Tropical Geometry of Phylogenetic Tree Space: A Statistical Perspective

Anthea Monod[1,†], Bo Lin[2], Ruriko Yoshida[3], and Qiwen Kang[4]

1 Department of Mathematics, Imperial College London, UK
2 School of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA
3 Department of Operations Research, Naval Postgraduate School, Monterey, CA, USA
4 Medpace, Inc. and Department of Statistics, University of Kentucky, Lexington, KY, USA

† **Corresponding e-mail: a.monod@imperial.ac.uk**

arXiv:1805.12400v7 [math.MG] 3 Nov 2021

Abstract

Phylogenetic trees are the fundamental mathematical representation of evolutionary processes in biology. They are also objects of interest in pure mathematics, such as algebraic geometry and combinatorics, due to their discrete geometry. Although they are important data structures, they face the significant challenge that sets of trees form a non-Euclidean phylogenetic tree space, which means that standard computational and statistical methods cannot be directly applied. In this work, we explore the statistical feasibility of a pure mathematical representation of the set of all phylogenetic trees based on tropical geometry. We show that the tropical geometric phylogenetic tree space endowed with a generalized Hilbert projective metric exhibits analytic, geometric, and topological properties that are desirable for theoretical studies in probability and statistics. Moreover, this approach exhibits increased computational efficiency and statistical performance over the current state-of-the-art, which we illustrate with a real data example on seasonal influenza. Our results demonstrate the viability of the tropical geometric setting for parametric statistical and probabilistic studies of sets of phylogenetic trees.

Keywords: BHV tree space; phylogenetic tree space; tree metric; tropical geometry; tropical metric.

1 Introduction

Evolutionary relationships describing how organisms are related by a common ancestor are represented in a branching diagram known as a *phylogenetic tree*. Phylogenetic trees model many important and diverse biological phenomena, such as speciation, the spread of pathogens, and cancer evolution. Methodology to analyze phylogenetic datasets has been under active research for several decades for two important reasons. First, explicit computations directly on collections of phylogenetic trees are challenging due to high dimensionality in terms of a large number of leaves, a long evolutionary history, and an intricate branching pattern. Second, standard statistical methodologies are not directly applicable due to the non-Euclidean nature of the trees themselves as well as the set that they make up. Significant previous work addresses various classical statistical interests, however a fundamental breakthrough for quantitative studies on sets of trees emerged through studying the geometry of the set of all phylogenetic trees (Billera et al., 2001).

Referred to in the literature as *BHV tree space*—after the authors Billera, Holmes, and Vogtmann—the set of all phylogenetic trees is studied in a setting where each tree is represented as an individual point. The geometry is characterized by a unique geodesic between any two points; its length defines a metric on the space. Since its introduction in 2001, it has been actively studied in various wide-reaching domains, including algebraic geometry (e.g., Devadoss and Morava, 2015), category theory (e.g., Baez and Otter, 2017), computational biology (e.g., Weyenberg and Yoshida, 2016), and statistical genetics (e.g., Nye et al., 2017). Despite its indisputable significance, the BHV geometry nevertheless poses significant data-analytic complications for both descriptive and inferential statistics.

Considering an alternative approach from pure mathematics based on *tropical geometry*—a variant of algebraic geometry—alleviates some of these complications and is a promising alternative approach for probability-based statistics on sets of phylogenetic trees. The first formal connection between tropical geometry and mathematical phylogenetics arises in the space of phylogenetic trees in relation to a particular

tropical algebraic variety (Speyer and Sturmfels, 2004). This coincidence has been further studied in theoretical research (e.g., Ardila and Klivans (2006); Manon (2011)), however, its implication and potential in applied work remain largely understudied and untapped.

In this paper, we explore the tropical geometric perspective of phylogenetic tree space with the aim of enabling exact distributional theory and parametric statistical inference. Specifically, we study the subspace of the *tropical projective torus* corresponding to the space of phylogenetic trees equipped with a generalized projective Hilbert metric, which we refer to as the *tropical metric*. We refer to this metric space as *palm tree space* (tropical tree space) and show that it satisfies fundamental assumptions to ensure that probabilistic and parametric statistical questions are valid and well-defined. Moreover, this setting exhibits improvements in computational efficiency and improved statistical performance over the BHV setting, which we demonstrate via a real-data application to seasonal influenza data.

The remainder of this paper is organized as follows. In Section 2, we provide background and motivation for our study. In Section 3, we discuss properties of the tropical metric on the space of phylogenetic trees and formally define palm tree space. We study its geometry, topology, and analytic properties in relation to BHV space. We also give some examples of theoretical probability measures used in statistics and that are important in probability theory. In Section 4, we give an example of a statistical analysis on real data in both palm tree and BHV space. We close in Section 5 with a discussion, and some directions for future research.

2 Background and Motivation

Phylogenetic trees are symbolic objects that model evolutionary divergence from a common ancestor. In computational biology, the reconstruction of a phylogenetic tree from an input of sequence alignment data (e.g., DNA and RNA) is a challenging problem; reconstruction methods are known to be highly sensitive to the input sequences (different genes or coding regions will give rise to different trees), measurement errors (alignment or sequencing errors), and noise typical to this type of biological data (e.g., Leaché and Rannala, 2010). This sensitivity naturally invites the question of how to compare trees, for example, arising from different reconstruction methods. Mathematically, comparing objects entails measuring the distance between them; in the context of phylogenetic trees, this gives rise to a *tree space* equipped with a *metric between trees*. One of the most significant challenges in computational work with phylogenetic trees as data objects is that their graphical structure gives rise to a non-Euclidean tree space.

In this section, we provide the mathematical background to phylogenetic trees from the pure mathematical perspective and the statistical motivation for studying this perspective.

2.1 Defining Phylogenetic Trees

In what follows, we consider $N \in \mathbb{N}_0$ and set $n := \binom{N}{2}$. A tree is an acyclic connected graph $T = (V, E)$, defined by a set of vertices V and a set of edges E. An *N-tree* is a tree with N labeled terminal nodes called *leaves*. Edges connecting to leaves are called *external edges*, other edges are called *internal edges*. A *binary N-tree* is an N-tree with the following conditions on the degree of a vertex $v \in V$: if $\deg(v) = 2$, then v is the *root* of the tree and it is unique; if $\deg(v) = 1$, then v is a leaf; and if $\deg(v) = 3$, then v is an internal vertex. A *tree topology* is the "shape" of the tree; it is a branching configuration of edges together with a leaf labelling scheme. There are $(2N-3)!!$ binary tree topologies on N leaves (Schröder, 1870). A *metric N-tree* is a tree with zero or positive lengths on all edges; metric N-trees are also known as *phylogenetic trees*. We denote the space of phylogenetic trees with N leaves by $\mathcal{T}_N$.

A phylogenetic tree may be represented by all pairwise distances between leaves. Let w_{ij} denote the distance between leaves i and j, given by the sums of edge lengths along the unique path between i and j. The $N \times N$ matrix W with entries w_{ij} then represents a phylogenetic tree. Since W is symmetric with zeros along the diagonal, the upper-triangular portion of the matrix contains all of the unique information needed to specify a phylogenetic tree in terms of its pairwise distances. Define the following map to vectorize this information:

$$\begin{aligned} &\mathcal{W} : \mathcal{T}_N \to \mathbb{R}^n, \\ &\quad W \mapsto w = (w_{12}, w_{13}, \ldots, w_{1N}, w_{23}, \ldots, w_{2N}, \ldots, w_{(N-1)N}). \end{aligned} \tag{1}$$

Notice that W is, in essence, a distance matrix or a metric. However, in order for the distance matrix W to represent a phylogenetic tree, the following additional condition must be satisfied.

Definition 1 (Four-Point Condition (Buneman, 1974; Maclagan and Sturmfels, 2015))**.** A distance matrix W represents a phylogenetic tree if it satisfies the conditions to be a metric and the maximum among the following *Plücker relations* is attained at least twice for $1 \leq i < j < k < \ell \leq N$:

$$w_{ij} + w_{k\ell}, \qquad w_{ik} + w_{j\ell}, \qquad w_{i\ell} + w_{jk}, \tag{2}$$

or equivalently, that

$$w_{ij} + w_{k\ell} \leq \max(w_{ik} + w_{j\ell},\, w_{i\ell} + w_{jk}) \tag{3}$$

for all distinct $i, j, k, \ell \in \{1, 2, \ldots, N\}$.

A distance matrix W satisfying the conditions of Definition 1 is known as a *tree metric*. Note that tree metrics represent phylogenetic trees; these differ from metrics between trees.

Example 2. The tree metric $w \in \mathbb{R}^6$ for the tree in Figure 1 expressed as a vector is $(w_{PQ},\, w_{PR},\, w_{PS},\, w_{QR},\, w_{QS},\, w_{RS})$. As a matrix W, it is

$$\begin{pmatrix} 0 & w_{PQ} & w_{PR} & w_{PS} \\ & 0 & w_{QR} & w_{QS} \\ & & 0 & w_{RS} \\ & & & 0 \end{pmatrix} = \begin{pmatrix} 0 & a+b & a+c+d & a+c+e \\ & 0 & b+c+d & b+c+e \\ & & 0 & d+e \\ & & & 0 \end{pmatrix}.$$

The Plücker relations (2) associated with W are

$$\begin{aligned} A &:= w_{PQ} + w_{RS} = a + b + d + e, \\ B &:= w_{QR} + w_{PS} = a + b + 2c + d + e, \\ C &:= w_{PR} + w_{QS} = a + b + 2c + d + e. \end{aligned}$$

The maximum $B = C$ is achieved exactly twice, and $B - A = 2c > 0$. Also, (3) holds: $A \leq \max\{B, C\} = B$.

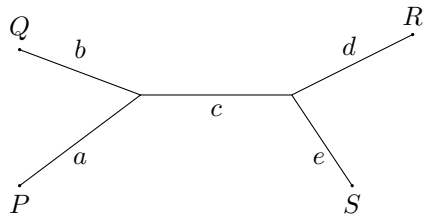

Figure 1: Example of an unrooted phylogenetic tree to illustrate the four-point condition.

2.2 Metrics on Tree Spaces: BHV Space

Various metrics between trees have been derived in biology. A notable class of metrics strives to retain the inner product property akin to Euclidean distance, which makes them popular due to their integrability into a wide range of statistical approaches, such as functional and nonparametric modeling. One metric from this class extensively used in biology is the *Robinson–Foulds metric* (Robinson and Foulds, 1981). This metric (and other inner-product distances between trees) is known to suffer from structural and interpretive errors. For example, many pairs of trees measure the same distance apart; also, large distances between trees, counterintuitively, do not necessarily indicate a disparity in ancestral heritage (Steel and Penny, 1993). Other commonly-occurring distances include the nearest neighbor interchange metric (Waterman et al., 1976), subtree transfer distance (Allen and Steel, 2001), and variational distance (Steel and Székely, 2006). For a detailed survey and review of metrics between trees, see Weyenberg and Yoshida (2016); St. John (2016). A pioneering approach that bypasses difficulties and limitations of these metrics focuses on the geometry of tree space (Billera et al., 2001).

Specifically, the space of phylogenetic trees is modeled as a *moduli space*, where each point in the space represents a phylogenetic tree. Trees are expressed only by the lengths of their internal edges, which are recorded as entries in a vector of dimension $N-2$ since in a binary tree, there are at most $N-2$ internal edges. External edges are not considered, since taking them into account does not affect the geometry of the space: including external edges simply amounts to taking the product of tree space with an N-dimensional Euclidean space. A nonnegative Euclidean orthant $\mathbb{R}^{N-2}_{\geq 0}$ is associated to each tree topology. BHV space may also be interpreted combinatorially: For each orthant, the *link of the origin*

$$\mathcal{L}_N := \left\{(x_1, \ldots, x_{N-2}) \mid \sum_i x_i = 1\right\} \tag{4}$$

gives rise to a simplicial complex of dimension $N-3$. BHV space is then an infinite cone over $\mathcal{L}_N$.

The $(2N-3)!!$ orthants are grafted at right-angles to make up the tree space, which gives rise to a property of nonpositive curvature known as CAT(0). In CAT(0) spaces, there is a unique shortest path between any two points; here, this is the *BHV geodesic*. To compute BHV geodesics, first, the geodesic distance between two trees is computed, and then the external branch lengths are considered to compute the overall geodesic distance between two trees, by taking the differences between external branch lengths. Since each orthant is locally viewed as a Euclidean space, the shortest path between two points within a single orthant is a straight Euclidean line. The difficulty appears in establishing which sequence of orthants joining the two topologies contains the geodesic. In the case of four leaves, this can be readily determined using a systematic grid search, but such a search is intractable with larger trees. Owen and Provan (2011) present a quartic-time algorithm (in the number of leaves N) for finding the geodesic path between any two points in this tree space, which is the currently the fastest available method with a time complexity of $O(N^4)$. The length structure of the BHV geodesic induces the *BHV metric* d_{BHV} on this space. This setup has come to be known as *BHV space* $\mathcal{T}_N^{\text{BHV}}$ and is ubiquitous even in non-biological fields, including computer vision, combinatorics, and category theory. It has also been proposed as the definitive setting for computational studies on sets of phylogenetic trees (Gavryushkin and Drummond, 2016).

It turns out that BHV space poses considerable limitations for classical descriptive and inferential statistics. On the descriptive front, the convex hull of finitely many points in tree space with edges given by BHV geodesics is unbounded in dimension (Lin et al., 2017), so there exists no obvious subspace for projections and no lower dimensional representations of data. This is restrictive for classical dimensionality reduction and data visualization methods, such as principal component analysis (PCA). On the inferential front, in BHV space, Fréchet means are *sticky*: the mean fails to be injective and "sticks" to lower dimensional strata (Hotz et al., 2013); see Example 3. Thus, perturbing points in a sample results in no change in the mean, meaning that exact parametric asymptotic results cannot be derived, which prohibits classical exact statistical inference.

Example 3. In Figure 2, we position three unit masses on the 3-spider, which is the stratified space of three $\mathbb{R}_{\geq 0}$ rays joined at the origin. This is precisely the BHV space of phylogenetic trees with three leaves and fixed external edge lengths. The position x of the barycenter (Fréchet mean) is calculated by minimizing $2(1+x)^2 + (a-x)^2$. The solution is $x = 0$ for $a < 2$, and $x = (a-2)/3$ for $a \geq 2$. The Fréchet mean tends to stick to lower-dimensional strata.

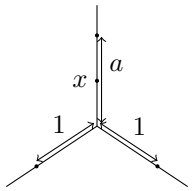

Figure 2: 3-spider to illustrate stickiness.

Sophisticated methods have been developed to bypass these difficulties. Nye et al. (2017) construct a locus of BHV Fréchet means and prove that its dimensionality is well-behaved and serves as a suitable lower dimensional projective space, while Barden et al. (2018) derive a central limit theorem for BHV Fréchet means by implementing a generalized delta method. Inferential techniques have also been proposed based

on this generalized delta method strategy; e.g., Willis (2019); Willis and Bell (2018). These, and other proposed methods, are largely approximate, rather than exact statistical methods; additionally, they tend to be nonparametric, rather than parametric. These statistical challenges have spurred recent proposals of alternative tree spaces (Garba et al., 2021).

2.3 Tropical Geometry and Phylogenetic Tree Space

In this work, we focus on the appearance of phylogenetic tree space in tropical geometry in the groundbreaking work of Speyer and Sturmfels (2004), who formally connect the space of phylogenetic trees and the *tropical Grassmannian*. We now outline the connection between tropical geometry and phylogenetic tree space. To do this, we return to the map $\mathcal{W}$ (1). Specifically, we would like to understand what the image of $\mathcal{W}$ is: if it is a linear space, then theory from linear algebra is applicable; if it is a manifold, then principles of Riemannian geometry may be applied. It turns out that the image of $\mathcal{W}$ is tropical geometric, so new tools for statistics are needed.

To see this, notice that the embedding (1) of trees into Euclidean space may be refined: if we do not wish to distinguish between phylogenetic trees differing by a constant on each external edge, we may consider the quotient space $\mathbb{R}^n/\mathbb{R}\mathbf{1}$, where $\mathbf{1}$ is the all-one vector $(1, 1, \ldots, 1)$, which gives a reduction in dimension. The quotient space $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is known as the *tropical projective torus* and it is generated by an equivalence relation $\sim$ specifying that for two points $x, y \in \mathbb{R}^n$, $x \sim y$ if and only if all coordinates of their difference $x - y$ are equal. In the context of trees, the quotient normalizes evolutionary time between trees. The tropical projective torus is the ambient space of the space of the space of phylogenetic trees; $\mathcal{T}_N$ is a proper subset of $\mathbb{R}^n/\mathbb{R}\mathbf{1}$. The tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ may also be generated by a group action: Let $G := \{(c, c, \ldots, c) \in \mathbb{R}^n \mid c \in \mathbb{R}\}$ with coordinate-wise addition, then G is an additive group. G acts on $\mathbb{R}^n$ as follows: for $g \in G$ and $x \in \mathbb{R}$,

$$g \circ x = (x_1 + g_1,\, x_2 + g_2,\, \ldots,\, x_n + g_n).$$

Each point in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is then exactly one orbit under the group action of G on $\mathbb{R}^n$.

Furthermore, if we disregard differences on external edges, we may consider the quotient space $\mathbb{R}^n/\operatorname{im}(\varphi)$ where the map $\varphi : \mathbb{R}^N \to \mathbb{R}^n$ is given by $\varphi(x_1, \ldots, x_N) = (x_1 + x_2,\, x_1 + x_3,\, \ldots,\, x_{N-1} + x_N)$. We thus obtain the following sequence of maps:

$$\mathcal{T}_{N-1} \to \mathbb{R}^n \to \mathbb{R}^n/\mathbb{R}\mathbf{1} \to \mathbb{R}^n/\operatorname{im}(\varphi). \tag{5}$$

In algebraic geometry, the solution sets of systems of polynomial equations—referred to as *algebraic varieties*—are studied. In *tropical geometry*, these polynomial equations are defined in the *tropical semiring*, $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$ where $a \oplus b := \min(a, b)$ and $a \odot b := a + b$. Tropical mathematics involves studying various mathematical objects and problems which are defined using these operations. For example, let a^N denote the tropical product of a with itself N times; let $\mathcal{A} \subset \mathbb{N}^N$. Tropical polynomials are piecewise linear functions:

$$f(x_1, \ldots, x_N) = \bigoplus_{a \in \mathcal{A}} c_a \odot x_1^{a_1} \odot \cdots \odot x_N^{a_N} = \min_{a \in \mathcal{A}}(c_a + a_1 x_1 + \cdots + a_N x_N).$$

A *tropical hypersurface* $\mathcal{H}(f)$ is the set of all $(x_1, \ldots, x_N) \in \mathbb{R}^N$ where f is attained at least twice as a runs over $\mathcal{A}$.

Notice that the Plücker relations (2) given in Definition 1 are tropical polynomials, and thus, the set of all phylogenetic trees constitutes a tropical hypersurface with min replaced by max. Note also that the max-plus semiring $(\mathbb{R} \cup \{-\infty\}, \boxplus, \odot)$, where $a \boxplus b := \max(a, b)$, is isomorphic to the tropical semiring. Thus, the four-point condition defining phylogenetic trees is tropical.

In algebraic geometry, the real Grassmannian $G_{2,n}$ is the following projective variety in the projective space $\mathbb{P}^{n-1}$:

$$G_{2,N} = \left\{(x_{12},\, x_{13},\, \ldots,\, x_{(N-1)N}) \ \in \mathbb{P}^{n-1} \mid x_{ij}x_{k\ell} - x_{ik}x_{j\ell} + x_{i\ell}x_{jk} = 0 \text{ for } 1 \le i < j < k < \ell \le N\right\}.$$

The *tropical Grassmannian* $\mathcal{G}_{2,N}$ is then obtained by replacing the polynomial by its tropicalization and the vanishing set by intersections of tropical hypersurfaces. In other words, $\mathcal{G}_{2,N}$ is given by the intersection of tropical hypersurfaces $\mathcal{H}(x_{ij} \odot x_{k\ell} \oplus x_{ik} \odot x_{j\ell} \oplus x_{i\ell} \odot x_{jk})$ for $1 \le i < j < k < \ell \le N$.

To visualize $\mathcal{G}_{2,N}$, we have the following behavior of images through the sequence of maps (5): the image of $\mathcal{G}_{2,N}$ in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is a fan $\mathcal{G}'_{2,N}$ of dimension $(2N-2)$; the image of $\mathcal{G}'_{2,N}$ in $\mathbb{R}^n/\operatorname{im}(\varphi)$ is a fan $\mathcal{G}''_{2,N}$ of dimension $N-3$; and intersecting $\mathcal{G}''_{2,N}$ with the unit sphere yields a polyhedral complex $\mathcal{G}'''_{2,N}$, where each facet $\mathcal{G}'''_{2,N}$ is a polytope of dimension $N-4$. The insightful result that Speyer and Sturmfels (2004) prove is that $\mathcal{G}''_{2,N}$ coincides with $\mathcal{T}^{\mathrm{BHV}}_{N-1}$, $\mathcal{G}'''_{2,N}$ coincides with $\mathcal{L}_{N-1}$ (4), and the image of $\mathcal{W}$ is precisely the tropical Grassmannian $\mathcal{G}_{2,N}$.

Example 4. As an illustrative example, we study the case of $N=4$ leaves: $\mathcal{G}_{2,4}$ is the hypersurface $\mathcal{H}(x_{12}\odot x_{34}\oplus x_{13}\odot x_{24}\oplus x_{14}\odot x_{23})$, which is the collection of points such that at least one of the following systems holds: $x_{12}+x_{34}=x_{13}+x_{24}\leq x_{14}+x_{23}$, $x_{12}+x_{34}=x_{14}+x_{23}\leq x_{13}+x_{24}$, $x_{14}+x_{23}=x_{13}+x_{24}\leq x_{12}+x_{34}$. For each system, equality determines a 5-dimensional hyperplane in $\mathbb{R}^6$, while inequality determines a closed

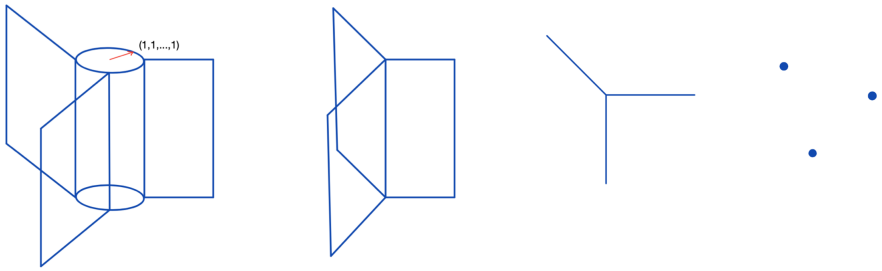

Figure 3: Visualizing the tropical Grassmannian $\mathcal{G}_{2,4}$. From left to right, we have the images of $\mathcal{G}_{2,4}$, $\mathcal{G}'_{2,4}$, $\mathcal{G}''_{2,4}$, and $\mathcal{G}'''_{2,4}$. Notice that $\mathcal{G}''_{2,4}$ is $\mathcal{T}^{\mathrm{BHV}}_3$ and $\mathcal{G}'''_{2,4}$ is $\mathcal{L}_3$.

half-space in $\mathbb{R}^6$. Their intersection is isomorphic to $\mathbb{R}^4\times\mathbb{R}_{\geq 0}$. Since there are three systems, $\mathcal{G}_{2,4}$ is the union of three copies of $\mathbb{R}^4\times\mathbb{R}_{\geq 0}$ glued along the space $x_{12}+x_{34}=x_{13}+x_{24}=x_{14}+x_{23}$, which is the image of $\varphi:\mathbb{R}^4\to\mathbb{R}^6$. $\mathcal{G}''_{2,4}$ then consists of three copies of $\mathbb{R}_{\geq 0}$ (i.e., $\mathcal{T}^{\mathrm{BHV}}_3$; see also Example 3) and $\mathcal{G}'''_{2,4}$ consists of three points (i.e., $\mathcal{L}_3$).

Other Applications of Tropical Geometry

Tropical geometry also arises in other applied settings, specifically in computer science, statistics, and economics: Maclagan and Sturmfels (2015) and Pachter and Sturmfels (2005) discuss the use of tropical mathematics to reinterpret the dynamic programming approach to the problem of sequence alignment for molecular data in computational biology. In the context of statistics, tropical geometry arises in the reinterpretation of various stochastic models. As a field of research, one of the core principles of algebraic statistics is the fact that algebraic varieties and semi-algebraic sets are statistical models (Améndola et al., 2018; Sullivant, 2018). For statistical models, such as graphical models, tropicalized statistical models (i.e., tropicalized algebraic varieties) are fundamental in parametric inference, which was specifically demonstrated on hidden Markov models and general Markov models on binary trees (Pachter and Sturmfels, 2004). In computational biology, this algebraic statistical approach was adapted to study invariants of joint probability distributions on leaf labels of phylogenetic trees (Sturmfels and Sullivant, 2005). In economics and finance, tropical geometry arises in game theoretic settings and max-linear models for financial data (Einmahl et al., 2018; Lin and Tran, 2019).

3 Palm Tree Space

A fundamental requirement to comparative and statistical studies on the tropical geometric interpretation of phylogenetic tree space is a metric. On the tropical projective torus, a *generalized Hilbert projective metric* has been used in other settings (e.g., Joswig et al., 2007; Akian et al., 2011; Cohen et al., 2004). We adapt this metric in our studies and refer to it as the *tropical metric*.

In this section, we review the tropical metric and study its properties, especially in relation to the BHV metric. We then present our main contribution, which is a formal and theoretical study of mathematical

properties of the metric space $(\mathcal{T}_N, d_{\mathrm{tr}})$ which we refer to as *palm tree space* (tropical tree space). We show that palm tree space possesses fundamental characteristics for studies in probability and statistics to be well-defined; namely, that it is a Polish space.

3.1 The Tropical Metric

Definition 5. For two points $[x], [y] \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, consider the distance between $[x]$ and $[y]$ given by

$$\begin{aligned} d_{\mathrm{tr}}([x],[y]) &:= \max_{1\le i<j\le n} \left|(x_i - y_i) - (x_j - y_j)\right| \\ &= \max_{1\le i\le n}(x_i - y_i) - \min_{1\le i\le n}(x_i - y_i). \end{aligned}$$

We refer to the function d_{tr} as the *tropical metric*.

Proposition 6. *The function d_{tr} is a well-defined metric function on $\mathbb{R}^n/\mathbb{R}\mathbf{1}$.*

Proof. We verify that the defining properties of metrics are satisfied. By definition, for $[u], [v] \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, $d_{\mathrm{tr}}([u],[v]) = d_{\mathrm{tr}}([v],[u])$, satisfying symmetry. The tropical metric is nonnegative, since $\left|(u_i - v_i) - (u_j - v_j)\right| \ge 0$, so is $d_{\mathrm{tr}}([u],[v]) \ge 0$. If $d_{\mathrm{tr}}([u],[v]) = 0$, then $u_i - v_i$ are equal for all $1 \le i \le n$, thus $[u] = [v]$, so indiscernibles are identifiable.

For $[u], [v], [w] \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, we now show that triangle inequality is satisfied: $d_{\mathrm{tr}}([u],[w]) \le d_{\mathrm{tr}}([u],[v]) + d_{\mathrm{tr}}([v],[w])$. Suppose $1 \le i' < j' \le n$ such that

$$\left|(u_{i'} - w_{i'}) - (u_{j'} - w_{j'})\right| = \max_{1\le i<j\le n} |u_i - w_i - u_j + w_j|,$$

then $d_{\mathrm{tr}}([u],[w]) = |u_{i'} - w_{i'} - u_{j'} + w_{j'}|$. Note that

$$u_{i'} - w_{i'} - u_{j'} + w_{j'} = (u_{i'} - v_{i'} - u_{j'} + v_{j'}) + (v_{i'} - w_{i'} - v_{j'} + w_{j'}).$$

Hence

$$\begin{aligned} d_{\mathrm{tr}}([u],[w]) = |u_{i'} - w_{i'} - u_{j'} + w_{j'}| &\le |u_{i'} - v_{i'} - u_{j'} + v_{j'}| + |v_{i'} - w_{i'} - v_{j'} + w_{j'}| \\ &\le d_{\mathrm{tr}}([u],[v]) + d_{\mathrm{tr}}([v],[w]). \end{aligned}$$

Thus, d_{tr} is a metric function on $\mathbb{R}^n/\mathbb{R}\mathbf{1}$. □

Notice that the metric space $(\mathbb{R}^n/\mathbb{R}\mathbf{1}, d_{\mathrm{tr}})$ can be identified with the normed linear space $\mathbb{R}^{n-1}$ via the linear isomorphism $\pi : \mathbb{R}^n/\mathbb{R}\mathbf{1} \to \mathbb{R}^{n-1}$ with $[x] \mapsto (x_2 - x_1, \ldots, x_n - x_1)$. π is in fact an isometry: define a norm on $\mathbb{R}^{n-1}$ by $\|x\|_{\mathrm{tr}} := \max(\max|x_i - x_j|, \max|x_i|)$ and denote the induced distance by $\hat{d}_{\mathrm{tr}}$, then

$$\begin{aligned} d_{\mathrm{tr}}([x],[y]) &= \max\left(\max_{2\le i<j\le n} |(x_i - y_i) - (x_j - y_j)|, \max_{2\le i\le n} |x_i - y_i|\right) \\ &= \|\pi([x]) - \pi([y])\|_{\mathrm{tr}} = \hat{d}_{\mathrm{tr}}(\pi([x]), \pi([y])). \end{aligned} \tag{6}$$

Restricting to the subspace of phylogenetic trees equipped with the tropical metric gives the following construction.

Definition 7. For a positive integer N, let $\mathcal{T}_N$ be the space of phylogenetic trees with N leaves. The metric space $\mathcal{P}_N := (\mathcal{T}_N, d_{\mathrm{tr}})$ is called the *palm tree space*.

The spaces $\mathcal{T}_N^{\mathrm{BHV}}$ and $\mathcal{P}_N$ are not isometric, meaning that absolute lengths measured by each metric are not consistent. To understand the variation in length discrepancy, we study the stability of the tropical metric d_{tr} and find that perturbations of points in BHV space, measured by the BHV metric d_{BHV}, correspond to bounded perturbations of their images in palm tree space, measured by the tropical metric. This stability property is desirable, since it allows for interpretable comparisons between the two spaces, and allows for "translations" in the widely-used BHV framework over to palm tree space.

The following lemma ensures coordinate-wise stability of the tropical metric in $\mathcal{P}_N$.

Lemma 8. *Let $u \in \mathbb{R}^n$. For $1 \le i \le n$, if we perturb the ith coordinate of u by ε to obtain another point $u' \in \mathbb{R}^n$, then in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ we have*

$$d_{\mathrm{tr}}([u],[u']) = |\varepsilon|.$$

Proof. For $1 \le j \le n$, the difference $u'_j - u_j = 0$ if $j \neq i$, and $u'_i - u_i = \pm\varepsilon$. The set of these differences is then either $\{0, \varepsilon\}$ or $\{0, -\varepsilon\}$. By Definition 5, $d_{\mathrm{tr}}([u],[u']) = |0 - \pm\varepsilon| = |\varepsilon|$. □

Theorem 9 (Stability)**.** *Let N be the number of leaves in palm tree space and BHV space. Let u and u' be two phylogenetic trees with N leaves. Then the following inequality holds:*

$$d_{\mathrm{tr}}(u,u') \le \sqrt{N+1} \cdot d_{\mathrm{BHV}}(u,u').$$

Moreover, the smallest possible constant is $\sqrt{N+1}$.

Proof. We first prove that for any two trees u, u' in vector representation (1) with N leaves, $d_{\mathrm{tr}}(u,u') \le \sqrt{N+1} \cdot d_{\mathrm{BHV}}(u,u')$. First, assume that u, u' belong to the same orthant in BHV space. Then no matter what the tree topology is, if we denote the differences of the lengths of the $N-3$ internal edges in u and u' (see (4)) by $d_1, d_2, \ldots, d_{N-3}$, and the differences of the length of the N external edges by $p_1, p_2, \ldots, p_N$, we always have

$$d_{\mathrm{BHV}}(u,u') = \sqrt{\sum_{i=1}^{N-3} d_i^2 + \sum_{i=1}^{N} p_i^2}$$

(e.g., Lin et al. (2017); Owen and Provan (2011)).

For every pair of leaves i, j in both trees, the distance between them is a sum of the length of some internal edges and two external edges. In other words, all differences $w_{ij}^u - w_{ij}^{u'}$ are of the form of the sum between some d_k, and $p_i + p_j$. Thus, the maximum of these differences is at most the sum of all positive d_i values, plus the two greatest p_i values (take these to be p_{i_1} and p_{i_2}), while the minimum of these differences is at least the sum of all negative d_i values, plus two smallest p_i values (take these to be p_{i_3} and p_{i_4}). By definition, $d_{\mathrm{tr}}(u,u')$ is the maximum minus the minimum of these differences, so we have

$$d_{\mathrm{tr}}(u,u') \le \sum_{i=1}^{N-3} |d_i| + |p_{i_1}| + |p_{i_2}| + |p_{i_3}| + |p_{i_4}|.$$

By the Cauchy–Schwarz inequality (Cauchy, 1821; Schwarz, 1890),

$$(N+1) \cdot \left(\sum_{i=1}^{N-3} |d_i|^2 + |p_{i_1}|^2 + |p_{i_2}|^2 + |p_{i_3}|^2 + |p_{i_4}|^2 \right) \ge \left(\sum_{i=1}^{N-3} |d_i| + \sum_{i=1}^{N} |p_i| \right)^2.$$

Hence

$$\begin{aligned} d_{\mathrm{tr}}(u,u') &\le \sum_{i=1}^{N-3} |d_i| + |p_{i_1}| + |p_{i_2}| + |p_{i_3}| + |p_{i_4}| \\ &\le \sqrt{N+1} \cdot \sqrt{\sum_{i=1}^{N-3} |d_i|^2 + |p_{i_1}|^2 + |p_{i_2}|^2 + |p_{i_3}|^2 + |p_{i_4}|^2} \\ &\le \sqrt{N+1} \cdot \left(\sum_{i=1}^{N-3} |d_i|^2 + \sum_{i=1}^{N} p_i^2 \right) \\ &= \sqrt{N+1} \cdot d_{\mathrm{BHV}}(u,u'). \end{aligned}$$

Now, for u, u' with distinct tree topologies, we consider the unique geodesic connecting them: there exist finitely many points $u^1, \ldots, u^{k-1}$ in BHV space such that u^i and u^{i+1} belong to the same orthant corresponding to a tree topology for $0 \le i \le k-1$, where $u^0 = u$ and $u^k = u'$, and $d_{\mathrm{BHV}}(u,u') = \sum_{i=0}^{k-1} d_{\mathrm{BHV}}(u^i, u^{i+1})$. For $1 \le i \le k-1$, by the proof above, we have that

$$d_{\mathrm{tr}}(u^i, u^{i+1}) \le \sqrt{N+1} \cdot d_{\mathrm{BHV}}(u^i, u^{i+1}) \quad \forall \quad 1 \le i \le k-1.$$

Thus,

$$d_{\mathrm{tr}}(u,u') \leq \sum_{i=0}^{k-1} d_{\mathrm{tr}}(u^i,u^{i+1}) \leq \sum_{i=0}^{k-1} \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u^i,u^{i+1}) = \sqrt{N+1}\cdot d_{\mathrm{BHV}}(u,u').$$

Next, we consider the case where the equality holds: consider two trees t and t' with N leaves and the same tree topology, given by the following nested sets

$$\big\{\{1,2\},\{1,2,3\},\ldots,\{1,2,\ldots,N-2\}\big\}.$$

Suppose in t, the internal edges have lengths

$$b^t(e_i) = \begin{cases} 2, & \text{if } 1 \leq i \leq N-4; \\ 1, & \text{if } i = N-3. \end{cases}$$

Similarly, in t', the internal edges have lengths

$$b^{t'}(e_i) = \begin{cases} 1, & \text{if } 1 \leq i \leq N-4; \\ 2, & \text{if } i = N-3. \end{cases}$$

The external edge lengths of t and t' are

$$p^i_j = \begin{cases} 1, & \text{if } (i,j) = (1,2),(1,N-2),(2,N-1),(2,N); \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$d_{\mathrm{BHV}}(t,t') = \sqrt{(N-4)\cdot(2-1)^2 + (1-2)^2 + 2\cdot(1-0)^2 + 2\cdot(0-1)^2} = \sqrt{N+1}.$$

For $1 \leq i < j \leq N$, in either tree the distance w_{ij} is the sum of the edge lengths of

$$p_i,\ e_{\max(i-1,1)},\ e_{\max(i-1,1)+1},\ \ldots,\ e_{\max(j-2,N-3)},\ p_j.$$

Since $b^t(e_i) > b^{t'}(e_i)$ for $i < N-3$ and $b^1(e_i) < b^2(e_i)$ for $i = N-3$, the maximum of all differences $w^t_{ij} - w^{t'}_{ij}$ is

$$w^t_{2(N-2)} - w^{t'}_{2(N-2)} = ((N-4)\cdot 2 + 2\cdot 1) - (N-4)\cdot 1 = N-2;$$

and the minimum of all differences $w^t_{ij} - w^{t'}_{ij}$ is

$$w^t_{(N-2)(N-1)} - w^{t'}_{(N-2)(N-1)} = 1 - (2+1+1) = -3.$$

By definition, $d_{\mathrm{tr}}(t,t') = (N-2) - (-3) = N+1 = \sqrt{N+1}\cdot d_{\mathrm{BHV}}(t,t')$ in this case. Thus, $\sqrt{N+1}$ is the smallest possible stability constant.

□

In general, and especially data applications, the number of leaves is fixed prior to the study so the stability constant $\sqrt{N+1}$ is indeed a constant.

We note that explicit calculations involving geodesics between trees in the original paper by Billera et al. (2001) do not include external edges, since these do not modify the geometry of the space. Indeed, their inclusion only amounts to an additional Euclidean factor, since the tree space then becomes the cross product of BHV space of trees with internal edges only, and $\mathbb{R}^N_{\geq 0}$. Geodesic distances, which depend directly on geodesic paths (the former is the length of the latter), considered in Billera et al. (2001) also do not include external edges. In the proof of Theorem 9, we follow the algorithm of Owen and Provan (2011) to compute BHV distances which includes external edge lengths, not only because it is the fastest algorithm to date but also necessary in this comparative setting, since the tropical distance is defined by external edge lengths.

In terms of interpretation, Theorem 9 provides an important comparative measure and guarantees that quantitative results from BHV space are bounded in palm tree space. For example, in single-linkage clustering, where clusters are fully determined by distance thresholds, the stability result means that a given clustering pattern in BHV space will be preserved in palm tree space, thus maintaining interpretability of clustering behavior.

3.2 Geometry of Palm Tree Space

The uniqueness property of geodesics in BHV space, used in the proof of Theorem 9, leads naturally to the study of similar geometric properties that characterize palm tree space as well as important differences between the two spaces. These characteristics will now be developed in this section.

3.2.1 Geodesics in Palm Tree Space

In palm tree space, geodesics are in general not unique, which is a common occurrence in various metric spaces. There exists, however, a unique path joining two points in palm tree space, which is also a geodesic—the tropical line segment.

Definition 10. Given $[x], [y] \in \mathbb{R}^n/\mathbb{R}\mathbf{1}$, the *tropical line segment* with endpoints $[x]$ and $[y]$ is the set

$$\{a \odot [x] \boxplus b \odot [y] \in \mathbb{R}^n/\mathbb{R}\mathbf{1} \mid a, b \in \mathbb{R}\},$$

where $\odot$ is tropical multiplication and max-plus addition $\boxplus$ for two vectors is performed coordinate-wise.

Proposition 11. *For two trees $t, t' \in \mathcal{P}_N$, the tropical line segment connecting t and t' is a geodesic.*

Proof. It suffices to show that for any $a, b \in \mathbb{R}$, we have that

$$d_{\mathrm{tr}}(z, t) + d_{\mathrm{tr}}(z, t') = d_{\mathrm{tr}}(t, t'),$$

where $z = a \odot t \boxplus b \odot t'$ is the tropical line segment. We may assume that $t_i - t'_i \leq t_{i+1} - t'_{i+1}$ for $1 \leq i \leq n-1$. Under this assumption, $d_{\mathrm{tr}}(t, t') = (t_n - t'_n) - (t_1 - t'_1)$. Now if $0 \leq j \leq n$ is the largest index such that $t_j - t'_j \leq b - a$, then for some $i \geq j+1$, $z_i = b + t'_i$ and, analogously, $z_i = a + t_i$. If $j = 0$ or $j = n$, then z is equal to either t or t' and the claim is apparent. We may thus assume $1 \leq j \leq n-1$.

The set of all differences $t_i - z_i$ contains $-a$ and the greater values $t_i - t'_i - b > -a$ for $i \geq j+1$. So,

$$d_{\mathrm{tr}}(z, t) = (t_n - t'_n - b) - (-a) = (t_n - t'_n) + (a - b).$$

Similarly, the set of all differences $z_i - t'_i$ contains b and the smaller values $(t_i - t'_i) + a < b$ for $i \leq j$. So,

$$d_{\mathrm{tr}}(z, t') = b - (t_1 - t'_1 + a) = (b - a) - (t_1 - t'_1).$$

Therefore, $d_{\mathrm{tr}}(z, t) + d_{\mathrm{tr}}(z, t') = d_{\mathrm{tr}}(t, t')$, and the tropical line segment connecting t and t' is a geodesic. □

In addition, it turns out that tropical line segments are easy and fast to compute. In particular, the time complexity to compute them is lower than that of Owen and Provan (2011).

Proposition 12. *(Maclagan and Sturmfels, 2015, Proposition 5.2.5) The time complexity to compute the tropical line segment connecting two points in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is $O(n \log n) = O(N^2 \log N)$.*

3.2.2 Structure of Palm Tree Space

In the same way that $\mathcal{T}_N^{\mathrm{BHV}}$ is constructed as the union of orthants, the geometry of $\mathcal{P}_N$ is also given by such a union.

Proposition 13. *(Maclagan and Sturmfels, 2015, Proposition 4.3.10) The space $\mathcal{T}_N$ is the union of $(2N-5)!!$ polyhedra in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ with dimension $N-3$.*

3.3 Topology of Palm Tree Space

The measure of a space is relevant in probabilistic studies; the measure of a topological space, in particular, results in a desirable compatibility where the topology of a space may be interpreted in terms of measures. For example, Radon measures may also be interpreted as linear functionals on the space of continuous functions with compact support, which is locally convex, by e.g., Bourbaki (2004), Chapter 3. This motivates our study of the topology of palm tree space.

The following two lemmas allow us to characterize the topology of palm tree space. Recall that for $x \in \mathbb{R}^n$, the set $B(x, r) = \{y \in \mathbb{R}^n \mid |y - x| < r\}$ is the open ball centered at x with radius r. By identifying $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ with $\mathbb{R}^{n-1}$ via (6), an equivalent set may be correspondingly defined in palm tree space as follows.

Definition 14. Under the tropical metric d_{tr}, we define $B_{\text{tr}}(x, r) = \{y \in \mathbb{R}^n \mid d_{\text{tr}}((0, y), (0, x)) < r\}$ to be the open *tropical ball* centered at $x \in \mathbb{R}^{n-1}$ with radius r.

Lemma 15. *For $x \in \mathbb{R}^{n-1}$ and $r > 0$, the open tropical ball $B_{\text{tr}}(x, r)$ is the open convex polytope defined by the following strict inequalities for $1 \leq i < j \leq n-1$:*

$$\begin{aligned} y_i &> x_i - r, \\ y_i &< x_i + r, \\ y_i - y_j &> x_i - x_j - r, \\ y_i - y_j &< x_i - x_j + r. \end{aligned} \tag{7}$$

Proof. For $y \in \mathbb{R}^{n-1}$, $y \in B_{\text{tr}}(x, r)$ if and only if $d_{\text{tr}}((0, \bar{x}), (0, \bar{y})) < r$. Definition 5 admits the strict inequalities in (7). □

Lemma 16. *For $r > 0$ and $x \in \mathbb{R}^{n-1}$, $B(x, r) \subseteq B_{\text{tr}}(x, 2r)$ and $B_{\text{tr}}(x, r) \subseteq B(x, \sqrt{n-1}r)$.*

Proof. By Lemma 15, if a point y lies in $B_{\text{tr}}(x, r)$, then for $1 \leq i \leq n-1$, $|y_i - x_i| < r$, thus $y \in B(x, \sqrt{n-1}r)$. Conversely, if a point y lies in $B(x, r)$, then for $1 \leq i \leq n-1$, we have that $|y_i - x_i| < r$. Therefore,

$$\left|(y_i - y_j) - (x_i - x_j)\right| = \left|(y_i - x_i) - (y_j - x_j)\right| < 2r.$$

Hence $y \in B_{\text{tr}}(x, 2r)$. □

Theorem 17. *On $\mathbb{R}^{n-1}$, the family of open balls $B(x, r)$ and the family of open tropical balls $B_{\text{tr}}(x, r)$ define the same topology.*

Proof. Suppose for all $r > 0$ and $x \in \mathbb{R}^{n-1}$ that the open balls $B(x, r)$ form a topological basis. For any $y \in \mathbb{R}^{n-1}$ and $s > 0$, we consider the ball $B_{\text{tr}}(y, s)$: For any point $z \in B_{\text{tr}}(y, s)$, we have that $d_{\text{tr}}(z, y) < s$. Let $\varepsilon = \dfrac{s - d_{\text{tr}}(z, y)}{2} > 0$. Then $B_{\text{tr}}(z, 2\varepsilon) \subseteq B_{\text{tr}}(y, s)$. By Lemma 16, we have $B(z, \varepsilon) \subseteq B_{\text{tr}}(z, 2\varepsilon) \subseteq B_{\text{tr}}(y, s)$. Therefore, $B_{\text{tr}}(y, s)$ is also an open set. The other direction is proved in the same manner. □

Example 18. Figure 4 illustrates the unit balls in Euclidean, BHV, and palm tree space. Here, the number of leaves is fixed to be 3. There are three 1-dimensional cones in BHV space, and they share the origin. The palm tree space $\mathcal{P}_3 = \{w = (w_{12}, w_{13}, w_{23}) \in \mathbb{R}^3/\mathbb{R}\mathbf{1} \mid \max(w) \text{ is attained at least twice}\}$ may be embedded in $\mathbb{R}^2$.

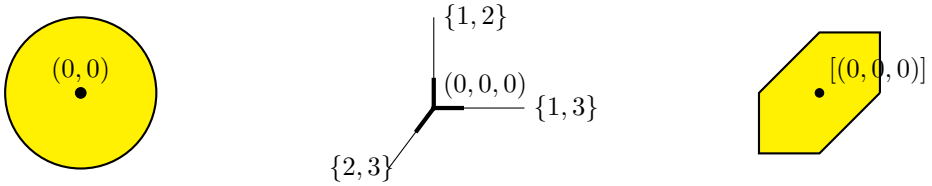

Figure 4: Comparison of unit balls in Euclidean, BHV, and palm tree space for $N = 3$ leaves. The leftmost figure is the unit ball $B((0, 0), 1)$ in $\mathbb{R}^2$; the center figure is the unit ball centered at the origin with radius 1 in a BHV space with 3 leaves; the rightmost figure is the unit ball $B_{\text{tr}}([(0, 0, 0)], 1)$ in $\mathcal{P}_3$.

3.4 Palm Tree Space is a Polish Space

We now show that additional analytic properties of palm tree space that are desirable for probabilistic and statistical analysis are satisfied. Specifically, we prove that palm tree space is a separable, completely metrizable topological space, and thus a Polish space, by definition.

Polish spaces are important settings for studies in probability due to the fact that classical results maybe formulated and generalized in a well-behaved manner; some examples are the construction of conditional expectations, Kolmogorov's extension theorem (which guarantees the definition of a stochastic process from a series of finite-dimensional distributions), and Prokhorov's theorem (which guarantees weak convergence by relating tightness of measures to compactness in a probability space) (Parthasarathy, 1967).

Proposition 19. *$\mathcal{P}_N$ is complete.*

Proof. For convenience, when considering points in $\mathcal{P}_N$, we always choose their unique preimage in $\mathbb{R}^n$ whose first coordinate is 0. Then, we may denote each point in $\mathcal{P}_N$ by an $(n-1)$-tuple in $\mathbb{R}^{n-1}$. Let $t_1, t_2, \ldots \in \mathbb{R}^{n-1}$ be a Cauchy sequence of points in $\mathcal{P}_N$. For $1 \leq i \leq n-1$, we claim that $(t_k^i)_{k\geq 1}$ also form a Cauchy sequence in $\mathbb{R}$: For any $\varepsilon > 0$, there exists M such that for $k_1, k_2 > M$, we have $d_{\mathrm{tr}}(t_{k_1}, t_{k_2}) < \varepsilon$. By Definition 5, $d_{\mathrm{tr}}(t_{k_1}, t_{k_2}) \geq |0 - 0 - t^i_{k_2} + {}^i_{k_1}| = |t^i_{k_1} - t^i_{k_2}|$. Thus, for $k_1, k_2 > M$, we have

$$\left|t^i_{k_1} - t^i_{k_2}\right| < \varepsilon.$$

Suppose now that the Cauchy sequence $(t_k^i)_{k\geq 1}$ converges to $t_0^i \in \mathbb{R}$. It suffices to show that

(i) $t_0 = (t_0^1, t_0^2, \ldots, t_0^{n-1})$ represents a point in $\mathcal{P}_N$;

(ii) $\lim_{k\to\infty} d_{\mathrm{tr}}(t_k, t_0) = 0$.

To show (ii), we argue that since $(t_k^i)_{k\geq 1}$ converges to t_0^i for all $1 \leq i \leq n-1$, then for any $\varepsilon > 0$ there exists M such that for $k > M$, we have $|t_k^i - t_0^i| < \frac{\varepsilon}{2}$ for all $1 \leq i \leq n-1$. Then by Definition 5,

$$\begin{aligned} d_{\mathrm{tr}}(t_k, t_0) &= \max_{1\leq i\leq n-1}\left(0, t_k^i - t_0^i\right) - \min_{1\leq i\leq n-1}\left(0, t_k^i - t_0^i\right) \\ &< \frac{\varepsilon}{2} - \left(-\frac{\varepsilon}{2}\right) = \varepsilon. \end{aligned}$$

So $\lim_{k\to\infty} d_{\mathrm{tr}}(t_k - t_0) = 0$.

To show (i), note that each coordinate of t_0, including the first, is the limit of the corresponding coordinates of $(t_k)_{k\geq 1}$. Suppose $t_0 \notin \mathcal{P}_N$, then there exists $1 \leq i < j < k < l \leq N$ such that one term of t_0 in (2) is strictly greater than the remaining two. Then there exists M_2 such that for all $k > M_2$, the one term of t_k in (2) is also strictly greater than the remaining two, thus $t_k \notin \mathcal{P}_N$—a contradiction. Hence (i) holds, and $\mathcal{P}_N$ is complete. □

Proposition 20. *$\mathcal{P}_N$ is separable.*

Proof. We claim that the set of all trees with all rational coordinates is dense in $\mathcal{P}_N$: Fix any tree $t = (w_{ij}) \in \mathcal{P}_N$. By Proposition 13, t belongs to a polyhedron and there exists a tree topology with $(N-3)$ internal edges. Then the distance between any two leaves is the sum of the lengths of the edges along the unique path connecting them. The number of edges along each path is at most $(N-1)$. For any $\varepsilon > 0$ and length b_k of each edge of the tree t, since $\mathbb{Q}$ is dense in $\mathbb{R}$, we can find a rational number q_k such that $|q_k - b_k| < \frac{1}{2(N-1)}\varepsilon$. Now, construct another tree $t' = (w'_{ij})$ with the same topology as t, and with corresponding edge lengths q_k. Then for any $1 \leq i < j \leq N$ we have that $|w'_{ij} - w_{ij}| < \frac{\varepsilon}{2}$. Thus

$$d_{\mathrm{tr}}(t', t) = \max_{1\leq i<j\leq n}\left(|w'_{ij} - w_{ij}|\right) - \min_{1\leq i<j\leq n}\left(|w'_{ij} - w_{ij}|\right) < \varepsilon,$$

and all coordinates of q_k are rational. Thus, $\mathcal{P}_N$ is separable. □

The above results on completeness and separability are proved by definition. An alternative approach follows the work of Ardila (2005), which has also been used by Bernstein and Long (2017); Bernstein (2020): Consider a linear mapping from $\mathbb{R}^N$ to $\mathbb{R}^n$ where $(x_1, \ldots, x_N) \mapsto (x_i - x_j)$ for all pairs $i < j$. The image of such a map is isomorphic to the tropical projective torus and the tropical metric is then the ℓ_∞ distance on $\mathbb{R}^n$ restricted to the image of this map. Palm tree space forms a closed subset of $\mathbb{R}^n$, since the four-point condition (Definition 1) defines a closed subset and $\mathbb{R}^n$ equipped with the ℓ_∞ distance is complete and separable. This formulation also provides insight into the topology of palm tree space described in Theorem 17.

Finally, we have the following characterization of compact subsets in $\mathbb{R}^n$.

Theorem 21 (Heine–Borel Theorem (Conway, 2014, Theorem 1.4.8))**.** *In the Euclidean space $\mathbb{R}^n$, a subset is compact if and only if it is closed and bounded.*

Thus, for palm tree space, there exist compact subsets in palm tree space.

Corollary 22. *In $\mathcal{P}_N$, a subset is compact if and only if it is closed and bounded.*

3.5 Probability Measures and Means in Palm Tree Space

We showed in Section 3.4 that palm tree space is a Polish space, and thus exhibits desirable properties for rigorous probability and statistics. Such properties ensure well-behaved measure-theoretic properties, and in particular, allow for classical probabilistic and statistical studies, such as convergence in various modes, as well as ensuring that stochastic processes are well defined. We now study the existence of probabilistic and statistical quantities for parametric data analysis, such as probability measures and Fréchet means and variances.

3.5.1 Tropical Measures of Central Tendency

For distributions in general metric spaces, there are various measures of central tendency. These may be framed in palm tree space as follows (and may be generalized by replacing the tropical metric d_{tr} with any well-defined metric).

Definition 23. Given a probability space $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N), \mathbb{P}_{\mathcal{T}_N})$, the quantity

$$\text{Var}_{\mathbb{P}_{\mathcal{T}_N}}(t) = \int_{\mathcal{T}_N} d_{\text{tr}}(t,t')^2 d\nu(t') < \infty \tag{8}$$

is known as the *tropical Fréchet variance*. The minimizer of the quantity (8) is the *tropical Fréchet population mean* or *barycenter* μ_F of a distribution ν:

$$\mu_{\text{tr}}^F = \arg\min_t \int_{\mathcal{T}_N} d_{\text{tr}}(t,t')^2 d\nu(t') < \infty. \tag{9}$$

Definition 24. The *tropical Fermat–Weber point* of a distribution is a similarly-defined measure of central tendency, and can be thought of as a generalized median of a distribution in a general metric space:

$$\mu_{\text{tr}}^{FW} = \arg\min_t \int_{\mathcal{T}_N} d_{\text{tr}}(t,t') d\nu(t'). \tag{10}$$

For general metric spaces, neither existence nor uniqueness of (9) nor (10) are guaranteed. Ohta (2012) proves a condition under which barycenters are guaranteed to exist.

Lemma 25. *(Ohta, 2012, Lemma 3.2) If (M,d) is a proper metric space, then any distribution ν where $\int_M d(t,t')^2 d\nu(t') < \infty$ has a barycenter.*

As a consequence of the Heine–Borel theorem (see Corollary 22), palm tree space is a proper metric space. Thus, (9) evaluated according to the tropical metric is guaranteed to exist. However, since geodesics are not unique in palm tree space, Fermat–Weber points and Fréchet means will also, in general, not be unique. It is known that the set of tropical Fermat–Weber points is a classical convex polytope; Lin and Yoshida (2018) present a formal and complete treatment of the Fermat–Weber point under the tropical metric.

3.5.2 Tropical Probability Measures

Probability measures on combinatorial and phylogenetic trees have been previously discussed, for example by Aldous (1996) and Billera et al. (2001). This section is dedicated to an analogous discussion on palm tree space. In $\mathcal{P}_N$, the Borel σ-algebra $\mathcal{B}(\mathcal{T}_N)$ is the σ-algebra generated by the open tropical balls B_{tr} of $\mathcal{T}_N$, given in Definition 14. We begin by providing the existence of probability measures on $\mathcal{P}_N$.

Definition 26. A *finite tropical Borel measure* on $\mathcal{T}_N$ is a map $\mu : \mathcal{B}(\mathcal{T}_N) \to [0,\infty)$ such that $\mu(\emptyset) = 0$, and for mutually disjoint Borel sets $A_1, A_2, \ldots \in \mathcal{B}(\mathcal{T}_N)$ implies that $\mu(\bigcup_{i=1}^\infty B_{\text{tr}}^{(i)}) = \sum_{i=1}^\infty \mu(B_{\text{tr}}^{(i)})$. If in addition $\mu(\mathcal{T}_N) = 1$, then μ is a *tropical Borel probability measure* on $\mathcal{T}_N$.

Since $\mathcal{T}_N$ is a finite union of polyhedra in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ (see Proposition 13), tropical Borel probability measures exist if finite tropical Borel measures μ exist on $\mathcal{T}_N$ by an appropriate scaling of the value of μ on each polyhedron.

Alternatively, the existence of probability measures on $\mathcal{T}_N$ can be seen by considering an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a measurable map $X : \Omega \to \mathcal{T}_N$. Such maps exist, since we have shown in Section 3 that $\mathcal{T}_N$ is a Polish space and thus $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ is a standard measurable space (e.g., Taylor, 2012). The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space by assumption, thus also measurable. In this case, the map X is a random variable taking values in $\mathcal{T}_N$. Then, X induces a probability measure $\mathbb{P}_{\mathcal{T}_N}$ on $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ by the pushforward measure $X_*\mathbb{P}$ of $\mathbb{P}$ under X, known as the *distribution*, for all Borel sets $A \in \mathcal{B}(\mathcal{T}_N)$:

$$X_*\mathbb{P}(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Example 27. If $\mathbb{P}_\vartheta$ is a probability measure parameterized by ϑ, then we may obtain parametric probability distributions on $(\mathcal{T}_N, \mathcal{B}(\mathcal{T}_N))$ induced by $\mathbb{P}_\vartheta$. In the Bayesian setting, if the measure λ gives the prior distribution of ϑ, then the joint probability measure $\mathbb{P}(\mathcal{T}_N, \vartheta)$ is given by the product measure $\mathbb{P}_{\mathcal{T}_N} \times \lambda$. Similarly, the conditional measure $\mathbb{P}(\vartheta \mid \mathcal{T}_N)$ is also proportional to this same product measure, which can be seen by Bayes' rule, where for events $B_1, B_2, \ldots$ that partition the sample space, then for any event A,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)}.$$

Additionally, for continuous random variables T and T', the conditional density $f_{T'|T}(t' \mid t)$ satisfies

$$f_{T'|T}(t' \mid t) = \frac{f_{T|T'}(t \mid t')f_{T'}(t')}{f_T(t).}$$

Specific probability measures analogous to classical probability distributions exist on $\mathcal{P}_N$, which we now exemplify. We note that these measures may be generalized to arbitrary compact subspaces of $\mathcal{T}_N$ by Corollary 22.

By Proposition 13, since $\mathcal{T}_N$ is a finite union of polyhedra in $\mathbb{R}^n/\mathbb{R}\mathbf{1}$, the *base measure* on palm tree space can be defined by assigning a probability of $1/(2N-5)!!$ to each polyhedron, and a uniform distribution within each polyhedron. Since each polyhedron is unbounded, the uniform distribution here would be an improper prior. A proper uniform distribution may be obtained by a rescaling of each polyhedron to be unitary.

An exponential family-type analog for palm tree space may be defined using the tropical metric as follows:

$$f(w) = C \cdot \exp\{d_{\mathrm{tr}}(w, \mu_{\mathrm{tr}}^0)\}, \tag{11}$$

where C is a normalizing constant, and μ_0 is taken to be a measure of central tendency, as discussed above in Section 3.5.1. Measures of the form (11) give rise to families of distributions concentrated on a tropical central tree, μ_{tr}^0.

4 Application to Data: Seasonal Influenza

We now provide an example of a statistical study in palm tree space following Yoshida et al. (2019). We compare the performance of tropical versus BHV principal component analysis (PCA) of the seasonal influenza virus by studying its diversity over twenty years of collected longitudinal data and see that the tropical approach outperforms the BHV approach.

Influenza is an RNA virus affecting up to 10% of adults and 30% of children worldwide and on an annual basis, resulting in more than half a million deaths (WHO, 2016). Because of the rapid evolution of the virus genome, the development of an effective vaccine critically relies on being able to effectively visualize, analyze, and predict the viral evolution patterns in a statistically rigorous setting.

4.1 Influenza Data

We focus on the influenza type A virus, which is an RNA virus that is classified by subtype according to the two proteins occurring on the surface of the virus: hemagglutinin (HA) and neuraminidase (NA). Here, we focus on HA, which tends to be the most variable protein in genomic evolution, in terms of changing the antigenic make-up of surface proteins. Such antigenic variability (known as *antigenic drift*) is

an important driving factor behind vaccine failure. We restrict our study to the subtype H3N2, which is becoming increasingly abundant, and a dominant factor studied in developing flu vaccines due to its recently increasing resistance to standard antiviral drugs (Altman, 2006). It was also the cause of an epidemic due to vaccine failure in 2002-2003 (CDC, 2004).

Genomic data for 1089 full length sequences of hemagglutinin (HA) for influenza A H3N2 from 1993 to 2017 in the state of New York were obtained from the GI-SAID EpiFlu^{TM} database (`www.gisaid.org`) and aligned with MUSCLE (Edgar, 2004) using default settings. HA sequences from each season were related to those of the preceding season. We then applied *tree dimensionality reduction* (Zairis et al., 2016) using temporal windows of 5 consecutive seasons to create 21 datasets. The date of each dataset corresponds to the first season; for example, the dataset dated 2013 consists of 5-leaved trees where the leaves come from seasons 2013 through 2017. Each unrooted tree in these datasets was constructed using the neighbor-joining method (Saitou and Nei, 1987) with Hamming distance. Outliers were then removed from each season using KDETREES (Schardl et al., 2014). On average, there were approximately 20,000 remaining trees in each dataset. Finally, PCA was performed under the tropical metric (Yoshida et al., 2019) and under the BHV metric (Nye et al., 2017).

4.1.1 Tree Dimensionality Reduction

The influenza virus is assumed to emerge and evolve from a common ancestor: although the virus mutates each season and within each patient, each of these seasonal and patient-specific mutations can be traced back to a single virus (e.g., Liu et al. (2009)). This evolutionary pattern is depicted in a large phylogenetic tree. Tree dimensionality reduction is a sampling method that generates a data set of smaller trees via a structural and systematic sampling of the larger tree (Zairis et al., 2016). In other words, the method produces a collection of smaller trees that faithfully represents the evolutionary behavior of the single large tree, and thus allows the evolutionary information to be treated as a dataset with multiple points akin to bootstrapping, rather than viewing the large tree as a single datum. Specifically, this is done by randomly sampling one individual from each season over the length of the desired temporal window; each individual then gives one leaf of the reduced tree.

In the applications presented in Zairis et al. (2016), smaller trees were constructed with three, four, and five leaves; we follow within this size scale for consistency of the original method, but choose the largest number of leaves, since the two tree PCA methods were implemented in Nye et al. (2017) and Yoshida et al. (2019) for trees with a larger number of leaves.

4.2 PCA in Tree Spaces

PCA is a fundamental technique in descriptive and exploratory statistics that visualizes relationships within the data by reducing their dimensionality. As such, PCA has many important implications—for example, it projects to the subspace of the solution of k-means clustering (Ding and He, 2004)—and may be interpreted in several different ways. One interpretation may be seen as searching for a lower-dimensional plane that minimizes the sum of squared distances from the data points to the plane, and then finding an orthogonal projection from the data points onto the plane to visualize them. More precisely, given a set X of data points $\{x_1, x_2, \ldots, x_n\}$ where each $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, we may consider a subset of $k+1$ of these points $V = \{v_0, \ldots, v_k\} \subset \mathbb{R}^m$ and define

$$\Pi(V) := \left\{ \sum_{i=0}^{k} p_i v_i \mid p_0, \ldots, p_k \in \mathbb{R} \text{ such that } p_0 + \cdots + p_k = 1 \right\},$$

where $\Pi(V)$ is the affine subspace of $\mathbb{R}^m$ containing V. The orthogonal L^2 distance between any point $y \in \mathbb{R}^m$ and $\Pi(V)$ is denoted by $d(y, \Pi(V))$; the squared of projected distances for the data X onto $\Pi(V)$ is then

$$D_X^2(\Pi(V)) = \sum_{i=1}^{n} d(x_i, \Pi(V))^2.$$

The kth principal component Π_k is the choice of V minimizing D_X^2. In Euclidean space, Π_0 corresponds to the sample mean, while Π_1 is the regression line passing through the mean, and so on, for higher dimensions.

These principal components are nested: $\Pi_0 \subset \Pi_1 \subset \Pi_2 \subset \cdots$. This interpretation relies heavily on the setting of a vector space, since $\Pi(V)$ is a linear combination of vectors, and visualizing the data on $\Pi(V)$ relies on orthogonal projection, both of which are inherently linear algebraic notions, and hence difficult to directly implement in tree spaces.

It is important to emphasize that PCA is a descriptive and exploratory technique for data, and a form of unsupervised learning. As such, there is no *a priori* assumption on the distribution of the data, despite the coincidence of the above mentioned description of the technique with classical linear regression, which is an inferential tool (and thus, classically, assumes some distributional properties). PCA can always be used to reduce the dimensionality of and visualize data, no matter what their distribution (and no matter what interpretation of PCA is adapted to implement the procedure).

Respective adaptations of the above-mentioned interpretation of PCA to palm tree space and BHV space are given by Nye et al. (2017) and Yoshida et al. (2019): convex, triangular regions—the tropical triangle and the *locus* of weighted Fréchet means computed with respect to the BHV metric, respectively—define notions of second principal components in the respective spaces. Following Nye (2014), in the setting of phylogenetic tree space, we seek a convex hull of $k+1$ points to represent the kth principal component, $\Pi(V)$ where now each v_i in $V = \{v_0, \ldots, v_k\}$ is a tree. Since any geodesic segment is the convex hull of its endpoints, a natural extension to the plane, or second principal component, would be to use the convex hull of three points. This also gives us a representation with projections that are readily visualizable.

In other words, we seek a 2-plane defined by three ultrametric trees and visualize the data by restricting and projecting onto the convex hull of these three points. In palm tree space, this is straightforward since tropical triangles are always 2-dimensional (Lin et al., 2017). In BHV space, however, this entails finding a locus, which is constructed from the set of discrete Fréchet means weighted by values on the probability simplex. As opposed to BHV triangles, the dimensionality of loci of weighted Fréchet means computed under the BHV metric is well behaved (Nye et al., 2017).

4.2.1 Tropical PCA

The (tropically-) convex hull we seek in palm tree space by implementing the method of Yoshida et al. (2019) represents the second principal component is the tropical triangle. Edges are given by tropical line segments between their three vertices. Projections of data points onto the tropical triangle are exact and unique.

To build the tropical triangles, we seek three points within the dataset; in this sense, the PCA method is approximate and not exact. For smaller datasets, such a search may be combinatorially tractable. For larger datasets, we implement a Markov chain Monte Carlo (MCMC) algorithm to find these three points (Kang and Yoshida, 2018). Briefly, the idea is to first choose three points from the dataset at random and calculate the sums of the tropical distances of the remaining points to the tropical triangle initially defined by the three points. One point among these vertices is then randomly chosen to be iteratively replaced by other points within the dataset to find a smaller sum of tropical distances to the resulting triangles until a minimal sum of tropical distances is found; the corresponding three vertices of the triangle are then retained to define the second tropical principal components. Unlike in classical Euclidean PCA where the components are mutually orthogonal, the vertices of the tropical triangle onto which the data points are projected are in general not. The projections of the data points onto the tropical triangle are tropically-convex combinations of the vertices. This procedure was run 5 times on each of the 21 datasets; each iteration was run in parallel on 18 CPU cores—Intel(R) Xeon(R) W-2155 CPU @ 3.30 GHz—and took approximately two hours to terminate, for a total of around 210 CPU hours to complete the tropical PCA procedure.

4.2.2 BHV PCA

Since geodesic triangles (and more generally, polytopes) in BHV space are not well-behaved convex hulls (Lin et al., 2017), they are thus problematic candidates for BHV second principal components. An alternative method proposed by Nye et al. (2017) searches for the *locus* of weighted, discrete, Fréchet means computed with respect to the BHV metric as a BHV kth principal component.

Briefly, for the weighted, discrete, BHV Fréchet mean,

$$\mu(V, w) = \arg\min_{y \in \mathcal{T}_N^{\mathrm{BHV}}} \sum_{i=0}^{k} w_i \cdot d_{\mathrm{BHV}}(y, v_i)^2,$$

if the weights are given by elements in the k-dimensional simplex of probability vectors,

$$\mathcal{S}^k := \left\{ (p_0, \ldots, p_k) \mid p_i \geq 0,\, i = 0, \ldots, k,\, \sum_{i=0}^{k} p_i = 1 \right\},$$

then $\Pi(V)$ defined by

$$\Pi(V) := \{\mu(V, p) \mid p \in \mathcal{S}^k\}$$

is the locus of the Fréchet mean of V. Geometric properties of the locus, such as convexity and dimensionality, are well-behaved, as opposed to BHV convex hulls. In our application, we implement the algorithm by Bačák to compute the weighted, discrete, BHV Fréchet means and thus construct the locus (Bačák, 2014). Projections of data points onto the locus are approximate and need not be unique.

Similar to the tropical PCA case described above, the procedure is approximate and uses trees from the dataset to construct the locus and its boundaries. We follow the algorithm proposed in the original reference, which is also a stochastic optimization algorithm. As in the tropical PCA case, the procedure was run 5 times on each of the 21 datasets; each iteration was run in parallel on 18 CPU cores—Intel(R) Xeon(R) W-2155 CPU @ 3.30 GHz—and also took approximately two hours to terminate, for a total of also around 210 CPU hours to complete the BHV PCA procedure.

Software and Data Availability

Software to compute both tropical and BHV PCA is publicly available in R and Java code. Their implementation to the influenza data described in this paper is located on the FluPCA GitHub repository at https://github.com/antheamonod/FluPCA.

The data used in this paper were obtained from publicly available sources and preprocessed, as detailed in Section 4.1. The final version used in the analyses in this paper are also publicly available on the FluPCA GitHub repository.

The resulting figures from both BHV and tropical PCA projections for all 21 data sets are also available on the FluPCA GitHub repository.

4.3 Interpretation of Tree PCA

In the tropical case, the second principal component represented by the tropical triangle—whose vertices are given by three ultrametric trees, and whose edges are given by the tropical line segments between them—divides into cells, which are determined by the tree topologies that an edge (tropical line segment) traverses. Trees in the dataset are then projected into the cells corresponding to their topologies in the PCA visualization. The simplest case in both tropical PCA is where all three vertices of the triangle are of the same tree topology, then there will only be one cell and all projections will be of the same topology; if two points are trees of the same tree topology, then every point on the tropical line segment connecting them will also be of the same tree topology. An example can be seen in the 1993 data set, available at https://github.com/antheamonod/FluPCA/Figures.

A more interesting example can be seen in the 2008 dataset in Figure 5. The tropical triangle divides into six cells; a complete discussion on tropical polytopes, their decomposition into cells and their self-duality can be found in Section 5.2 of Maclagan and Sturmfels (2015), Theorem 5.2.21 in particular gives the isomorphism from a tropical polytope to its self-dual. As conjectured in Yoshida et al. (2019), one possible biological interpretation of the cells is that neighboring cells represent similar yet distinct tree structures differing by tree rearrangements of one move. Tree rearrangements are used in algorithms where the goal is to search for the optimal tree structure in statistical tree reconstruction methods (see e.g., Bordewich and Semple (2005); Felsenstein (2004)). The tropical line segment forming the topmost edge of the triangle traverses five different tree topologies (indicated by the pink, green, black, blue, and red points along the topmost line segment), indicating that the three vertices of the triangle are quite different from one another in terms of tree topology.

The BHV locus, which represents the second principal component in BHV space, is also generated by three vertices (ultrametric trees), and depicted in the BHV PCA plots by a triangle. Here, varying tree

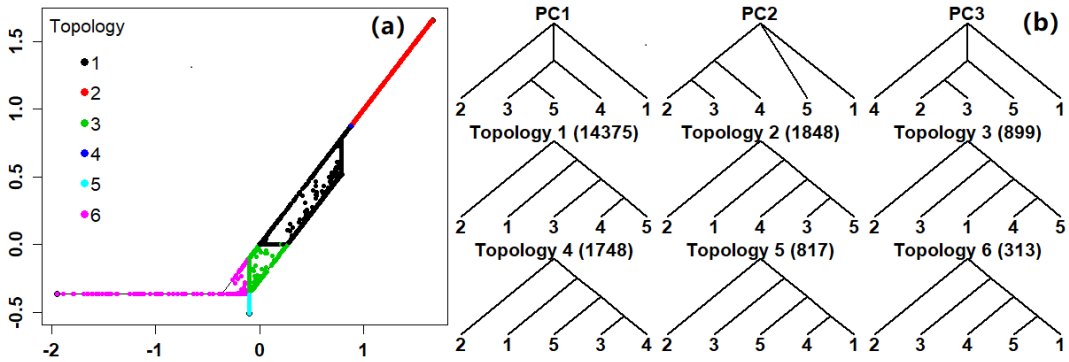

Figure 5: 2008: The tropical triangle as the second tropical principal component. (a) Tropical triangle and projected data points; (b) Vertices of the tropical triangle and projected tree topologies.

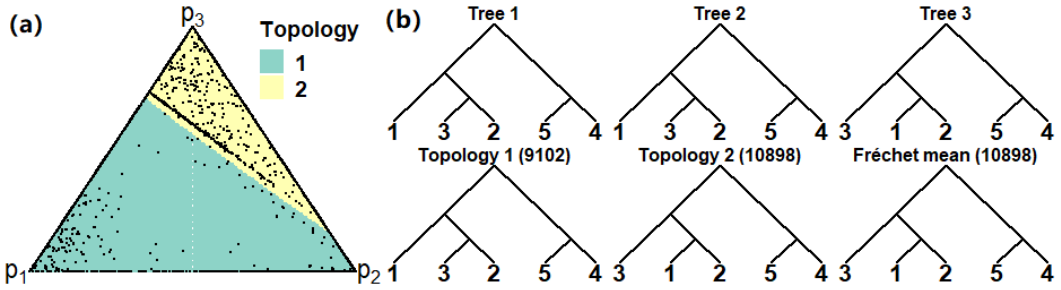

Figure 6: 2008: The locus of BHV Fréchet means as the second principal component. (a) The simplex shaded by topology of corresponding points on the affine subspace; (b) Trees 1, 2, and 3 correspond to three weighted Fréchet means.

topologies are depicted by the multicolored patches within the triangular region, indicating that the locus straddles several orthants in BHV space. In the 2008 dataset in Figure 6, the locus straddles two BHV orthants and we see two tree topologies occurring among the projected points.

In the 2008 dataset, palm tree space and the tropical geometric approach to computations in phylogenetic tree space appears to allow for occurrences of richer and more subtle structures and methods: in these examples, we see tropical PCA projections of six different topologies, versus two in the BHV case.

4.4 Results: Proportion of Explained Variance R^2

In terms of explained variance given in Table 1, we see that in general, tropical PCA is able to explain more of the variance in the data than does BHV PCA. BHV PCA results also have a higher variability over a wider range than tropical PCA: BHV PCA explains between ∼0.0% and 95% of the variance, while tropical PCA explains between 46% and 96%.

5 Discussion

In this paper, we defined palm tree space as the space of phylogenetic trees with N leaves, endowed with the tropical metric. We gave results on its analytic, topological, geometric, and combinatorial properties and showed that they are conducive to rigorous treatments and studies in probability. We performed a descriptive statistical analysis on real data and showed that the tropical approach outperforms the BHV approach. We showed that in certain respects, palm tree space is more natural and amenable to statistical analyses than BHV space. An important difference, however, is that geodesics in palm tree space are not unique. Our work invites the reinterpretation of existing statistical methods in terms of the tropical metric to make a wider array of exact analyses readily available and interpretable to phylogenetic research with

Table 1: Proportion of Explained Variance for Tropical and BHV PCA

	1993	1994	1995	1996	1997	1998	1999
Tropical	0.7269	0.8505	**0.9577**	0.7482	0.8437	0.8790	0.8564
BHV	0.3019	0.4347	0.3151	0.5025	0.0505	0.6408	**0.9524**
	2000	2001	2002	2003	2004	2005	2006
Tropical	0.7942	0.8302	0.9525	0.8622	0.7931	0.8304	0.7300
BHV	**0.0014**	0.9488	0.8962	0.4927	0.3651	0.3634	0.2383
	2007	2008	2009	2010	2011	2012	2013
Tropical	0.6995	**0.4637**	0.6289	0.6665	0.5920	0.5568	0.5624
BHV	0.2727	0.0460	0.1563	0.1935	0.2771	0.1998	0.1279

potential impact for biological discoveries.

Biological and Statistical Implications

Despite the statistical challenges of arbitrariness of dimension of BHV polytopes and stickiness which affect descriptive and inferential statistics, the BHV parameterization has been successfully implemented to reveal important biological findings (e.g., Zairis et al., 2014). In terms of interpretation, the unresolved singularities of BHV Fréchet means translate to "indecisiveness" of which branching patterns or tree topologies are "preferred," which is consistent with what is often seen in some biological settings where the trees arise from sequence alignment. However, mathematically, trees are used to model other biological phenomena, such as pulmonary paths as airway trees (e.g., Feragen et al., 2013, 2015)); brain growth and structure (e.g., Yan and Yan, 2013); and neuronal morphologies (e.g., Kanari et al., 2018). Such a probabilistic assumption may not be reasonable in these other settings. Given recent research interest in developing methods to bypass these difficulties support the goal of our work, which is that exploring alternative representations is an important research direction (e.g., Anaya et al., 2020; Skwerer et al., 2018). An important and interesting direction for future research is the identification of non-uniform probability distributions in the tropical setting, which is challenging yet promising: Tran (2018) outlines various ways in which tropical Gaussian distributions may be constructed.

In the context of shape statistics (Kendall, 1989) and computational anatomy (Grenander and Miller, 1998), the data objects of interest are often modeled as elements of algebraic spaces. In particular, these algebraic spaces are quotient spaces generated by group actions. Recent work has studied the behavior of estimators on such spaces, uncovering undesirable properties, such as biasedness and inconsistency when the group actions are random (that is, when the quotient spaces are generated by elements of the group are chosen at random to act on the topological space) or continuous (as in the case of Lie groups acting on Riemannian manifolds) (Devilliers et al., 2017; Miolane and Pennec, 2015). Nonparametric methods have been developed to bypass the problem of inconsistency by Bhattacharya and Patrangenaru (2014). As previously mentioned, the tropical projective torus $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ is a quotient space that may be generated by a group action, however the biasedness and inconsistency in previous work arise due to the poor behavior of the transformed metric after it is mapped into the quotient space, which results in a pseudometric. In our case, the tropical metric is well-behaved and defined directly on the quotient space, therefore differing in setting to previous work.

It should also be noted that the non-uniqueness property of geodesics in palm tree space poses computational difficulties, but does not prohibit statistical analysis and can still yield useful descriptive as well as inferential information, for example, on clustering behavior. Another important setting where geodesics are not unique is that of positively-curved spaces. Bhattacharya and Bhattacharya (2008) study asymptotic behavior and distributions on Riemannian manifolds, including positively curved manifolds. Recent work such as that by Kobayashi and Wynn (2019) develops techniques for data analysis on curved spaces by tuning the geodesic metrics accordingly. In particular, a general Fréchet function is defined, and its parameters are chosen accordingly, depending on the goal: for example, one geodesic metric may be transformed into another to control the curvature of the space for data analysis. There are large bodies of existing work in related areas on curved spaces, for example, in the case of manifold learning; shape statistics; Wasserstein spaces for

probability measures; and information geometry. Though these domains each have their own specific goals and studies, data analysis and computation play a central role in these settings. Moreover, there are known settings in these related areas where positive curvature, and thus non-uniqueness of geodesics, arises (for example, the 2-Wasserstein space for Gaussian measures is positively curved (Takatsu, 2011)). Adapting existing techniques in these settings to statistical analysis in palm tree space is an important direction of research that merits exploration.

Acknowledgments

The authors are grateful to Robert J. Adler, Omer Bobrowski, Yueqi Cao, Ioan Filip, Maia Fraser, Stephan Huckemann, Elliot Paquette, Juan Ángel Patiño-Galindo, Yitzchak (Elchanan) Solomon, and Bernd Sturmfels for helpful discussions. We are especially indebted to Carlos Améndola and Hossein Khiabanian for their extensive commentary, suggestions, and support throughout the course of this project.

A.M. and R.Y. wish to acknowledge the Mathematisches Forschunginstitut Oberwolfach (MFO) for hosting their visit in January 2018, which inspired this work. B.L. wishes to acknowledge support by the Simons Foundation and the hospitality of the Institut Mittag-Leffler during the spring of 2018, where a large part of this research was carried out.

The authors would also like to acknowledge the GI-SAID EpiFlu^{TM} initiative for making the data used in this study publicly available.

R.Y. is supported in part by NSF DMS #1622369 and #1916037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author Contributions Statement

A.M. conceived the study and designed the research. A.M., B.L., and R.Y. performed research. A.M. and R.Y. developed the algorithms; Q.K. implemented the software and performed the analyses. A.M. wrote the manuscript.

Competing Financial Interests Statement

The authors declare that no competing financial interests exist.

References

Akian, M., S. Gaubert, V. Niţică, and I. Singer (2011). Best Approximation in Max-plus Semimodules. *Linear Algebra and its Applications 435*(12), 3261–3296.

Aldous, D. (1996). Probability Distributions on Cladograms. In D. Aldous and R. Pemantle (Eds.), *Random Discrete Structures*, New York, NY, pp. 1–18. Springer New York.

Allen, B. L. and M. Steel (2001). Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics 5*(1), 1–15.

Altman, L. K. (2006). This Season's Flu Virus Is Resistant to 2 Standard Drugs. https://www.nytimes.com/2006/01/15/health/this-seasons-flu-virus-is-resistant-to-2-standard-drugs.html.

Améndola, C., M. Casanellas, and L. D. G. Puente (2018). Tapas of Algebraic Statistics. *Notices of the AMS 65*(8).

Anaya, M., O. Anipchenko-Ulaj, A. Ashfaq, J. Chiu, M. Kaiser, M. S. Ohsawa, M. Owen, E. Pavlechko, K. S. John, S. Suleria, et al. (2020). Properties for the Fréchet Mean in Billera–Holmes–Vogtmann Treespace. *Advances in Applied Mathematics 120*, 102072.

Ardila, F. (2005). Subdominant matroid ultrametrics. *Annals of Combinatorics 8*(4), 379–389.

Ardila, F. and C. J. Klivans (2006). The Bergman Complex of a Matroid and Phylogenetic Trees. *Journal of Combinatorial Theory, Series B 96*(1), 38–49.

Baez, J. C. and N. Otter (2017). Operads and phylogenetic trees. *Theory and Applications of Categories 32*(40), 1397–1453.

Barden, D., H. Le, and M. Owen (2018). Limiting Behaviour of Fréchet Means in the Space of Phylogenetic Trees. *Annals of the Institute of Statistical Mathematics 70*(1), 99–129.

Bačák, M. (2014). Computing Medians and Means in Hadamard Spaces. *SIAM Journal on Optimization 24*(3), 1542–1566.

Bernstein, D. I. (2020). L-infinity optimization to Bergman fans of matroids with an application to phylogenetics. *SIAM Journal on Discrete Mathematics 34*(1), 701–720.

Bernstein, D. I. and C. Long (2017). L-infinity optimization to linear spaces and phylogenetic trees. *SIAM Journal on Discrete Mathematics 31*(2), 875–889.

Bhattacharya, A. and R. Bhattacharya (2008). Statistics on Riemannian Manifolds: Asymptotic Distribution and Curvature. *Proceedings of the American Mathematical Society 136*(8), 2959–2967.

Bhattacharya, R. and V. Patrangenaru (2014). Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications. *Journal of Statistical Planning and Inference 145*, 1–22.

Billera, L. J., S. P. Holmes, and K. Vogtmann (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics 27*(4), 733–767.

Bordewich, M. and C. Semple (2005, Jan). On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance. *Annals of Combinatorics 8*(4), 409–423.

Bourbaki, N. (2004). Elements of Mathematics: Integration I.

Buneman, P. (1974). A Note on the Metric Properties of Trees. *Journal of Combinatorial Theory, Series B 17*(1), 48–50.

Cauchy, A.-L. (1821). Sur les formules qui résultent de l'emploie du signe et sur $>$ ou $<$, et sur les moyennes entre plusieurs quantités. *Cours d'Analyse, 1er Partie: Analyse algébrique*, 373–377.

CDC (2004). Preliminary assessment of the effectiveness of the 2003-04 inactivated influenza vaccine–Colorado, December 2003. *MMWR. Morbidity and mortality weekly report 53*(1), 8.

Cohen, G., S. Gaubert, and J.-P. Quadrat (2004). Duality and Separation Theorems in Idempotent Semimodules. *Linear Algebra and its Applications 379*, 395–422. Special Issue on the Tenth ILAS Conference (Auburn, 2002).

Conway, J. B. (2014). *A Course in Point Set Topology*. Undergraduate Texts in Mathematics. Springer, Cham.

Devadoss, S. L. and J. Morava (2015). Navigation in Tree Spaces. *Advances in Applied Mathematics 67*, 75–95.

Devilliers, L., S. Allassonnière, A. Trouvé, and X. Pennec (2017). Template Estimation in Computational Anatomy: Fréchet Means Top and Quotient Spaces Are Not Consistent. *SIAM Journal on Imaging Sciences 10*(3), 1139–1169.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 29. ACM.

Edgar, R. C. (2004, 03). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*(5), 1792–1797.

Einmahl, J. H. J., A. Kiriliouk, and J. Segers (2018, Jun). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes 21*(2), 205–233.

Felsenstein, J. (2004). *Inferring phylogenies*, Volume 2. Sinauer associates Sunderland, MA.

Feragen, A., P. Lo, M. de Bruijne, M. Nielsen, and F. Lauze (2013, Aug). Toward a Theory of Statistical Tree-Shape Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 2008–2021.

Feragen, A., J. Petersen, M. Owen, P. Lo, L. H. Thomsen, M. M. W. Wille, A. Dirksen, and M. de Bruijne (2015, June). Geodesic Atlas-Based Labeling of Anatomical Trees: Application and Evaluation on Airways Extracted From CT. *IEEE Transactions on Medical Imaging 34*(6), 1212–1226.

Garba, M. K., T. M. Nye, J. Lueg, and S. F. Huckemann (2021). Information geometry for phylogenetic trees. *Journal of Mathematical Biology 82*(3), 1–39.

Gavryushkin, A. and A. J. Drummond (2016). The Space of Ultrametric Phylogenetic Trees. *Journal of Theoretical Biology 403*, 197–208.

Grenander, U. and M. I. Miller (1998). Computational anatomy: An emerging discipline. *Quarterly of applied mathematics 56*(4), 617–694.

Hotz, T., S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer (2013). Sticky Central Limit Theorems on Open Books. *The Annals of Applied Probability 23*(6), 2238–2258.

Joswig, M., B. Sturmfels, and J. Yu (2007). Affine buildings and tropical convexityyu. *Albanian J. Math. 1*(4), 187–211.

Kanari, L., P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram (2018, Jan). A Topological Representation of Branching Neuronal Morphologies. *Neuroinformatics 16*(1), 3–13.

Kang, Q. and R. Yoshida (2018). Estimating Tropical Principal Components Using Metropolis Hastings Algorithm. In *International Congress on Mathematical Software*, pp. 272–279. Springer.

Kendall, D. G. (1989). A Survey of the Statistical Theory of Shape. *Statistical Science 4*(2), 87–99.

Kobayashi, K. and H. P. Wynn (2019, Feb). Empirical geodesic graphs and CAT(k) metrics for data analysis. *Statistics and Computing*.

Leaché, A. D. and B. Rannala (2010, 11). The Accuracy of Species Tree Estimation under Simulation: A Comparison of Methods. *Systematic Biology 60*(2), 126–137.

Lin, B., B. Sturmfels, X. Tang, and R. Yoshida (2017). Convexity in Tree Spaces. *SIAM Journal on Discrete Mathematics 31*(3), 2015–2038.

Lin, B. and N. M. Tran (2019). Two-player incentive compatible outcome functions are affine maximizers. *Linear Algebra and its Applications 578*, 133–152.

Lin, B. and R. Yoshida (2018). Tropical Fermat–Weber Points. *SIAM Journal on Discrete Mathematics*. To appear. Available at arXiv:1604.04674.

Liu, S., K. Ji, J. Chen, D. Tai, W. Jiang, G. Hou, J. Chen, J. Li, and B. Huang (2009, 03). Panorama Phylogenetic Diversity and Distribution of Type A Influenza Virus. *PLOS ONE 4*(3), 1–20.

Maclagan, D. and B. Sturmfels (2015). *Introduction to Tropical Geometry (Graduate Studies in Mathematics)*. American Mathematical Society.

Manon, C. (2011). Dissimilarity Maps on Trees and the Representation Theory of $\mathrm{SL}_m(\mathbb{C})$. *Journal of Algebraic Combinatorics 33*(2), 199–213.

Miolane, N. and X. Pennec (2015). Biased Estimators on Quotient Spaces. In F. Nielsen and F. Barbaresco (Eds.), *Geometric Science of Information*, Cham, pp. 130–139. Springer International Publishing.

Nye, T. M. W. (2014). An Algorithm for Constructing Principal Geodesics in Phylogenetic Treespace. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 11*(2), 304–315.

Nye, T. M. W., X. Tang, G. Weyenberg, and R. Yoshida (2017). Principal Component Analysis and the Locus of the Fréchet Mean in the Space of Phylogenetic Trees. *Biometrika 104*(4), 901–922.

Ohta, S.-I. (2012). Barycenters in Alexandrov Spaces of Curvature Bounded Below. *Advances in Geometry 14*(4).

Owen, M. and J. S. Provan (2011). A Fast Algorithm for Computing Geodesic Distances in Tree Space. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 8*(1), 2–13.

Pachter, L. and B. Sturmfels (2004). Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences 101*(46), 16132–16137.

Pachter, L. and B. Sturmfels (2005). *Algebraic Statistics for Computational Biology*, Volume 13. Cambridge University Press.

Parthasarathy, K. (1967). Probability and Mathematical Statistics: A Series of Monographs and Textbooks. In *Probability Measures on Metric Spaces*, Probability and Mathematical Statistics: A Series of Monographs and Textbooks, pp. ii. Academic Press.

Robinson, D. F. and L. R. Foulds (1981). Comparison of Phylogenetic Trees. *Mathematical Biosciences 53*(1), 131–147.

Saitou, N. and M. Nei (1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution 4*(4), 406–425.

Schardl, C. L., D. K. Howe, G. Weyenberg, P. M. Huggins, and R. Yoshida (2014, 04). KDETREES: Non-parametric estimation of phylogenetic tree distributions. *Bioinformatics 30*(16), 2280–2287.

Schröder, E. (1870). Vier kombinatorische Probleme. *Z. Math. Phys 15*, 361–376.

Schwarz, H. A. (1890). *Ueber ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung*, pp. 223–269. Berlin, Heidelberg: Springer Berlin Heidelberg.

Skwerer, S., S. Provan, and J. Marron (2018). Relative Optimality Conditions and Algorithms for Treespace Fréchet Means. *SIAM Journal on Optimization 28*(2), 959–988.

Speyer, D. and B. Sturmfels (2004). The Tropical Grassmannian. *Advances in Geometry 4*(3).

St. John, K. (2016, 06). Review Paper: The Shape of Phylogenetic Treespace. *Systematic Biology 66*(1), e83–e94.

Steel, M. A. and D. Penny (1993). Distributions of tree comparison metrics: Some new results. *Systematic biology 42*(2), 126–141.

Steel, M. A. and L. A. Székely (2006, 08). On the variational distance of two trees. *Ann. Appl. Probab. 16*(3), 1563–1575.

Sturmfels, B. and S. Sullivant (2005). Toric ideals of phylogenetic invariants. *Journal of Computational Biology 12*(2), 204–228.

Sullivant, S. (2018). *Algebraic Statistics*, Volume 194. American Mathematical Soc.

Takatsu, A. (2011, 12). Wasserstein geometry of Gaussian measures. *Osaka J. Math. 48*(4), 1005–1026.

Taylor, J. C. (2012). *An Introduction to Measure and Probability*. Springer Science & Business Media.

Tran, N. M. (2018). Tropical Gaussians: A Brief Survey. *arXiv preprint arXiv:1808.10843*.

Waterman, M. S., T. F. Smith, and W. A. Beyer (1976). Some biological sequence metrics. *Advances in Mathematics 20*(3), 367–387.

Weyenberg, G. and R. Yoshida (2016). Phylogenetic Tree Distances. In *Encyclopedia of Evolutionary Biology*, pp. 285–290. Elsevier.

WHO (2016). World Health Organization — Influenza. https://www.who.int/biologicals/vaccines/influenza/en/.

Willis, A. (2019). Confidence sets for phylogenetic trees. *Journal of the American Statistical Association 114*(525), 235–244.

Willis, A. and R. Bell (2018). Uncertainty in phylogenetic tree estimates. *Journal of Computational and Graphical Statistics 27*(3), 542–552.

Yan, B. C. and J. F. Yan (2013). A tree-like model for brain growth and structure. *J Biophys 2013*, 241612.

Yoshida, R., L. Zhang, and X. Zhang (2019, Feb). Tropical Principal Component Analysis and Its Application to Phylogenetics. *Bulletin of Mathematical Biology 81*(2), 568–597.

Zairis, S., H. Khiabanian, A. J. Blumberg, and R. Rabadan (2014). Moduli spaces of phylogenetic trees describing tumor evolutionary patterns. In *International Conference on Brain Informatics and Health*, pp. 528–539. Springer.

Zairis, S., H. Khiabanian, A. J. Blumberg, and R. Rabadán (2016). Genomic Data Analysis in Tree Spaces. *arXiv:1607.07503*.