths go negative, we reset the length to a small positive number. However

1. If a single internal edge is negative or zero, we look at both nearest neighbor interchange (NNI) orthants and step into the orthant that is closest in covariance distance to Σ. The new edge lengths satisfy (6.13) apart from the single NNI edge that went negative which is reset to a small positive number.

2. If two or more internal edge lengths go zero or negative, we rescale the increment

so that exactly one edge go zero or negative and we go to step 1.

6.4.3 Projection of the extrinsic mean

The extrinsic mean associated with N given points $\Sigma_1, \ldots, \Sigma_N \in S^+$ is defined as

$$\mu(\Sigma_1, \ldots, \Sigma_N) = \underset{\Sigma \in S^+}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{N} D^2(\Sigma, \Sigma_k).$$

In other words, the extrinsic mean locally minimizes the variance which is given as the expectation of squared distance. The term "extrinsic" is used since if $\Sigma_1, \ldots, \Sigma_N \in \mathcal{E}_n \subset S^+$, the mean does not necessarily lie in $\mathcal{E}_n$. Karcher (1977) proved the existence and uniqueness of the extrinsic mean on manifolds of non-positive sectional curvature and hence on S^+. In general, the extrinsic mean has no analytical solution when $N > 2$ (Moakher, 2005). However, Lenglet *et al.* (2006) provided a gradient descent algorithm to compute the extrinsic mean and we implemented this algorithm.

To test the calculation of the mean and the projection algorithm, we used the following procedure. We consider a tree $T \in \mathcal{E}_n$ such that $T = (\tau, \boldsymbol{\ell})$ with $n = 10$ taxa, where the topology τ is sampled from a Yule distribution and edge lengths $\boldsymbol{\ell}$ sampled from a Gamma distribution with mean 0.1 and variance 0.005. First we perturb $\Sigma(T)$ using the Wishart distribution in the following way: for $k = 1, \ldots, N$, draw random sample $X_k \sim W_n\big(v, \Sigma(T)\big)$, with mean $v\Sigma(T)$ and v degrees of freedom. Rescale X_k by $1/v$ in order to obtain Σ_k. This gives a set of points Σ_k, $k = 1, \ldots, N$ ($N = 1000$) in the extrinsic space S^+ which are perturbations of $\Sigma(T)$ but may not necessarily be the image of some trees in $\mathcal{E}_n$. We compute the extrinsic mean $\mu \in S^+$ of these points. We then project μ on to tree space to obtain $P(\mu) \in \mathcal{E}_n$ which we compare with $\Sigma(T)$. Figure 6.4 shows the squared covariance distance $D^2\big(\Sigma(T), P(\mu)\big)$ for different degrees of freedom. We see that as the degrees of freedom increases, the distance converges to zero due to the fact that $P(\mu)$ resembles $\Sigma(T)$ when the degree of freedom is large.

A nice property of the covariance matrix geometry is that it can be applied to situations when trees have missing taxa. In this case, the elements of the covariance matrix associated with any missing taxa is zero as described in (6.3) because a path from that taxa does not exist. This is similar to the approach taken in Section 4.6.2 where missing taxa on the tree are attached with infinitely long edges, which means that they are independent of the other taxa on the tree. Again, we consider the data set of 100 bootstrap replicates of trees obtained from primate DNA data (Huelsenbeck & Ronquist, 2001) with

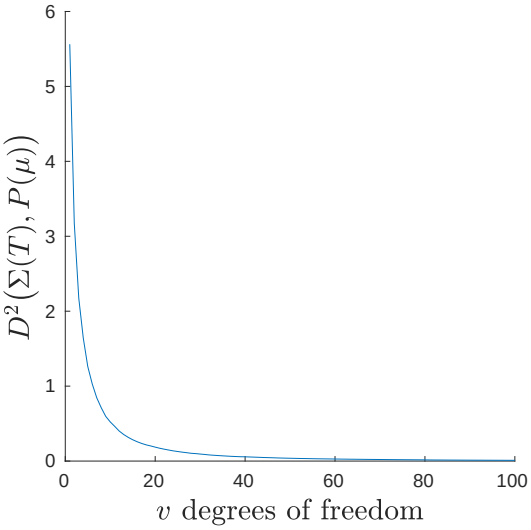


Figure 6.4: Squared covariance distance between a covariance matrix $\Sigma(T)$ associated with a 10 taxa tree and a projection matrix $P(\mu)$ for different degrees of freedom v. For each v, we obtain pertubations Σ_k, $k = 1, \ldots, 1000$ of $\Sigma(T)$ from a Wishart distribution $X_k \sim W_n\big(v, \Sigma(T)\big)$ and compute their extrinsic mean μ.

their associated maximum likelihood (ML) tree (see Section 4.7.4 for details). Firstly we compute the covariance matrix associated with each tree, then we estimate the extrinsic mean of the collection of covariance matrices and project it onto tree space. The resultant tree is shown in Figuire 6.5a. Secondly each of the 100 trees was subjected to a random deletion of the same number of taxa and we compute the covariance matrix associated with each tree with taxa removed. We then estimate the extrinsic mean of the collection of covariance matrices and project it onto tree space. Figure 6.5b-d show the results obtained for different number of missing taxa. It can be seen that the projected extrinsic mean closely approximates the maximum likelihood tree (Figure 6.6) for all the three levels of missingness as well as when no taxa are missing (Fig. 6.5a).

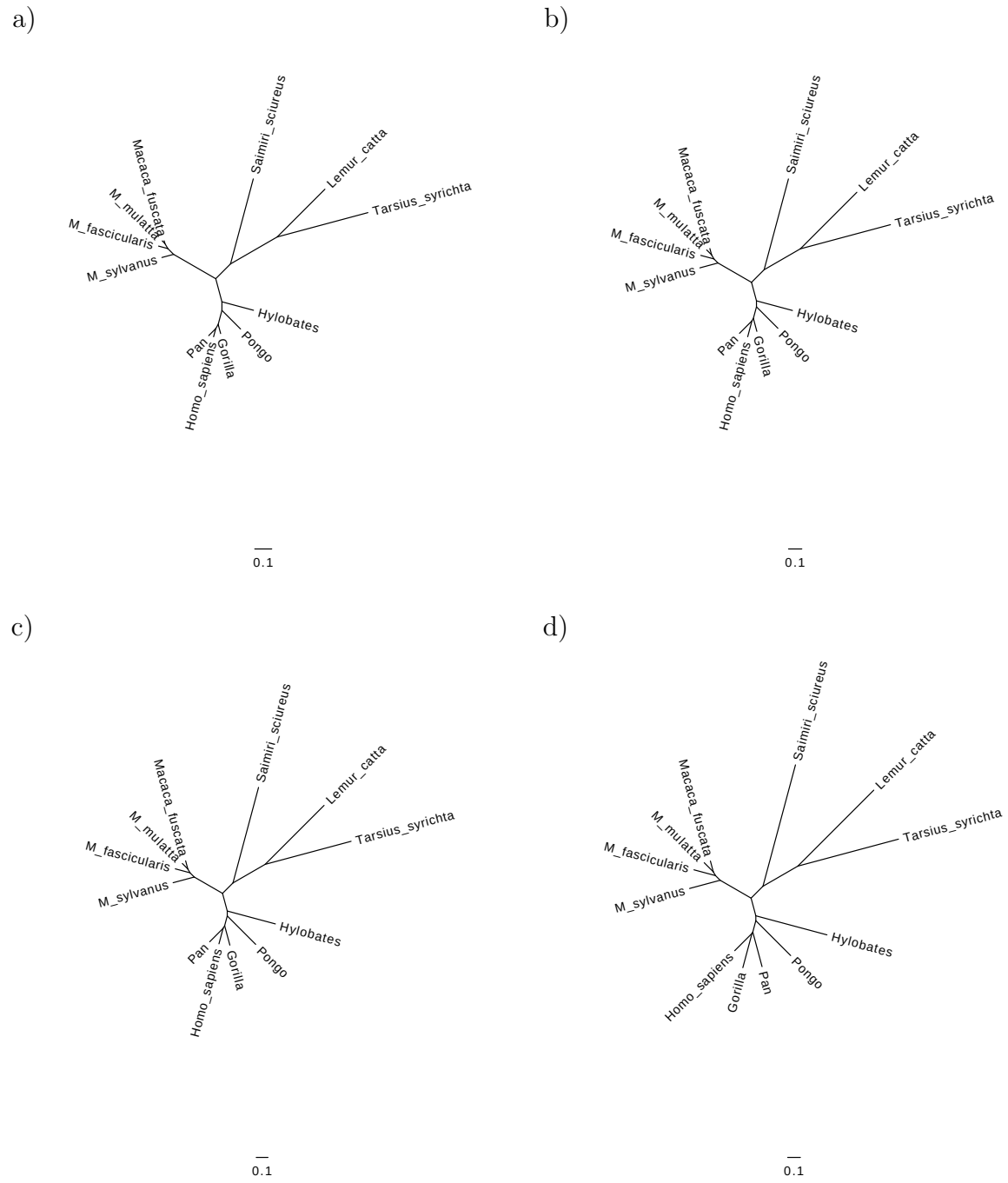


Figure 6.5: Projection of the extrinsic mean of a collection of covariance matrices computed from 100 bootstrap replicates of trees obtained from primate DNA data due to Huelsenbeck & Ronquist (2001) with a) No missing taxa, b) 2 missing taxa, c) 3 missing taxa and d) 4 missing taxa. The results are very similar to the Maximum likelihood tree shown in Figure 6.6. In (d) the topology differs in the placement of Pan (chimpanzee).

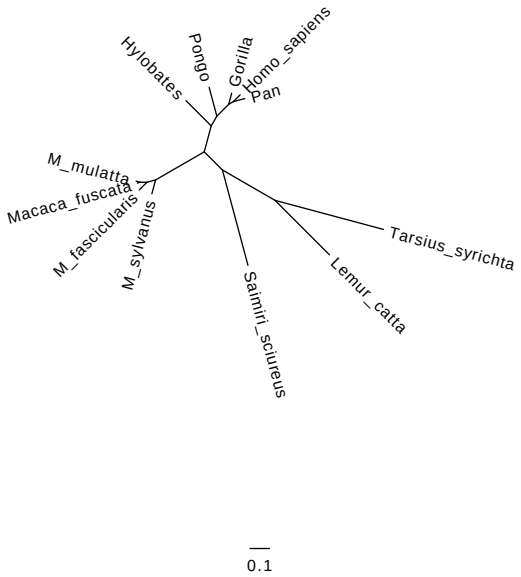


Figure 6.6: Maximum likelihood tree of a data set of 100 bootstrap replicates of trees obtained from primate DNA data due to Huelsenbeck & Ronquist (2001).

6.5 Firing geodesics in the induced geometry

It is possible to write down the induced Riemannian metric within each orthant of $\mathcal{E}_n$, and then solve the geodesic equation numerically, much like we did for the information metric in Chapter 5. The metric on $\mathcal{E}_n$ is detailed as follows. For fixed topology τ, the parameters for a tree $T = (\tau, \boldsymbol{\ell}) \in \mathcal{E}_n$ are the edge lengths ℓ_i, $i = 1, \ldots, 2n-3$ on the tree. Therefore, the Riemannian metric at T is given by

$$g_{ij}(\boldsymbol{\ell}) = \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\{\partial_i \Sigma\}\Sigma^{-1}\{\partial_j \Sigma\}\right), \quad i, j = 1, \ldots, 2n-3$$

where $\partial_i \stackrel{\text{def}}{=} \frac{\partial}{\partial \ell_i}$ and $\Sigma = \Sigma(T)$ as in (6.4). We solve the geodesic equation (5.1) using similar procedure outlined in Section 5.2 in order to obtain geodesics in the space $\mathcal{E}_n$. However, in this case, the Christoffel symbols (6.7) depend on a more complicated partial derivative of the Riemannian metric given as

$$\begin{aligned}\partial_k g_{ij}(\boldsymbol{\ell}) \;=\; & -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\{\partial_k \Sigma\}\Sigma^{-1}\{\partial_i \Sigma\}\Sigma^{-1}\{\partial_j \Sigma\}\right) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\{\partial_k \partial_i \Sigma\}\Sigma^{-1}\{\partial_j \Sigma\}\right) \\ & -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\{\partial_i \Sigma\}\Sigma^{-1}\{\partial_k \Sigma\}\Sigma^{-1}\{\partial_j \Sigma\}\right) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\{\partial_i \Sigma\}\Sigma^{-1}\{\partial_k \partial_j \Sigma\}\right).\end{aligned}$$

Figure 6.7 shows geodesics (black lines) in single orthants each corresponding to a fixed topology (Figure 5.1) for different initial conditions for a 5-taxon unrooted tree. Geodesics are fired in different directions from a fixed initial tree (at the centre). Pendant edge lengths on the initial tree are all the same and fixed at $0.1, 0.25, 0.5, 1.0$ respectively (row-wise), and all their initial directions are zero. The red lines are contours of distance travelled. Geodesics under this geometry behave similarly to information geometry geodesics (Figure 5.2). It is easy to see that plots in Figure 6.7 are not different from their counterparts in Figure 5.2. This result holds across range of conditions in $\mathcal{E}_5$. Although the information metric is defined in a different way, it appears that it is the covariance structure induced by the symmetric 2-state substitution model which primarily determines the shape of the information geodesics. The result implies that the geometry induced from S^+ onto $\mathcal{E}_n$ closely approximates the information geometry.

6.6 Algorithms for constructing approximate geodesics

6.6.1 Definition of algorithms

In this section we describe some algorithms for constructing an approximate geodesic between two given trees T_1, $T_2 \in \mathcal{E}_n$, all of which are based on the idea of projection from covariance matrix space S^+ onto the embedded tree space $\mathcal{E}_n$. These algorithms all rely on the projection algorithm described in Section 6.4.2. They differ from the algorithm in Section 6.5 in that they are "joining" algorithms.

Algorithm 1: naive projection

Consider constructing the extrinsic geodesic between T_1 and T_2 in S^+ and simply projecting each point on the geodesic into $\mathcal{E}_n$. This algorithm produces discontinuous paths in $\mathcal{E}_n$, as shown in Figure 6.8a-c, due to $\mathcal{E}_n$ not being convex. Hence, the projection of a point $r \in S^+$ into $\mathcal{E}_n$ can be caught in a local minimum before reaching the global minimum; see Figure 6.9 for a simple illustration. As a result, this is a very poor algorithm.

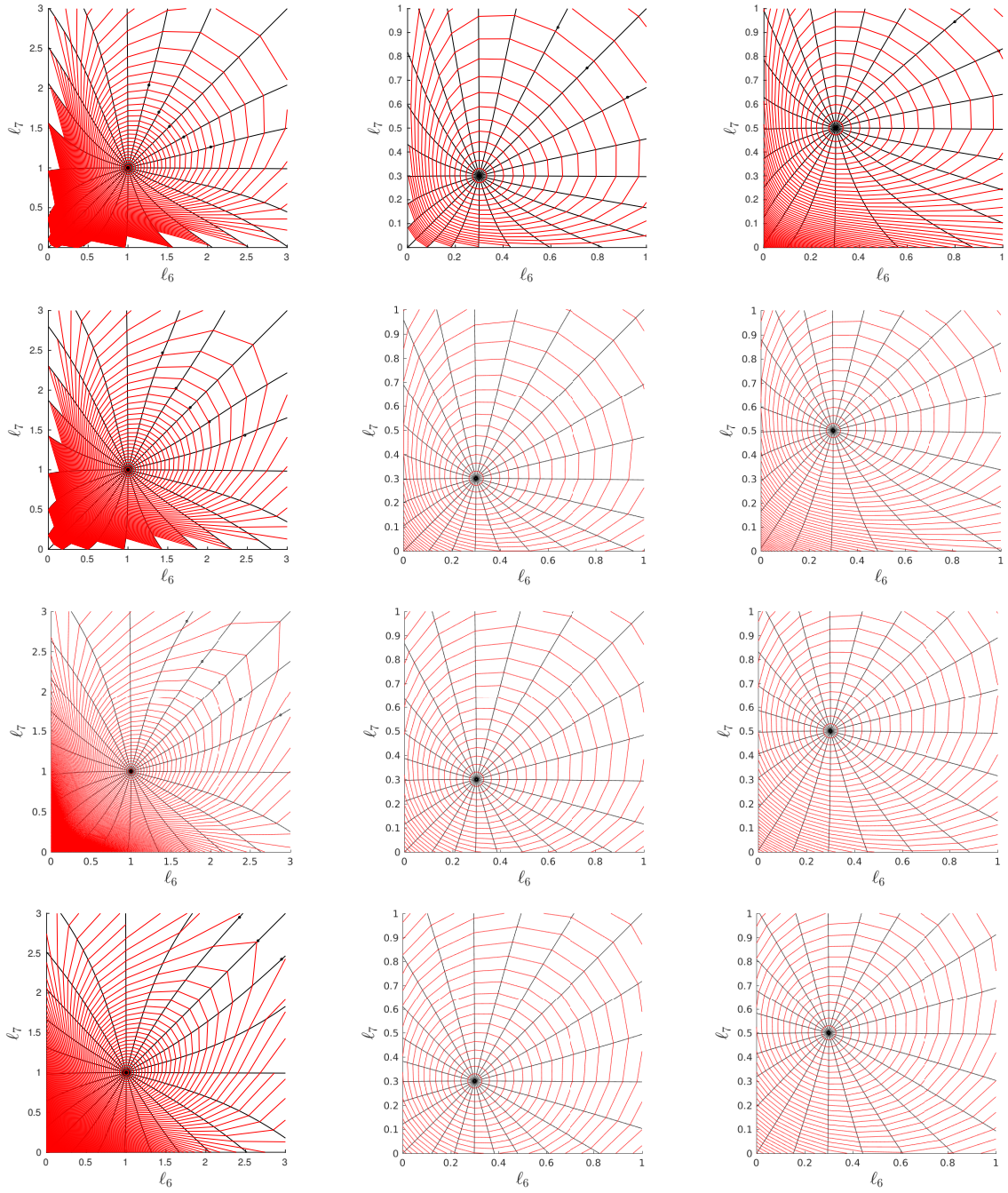


Figure 6.7: Geodesics (black) and contours of distance (red) within a single orthant. Geodesics were fired from the central tree in each case. Rows correspond to different initial pendant edge lengths: $\ell_i = 0.1, 0.25, 0.5, 1.0$ (each fixed for all pendants $i = 1, \ldots, 5$) respectively. Columns correspond to different initial values for the internal edge lengths ℓ_6 and ℓ_7. The black lines show the initial velocity vector at the starting point, which was in each case zero for the pendant edges.

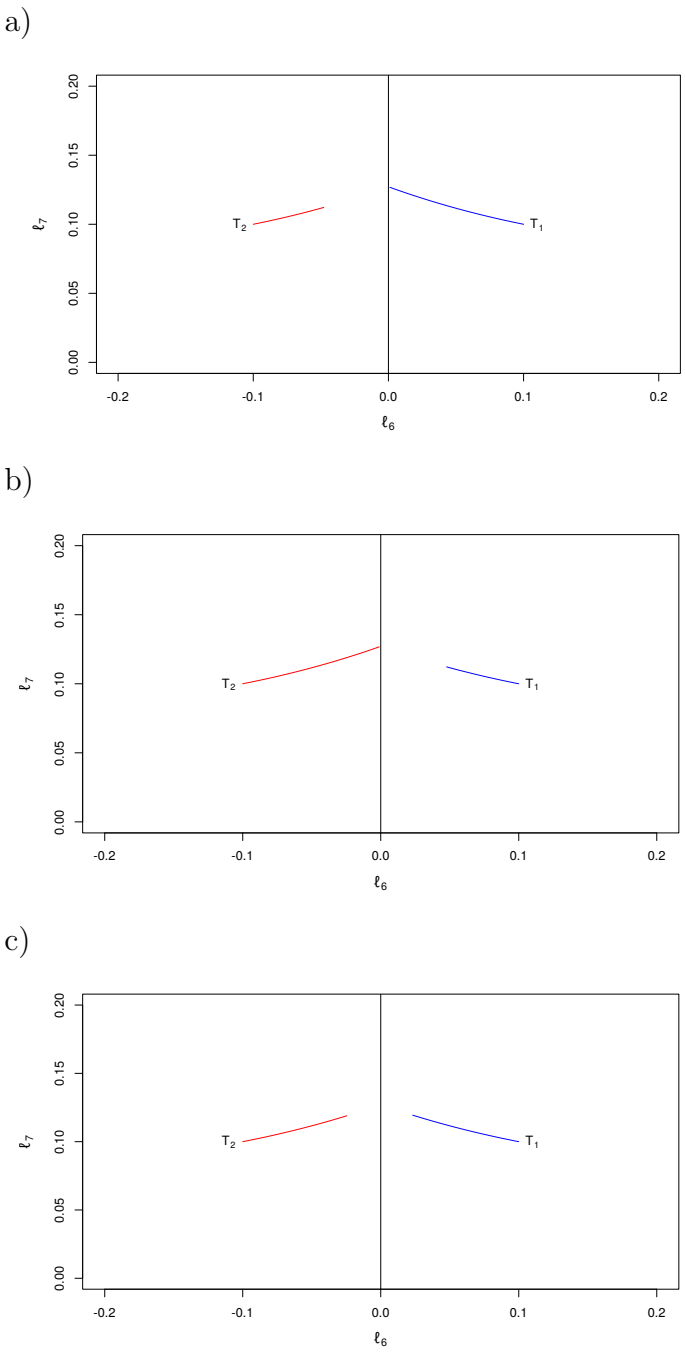


Figure 6.8: Approximate geodesic between two trees T_1 and T_2 in neighboring orthants in $\mathcal{E}_n$ constructed using Algorithm 1. Pendant edge lengths on both trees are 0.1. Projection is performed starting the algorithm from a) tree T_1 and b) tree T_2. In c), we choose the closest (in terms of covariance distance) to the point in the extrinsic geodesic between the projected points obtained in a) and b). Negative coordinates are due to graphical representation but they actually refer to positive edge lengths in different orthants.

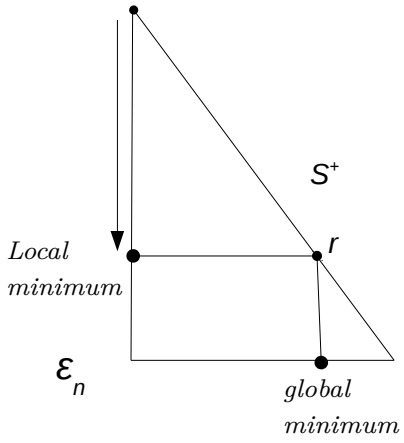


Figure 6.9: Illustration of the projection algorithm converging to a local minimum. The projection of $r \in S^+$ into tree space $\mathcal{E}_n$ can be trapped in a local minimum due to the non-convexity of the space, assuming the algorithm starts from the point s.

Algorithm 2: recursive construction

Let $T_{X_0} = T_1$. For $i = 1, \ldots, k$ where k is the number of steps, do the following

1. Compute extrinsic geodesic $\gamma(T_{X_{i-1}}, T_2)$ using (6.10).
2. Find the point $s \in S^+$ a proportion $1/(k-i+1)$ along $\gamma(T_{X_{i-1}}, T_2)$.
3. Let T_{X_i} be the projection of s into tree space $\mathcal{E}_n$. Projection is performed starting the algorithm from $T_{X_{i-1}}$.

The output of this algorithm is the set of projected points (trees) $T_{X_0}, \ldots, T_{X_k}$ in the space $\mathcal{E}_n$ which together form an approximate geodesic path between T_1 and T_2 in $\mathcal{E}_n$. However, experimentation shows that the path produced by Algorithm 2 is not symmetric in the sense that the path from T_1 to T_2 is different to the path from T_2 to T_1. This motivates the formulation of the symmetric algorithm below.

Algorithm 3: symmetric construction

Assume the number of steps k is even. Let $T_{X_0} = T_1$ and $T_{Y_0} = T_2$. For $i = 1, \ldots, k/2$, do

1. Compute extrinsic geodesic $\gamma(T_{X_{i-1}}, T_{Y_{i-1}})$ using (6.10).
2. Find points $r, s \in S^+$ proportions $1/(k-i+1)$ and $1 - 1/(k-i+1)$ along $\gamma(T_{X_{i-1}}, T_{Y_{i-1}})$.

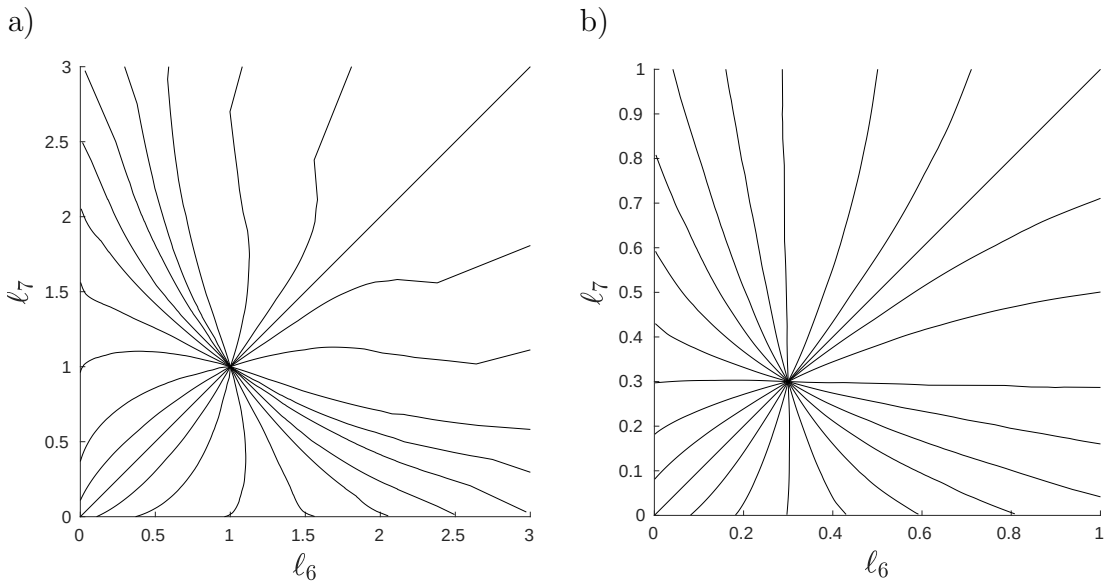


Figure 6.10: Approximate geodesics within single orthants in $\mathcal{E}_5$ constructed using Algorithm 2. Algorithm 3 produces very similar geodesic paths and these are similar to the paths produced by the "firing" algorithm in Section 6.5 (see Figure 6.7).

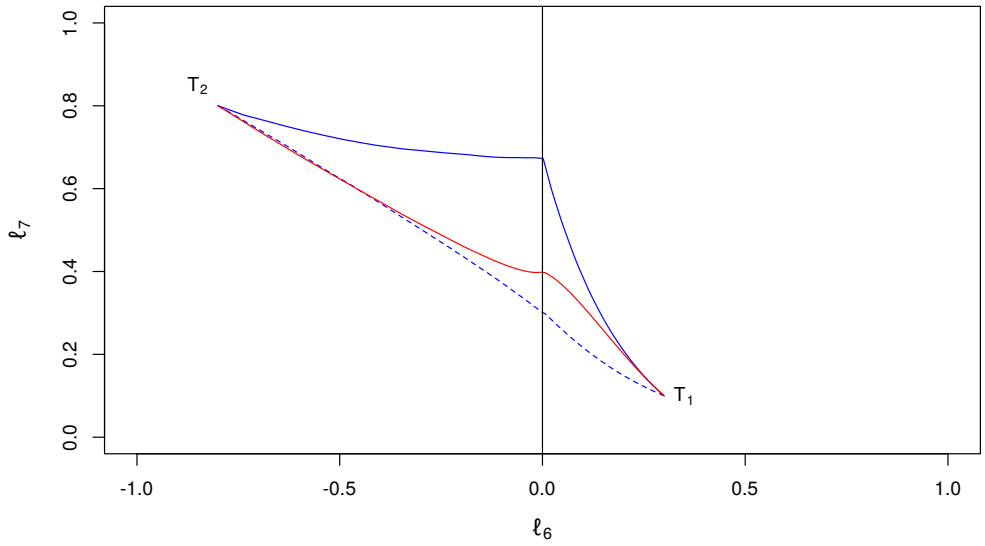
3. Let T_{X_i} be the projection of r into tree space $\mathcal{E}_n$ (starting the projection algorithm from $T_{X_{i-1}}$) and let T_{Y_i} be the projection of s into $\mathcal{E}_n$ (starting from $T_{Y_{i-1}}$).

The output is the collection of points $T_{X_0}, \ldots, T_{X_k}, T_{Y_k}, \ldots, T_{Y_0} \in \mathcal{E}_n$.

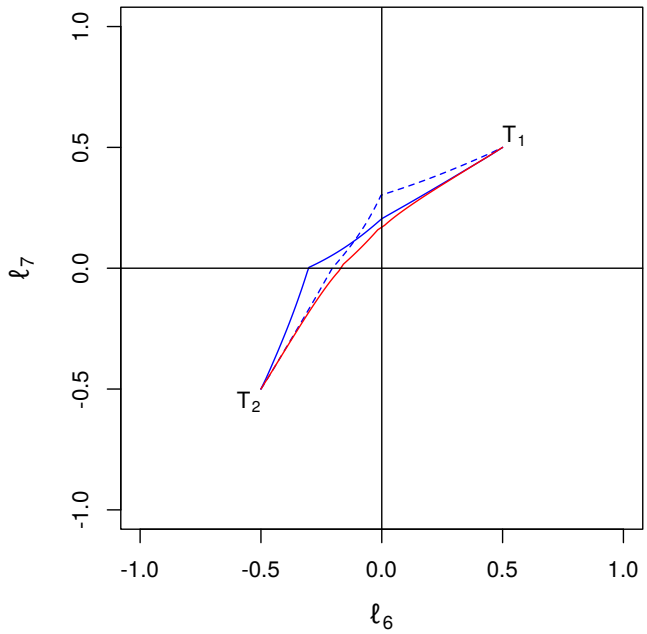
6.6.2 Results

We apply algorithms 2 and 3 to some small examples in $\mathcal{E}_5$. Figures 6.10 and 6.11 show geodesics between pairs of trees T_1 and T_2 in single orthants and across different orthants. The trees are determined by their internal edge lengths (ℓ_6, ℓ_7) at both end of the geodesics. All pendent edge lengths in Figure 6.11 are fixed at 0.1. Figure 6.10 is obtained in the following way. We deduce the end point (tree T_2) of each geodesic in Figure 6.7, which is a point at the boundary of the orthant. We use Algorithm 2 to construct a geodesic path from the central tree T_1 to T_2 for each geodesic in the figure, and the "reverse" path from T_2 to T_1. The exact geodesic should look the same under this operation (assuming the geodesic is unique). Within a single orthant, both algorithms produce very similar paths and these are similar to the paths produced by the "firing" algorithm in Section 6.5 (Figure 6.7). Some geodesics in Fig. 6.10a are not smooth which is probably due to the fact that the edge lengths are already very long.

However, when the end points T_1 and T_2 lie in different orthants, the forward path from Algorithm 2 does not match the backward path, which is in turn different from the symmetric path produced by Algorithm 3. See figures 6.11a and 6.11b. For these examples, we computed the length of each approximate geodesic by summing the extrinsic distance between each successive pair of points along the paths: 1.6097 (solid blue), 1.6163 (dashed blue), 1.6025 (red) (Fig. 6.11a) and 2.1505 (solid blue), 2.1505 (dashed blue), 2.1424 (red) (Fig. 6.11b). These results suggest that Algorithm 3 produces the shortest paths and is therefore our preferred algorithm. The computation time for both algorithms is fairly fast. For example, the time taken to construct all the geodesic paths in Figure 6.10b is 2 seconds using a desktop computer with an Intel(R) Core i7-4790S processor running at 3.20HGz.



a)



b)

Figure 6.11: Approximate geodesics in $\mathcal{E}_5$ across a) two neighboring orthants and b) three orthants. The blue (solid and dashed) geodesics were constructed using Algorithm 2: from T_1 to T_2 (solid) and from T_2 to T_1 (dashed), while the red geodesics are constructed with Algorithm 3. Negative coordinates are due to graphical representation but they actually refer to positive edge lengths in different orthants.

Chapter 7

Conclusion

7.1 Conclusion

In this thesis, we presented probability distances between phylogenetic trees when they are regarded as sequence models. These were defined in terms of the Hellinger distance, total variation distance, Kullback-Leibler divergence and Jensen-Shannon distance between distributions. These distances can be calculated exactly for small phylogenies. However, in general, we used Monte Carlo simulation methods to compute approximate distances. We have shown that the definition of these distances can be extended to account for missing taxa in the phylogenies. Through simulation, we devised an appropriate method to estimate an adequate sample size for our simulation methods. In order to reduce computational variability, we improved the sample size method using the control variate method of variance reduction. We demonstrated interesting properties of our probabilistic distances through several applications, and showed how they differ from existing metrics.

Furthermore, we constructed geodesics between phylogenies using information geometry techniques. Working exclusively on 5-taxon phylogenies, we solved the geodesic equation numerically within orthants. We discovered curved geodesics with irregularly spaced contours of distance travelled in contrast to straight lines with equally spaced distance contours in the well-known Billera-Holmes-Vogtmann (BHV) tree space. Geodesics in the BHV tree space are attracted to the origin, but geodesics here appear to be more inclined towards the boundary at infinity.

However, calculating geodesics in this way is not only computationally slow, but also based on shooting geodesics in different directions. In practice, we are usually given two end points between which we want to find the geodesic. To solve this problem, we showed a natural way to embed tree space in the space of positive semi-definite matrices,

or covariance matrices. Geodesics in this space can be computed analytically. Through this embedding, we constructed approximate geodesics in tree space with respect to the induced geometry in the space. We considered the boundary at infinity formally by defining the edge-product space which can be viewed as a compactification of the BHV tree space. Doing geometry in the edge-product space has been a long-standing problem in phylogenetics, and we have provided the first steps to solving this problem.

7.2 Future work

In this thesis, missing taxa on a phylogeny are joined with infinitely long edges in probability distances or equivalently, are uncorrelated with the remaining taxa in covariance matrix geometry. However, since all species evolve from the same common ancestor, it is more realistic to have missing taxa equally correlated with all other taxa. For example, in the covariance matrix, rather than assigning zero to the covariance associated with any missing taxon, we could assign a constant c. This is very likely to affect both the results of the probability distances and the covariance matrix geometry.

When exploring the concept of extrinsic mean in the space of covariance matrices, we considered a very brief example. There is need for further study of extrinsic mean compared to existing means, for example, the Fréchet mean, on real experimental data. With missing taxa, the extrinsic mean method is a so called "supertree" method (Baum, 1992; Ragan, 1992), or in other words a method that combines a collection of phylogenetic trees with different taxa into a single phylogeny (called supertree) on all the taxa. The extrinsic mean method for missing taxa takes a collection of phylogenetic trees with missing taxa, turns each phylogeny into a covariance matrix but assigns zero covariance to missing taxa, then calculates the extrinsic average of the resulting covariance matrices and projects it into tree space. This produces a phylogeny containing all the taxa. Therefore, this method could be compared with existing supertree methods and its properties explored further. This could be a basis of another paper.

The induced covariance geometry on tree space poses a number of open questions: (i) what can be said about the existence and uniqueness of geodesics in tree space with respect to the induced metric? Whilst we focus on computation, we don't know much about some basic issues, such as the curvature of the induced metric in tree space. The metric in the extrinsic space has a non-constant and non-positive sectional curvature which guarantees the existence and uniqueness of geodesic paths as well as centroid in the space (Lenglet *et al.*, 2006). This has not been investigated in the embedded tree space.

(ii) do geodesics ever go through infinity? A potential way to address this is to construct an explicit example of a geodesic path in tree space that goes through infinity. However, it would be interesting to have a theoretical justification of the idea of geodesics through infinity, in terms of curvature at the boundary at infinity.

Furthermore, the development of efficient variational methods for geodesic computation (see Schmidt *et al.* (2006) for example) in tree space could be applied to the construction of geodesics in the induced covariance geometry. The algorithm in Section 6.5, while exact, is not only computationally slow but also based on firing in a specified direction. A potential way of improving the algorithm is to incorporate a continuous variation in the initial direction until the target point is reached and a geodesic path is obtained. The problem of which direction should be taken at orthant boundaries would remain, but solutions might emerge using projections of geodesics in the extrinsic covariance matrix space.

Bibliography

Abecasis, A. B., Pingarilho, M. & Vandamme, A.-M. 2018 Phylogenetic analysis as a forensic tool in HIV transmission investigations. *AIDS* **32** (5), 543–554.

Ali, S. M. & Silvey, S. D. 1966 A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)* **28** (1), 131–142.

Allman, E. S., Ané, C. & Rhodes, J. A. 2008 Identifiability of a markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability* **40** (1), 229–249.

Amari, S. & Nagaoka, H. 2000 *Methods of Information Geometry*. American Mathematical Society.

Arnaoudova, E., Haws, D. C., Huggins, P., Jaromczyk, J. W., Moore, N., Schardl, C. L. & Yoshida, R. 2010 Statistical phylogenetic tree analysis using differences of means. *Frontiers in Neuroscience* **4**, 204.

Barzilai, J. & Borwein, J. M. 1988 Two-point step size gradient methods. *IMA Journal of Numerical Analysis* **8** (1), 141–148.

Basu, A., Mandal, A. & Pardo, L. 2010 Hypothesis testing for two discrete populations based on the Hellinger distance. *Statistics and Probability Letters* **80** (3-4), 206–214.

Baum, B. R. 1992 Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41** (1), 3–10.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D. & Sayers, E. W. 2018 GenBank. *Nucleic Acids Research* **46** (Database), D41–D47.

Bergsten, J. 2005 A review of long-branch attraction. *Cladistics* **21** (2), 163–193.

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

Bofkin, L. & Goldman, N. 2006 Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution* **24** (2), 513–521.

Bordewich, M. & Semple, C. 2005 On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics* **8** (4), 409–423.

Bröcker, T. & Jänich, K. 1982 *Introduction to differential topology*. Cambridge University Press.

Brodal, G. S., Fagerberg, R. & Pedersen, C. N. 2004 Computing the quartet distance between evolutionary trees in time O(n log n). *Algorithmica* **38** (2), 377–395.

Burbea, J. 1984 Information geometry of probability spaces. In *Technical report 84-52*.

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. 1999 Predicting the evolution of human influenza A. *Science* **286** (5446), 1921–1925.

Calvo, M. & Oller, J. M. 1991 An explicit solution of information geodesic equations for the multivariate normal model. *Statistics and Decisions* **9**, 119–138.

Cavalli-Sforza, L. L. & Edwards, A. W. F. 1967 Phylogenetic analysis: Models and estimation procedures. *Evolution* **21**, 550–570.

Chakerian, J. & Holmes, S. 2012 Computational tools for evaluating phylogenetic and hierarchical clustering trees. *Journal of Computational and Graphical Statistics* **21** (3), 581–599.

Chong, E. & Zak, S. 2013 *An Introduction to Optimization*. Wiley.

Clucas, B., Ord, T. J. & Owings, D. H. 2010 Fossils and phylogeny uncover the evolutionary history of a unique antipredator behaviour. *Journal of Evolutionary Biology* **23** (10), 2197–2211.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N. *et al.* 2009 Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Research* **37** (Database), D19–D25.

Costa, S. I., Santos, S. A. & Strapasson, J. E. 2015 Fisher information distance: A geometrical reading. *Discrete Applied Mathematics* **197**, 59–69.

Cover, T. M. & Thomas, J. A. 2006 *Elements of Information Theory*. New Jersey: John Wiley & Sons Inc.

Critchlow, D. E., Pearl, D. K. & Qian, C. 1996 The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology* **45** (3), 323–334.

Csiszár, I. 1963 Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markhoffschen ketten. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* **8**, 85–108.

Davenport, T. R., Stanley, W. T., Sargis, E. J., De Luca, D. W., Mpunga, N. E., Machaga, S. J. & Olson, L. E. 2006 A new genus of African monkey, Rungwecebus: Morphology, ecology, and molecular phylogenetics. *Science* **312** (5778), 1378–1381.

De Bruyn, A., Martin, D. P. & Lefeuvre, P. 2014 Phylogenetic reconstruction methods: An overview. *In Methods in Molecular Biology (Clifton, N.J.)* **1115**, 257–277.

Doolittle, W. F. & Brunet, T. D. P. 2016 What is the tree of life? *PLoS Genetics* **12** (4), e1005912.

Drummond, A. J. & Rambaut, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.

Eisen, J. A. & Wu, M. 2002 Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theoretical Population Biology* **61** (4), 481–487.

Endres, D. M. & Schindelin, J. E. 2003 A new metric for probability distributions. *IEEE Transactions on Information Theory* **49** (7), 1858–1860.

Eschenburg, J. H. 2018 Lecture notes on symmetric spaces. `http://myweb.rz.uni-augsburg.de/~eschenbu/symspace.pdf`, [Online; accessed 12 July 2018].

Estabrook, G. F., McMorris, F. R. & Meacham, C. A. 1985 Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Biology* **34** (2), 193–200.

Excoffier, L. & Yang, Z. 1999 Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution* **16** (10), 1357–1368.

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

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

Felsenstein, J. 2004 *Inferring phylogenies*. Sinauer Associates, Inc.

Fitch, W. M. 1967 Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of Molecular Biology* **26** (3), 499–507.

Fitch, W. M. & Margoliash, E. 1967 A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochemical Genetics* **1** (1), 65–71.

Fletcher, P. T. & Joshi, S. 2007 Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing* **87** (2), 250–262.

Förstner, W. & Moonen, B. 2003 A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium* (ed. E. W. Grafarend, F. W. Krumm & V. S. Schwarze), pp. 299–309. Springer.

Garba, M. K., Nye, T. M. W. & Boys, R. J. 2018 Probabilistic distances between trees. *Systematic Biology* **67** (2), 320–327.

Gascuel, O. 2005 *Mathematics of evolution and phylogeny*. Oxford University Press.

Gibbs, A. L. & Su, F. E. 2002 On choosing and bounding probability metrics. *International Statistical Review* **70** (3), 419–435.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gori, K., Suchan, T., Alvarez, N., Goldman, N. & Dessimoz, C. 2016 Clustering genes of common evolutionary history. *Molecular Biology and Evolution* **33** (6), 1590–1605.

Goujon, C. P., Schneider, V. M., Grofti, J., Montigny, J., Jeantils, V., Astagneau, P. *et al.* 2000 Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. *Journal of Virology* **74** (6), 2525–2532.

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

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

Hedges, S. B., Moberg, K. D. & Maxson, L. R. 1990 Tetrapod phylogeny inferred from 18s and 28s ribosomal rna sequences and a review of the evidence for amniote relationships. *Molecular Biology and Evolution* **7** (6), 607–633.

Hickey, G., Dehne, F., Rau-Chaplin, A. & Blouin, C. 2008 SPR distance computation for unrooted trees. *Evolutionary Bioinformatics* **4**, 17–27.

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

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

Jost, J. 2017 *Riemannian Geometry and Geometric Analysis*. Springer International Publishing.

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. N. Munro), pp. 21–132. Academic Press.

Karcher, H. 1977 Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics* **30** (5), 509–541.

Kellogg, E. A. 2001 Evolutionary history of the grasses. *Plant Physiology* **125** (3), 1198–1205.

Kenah, E., Britton, T., Halloran, M. E. & Longini, Jr., I. M. 2016 Molecular infectious disease epidemiology: Survival analysis and algorithms linking phylogenies to transmission trees. *PLOS Computational Biology* **12** (4), e1004869.

Kendall, M. & Colijn, C. 2015 A tree metric using structure and length to capture distinct phylogenetic signals. ArXiv e-prints, 1507.05211, `http://adsabs.harvard.edu/abs/2015arXiv150705211K`.

Kim, J. 2000 Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Molecular Phylogenetics and Evolution* **17** (1), 58–75.

Kuhner, M. K. & Felsenstein, J. 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11** (3), 459–468.

Kwak, S. G. & Kim, J. H. 2017 Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology* **70** (2).

Lam, T. T.-Y., Hon, C.-C. & Tang, J. W.-T. 2010 Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical Reviews in Clinical Laboratory Sciences* **47** (1), 5–49.

Lenglet, C., Rousson, M., Deriche, R. & Faugeras, O. 2006 Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *Journal of Mathematical Imaging and Vision* **25** (3), 423–444.

Li, M. & Zhang, L. 1999 Twist-rotation transformations of binary trees and arithmetic expressions. *Journal of Algorithms* **32** (2), 155–166.

Lin, Y., Rajan, V. & Moret, B. M. E. 2012 A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9** (4), 1014–1022.

Lió, P. & Goldman, N. 1998 Models of molecular evolution and phylogeny. *Genome Research* **8**, 1233–1244.

Miller, E., Owen, M. & Provan, J. S. 2012 Averaging metric phylogenetic trees. `http://comet.lehman.cuny.edu/owen/pub/mean_trees.pdf`, unpublished manuscript.

Moakher, M. 2005 A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* **26** (3), 735–747.

Mooers, A. O. & Heard, S. B. 1997 Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology* **72** (1), 31–54.

Morgan, B. J. T. 1984 *Elements of simulation*. Chapman and Hall.

Moulton, V. & Steel, M. 2004 Peeling phylogenetic oranges. *Advances in Applied Mathematics* **33** (4), 710–727.

Ng, A. Y., Jordan, M. I. & Weiss, Y. 2001 On spectral clustering: Analysis and an algorithm. *Neural Information Processing Symposium* .

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

Nye, T. M. W., Tang, X., Weyenberg, G. & Yoshida, R. 2017 Principal component analysis and the locus of the frchet mean in the space of phylogenetic trees. *Biometrika* **104** (4), 901–922.

Ou, C.-Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C.-C., Korber, B. T. *et al.* 1992 Molecular epidemiology of HIV transmission in a dental practice. *Science* **256** (5060), 1165–1171.

Owen, M. & 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.

Pearson, W. R. 2013 An introduction to sequence similarity (homology) searching. *Current Protocols in Bioinformatics* **42**, 3.1.1–3.1.8.

Penny, D. & Hendy, M. D. 1985 The use of tree comparison metrics. *Systematic Zoology* **34** (1), 75–82.

Penny, D., Watson, E. E. & Steel, M. A. 1993 Trees from languages and genes are very similar. *Systematic Biology* **42** (3), 382–384.

Peterson, P. 2006 *Riemannian Geometry*. Springer.

Phillips, A., Janies, D. & Wheeler, W. 2000 Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution* **16** (3), 317–330.

Ragan, M. A. 1992 Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1** (1), 53–58.

Robinson, D. F. & Foulds, L. R. 1979 Comparison of weighted labelled trees. In *Combinatorial Mathematics VI* (ed. A. F. Horadam & W. D. Wallis), pp. 119–126. Berlin, Heidelberg: Springer Berlin Heidelberg.

Robinson, D. F. & Foulds, L. R. 1981 Comparison of phylogenetic trees. *Mathematical Biosciences* **53** (1-2), 131–147.

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. 2003 Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804.

Rosenthal, J. S. 1995 Convergence rates for markov chains. *SIAM Review* **37** (3), 387–405.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4** (4), 406–425.

Salicru, M., Morales, D., Menendez, M. L. & Pardo, L. 1994 On the applications of divergence type measures in testing statistical hypotheses. *Journal of Multivariate Analysis* **51** (2), 372–391.

Sarkar, S. & Basu, A. 1995 On disparity based robust tests for two discrete populations. *The Indian Journal of Statistics* **57** (3), 353–364.

Saslis-Lagoudakis, C. H., Savolainen, V., Williamson, E. M., Forest, F., Wagstaff, S. J., Baral, S. R. *et al.* 2012 Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences* **109** (39), 15835–15840.

Sason, I. & Verdú, S. 2016 f-divergence inequalities. *IEEE Transactions on Information Theory* **62** (11), 5973–6006.

Schmidt, F. R., Clausen, M. & Cremers, D. 2006 Shape matching by variational computation of geodesics on a manifold. In *Pattern Recognition* (ed. K. Franke, K.-R. Müller, B. Nickolay & R. Schäfer), pp. 142–151. Berlin, Heidelberg: Springer Berlin Heidelberg.

Siljic, M., Salemovic, D., Jevtovic, D., Pesic-Pavlovic, I., Zerjav, S., Nikolic, V., Ranin, J. & Stanojevic, M. 2018 Forensic application of phylogenetic analysis - exploration of suspected epidemiological linkage. *BMC Infectious Diseases* **14** (Suppl 4), O21.

Skovgaard, L. T. 1984 A riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics* **11** (4), 211–223.

Smith, W. L. & Wheeler, W. C. 2006 Venom evolution widespread in fishes: A phylogenetic road map for the bioprospecting of piscine venoms. *Journal of Heredity* **97** (3), 206–217.

Squartini, F. & Arndt, P. F. 2008 Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular Biology and Evolution* **25** (12), 2525–2535.

Stamatakis, A. 2006 RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22** (21), 2688–2690.

Stockham, C., Wang, L.-S. & Warnow, T. 2002 Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics* **18** (Suppl 1), S285–S293.

Tavaré, S. 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. In *American Mathematical Society: Lectures on Mathematics in the Life Sciences* (ed. R. M. Miura), pp. 57–86. Amer Mathematical Society.

Tierney, L. 1994 Markov chains for exploring posterior distributions. *The Annals of Statistics* **22** (4), 1701–1728.

Uzzell, T. & Corbin, K. W. 1971 Fitting discrete probability distributions to evolutionary events. *Science* **172** (3988), 1089–1096.

Vézquez, D. P. & Gittleman, J. L. 1998 Biodiversity conservation: Does phylogeny matter? *Current Biology* **8** (11), R379–R381.

Wakeley, J. 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution* **37** (6), 613–623.

Waterman, M. S. & Smith, T. F. 1978 On the similarity of dendograms. *Journal of Theoretical Biology* **73** (4), 789–800.

Whidden, C. & Matsen, F. A. 2015 Quantifying MCMC exploration of phylogenetic tree space. *Systematic Biology* **64** (3), 472–491.

Yang, Z. 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10** (6), 1396–1401.

Yang, Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39** (3), 306–314.

Yang, Z. 1997 PAML: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13** (5), 555–556.