The Dissertation Committee for Gillian Roxanne Grindstaff
certifies that this is the approved version of the following dissertation:

Geometric Data Analysis for Phylogenetic Trees and Non-contractible Manifolds

Committee:

Andrew Blumberg, Co-Supervisor

David Ben-Zvi, Co-Supervisor

Lewis Bowen

Megan Owen

Ngoc Tran

Geometric Data Analysis for Phylogenetic Trees and Non-contractible Manifolds

by

Gillian Roxanne Grindstaff

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2021

Dedicated to my father, Chuck.

Acknowledgments

I would like to thank my committee members for their mentorship, encouragement, and teaching. In particular, the content of Chapter 3 was developed in collaboration with Megan Owen, who was extremely patient with me in the process of writing and submitting my first paper.

I am deeply grateful for the camaraderie and support of all my fellow grad students at UT, especially my academic siblings, MGMN, and the cohort of 2015 - you made it joyful, when it didn't have to be. I'd also like to thank my real siblings, Russell and Abby, for being stellar roommates. And I could not have made it without Eliza, Katie, Mike, and Hadrien, who supported me through countless personal and professional struggles.

Most of all, I owe a profound debt of gratitude to my advisor, Andrew Blumberg. His unwavering encouragement and enthusiasm for my success carried me through grad school - I would not have finished this degree without him.

Geometric Data Analysis for Phylogenetic Trees and Non-contractible Manifolds

Publication No. _____

Gillian Roxanne Grindstaff, Ph.D.
The University of Texas at Austin, 2021

Supervisors: Andrew Blumberg
David Ben-Zvi

A phylogenetic tree is an acyclic graph with distinctly labeled leaves, whose internal edges have a positive weight. Given a set $\{1, 2, \ldots, n\}$ of n leaves, the collection of all phylogenetic trees with this leaf set can be assembled into a metric cube complex known as phylogenetic tree space, or Billera-Holmes-Vogtmann tree space, after [9]. In Chapter 2, we show that the isometry group of this space is the symmetric group S_n. This fact is relevant to the analysis of some statistical tests of phylogenetic trees, such as those introduced in [11]. In Chapter 3, co-authored with Megan Owen, we give a rigorous framework for comparing trees in different moduli spaces of phylogenetic trees, and apply this to define extension spaces of trees, a conservative split-based supertree construction method, and two measures of compatibility between tree fragments.

In Chapter 4, we discuss some techniques in manifold learning, and outline a new topologically-constrained nonlinear dimensionality reduction al-

gorithm, which quickly reduces a nerve complex build on local tangent space approximations to produce a small number of manifold charts, visualized by a collection of least squares alignments of contractible components. We also give a method to optimize tangent space alignment on a sphere, and a template for using local tensor decomposition of higher-order moments to extend this technique to intersecting and stratified manifolds.

Table of Contents

List of Figures

Chapter 1

Phylogenetic tree space

In the context of evolutionary biology, given a set of organisms referred to as taxa, a phylogenetic tree is a semi-labeled, weighted acyclic graph representing a possible evolutionary relationship between the taxa, using genotypic or phenotypic data. Such trees typically have a root which represents the common ancestor of the taxa, with a branch point at each speciation event, and a leaf for each taxon, such that the taxa which share more features are "nearer" to each other in the tree. The phylogenetic tree itself represents a finite metric space, with metric given by shortest weighted path length: a sequence of edges without repetition gives a unique path from one leaf to another, and the sum of their lengths is the distance, quantifying the genetic or phenotypic changes and differences between the taxa.

In addition to the distances between the taxa that a single phylogenetic tree represents, a distance between distinct phylogenetic trees with the same set of taxa can also be defined through the construction of *phylogenetic tree space* $\mathfrak{T}^n$ and BHV_n, for taxa labels $\{1, \ldots, n\}$[9]. Each tree is represented by a point in tree space, with location determined by the topology (shape) of the tree and its vector of edge lengths. The BHV distance between two trees is

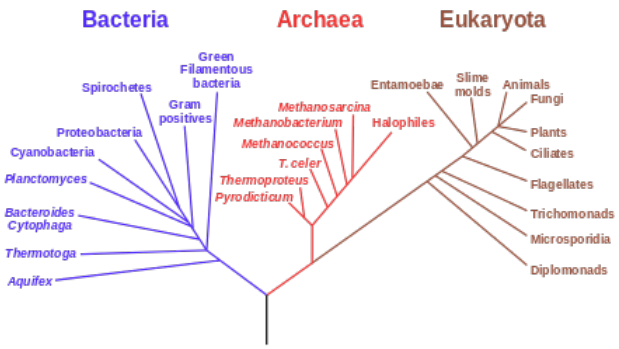


Figure 1.1: Phylogenetic Tree of Life. Image credit Wikimedia Commons.

the length of the shortest path between the two points in tree space.

1.1 Notation and Definitions

1.1.1 Phylogenetic trees

Definition 1.1.1. A **phylogenetic tree** T is an acyclic connected graph (a **tree**) with

- No degree 2 vertices.
- Degree 1 vertices each have a unique label. Such vertices are called **leaves** of T. The set of leaf labels is denoted $\mathcal{L}(T)$.
- There is a positive weight w_e for each edge e, and the set of edges is denoted $\mathcal{E}(T)$.

Unless indicated otherwise, $\mathcal{L}(T) = [n] = \{1, 2, \ldots, n\}$ for n the number of leaves. Phylogenetic trees are sometimes **rooted**, meaning the tree has a

distinguished leaf, the **root**, often an ancestor. The **topology** of a tree is the unweighted underlying tree with leaf labels.

Because phylogenetic trees are acyclic, the removal of an edge e separates T into two connected components. Since leaves are vertices in one component or the other, each edge e induces a partition of $\mathcal{L}(T)$ into the two components P_e and $P_e^c = \mathcal{L}(T) \setminus P_e$, called a **split** and represented as $P_e|P_e^c$. The set of all splits of T is denoted $S(T)$. When the ground set is obvious, we will suppress the complement and give a split by the smaller of its two partition sets, or if the two partitions are the same size, with the partition containing the lexicographically first leaf. There are two types of splits: a split is called **thick** (corresponding to an **internal** edge e) if P_e and P_e^c both have cardinality greater than 1, or equivalently if neither endpoint of e is a leaf, otherwise it is a **leaf** split (corresponding to a leaf edge). We will alternately refer to an edge $e \in T$ and the partition P_e it induces; for both, the weight is denoted w_e.

Definition 1.1.2. Two splits $P|P^c$ and $Q|Q^c$ are called **compatible** if one of: $P \cap Q, P \cap Q^c, P^c \cap Q, P^c \cap Q^c$ is empty. Two splits that are not compatible are called **incompatible**.

At most one of the intersections in Definition 1.1.2 can be empty. Compatibility of different splits P and Q is equivalent to the existence of a tree T containing two corresponding edges. In fact, tree topologies are in direct correspondence with pairwise-compatible sets of splits: given a set of i different

splits on leaf set $\mathcal{L}$ which are pairwise compatible, and weights for each, there is a unique phylogenetic tree (with i edges) realizing them [16, Theorem 1]. Conversely, for a phylogenetic tree T, the collection of all splits $S(T) = \{P_e\}$ (one for each edge e) is pairwise compatible. A phylogenetic tree contains at most $2|\mathcal{L}(T)| - 3$ splits, and $|\mathcal{L}(T)| - 3$ thick splits.

If the external (leaf) edges of T are also endowed with weights, then T is equivalent to an additive metric space, whose points are leaves with the weighted path metric on T. This correspondence is discussed further in Section 2.4.

1.1.2 Tree Space

For a fixed leaf set $\mathcal{L}$ and a set of compatible thick splits S on $\mathcal{L}$, there exists a unique tree topology realizing S, as discussed in the previous section. We can then organize the set of all phylogenetic trees with this topology by their weight sets, ordered lexicographically by the corresponding split of each weight, in a space isometric to $\mathbb{R}_+^{|S|}$. If we include the boundary, by allowing weights to be 0, then this space is isometric to $\mathbb{R}_{\geq 0}^{|S|}$ and called an **orthant**. Maximal orthants have dimension $|\mathcal{L}| - 3$. See Figure 1.2. We will denote the lowest-dimensional orthant containing tree T by $\mathcal{O}(T)$, and the lowest-dimensional orthant containing all trees with exactly the splits S by $\mathcal{O}(S)$. Conversely, the set of splits contained in all trees in the interior of orthant $\mathcal{O}$ is denoted by $S(\mathcal{O})$.

If two sets of compatible thick splits, S_1 and S_2, have splits in common,

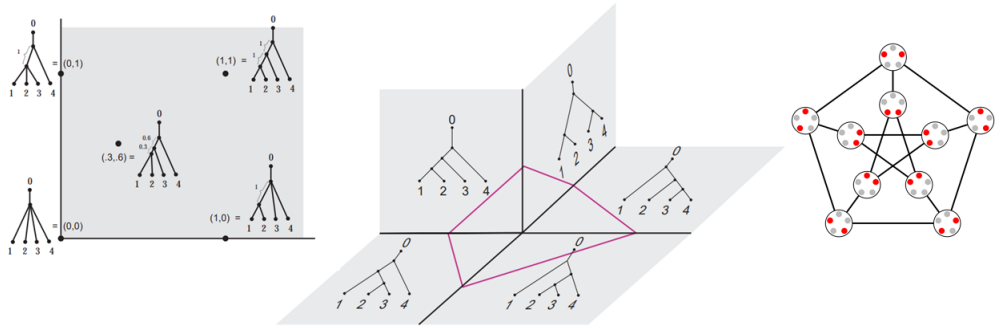


Figure 1.2: Left, a single orthant. Center, five orthants identified along common split sets. Right, the link L_5 of the origin, isomorphic to the Petersen graph. Image credit [9] and Wikimedia commons.

$C = S_1 \cap S_2$, then the orthants corresponding to S_1 and S_2 each have a boundary orthant $\mathbb{R}^{|C|}_{\geq 0}$ that contains the same trees. We identify all such common boundary orthants to produce a single space, called the **Billera-Holmes-Vogtmann (BHV) treespace** and denoted $\mathrm{BHV}_{\mathcal{L}}$, where $\mathcal{L}$ is the leaf set of all trees. When $\mathcal{L} = [n]$, we will alternatively write BHV_n for the space. The empty split set $S = \emptyset$ produces a single point, called the **cone point**, $\mathbf{0}$, which represents the unique star-shaped tree with no internal edges. The cone point is contained in each orthant at the origin, so the identified space is path-connected. We define the distance $d_{\mathrm{BHV}}(T, T')$ between points T and T' in this space to be the infimum of the lengths of all piecewise smooth paths from T to T', where path length is calculated by summing the L^2 distances of the path restricted to each orthant it passes through.

The BHV treespace was first proposed by Billera, Holmes, and Vogtmann in [9], where they showed that it is a contractible, complete, and globally non-positively curved, or CAT(0), cube complex. Global non-positive curva-

ture implies that there is a unique shortest path, or geodesic, between each pair of trees in the space. There exists a polynomial time algorithm to calculate this path and its length, given by Owen and Provan in [45].

1.1.3 Link graph

Definition 1.1.3. The **link** $L_{\mathcal{L}} := L_{\mathcal{L}}(\mathbf{0})$ of the cone point $\mathbf{0}$ is the set of all trees in $BHV_{\mathcal{L}}$ which have internal edge lengths summing to 1. Homeomorphically, $L_{\mathcal{L}}$ is the set of trees in $BHV_{\mathcal{L}}$ at fixed L_1 distance from $\mathbf{0}$.

Because $\mathrm{BHV}_{\mathcal{L}}$ is a cube complex, $L_{\mathcal{L}}$ is a simplicial complex; the face maps are restrictions of face maps of the cube complex, and every k-face of the cube complex intersects the link in a $(k-1)$-simplex. In particular, the 0-simplices correspond to single splits, the 1-simplices correspond to compatible split pairs, and k-simplices correspond to trees sharing the same k non-zero splits which have edge lengths summing to 1.

Chapter 2

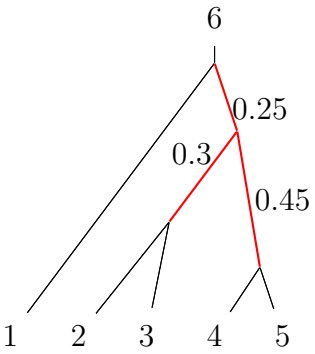
Isometries of phylogenetic tree space

BHV space, with geodesic metric, can be used to give precise geometric characterizations of collections of phylogenies, and to perform various statistical tests, such as those defined in [31], [59], and [5]. In [11], the matrix of pairwise distances between trees in a set is used as a signature to perform statistical inference. With techniques like this, which operate on the distance matrix instead of the trees themselves, the results are insensitive to isometry; this renders the classification of isometries of BHV_n extremely relevant.

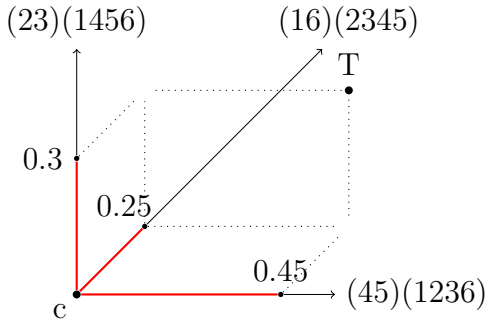
In Theorem 2.2.1, previously published in [27], we show that the group of isometries of BHV space is the symmetric group S_n, for n the total number of leaves including root. These isometries correspond to simple permutations of the leaves.

2.1 Background

An orthant boundary component of codimension k corresponds to a "degenerate" tree topology: trees on the boundary are 0 along k axes, so k of the edges in the orthant tree topology have length zero. This leaves a non-binary tree topology with $n-k-3$ non-trivial internal edges, and this



(a) A phylogenetic tree T with 6 leaves, 6 external ("leaf") edges, and 3 internal edges, weights as labeled.



(b) The orthant of BHV_6 [$\cong (\mathbb{R}^3)^{\geq 0}$] containing T, with an axis for each unweighted edge ("partition") of T. The axes are parametrized by edge length, so the point T is graphed above in relation to the other trees of identical topology.

topology appears on the boundary of a number of other orthants. This number is bounded in Lemma 2.2.6, which may be of independent interest. We then identify the orthant boundaries according to this (weighted labeled graph) equivalence. In particular, at the "origin" (the preimage of $(0, 0, \ldots, 0) \in \mathbb{R}^{n-3}$ under the parametrizing homeomorphism), every orthant exhibits the star-shaped tree having no internal edges of positive length. Under equivalence, then, the point $(0, 0, \ldots, 0)$, regardless of orthant, is shared and unique in BHV_n. Its image under identification is called the *cone point* c (see Figure 2.1b), well-named because for a particular simplicial complex L_n, it is the image of the quotient $\mathrm{BHV}_n = L_n \times [0, \infty)/(L_n \times 0)$. [9]

A metric on BHV_n is generated by the Euclidean metric within each

orthant: a path γ between trees T and T' has length

$$\ell(\gamma) = \sum_{S \in \mathcal{O}} |\gamma \cap S|,$$

where $|\cdot|$ is Euclidean path length via restriction to an orthant, and $\mathcal{O}$ is the set of all orthants in BHV_n. Then

$$d(T, T') := \inf_{\gamma:\gamma(0)=T,\gamma(1)=T'} \ell(\gamma)$$

is a complete metric, which is realized by a unique geodesic γ with $\ell(\gamma) = d(T, T')$ [9]. The natural Lebesgue measure for open sets in BHV is described analogously in Section 2.2.2 in order to give the volume of small neighborhoods of points in BHV_n; we suspect this might also be of independent interest.

2.1.1 Automorphisms versus isometries

It might seem natural to classify isometries of BHV_n, which is a CAT(0) cube complex (see [48]), via natural isomorphisms of that structure. However, it is important to note that in general, isometries of cube complexes can exceed their cube complex automorphisms, and if the cubes are endowed with a different metric, an automorphism may not be an isometry at all. As a trivial example, one can consider the integer cubulation of $\mathbb{R}^2$, which in addition to the $D_4 \times \mathbb{Z}^2$ lattice isometries, retains the $O(2) \times \mathbb{R}^2$ real isometries, which do not preserve the cube complex structure. This discrepancy was addressed recently in [14] - Bregman shows that for a CAT(0) cube complex C with unit euclidean metric on each cube and global metric given by minimal path length,

if $Isom(C) \neq Aut(C)$, then there is a full subcomplex D of C admitting a decomposition into a product $E \times \mathbb{R}^n$, where E is a full subcomplex of D. This shows that in some sense, the only additional isometries come from an $\mathbb{R}^n$-type subcomplex, possibly with non-flat curvature. We note that our result gives a counterexample to the converse: the full subcomplex of BHV_5 given by any 5-cycle in the link is $\mathbb{R}^2$ with the singular cone metric $\mathrm{Cone}(\mathbb{R}^2, 5)$, but we do not gain any additional isometries.

Besides the proof given in Section 2.2.1 of this chapter, $Aut(\mathrm{BHV}_n)$ is known from the work of Abreu and Pacini classifying cone complex automorphisms of the moduli space $M_{0,n}^{\mathrm{trop}}$ of tropical genus 0 curves with n marked points[1]. Their result is closely related to our Proposition 2.2.3. Inspection of the argument suggests that they are proving the same essential combinatorial fact, through an inductive technique. In fact, our main result could be proved via theirs through a direct application of Lemma 2.2.6 to the interior of top-dimensional orthants, analogously to our proof in Section 2.2.3 that $Aut(L_n) = Isom(L_n)$.

2.2 Main Theorem

Theorem 2.2.1. *For $n \geq 3$, the isometry group of BHV_n is isomorphic to S_n. These isometries correspond to permutation of leaf labels.*

It is clear that a permutation of the leaf labels induces an isometry from BHV_n to itself, so the following lemmas will build to the converse. This

will involve two stages.

First, in Section 2.2.1 we will use the Erdős-Ko-Rado theorem to give a new proof that the automorphism group of L_n, the spherical simplicial complex of points at distance 1 from the origin, is S_n. As we've remarked already, this fact is implied by recent work of [1], who computed the automorphisms of BHV_n as a cone complex.

In Section 2.2.2, we will then give local bounds on the natural volume measure in BHV_n to show that any isometry of BHV_n induces a self-map of the unit sphere L_n, and any isometry of the unit sphere to itself is an automorphism of simplicial complexes. Having classified these in the previous section, we conclude in Section 2.2.3 that any isometric automorphism of BHV_n must be a relabeling.

2.2.1 Link Automorphisms

Following [9], BHV_n can be expressed as a cone on a simplicial complex L_n, constructed:

- A 0-simplex (vertex) v for each subset $P_v \subset \{1, 2, \ldots, n\}$ such that $2 \leq |P_v| < n/2$. The size $|P_v|$ will often be denoted k. Each P_v determines a partition P_v, P_v^c of $[n]$, unique for $k < n/2$. If n is even, we also include a vertex for each pair P, P^c with $|P| = |P^c| = n/2$.

- A 1-simplex (edge) (v, w) for each *compatible* pair (P_v, P_v^c) and (P_w, P_w^c). P_v and P_w are said to be compatible if one of the sets $[P_v \cap P_w, P_v \cap$

$P_w^c, P_v^c \cap P_w, P_v^c \cap P_w^c]$ is empty. We will simplify this condition in Lemma 2.2.2.

- The complex (graph) constructed up to this point is denoted L_n^1, the 1-skeleton of L_n.

- L_n is the simplicial complex with a k-simplex, $k > 1$, for each $(k+1)$-clique present in L_n^1 (i.e. L_n is a *flag* simplicial complex).

- L_n is realized geometrically as a right-angled spherical simplicial complex: for S^k the unit sphere in $\mathbb{R}^k$, each simplex is isometric to

$$\{(x_1, \ldots, x_{k+1}) \in S^k : x_i \geq 0 \text{ for all } i\}$$

  with the spherical metric.

- Finally, BHV_n is a right-angled spherical metric cone on L_n, as described in [17]. Practically, this means that each tree topology is parametrized by $n-3$ non-negative, real coordinates, with the local standard metric in $\mathbb{R}^{n-3}$, as shown in the introduction.

We begin with some facts about L_n^1, and then show the automorphism group of L_n^1 in Proposition 2.2.3. This gives the automorphisms of L_n via the flag property in Corollary 2.2.4.

Lemma 2.2.1. *The degree of a vertex v of partition size k in L_n^1 is given by:*

$$\deg(v) = 2^k + 2^{n-k} - n - 4$$

Proof. The degree of v is the number of partitions (of size at least 2) compatible with P_v, P_v^c. For A, A^c distinct from P_v, we have four compatibility conditions: (1) $A \cap P_v^c = \emptyset$, or equivalently, $A \subset P_v$; (2) $A \cap P_v = \emptyset$, so $A \subset P_v^c$; (3) $A^c \cap P_v = \emptyset$, so $A^c \subset P_v^c$, and (4) $A^c \cap P_v^c = \emptyset$, so $A^c \subset P_v$.

If we have a subset of $[n]$, such that it or its complement satisfies one of these conditions, it can be labeled (A or A^c) so that in fact it satisfies (1) or (2). Therefore to count the number of total compatible partitions, we will count subsets $A \subset [n]$ satisfying (1) or (2); that is, nontrivial subsets of sufficient size of P_v or P_v^c:

$$\overbrace{\sum_{x=2}^{k-1} \binom{k}{x}}^{(1)} + \overbrace{\sum_{x=2}^{n-k-1} \binom{n-k}{x}}^{(2)} = (2^k - k - 2) + (2^{n-k} - (n-k) - 2) = 2^k + 2^{n-k} - n - 4.$$

□

Lemma 2.2.2. *For two distinct partitions (A, A^c), (B, B^c), of size $|A| = k_1$, $|B| = k_2$, $2 \leq k_1 \leq k_2 \leq n/2$, $(A, A^c), (B, B^c)$ are compatible iff $A \cap B = \emptyset$ or $A \subset B$. If $k_1 = k_2$, $A \cap B = \emptyset$ is equivalent to compatibility of distinct partitions.*

Proof. By the pigeonhole principle, $A^c \cap B^c$ is nonempty. If $B \cap A^c$ is empty, then $B \subseteq A$, which implies by size considerations that $B = A$. For distinct partitions this will not occur. On the other hand, we can have $A \cap B$ or $A \cap B^c$ empty. In the latter case, it is implied that $A \subseteq B$. If $k_1 = k_2 < n/2$, then $A \subseteq B$ implies $A = B$. □

Remark 2.2.1. The **Kneser graph** $KG_{n,k}$ is the graph whose vertices correspond to the k-element subsets of a set of n elements, and where two vertices are adjacent if and only if the two corresponding sets are disjoint. Labeling the vertices of L_n^1 by the smaller of the two partitions, and sorting by size, it follows immediately that L_n^1 contains a unique subgraph G_k isomorphic to $KG_{n,k}$ for each partition size $k = 2, 3, \ldots, \lceil n/2 \rceil - 1$. These subgraphs have disjoint vertex sets. If n is even, then there are an additional $\frac{1}{2}\binom{n}{n/2}$ vertices, pairwise disjoint from each other.

Proposition 2.2.3. *The automorphism group* $Aut(L_n^1) \cong S_n$.

Proof. To see that S_n is a subgroup of $Aut(L_n^1)$, we recall that L_n^1 is constructed via combinatorial conditions (compatibility) that are independent of choice of label. So any permutation of $\{1, \ldots, n\}$ gives an identical graph when constructed with the same notion of compatibility of partitions. Therefore given $\sigma \in S_n$, we can map $P = (x_1, x_2, \ldots x_k) \mapsto \sigma(P) = (\sigma(x_1), \ldots, \sigma(x_k))$, and this preserves adjacency.

It remains then to show that $Aut(L_n^1) \leq S_n$, which we will do by defining an injective group homomorphism $Aut(L_n^1) \to S_n$.

Let $\sigma \in Aut(L_n^1)$, and denote by G_k the induced subgraph on the k-vertex set $\{v \in V(L_n^1) : |P_v| = k\}$. By Lemma 2.2.1, the degree of a vertex v is completely determined by its size k. Since the expression $2^k + 2^{n-k} - n - 4$ is monotonically increasing (in k) for $k < n/2$, the degree of v is also unique

to vertices of the same partition size. This means that $\sigma(v)$ must be contained in G_k, so σ restricts to a graph automorphism on G_k.

We now show that this restriction map $Aut(L_n^1) \to Aut(G_k)$ is injective for $2 \leq k < n/2$. Let σ_{id_k} be an automorphism of L_n^1 which acts as the identity on G_k. Then we show that G_{k+1} is fixed as well, using the fact that adjacencies to G_k are preserved under automorphism.

Let $N(P_v)^{+1}$ denote the set of neighbors of $v \in G_k \subset L_n^1$ of size $k+1$, i.e.

$$N(P_v)^{+1} = \{P_w \in G_{k+1} : P_v \subset P_w \text{ or } P_v \cap P_w = \emptyset\}$$

by Lemma 2.2.2. Similarly, we denote by $N(P_v)^{-1}$ the set of neighbors of v with partitions one size lower: $N(P_v)^{-1} = \{P_w \in G_{k-1} : P_w \subset P_v \text{ or } P_v \cap P_w = \emptyset\}$. Let $P_z = (x_1, x_2, \ldots, x_{k+1}) \in G_{k+1}$. Then $(x_1, x_2, \ldots, x_{k+1})$ is the unique partition of size $k+1$ which is compatible with all of its size-k neighbors:

$$\{P_z\} = \bigcap_{P_v \in N(P_z)^{-1}} N(P_v)^{+1}$$

To show this, we note that for two distinct $(k+1)$-partitions of the same size, there exists at least one set of k labels which is compatible with one and not the other: for $P_w \neq P_z$, there is a label $i \in P_w, i \notin P_z$ and there is a $j \in P_w^c$, $j \notin P_z$ (by size considerations), so that any k-subset of P_z^c containing both i and j is compatible with P_z, but cannot be compatible with P_w, which excludes P_w from this intersection.

Now, since adjacencies and G_k are preserved by any automorphism, $N(P_v)^{+1}$ is preserved by σ_{id_k} for $v \in G_k$. So we can conclude by the set equivalence above that P_z is preserved as well, which gives the desired result that $\sigma_{id_k}(G_{k+1}) = G_{k+1}$, which implies that G_j for $j > k$ is preserved under σ, by repetition of the same argument. We have

$$P_z = \bigcap_{\alpha \in P_z^c} N(x_1 \dots x_k, \alpha)^{-1},$$

which shows $\sigma_{id_k}(G_j) = G_j$ for $j < k$ in the same manner. Since $V(L_n^1) = \bigsqcup_{k=1}^{\lfloor n/2 \rfloor} V(G_k)$, we have shown that $\sigma_{id_k} \in \ker(Aut(L_n^1) \to Aut(G_k))$ acts trivially on the vertices of L_n^1, so must be the trivial automorphism.

Now following [24], we show that $Aut(G_k) \cong S_n$ for $2 \leq k < n/2$. By the Erdős-Ko-Rado Theorem, any family of subsets of $\{1, 2, \dots, n\}$ of uniform size k having pairwise-nonempty intersection has size $\leq \binom{n-1}{k-1}$, and the subsets achieving equality are of the form

$$G_k^{(i)} = \{v \in G_k : i \in P_v\}$$

for $i \in [n]$.[22] Since these partitions pairwise-intersect, they are pairwise disjoint in G_k, and by definition form a maximum-size independent set in G_k. Correspondingly, $\sigma \in Aut(G_k)$ must induce a permutation on these maximum independent sets, which determines a (surjective) homomorphism $Aut(G_k) \to S_n$. To see that this is an isomorphism, note that if σ fixes the $G_k^{(i)}$, it must be the identity: suppose $\sigma(v) \neq v$. Then there exists some $j \in P_v$ such that $j \notin P_{\sigma(v)}$. This would imply that $\sigma(G_k^{(j)}) \neq G_k^{(j)}$, a contradiction.

Now we see that $Aut(L_n^1) \cong Aut(G_k) \cong S_n$ (for any/all $2 \leq k < n/2$, we really only needed one), which completes the proof. □

Corollary 2.2.4. *The group of simplicial automorphisms of L_n is isomorphic to $Aut(L_n^1)$.*

Proof. Let $n \geq 3$ be given. First we note that $Aut(L_n) = Aut(L_n^1)$: each simplicial automorphism induces an automorphism of the 1-skeleton, and since L_n contains no simplices with the same 1-skeleton, this map is injective. Then since L_n is a flag complex ([9]), given a graph automorphism of L_n^1, we can define a canonical extension by sending a k-simplex to the k-simplex determined by the image of its 1-skeleton k-clique. □

2.2.2 Measure and Isometry

We will now consider the entire metric space BHV_n, and show that the standard embedding of L_n into the unit sphere is invariant under isometry.

There is a natural volume measure μ on $\mathcal{B}(\mathrm{BHV}_n)$, which is given by the local Lebesgue measure in each orthant. Explicitly, for $A \in \mathcal{B}(\mathrm{BHV}_n)$,

$$\mu(A) = \sum_S |A \cap S|$$

where $S \cong (\mathbb{R}^+)^{n-3}$ is an orthant of BHV_n and $|\cdot|$ is the real Lebesgue measure. As we will see in the following lemmas, the volume of small neighborhoods can vary exponentially under translation; this fact is one of the major impediments to statistical techniques in tree space.

Lemma 2.2.5. *For $\sigma \in Isom(BHV_n)$, σ preserves the volume measure μ on BHV_n.*

Proof. Let B_x be a ball of radius 1 centered at a point $x \in \mathrm{BHV}^n$. For a fixed orthant S, σ induces an isometry of S into BHV_n, so $\mu(\sigma(B_x \cap S)) = |B_x \cap S| = \mu(B_x \cap S)$. For a measure zero set Z on the boundary components of tree space, B_x can be written as a disjoint union:

$$B_x = Z \bigsqcup_S (B_x \cap \mathrm{int}(S)),$$

$$\begin{aligned}\sigma(B_x) &= \sigma\left(Z \bigsqcup_S (B_x \cap \mathrm{int}(S))\right) \\ &= \sigma(Z) \bigsqcup_S \sigma(B_x \cap \mathrm{int}(S)),\end{aligned}$$

since σ is injective. Therefore we conclude that $\mu(\sigma(B_x)) = \bigsqcup_S \mu(B_x \cap S) = \mu(B_x)$. □

Lemma 2.2.6. *Let $x \in BHV_n$, with $\{e_1, e_2, \ldots, e_p\}$ the set of positive-length edges in x, then $0 \leq p \leq n-3$. Let $\epsilon > 0$ be smaller than the length of e_i for each $i \in \{1, 2, \ldots, p\}$. Then for $B_x(\epsilon)$ the ball of radius ϵ centered at x,*

$$A_{n-3}(\epsilon) \;\leq\; \mu(B_x(\epsilon)) \;\leq\; (2n-2p-5)!!\frac{2^p}{2^{n-3}} A_{n-3}(\epsilon), \tag{2.1}$$

where $A_m(\epsilon)$ is the volume of a ball of radius ϵ in $\mathbb{R}^m$. Furthermore, the lower bound is achieved if and only if $p = n-3$, which means x is binary.

Proof. First, we note that x is contained in a cubical face F of dimension p in BHV_n. Then F is contained in some number $s(F)$ of top-dimensional

orthants, each representing a binary tree topology whose partition set contains the partition set of x. The restriction on ϵ ensures that $B_x(\epsilon)$ intersects no lower-dimensional faces, so just as a neighborhood of a point contained in a p-face in an $(n-3)$-cube, the restriction of $B_x(\epsilon)$ to each orthant is isometric to $\left(\frac{1}{2^{codim(F)}}\right)$-th of a Euclidean ϵ-ball. So we have that

$$\mu(B_x(\epsilon)) = \frac{s(F)}{2^{n-3-p}} A_{n-3}(\epsilon). \tag{2.2}$$

While $s(F)$ is highly dependent on the topology of F, we will show that $s(F) \leq (2n-2p-5)!!$, which gives (2.1).

Instead of describing the topology of F as a list of p internal partitions, we will now consider the internal nodes $y_1, \ldots, y_{p+1}$, with degree sequence $d_1, d_2, \ldots, d_{p+1}$. Note that

$$\sum_{i=1}^{p+1} (d_i - 3) = n - p - 3, \tag{2.3}$$

by the fact that the sum of the full degree sequence of a tree is twice the number of edges, so $\sum d_i + n = 2(n+p)$, from which the equality follows. Then

$$s(F) = \prod (2d_i - 5)!! \tag{2.4}$$

because locally, each vertex of degree d_i forms the interior node of a star tree with d_i "leaves" representing the subtrees. So to find the number of binary tree topologies with the same subtrees as leaves, we count the orthants in BHV_{d_i}, that is, $(2d_i-5)!!$. This choice fixes all other nodes of F, so an element of $s(F)$ is specified uniquely by freely choosing a binary tree at each interior node.

Next we note that $(2d_i - 5)!!$ has $d_i - 3$ terms greater than 1. For any degree sequence d_i, we then have by (2.3) that the product (2.4) has $(n-p-3)$ non-trivial terms, each of which is at least 3, which gives the lower bound. This product is maximized with the degree sequence $n-p, 3, 3, \ldots, 3$, for which $s(F) = (2(n-p)-5)!!$, which gives the upper bound. For $p < n-3$, $s(F)$ is strictly greater than 2^{n-3-p}. For $p = n-3$, we have a coefficient of 1. These two facts show that the lower bound is achieved only for binary trees. □

Corollary 2.2.7. *Let $n \geq 4$, c the cone point in BHV_n, $x \neq c \in BHV_n$. Then $\mu(B_c(\epsilon)) > \mu(B_x(\epsilon))$ for $\epsilon < \min_{e \in E(x)} w_e$, where $E(x)$ is the set of edges of x as a graph, and w_e their respective weight in x, so that ϵ is smaller than the length of the smallest non-zero edge of x.*

Proof. First note that $\mu(B_c(\epsilon)) = \frac{(2n-5)!!}{2^{n-3}} A_{n-3}(\epsilon)$ for any $\epsilon > 0$, where $A_m(\epsilon)$ is the volume of a ball of radius ϵ in $\mathbb{R}^m$. Then for $x \neq c$, $p \geq 1$, so by Lemma 2.2.6,

$$\mu(B_x(\epsilon)) \leq (2n-7)!! \frac{2}{2^{n-3}} A_{n-3}(\epsilon).$$

But since $2 < 2n - 5$, $\mu(B_x(\epsilon)) < \mu(B_c(\epsilon))$. □

2.2.3 Proof of Main Theorem

Proof. Let $n \geq 4$ be given.

Each of the relabeling automorphisms of L_n is an isometry, and it extends in the obvious way to an isometry of BHV^n by relabeling the leaves of

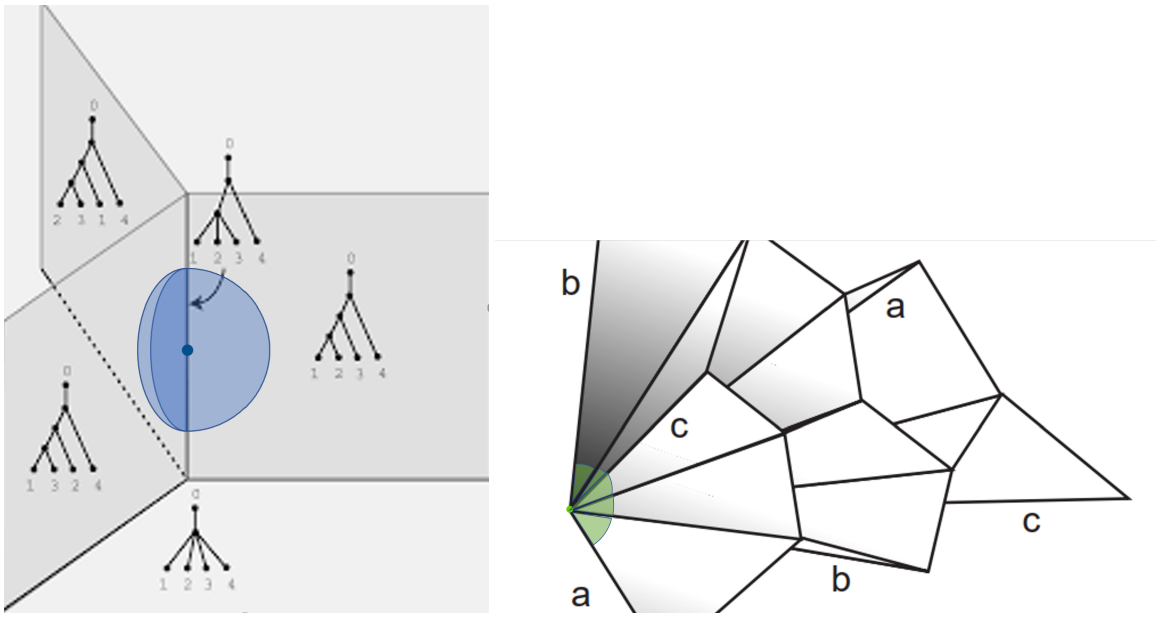


Figure 2.2: Left, a neighborhood in BHV^5 with volume $(3/2)\pi\epsilon^2$; Right, a neighborhood of c, with volume $15/4\pi\epsilon^2$.

an arbitrary tree, so we can conclude that $S_n \leq \mathrm{Isom}(\mathrm{BHV}_n)$.[1] Conversely, it remains to be shown that $\mathrm{Isom}(\mathrm{BHV}^n) \leq S_n$. Let $\sigma \in \mathrm{Isom}(\mathrm{BHV}^n)$ be given.

1. Let $B_x(\epsilon)$ denote the set of points at distance at most ϵ from x. Then by definition of an isometry, $\sigma(B_x(\epsilon)) = B_{\sigma(x)}(\epsilon)$ for all ϵ.

2. For $x \neq c$, $\epsilon < \min_{e \in E(x)} w_e$, the measure $\mu(B_x(\epsilon)) < \mu(B_c(\epsilon))$ by Cor. 2.2.7.

3. We conclude that $\sigma(c) = c$ by Lemma 2.2.5, so $\sigma(B_c(1)) = B_c(1)$.

[1] Equivalently, an automorphism of a cube complex with uniform euclidean metric is automatically an isometry.

4. Since $L_n = \partial(\overline{B_c(1)})$ is the set of points at distance 1 from c, we conclude that $\sigma(L_n) = L_n$.

5. In the remainder of the proof, we will show that $Isom(L_n) = Aut(L_n) \cong S_n$, and this will give the titular result.

Let $\sigma \in Isom(L_n)$be given. Let $x \in L_n$ be a binary tree, so x is contained in the interior of an $(n-4)$-simplex. Then by Lemma 2.2.6 and Lemma 2.2.5, $\sigma(x)$ is also necessarily a binary tree, and so contained in the interior of an $(n-4)$-simplex in L_n. An isometry which restricts to $\tau : \text{int}(\Delta^{n-4}) \to \text{int}(\Delta^{n-4})$ on the interior of an $(n-4)$-simplex must extend by continuity to an isometry $\bar{\tau} : \Delta^{n-4} \to \Delta^{n-4}$. Such an isometry is a simplicial map, sending k-simplices to k-simplices. But every k-simplex in L_n is on the boundary of a maximal simplex (equivalently, every non-binary tree has a choice of additional edges making it binary), so we conclude that σ is a simplicial map from L_n to L_n, i.e. $\sigma \in Aut(L_n)$. Since every automorphism is an isometry, we conclude $Isom(L_n) \cong Aut(L_n)$, and by Corollary 2.2.4, $Aut(L_n^1) \cong Aut(L_n) \cong S_n$. $\square$

Chapter 3

Representations of Partial Leaf Sets

Phylogenetic tree space allows for direct comparison and summary of trees that have different shape and size. However, it is sometimes necessary to analyze collections of trees on nonidentical taxa sets (i.e., with different numbers of leaves), and in this context it is not evident how to apply BHV space. Ren et al. [46] approach this problem by describing a combinatorial algorithm extending tree topologies to regions in higher-dimensional tree spaces, so that one can quickly compute which topologies contain a given tree as partial data. In this work, joint with Megan Owen, and previously published in [28], we refine and adapt their algorithm to work for metric trees to give a full characterization of the subspace of extensions of a subtree (see Algorithm 1 and Equation 3.1). We describe how to apply our algorithm to define and search a space of possible supertrees and, for a collection of tree fragments with different leaf sets, to measure their compatibility. We give theoretical guarantees on computation speed and accuracy for each procedure.

3.1 Introduction

To combine the data of more than two trees, e.g. if $\mathbf{T} = \{T_i\}$ is a set of phylogenetic trees describing different evolutionary relationships between the taxa (leaf set) $\mathcal{L}$, $\mathbf{T}$ is represented as a set of points in $\mathcal{T}^n$. By taking the mean of $\mathbf{T}$ [7, 8, 15, 40], or clustering the points [26], or constructing confidence regions [59], we can describe $\mathbf{T}$ in a way which incorporates the range of metric and combinatorial shape differences.

However, there are situations in which one of the assumptions of this model, that each tree in $\mathbf{T}$ has a fixed leaf set $\mathcal{L}$, is not reasonable. For example, with improvements in sequencing technology, many phylogenetic datasets now consist of thousands of gene trees, each of which represents the evolutionary history of a single gene in the species set of interest [39]. However, not all genes appear in all species, and currently genes with an incomplete leaf set are often discarded before beginning the analysis. A second example is comparing parallel evolutionary chains in viruses or tumors, where some strains are comparably similar across samples (and therefore can be considered the same leaf) but are not necessarily all present in every sample [62], i.e. each $T_i \in \mathbf{T}$ has its own leaf set $\mathcal{L}_i$ which is contained in some common larger set $[N]$. The fact that the trees T_i belong to different parametrized spaces prevents us from using the techniques of BHV analysis described previously, but as we will show, tree sets with some "combinatorial compatibility" will admit a fairly precise notion of distance which is based on the BHV metric in $\mathcal{T}^N$, with no loss of data.

Our approach to this problem uses the **tree dimensionality reduction** map Ψ defined in Zairis et al. [62], which gives a map from a tree space $\mathcal{T}^N$ to the lower-dimensional tree space $\mathcal{T}^{\mathcal{L}}$ that contains all trees with a subset of the leaves $\mathcal{L} \subset [N]$. This map is induced by the natural subspace projection. We will first construct the pre-image Ψ^{-1} of this map, which can be used to recover information about the original tree T from the images $\{\Psi_{\mathcal{L}}(T)\}$ for varying $\mathcal{L}$. This map Ψ is also fundamental to the previous applications, which we solve by mapping T_i to their preimages $\Psi^{-1}(T_i)$ in the common domain space $\mathcal{T}^N$, and comparing the sets.

This precise problem, of analyzing trees with different numbers of taxa collectively in BHV tree space, was first approached by Ren et al. [46]. They developed the theory behind the combinatorial step in Section 3.3.3.1, toward the goal of comparing trees with different taxa sets. The algorithm presented in that section, together with Proposition 3.3.4, clarifies their results and shows their implications for the computation of tree dimensionality reduction and its preimage.

Analysis in BHV space is, of course, not the only way to approach problems of this type. Given the set $\{T_i\}$, it is sometimes efficient to "prune" the trees to their common taxa $\cap_i \mathcal{L}_i$ for comparison, if such a set $\cap_i \mathcal{L}_i$ is sufficiently large to preserve important data. In this case, any tool for analyzing sets of trees with identical taxa can then be used. In the context of reconstructing a species tree from gene trees, the relationship between these trees is modeled by the coalescent process, and algorithms and approaches specific

to this situation can take advantage of this model [41, 47]. To avoid making simplifying assumptions, there are also some software packages currently available which use Bayesian coalescent-based techniques, from the original data rather than trees, to assemble multiple parallel, incomplete data samples into a single tree [21, 30, 38]. There are also algorithms, based on the (often reasonable) assumption that differences in topology arise from recombination events, that aggregate metric data into phylogenetic networks [52]. These algorithms can often accommodate non-uniform data as well. However, they share the same drawback as most classical phylogenetic tree algorithms, in that they produce a single tree or tree-like object, rather than a region of possible trees in tree space. Finally, there are approaches that instead estimate the distances from the missing leaves to the existing leaves using the existing entries in the trees' distance matrices [19, 57, 60]. None of these methods guarantee that the completed distance matrix is additive, and thus while the matrix can be successfully used in further analysis, it may not directly correspond to a completed tree, as in our framework.

There is also the problem of supertree reconstruction, which aims to combine partially overlapping phylogenies into a common tree. Summaries and selected supertree methods can be found in Bininda-Emonds [10], Akanni et al. [2], Warnow [55], and Wilkinson et al. [58]. The techniques in this chapter give a conservative (low tolerance for topological error), split-based supertree method for BHV space, which does not necessarily represent an improvement on the search for a maximum-likelihood supertree; rather, we can rigorously

(rather than heuristically) define the space of possible supertrees, in a manner amenable to search, and expand the possible analyses available.

With the geometric framework established in this chapter, we can define and compute some useful objects. First, in Section 3, we show how to efficiently compute $\Psi^{-1}(T)$, the preimage of tree T under the tree reduction map, which gives all trees with the full set of leaves N that map onto T. The algorithm, given in two parts, calculates the extension space E_T^N, which represents the set of all phylogenetic trees in $\mathcal{T}^N$ which can result from adding $N-|\mathcal{L}|$ additional leaves to tree T with leaves $\mathcal{L}$. Theorem 3.3.1 shows that this construction, which extends the results and definitions of [46], coincides with $\Psi_n^{-1}(T)$ in $\mathcal{T}^N$.

This fact immediately gives a method of finding the set of trees X which satisfy the system $\{\Psi_{\mathcal{L}_i}(X) = T_i\}$ for some collection of trees $\mathbf{T} = \{T_i\}$, and we suggest some shortcuts to speed up the process. This solution space $E_{\mathbf{T}}$ is computed efficiently in Section 4 in a method similar to the one presented in Section 3, and is shown in Proposition 3.4.4 to be the intersection of sets $\Psi_{\mathcal{L}_i}^{-1}(T_i)$ in a common domain.

Stability concerns lead us to Section 5, which first defines an approximate solution space to $\{\Psi_{\mathcal{L}_i}(X) = T_i\}$ with some parameter α of constant error tolerance, or p_α of error tolerance proportional to local size. These relaxations will be the products of Sections 5.1 and 5.2, and will allow for the stability results in Proposition 3.5.4 and Lemma 3.5.5. Proposition (3.5.4) implies an additional non-trivial fact about a set $\Psi^{-1}(T)$, that if it intersects a cubical face $\sigma \subset \mathcal{T}^N$, it intersects all cubes $\tau \supset \sigma$.

We use these error tolerance parameters for single trees, α and p_α, to define two parameters $\alpha_{\mathbf{T}}$ and $p_{\mathbf{T}}$ measuring the degree of metric distortion for a collection of trees $\mathbf{T} = \{T_i\}$ satisfying a combinatorial compatibility condition. The parameters represent the minimum error tolerance (uniform or proportional) necessary to construct a supertree from the $\{T_i\}$. These parameters will result from linear optimization problems related to the equations defining the approximate solutions spaces, and can be directly computed using the most efficient linear programming methods available.

3.2 Background

Unlike the previous chapter, the algorithm and results presented apply to the space $\mathcal{T}^n$, or $\mathcal{T}^{\mathcal{L}}$, for any set of leaves $\mathcal{L}$. This space embeds a phylogenetic tree according to the partition and weights of all of its edges, including leaf edges as well as the internal edges that parametrize BHV. Since all trees in $\mathrm{BHV}_{\mathcal{L}}$ have the same leaves, and therefore the same leaf partitions, we can represent these leaf edge lengths globally with non-negative coordinates $(\mathbb{R}_{\geq 0})^{|\mathcal{L}|}$, and define **tree space** $\mathcal{T}^{\mathcal{L}}$ with this product

$$\mathcal{T}^{\mathcal{L}} := \mathrm{BHV}_{\mathcal{L}} \times (\mathbb{R}_{\geq 0})^{|\mathcal{L}|}$$

In this case, the cone point is the tree with no edges and all leaves identified into a single point. Importantly, $\mathcal{T}^{\mathcal{L}}$ has all of the important features of $\mathrm{BHV}_{\mathcal{L}}$: it remains connected, globally non-positively curved, and contractible. As above, when $\mathcal{L} = [n]$, we may alternatively write $\mathcal{T}^n$ for the space. The distance

$d_{\mathcal{T}^{\mathcal{L}}}(\cdot,\cdot)$: $\mathcal{T}^{\mathcal{L}} \times \mathcal{T}^{\mathcal{L}} \to \mathbb{R}$ can also be computed by a version of the algorithm of Owen and Provan [45].

$\mathrm{BHV}_{\mathcal{L}}$ can then be expressed as a cone on $L_{\mathcal{L}}$ based at $\mathbf{0}$ (hence the name "cone point"), with the cone dimension parametrizing magnitude. Denote the 1-skeleton of the link $L^1_{\mathcal{L}}$. The global non-positive curvature condition on $\mathrm{BHV}_{\mathcal{L}}$ implies that $L_{\mathcal{L}}$ is a flag complex, meaning that each k-clique in $L^1_{\mathcal{L}}$ bounds a k-simplex in $L_{\mathcal{L}}$, which corresponds uniquely to the orthant of dimension k spanned by the k splits. Thus, $L_{\mathcal{L}}$ is recoverable from $L^1_{\mathcal{L}}$, which together encode all of the non-linearity of $\mathrm{BHV}_{\mathcal{L}}$. In [46], and in the algorithm presented in Section 3.3, $L^1_{\mathcal{L}}$ is used to calculate the (combinatorial) extension objects $G_{T_s,n,\ell}$ and $S_{T_s,n,\ell}$.

3.2.1 Tree dimensionality reduction

A weighted graph, endowed with the shortest path metric, is a metric space whose underlying set is the vertices of the graph. Acyclic graphs have unique geodesics, and so a metric tree with n leaves can be equivalently considered as a metric on the set of n leaves, with distance between two leaves given by the length of the unique path between them. A metric δ which arises from a tree in this way is called an **additive** metric, and satisfies the four point condition:

$$\delta(a,b) + \delta(c,d) \leq \max\{\delta(a,c) + \delta(b,d), \delta(a,d) + \delta(b,c)\}$$

for all leaves a, b, c, d.

The four point condition is also sufficient to determine additivity, which in turn implies the existence of a unique tree realizing this metric [16]. The **additive distance matrix** of a tree T with leaf set $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_n\}$ is denoted A_T and is an $n \times n$ matrix where the (i, j)-th entry is $\delta(\ell_i, \ell_j)$, the distance between leaves ℓ_i and ℓ_j in tree T.

A subspace of an additive metric space is additive, and additive subspaces can be seen as forming subtrees. **Tree dimensionality reduction (TDR)**, as defined in [62], is a method of generating the tree for a subspace of an additive metric space from the original metric tree, and for a more general class of metric spaces called "nearly" additive. This work concerns strictly additive metric spaces, although many algorithms exist to project nearly additive spaces to tree approximations.

Definition 3.2.1. Let T be a tree with leaf set $[N] = \{1, 2, \ldots, N\}$, and let $\mathcal{L} \subset [N]$. The **tree dimensionality reduction map** $\Psi_{\mathcal{L}} : \mathcal{T}^{[N]} \to \mathcal{T}^{\mathcal{L}}$ is the map sending $T \in \mathcal{T}^N$ to the induced subtree spanned by the leaves $\mathcal{L}$, where the induced subtree contains the vertices and edges on the shortest paths through T between the leaves in $\mathcal{L}$, with each resulting degree 2 vertex v and its incident edges $(v, u_1), (v, u_2)$ with lengths ℓ_1 and ℓ_2 respectively, being replaced by a single edge (u_1, u_2) with length $\ell_1 + \ell_2$. We refer to this process as **concatenation** of (v, u_1) and (v, u_2).

Example 3.2.1. *Starting with the tree on the left in Figure 3.1, tree dimensionality reduction to the leaf set* $\{1, 2, 3, 4\}$ *is performed by first pruning the*

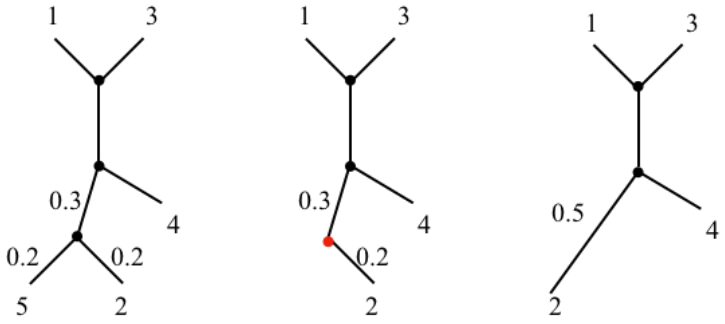


Figure 3.1: Left, a tree with 5 leaves. Center, the tree with leaf 5 and its edge deleted, resulting in a degree two vertex (in red). Right, the tree after concatenating the two edges adjacent to the degree two vertex.

5th leaf and its leaf edge, which gives the center tree. This tree has a degree 2 vertex, in red, which is removed, its boundary edges concatenated, to produce the final tree on the right.

We will also consider the related dimensionality reduction map on splits, which we will refer to as **projection**. For a split $P|P^c$ on leaf set $[N]$, the projection onto the leaf set $\mathcal{L} \subset [N]$ is the split $(P \cap \mathcal{L})|(P^c \cap \mathcal{L})$. Note that one of $P \cap \mathcal{L}$ or $P^c \cap \mathcal{L}$ may be empty, in which case the image is trivial. Since the tree dimensionality map $\Psi_{\mathcal{L}}$ operating on tree $T \in \mathcal{T}^N$ has the effect of projecting all splits $S = S(T)$ onto the leaf set $\mathcal{L}$, we will abuse notation and use $\Psi_{\mathcal{L}}(S)$ to represent this combinatorial projection.

The following result states that the dimensionality reduction will act on a tree naturally, when considered as an additive metric space.

Proposition 3.2.2 ([62, Proposition 4.4])**.** *Let T be a tree with leaf set $[N] = \{1, 2, \ldots, N\}$, and additive distance matrix A_T. Let $\mathcal{L} \subset [N]$, and define*

$(A_T)_{\mathcal{L}}$ to be the submatrix of A_T with rows and columns indexed by $\mathcal{L}$. Then $A_{\Psi_{\mathcal{L}}(T)} = (A_T)_{\mathcal{L}}$.

Note that Proposition 3.2.2 implies that if $\mathcal{L} \subset \mathcal{L}' \subset [N]$, then $\Psi_{\mathcal{L}} \circ \Psi_{\mathcal{L}'} = \Psi_{\mathcal{L}}$ on $\mathcal{T}^N$.

3.3 The Pre-Image of the Tree Dimensionality Reduction Map

The aim of this section will be to algorithmically construct the preimage of the tree dimensionality reduction map $\Psi_{\mathcal{L}} : \mathcal{T}^N \to \mathcal{T}^{\mathcal{L}}$, for $\mathcal{L} \subset [N]$, $|\mathcal{L}| = n$. We start with a binary tree $T \in \mathcal{T}^{\mathcal{L}}$ with edge lengths w_e for $e \in \mathcal{E}(T)$, and want to describe and compute the set of all trees $\bar{T} \in \mathcal{T}^N$ such that $\Psi_{\mathcal{L}}(\bar{T}) = T$. Since by Proposition 3.2.2 the distance of the leaves $N \backslash \mathcal{L}$ to each other and to the leaves $\mathcal{L}$ does not affect the distance between the leaves $\mathcal{L}$, many different tree topologies can map to T under $\Psi_{\mathcal{L}}$. Thus it is not immediately obvious how this set $\Psi_{\mathcal{L}}^{-1}(T)$ should be described.

As this section demonstrates, one effective approach, which we call the **extension algorithm**, is to:

1. Note that for any $\bar{T} \in \mathcal{T}^N$, the topology of the image $\Psi_{\mathcal{L}}(\bar{T})$ is completely determined by the topology of $\bar{T}$, and $\Psi_{\mathcal{L}}$ acts linearly on the $\mathcal{E}(\bar{T})$ edge weights in the orthant $\mathcal{O}(\bar{T})$ in $\mathcal{T}^N$. Thus, for a fixed maximal orthant of $\mathcal{T}^N$, $\Psi_{\mathcal{L}}$ restricts to a linear map $M : \mathbb{R}^{2N-3} \to \mathbb{R}^{2n-3}$. Any non-maximal orthant is on the boundary of at least three maximal orthants, and the

linear map of any of these maximal orthants can be used.

2. Find the orthants with a topology $\bar{T}$ such that $\Psi_{\mathcal{L}}(\bar{T})$ has the same topology as T. By Proposition 3.3.4, these orthants can be determined by individual and pairwise properties of their splits.

3. For a fixed orthant $\mathcal{O}$, form the matrix $M_T^{\mathcal{O}}$ which encodes the way the edges of trees in $\mathcal{O}$ concatenate under $\Psi_{\mathcal{L}}$.

4. Find the positive solutions of the linear system of equations $M_T^{\mathcal{O}}\mathbf{x}^{\mathcal{O}} = \mathbf{w}$, where $\mathbf{w}$ is the vector of edge weights in T, to determine the points $\bar{T} \sim \mathbf{x}^{\mathcal{O}} \in \mathcal{O}$ such that when $\Psi_{\mathcal{L}}$ is performed, all of the edges of $\bar{T} \in \mathcal{O}$ which concatenate to form an edge $e \in T$ have weights summing to w_e.

5. Take the union of all of the orthant-wise solutions, which we call the **extension space** E_T^N.

We will show that $E_T^N = \Psi_{\mathcal{L}}^{-1}(T) \subset \mathcal{T}^N$, and that the resulting space is connected, continuous, piecewise linear, of local dimension $2(N-n)$, and computable in cubic time relative to its size.

Note that we will assume that T is binary, since an unresolved tree is often used in biology when the underlying relationship of certain leaves or subtrees is not known. In such cases, the edge lengths near the unresolved vertex would not necessarily represent the expected length of their corresponding split in the true tree, which is the main assumption we are using. Thus we focus

on binary trees, and leave incorporating unresolved trees into this framework for future work.

3.3.1 Extension by one leaf

To give some intuition for how the extension space relates to the original tree, and to show the mechanics of the base case for later results, we first examine the case where $N = |\mathcal{L}| + 1$. That is, we want to find the set of trees $\Psi_{\mathcal{L}}^{-1}(T)$ which have one additional leaf, labeled g.

Definition 3.3.1. Let $\Psi_{\bar{g}} : \mathcal{T}^N \to \mathcal{T}^{N \backslash g}$ be the tree dimensionality reduction map which deletes leaf $g \in [N]$ and its adjacent edge, and concatenates the two edges at leaf g's attachment point. We will refer to this reduction as an g-**pruning**.

The reverse of pruning a leaf g is attaching a new leaf g to the tree with a new edge. We call this attachment operation grafting.

Definition 3.3.2. For a tree $T \in \mathcal{T}^{\mathcal{L}}$, the tree $\bar{T}$ is a g-**grafting** of T if $\mathcal{L}(\bar{T}) \backslash \mathcal{L}(T) = \{g\}$, and $\Psi_{\bar{g}}(\bar{T}) = T$.

In other words, a grafting of T consists of a tree identical to T, but with one additional leaf g and its leaf edge e_g. In considering the possibilities for such a grafting, there are two independent choices: the non-negative length of e_g, and a point on T at which to graft the non-leaf end. The next lemma shows the consequences of these two choices, and a bit more.

Lemma 3.3.1. *For tree $T \in \mathcal{T}^{\mathcal{L}}$ and leaf $g \notin \mathcal{L}$, the space of g-graftings of T, denoted $\Psi_{\bar{g}}^{-1}(T)$, is the direct product of $\mathbb{R}_{\geq 0}$ and a piecewise-linear connected curve which is graph-isomorphic to T and which intersects a strict subset of orthants each in a 1-dimensional linear curve.*

Proof. Consider any tree $T \in \mathcal{T}^{\mathcal{L}}$, leaf $g \notin \mathcal{L}$ and length $x \geq 0$. Recall that $\mathcal{E}(T)$ is the set of edges of tree $T \in \mathcal{T}^{\mathcal{L}}$, with each edge $e \in \mathcal{E}(T)$ having split P_e and length w_e.

We can attach a new edge e_g of length w_g ending in leaf g to any point, including an endpoint, on any edge of T to get a g-grafting of T. Thus the set of g-graftings of T, $\Psi_{\bar{g}}^{-1}(T)$, is not empty. For any $\bar{T} \in \Psi_{\bar{g}}^{-1}(T)$, its additive metric $A_{\bar{T}}$ restricted to the leaves $\mathcal{L}$ is just the additive metric of T, A_T. It follows $\bar{T}$ can be completely characterized by two independent choices: the choice of point on T for grafting, the space of which is graph-isomorphic to T, and a choice of length for the grafted leaf edge, which can be any non-negative real number.

Let $e \in \mathcal{E}(T)$ be the edge to which e_g, which has split $\bar{P}_g = g|\mathcal{L}$, will be grafted to form $\bar{T}$. If we are grafting g to a vertex of T, then choose e to be one of the edges adjacent to this vertex. For each edge $f \in \mathcal{E}(T)\backslash e$, the two partitions of the leaves in the corresponding split P_f induce two subtrees of T, and edge e is completely contained in one of these subtrees. Add leaf g to the partition of P_f corresponding to this subtree to get $\bar{P}_f$, the corresponding split in $\bar{T}$. The split P_e becomes the splits $\bar{P}_e{}^L = P_e|(P_e^c \cup g)$ and $\bar{P}_e{}^R = (P_e \cup g)|P_e^c$

in $\bar{T}$. If e_g was grafted to an endpoint of e, then one of $\bar{P_e}^L, \bar{P_e}^R$ will have zero weight, but we will still include it here as a split for consistency. Thus $\bar{T}$ has precisely the splits $\{\bar{P}_f : f \in \mathcal{E}(T)\backslash e\} \cup \bar{P}_g \cup \bar{P_e}^L \cup \bar{P_e}^R$.

For each edge $f \in \mathcal{E}\backslash e$, the weight of split $\bar{P}_f$ in $\bar{T}$ is the same as the weight of split P_f in T, since the edge corresponding to $\bar{P}_f$ projects to the edge corresponding to P_f without distortion. Thus, we will represent the weight of edge f in $\bar{T}$ by w_f as well. Split $\bar{P}_g$ has weight w_g, and let splits $\bar{P_e}^L$ and $\bar{P_e}^R$ have weights w_e^L and w_e^R, respectively. Then the space of all $\bar{T}$ formed by grafting leaf g to edge e is a two-parameter family satisfying $w_e = w_e^L + w_e^R$, and $w_g, w_e^L, w_e^R \geq 0$. Note that w_g is a free parameter, and $w_e = w_e^L + w_e^R$ is the equation of a line. Thus this solution space in this orthant is the direct product of $\mathbb{R}_{\geq 0}$ with the line that intersects the orthant boundaries at $w_e^L = 0, w_e^R = w_e$ and at $w_e^L = w_e, w_e^R = 0$.

It remains to show that the lines given by $w_e^L + w_e^R = w_e$ in each orthant are connected and graph isomorphic to tree T. Let e and e' be two adjacent edges in T, separated by vertex v. Edges e and e' are compatible because they exist in the same tree, and thus the intersection of one partition from each split is empty. Without loss of generality (by temporarily renaming the partitions if necessary), assume that $P_e \cap P_{e'} = \emptyset$. Then the case $w_e^L = w_e, w_e^R = 0$ corresponds to a tree with splits $\bar{P_e}^L = P_e|(P_e^c \cup g)$, with weight w_e, and $\bar{P_{e'}} = P_{e'}|(P_{e'}^c \cup g)$, with weight $w_{e'}$, as well as splits $\bar{P}_f$, with weight w_f, for all $f \in \mathcal{E}(T)\backslash\{e, e'\}$, and $\bar{P}_g$, with weight e_g. The case $w_{e'}^L = w_{e'}, w_{e'}^R = 0$ corresponds to a tree with splits $\bar{P_{e'}}^L = P_{e'}|(P_{e'}^c \cup g)$, with weight $w_{e'}$, and

$\bar{P}_e = P_e|(P_e^c \cup g)$, with weight w_e, as well as splits $\bar{P}_f$, with weight w_f, for all $f \in \mathcal{E}(T)\backslash\{e, e'\}$, and $\bar{P}_g$, with weight e_g. But these split and weight sets are identical, and thus the two line endpoints coincide. Since the two of these line segments meet if and only if they correspond to attaching leaf g to adjacent edges in e, we get that the piecewise-linear connected curve is graph-isomorhpic to T. □

Example 3.3.2. *Suppose we have a tree T with labels $\{1, 2, 3, 5\}$ as depicted in Figure 3.2, with leaf edges having length $\{0.15, 0.3, 0.2, 0.25\}$ respectively, and interior edge length 0.2. The corresponding additive distance matrix (indexed respectively) is given by*

$$A_T = \begin{pmatrix} 0 & .65 & .35 & .6 \\ .65 & 0 & .7 & .55 \\ .35 & .7 & 0 & .65 \\ .6 & .55 & .65 & 0 \end{pmatrix}$$

Then the preimage $\Psi_{\bar{4}}^{-1}(T)$ is the product of the subspace of $\mathfrak{T}^5$ depicted on the right in Figure 3.2 (with leaf edge length for $1, 2, 3, 5$ determined uniquely by the point on $\Psi_{\bar{4}}(T)$ below) and the copy of $\mathbb{R}_{\geq 0}$ (not shown) representing the "4"-leaf edge length. If we fix the length y of the 4 leaf, the $(4, y)$-grafting of T is the subspace shown by a thick line, together with unique local leaf coordinates

$$(w_1, w_2, w_3, w_4, w_5) = (0.15 - x_{(14)}, 0.3 - x_{(24)}, 0.2 - x_{(34)}, y, 0.25 - x_{(45)})$$

where $x_{(14)}, x_{(24)}, x_{(34)}, x_{(45)}$ are the weights of splits $(14), (24), (34), (45)$, respectively, if that split exists in the tree, and 0 otherwise.

Because Figure 3.2 omits the dimensions for the leaf edges, the four line segments corresponding to grafting g to a leaf edge appear to end mid-orthant.

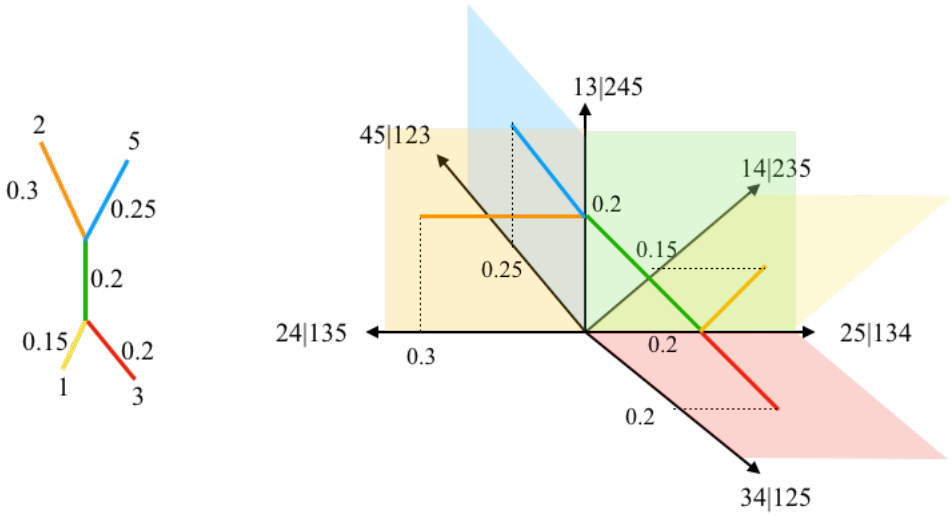


Figure 3.2: Left, a tree T with 4 leaves, $\{1, 2, 3, 5\}$. Right, the orthants of $\mathcal{T}^5$ containing the preimage $\Psi_4^{-1}(T)$, with the subspace corresponding to the preimage shown with the thick solid lines. Note that the dimensions corresponding to the 4 leaf edges lengths were not included for clarity.

In the full-dimensional space, the line segments end on boundaries where the respective leaf edge lengths are 0.

3.3.2 Extension by Multiple Leaves

As defined in [46], the **connection cluster** $C_{S(T),n,\ell}$ of a tree topology $S(T)$ on leaf set $[n] = \{1, 2, \ldots, n\}$ is the set of binary tree topologies with $n + \ell$ leaves obtained from adding ℓ leaves to arbitrary edges of T. We will generalize the definition of a connection cluster to allow the leaf set $\mathcal{L}$ of T to be any subset of $[N] = \{1, 2, ..., N\}$, and use the notation C_T^N, where $T \in \mathcal{T}^{\mathcal{L}}$ and $\mathcal{L} \subset [N]$. Throughout this section, we will still assume that $|\mathcal{L}| = n$, and $N = n + \ell$. The **connection space** $S_{S(T),n,\ell}$ in the notation of [46], or S_T^N

in our notation, is the union of the closed orthants in $\mathcal{T}^N$ that represent the elements of C_N^T, i.e. a non-negative real orthant for every unweighted tree in C_N^T under the normal identification of faces. The **connection graph** $G_{S(T),n,\ell}$, or with a change of notation, G_T^N, is the intersection of S_T^N with the link L_N^1, in which maximal cliques give elements of C_T^N. Ren et al. [46] and Lemma 3.3.5 below show that the edges of a connection graph are determined by normal pairwise compatibility of splits in $\mathcal{T}^N$, which allows for quick computation of C_N^T.

The connection space S_T^N can also be seen as the preimage in $\mathcal{T}^N$ under $\Psi_{\mathcal{L}}$ of the entire orthant represented by $S(T)$, namely $\Psi_{\mathcal{L}}^{-1}(\mathcal{O}(T))$. Similarly, the connection graph G_T^N is the corresponding preimage of the complete n-graph on $S(T)$. We are then interested in the subspace of S_T^N, restricted by the edge lengths of T, which projects under tree dimensionality reduction to T. This subspace will be a 2ℓ-dimensional linear submanifold supported in S_T^N. In other words, once the combinatorics of the extended trees are calculated through the connection cluster, we can use a set of $(2n-3)$ linear equations parametrized by the edge lengths in T to constrain sums of fixed edges in $\mathcal{T}^N$, and give the complete preimage $\Psi_{\mathcal{L}}^{-1}(T)$.

3.3.3 Calculating the Metric Extension Space

In this section we will construct, for phylogenetic tree $T \in \mathcal{T}^n$, the subset $E_T^N \subset S_N^T \subset \mathcal{T}^N$ which results from gluing ℓ leaves of arbitrary length to the metric tree T. The computation of the extension space E_T^N has two

steps:

The first step is the computation of S_T^N, via the method in [46] for constructing G_T^N and C_T^N. We will see that S_T^N is the preimage under $\Psi_{\mathcal{L}}$ of the orthant containing T.

The second step introduces the constraint that under the action of $\Psi_{\mathcal{L}}$ on S_T^N, the process of deleting and concatenating edge lengths as described in Definition 3.2.1 yields T precisely. To find the trees which satisfy this constraint, we solve a system of linear equations separately for each orthant in S_T^N.

3.3.3.1 Combinatorial Step

As in the previous section, we let $\{P_e\}_{e\in\mathcal{E}(T)}$ be the splits of T (including the leaf edges), with corresponding lengths $\{w_e\}_{e\in\mathcal{E}(T)}$. We will first state the algorithm for computing the connection cluster C_T^N and give an example, before proving correctness.

Algorithm 1 Computation of Connection Cluster

1: For each P_e, construct the set $\mathbf{Q}_e$ of splits projecting to P_e by adding the ℓ labels $N \backslash \mathcal{L}$ to P_e or P_e^c in all possible 2^ℓ ways.
2: Take the union $\mathbf{Q} = \cup_{e \in \mathcal{E}(T)} \mathbf{Q}_e$ to get the vertices of the connection graph G_T^N. Add an edge between each pair of vertices if and only if the two splits are compatible, which can be checked by the condition given in Definition 1.1.2.
3: Find all maximal $(n + \ell - 3)$ cliques in the subgraph of thick partitions, which is found by removing the leaf splits. Extend each maximal clique to include the leaf partitions, which are compatible with all other partitions, and return the corresponding set of cliques C_T^N.

Example 3.3.3. *Returning to the tree in Example 3.3.2, we find C_T^5 using Algorithm 1. The set of splits $S(T) = \{25|13, 1|235, 2|135, 3|125, 5|123\}$, so in Step 1, we find the set*

$$\mathbf{Q} = \{13|245, 25|134, 14|235, 24|125, 34|125, 45|123, 1|2345, 2|1345, 3|1245, 4|1235, 5|1234\}$$

In the second step, we form the graph G_T^5, which is shown in Figure 3.3.

In Step 3, we find maximal $(4 + 1 - 3)$-cliques in the thick subgraph. The 2-cliques are edges, and for each edge, add all of the leaf edges to obtain a unique topology of $\mathfrak{T}^5$. All such topologies form the connection cluster C_T^5. The orthants corresponding to these topologies are precisely those pictured in Example 3.3.2, and form S_T^5, the connection space, which is shown again in Figure 3.4 without the leaf dimensions.

The following proposition shows that the set of cliques returned in the final step of Algorithm 1 is indeed the connection cluster C_T^N, justifying the notation.

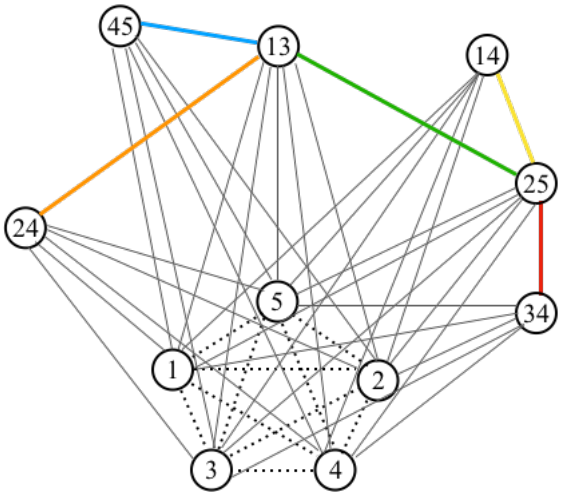


Figure 3.3: The connection graph G_T^5 for tree T from Example 3.3.2. The vertices corresponding to elements of $\mathbf{Q}$ are labeled by the smaller of the two pieces of the partition. The leaf partitions have automatic compatibility - these edges are shown dotted, while compatible thick partitions have colored edges.

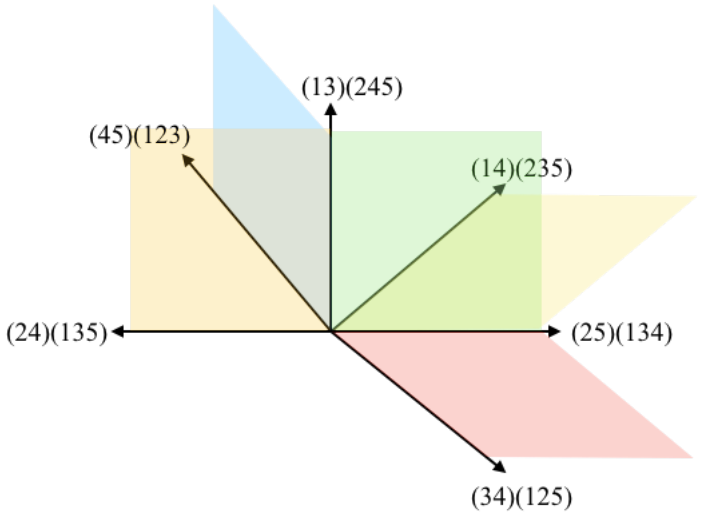


Figure 3.4: The connection space S_T^5 for tree T from Example 3.3.2.

Proposition 3.3.4. *For $T \in \mathcal{T}^{\mathcal{L}}$ with $\mathcal{L} \subset [N]$, Algorithm 1 returns the cliques C_T^N, which correspond to the orthant support of $\Psi_{\mathcal{L}}^{-1}(T) \subset \mathcal{T}^N$.*

Before proving Proposition 3.3.4, we show a preliminary result allowing us to reduce to conditions on the vertices of the extension graph.

Lemma 3.3.5. *For tree $T \in T^{\mathcal{L}}$ with $\mathcal{L} \subset [N]$, an orthant $\mathcal{O} \subset \mathcal{T}^N$ contains an element of $\Psi_{\mathcal{L}}^{-1}(T)$ if and only if $\Psi_{\mathcal{L}}(S(\mathcal{O})) = S(T)$. That is, $\mathcal{O}$ contains a tree in the extension space of T if and only if removing the labels $N \backslash \mathcal{L}$ from the splits $S(\mathcal{O})$ yields precisely the split set of T (with multiplicity).*

Proof. We proceed by induction on $\ell = |N \backslash \mathcal{L}|$.

If $\ell = 1$ and $\bar{T}$ is an extension of $T \in \mathcal{T}^{\mathcal{L}}$ by grafting leaf g to edge $e \in \mathcal{E}(T)$, then from the proof of Lemma 3.3.1, $\bar{T}$ has split set $S(\bar{T}) = \{\bar{P}_f : f \in \mathcal{E}(T) \backslash e\} \cup \bar{P}_g \cup \bar{P}_e{}^L \cup \bar{P}_e{}^R$. Recall that removing edge f from T induces two subtrees, the vertices of which become the two parts of splits P_f, and that $\bar{P}_f$ was constructed from P_f by adding leaf g to the partition corresponding to the subtree to which g was grafted. Thus $\bar{P}_f$ projects to P_f by construction for all f. Similarly, $\bar{P}_e{}^L$ and $\bar{P}_e{}^R$ were constructed such that they project unto P_e. Finally $\bar{P}_g$ projects onto a split with one partition empty, which we delete.

Conversely, if a set S of pairwise-compatible splits on $[N]$ projects to $S(T)$ under deletion of some leaf $g = N \backslash \mathcal{L}$, then we claim there exists a unique split $P|P^c \in S(T)$ which has two preimages. Suppose not. That is, suppose for $P|P^c$ and $Q|Q^c$ splits in T, the collective split preimages are $(P \cup g)|P^c$,

$P|(P^c \cup g)$, $(Q \cup g)|Q^c$, and $Q|(Q^c \cup g)$. Then compatibility of P and Q in T guarantees that precisely one of $Q \cap P, Q^c \cap P, Q \cap P^c, Q^c \cap P^c$ is empty, say without loss of generality $Q \cap P$. Then $(Q \cup g)|Q^c$ and $(P \cup g)|P^c$ are not compatible, because none of the four intersections of their partitions are empty. Thus S contains only one of them. So for any pair of splits in T, there are at most 3 preimage splits in S, and unique splits have distinct preimages, so we conclude that there is a unique split in T with both preimages, i.e. the set S must look precisely as above, $\{\bar{P}_f : f \in \mathcal{E}(T)\backslash e\} \cup \bar{P}_g \cup \bar{P}_e{}^L \cup \bar{P}_e{}^R$, and therefore we can construct $\bar{T} \in \Psi_{\mathcal{L}}^{-1}(T)$ uniquely by grafting the g-leaf edge to the middle of edge e.

So we have the result for the $\ell = 1$ case.

Then assume for induction that there exists $\bar{T} \in \mathcal{O} \subset \mathcal{T}^{n+\ell}$ such that $\Psi_{\mathcal{L}}(\bar{T}) = T$, if and only if $\Psi_{\mathcal{L}}(S(\mathcal{O})) = S(T)$. Then let $\mathcal{O}'$ be an orthant in $\mathcal{T}^{n+\ell+1}$. So then $\Psi_{n+\ell}(\mathcal{O}')$ is an orthant in $\mathcal{T}^{n+\ell}$, and applying the inductive hypothesis, there exists $\bar{T}' \in \Psi_{n+\ell}(\mathcal{O}')$ with $\Psi_{\mathcal{L}}(\bar{T}') = T$ if and only if $\Psi_{\mathcal{L}}(S(\Psi_{n+\ell}(\mathcal{O}'))) = S(T)$. Since $S(\Psi_{n+\ell}(\mathcal{O}')) = \Psi_{n+\ell}(S(\mathcal{O}'))$ from the one-step case, and $\Psi_{\mathcal{L}}(\Psi_{n+\ell}(S(\mathcal{O}'))) = \Psi_{\mathcal{L}}(S(\mathcal{O}'))$, giving us the forward direction. For the reverse direction, we know that $\bar{T}' \in \Psi_{n+\ell}(\mathcal{O}')$, which means that there is some tree $\bar{T} \in \mathcal{O}$ such that $\Psi_{n+\ell}(\bar{T}) = \bar{T}'$ by the base case. For $\bar{T}$ then, $\Psi_{\mathcal{L}}(\bar{T}) = \Psi_{\mathcal{L}}\Psi_{n+\ell}\bar{T} = \Psi_{\mathcal{L}}\bar{T}' = T$, and the proof is complete. □

Proof. (of Proposition 3.3.4) Suppose we have a maximal clique in G_N^T. Then this clique represents a set of pairwise compatible splits. Since L_n^1 is a flag

complex, these splits represents an orthant $\mathcal{O}$ in $\mathcal{T}^N$, of dimension corresponding to the size of the clique. By Lemma 3.3.5, these splits projects to the splits of T, so the orthant $\mathcal{O}$ contains elements of the extension space.

Conversely, suppose a tree $\bar{T}$ is in the extension space. Then by Lemma 3.3.5, the splits of $\bar{T}$ are among the vertex set of G_N^T, and since $\bar{T}$ is a tree in $\mathcal{T}^N$, its splits are compatible. Since compatibility is the condition for connectivity in G_N^T as well as L_n^1, $\bar{T}$ maps to a clique in G_N^T. □

Proposition 3.3.6. *The complexity of Algorithm 1 is* $O(2^{3\ell}n^3)$.

Proof. In the first step of the algorithm, we do a simple enumeration, with run time $(2n-3)2^\ell$. The second step of removing duplicates and initializing the graph is then $O(2^{2\ell}n^2)$, and to check compatibility is $O(2n-3+\ell)$ in each pair, so has $O(2^{2\ell}n^3)$. By [54], the run time of maximal clique enumeration is $O(|E|*|V|)$, and from [46] we have that the vertex set has size $2^\ell(2n-2)-\ell-n-1$, and the edge set size being at most the square of the size of the vertex set, we have a $O(2^{3\ell}n^3)$ run time for clique enumeration. Thus step 3 dominates the other steps, which gives the result. □

Note that while Algorithm 1 is fairly quick in n, it may be the case that we have small fragments of large trees, implying a very dominant ℓ term. In this case, Algorithm 1 is essentially reconstructing a large portion of $\mathcal{T}^{n+\ell}$, and so there is not much improvement which can be made, since the solution space itself is large. In the next section we will address a method for handling small tree fragments among a set of tree fragments.

3.3.3.2 Metric Step

Consider an orthant $\mathcal{O} \subset S_T^N \subset \mathcal{T}^N$, and index its corresponding splits by $Q_1, Q_2, \ldots, Q_{2N-3}$ (for example, in lexicographical order). By construction, $\Psi_{\mathcal{L}}(Q_j) = P_i$ for some $i \in \{1, \ldots, 2n-3\}$. We represent this assignment with a $(2n-3) \times (2N-3)$ **projection matrix** $M_T^{\mathcal{O}} = (m_{ij})$, where

$$m_{ij} = \begin{cases} 1 & \text{if } \Psi_{\mathcal{L}}(Q_j) = P_i \\ 0 & \text{otherwise} \end{cases}$$

Since $\Psi_{\mathcal{L}}$ is a well-defined map from $\{Q_j\}$ to $S(T) = \{P_i\}$, columns each have a unique non-zero entry. We then set up the real system of equations:

$$\begin{aligned} M_T^{\mathcal{O}} \cdot \mathbf{x}^{\mathcal{O}} &= \mathbf{w} \\ \mathbf{x}^{\mathcal{O}} &\geq 0 \end{aligned} \tag{3.1}$$

for $\mathbf{x}^{\mathcal{O}}$ the vector of non-negative edge weights in $\mathcal{O}$ (x_j the weight of split Q_j), and $\mathbf{w}$ the vector of edge weights in T.

Notice that (3.1) specifies, for each split P_i in T with weight w_i, the equation

$$x_{j_1} + x_{j_2} + \cdots + x_{j_{a_i}} = w_i$$

for $Q_{j_1}, \ldots, Q_{j_{a_i}} \in S(\mathcal{O})$ projecting to P_i, so that under tree dimensionality reduction $\Psi_{\mathcal{L}}$, the (non-negative) lengths of the edges $e'_{j_1}, e'_{j_2}, ..., e'_{j_{a_i}}$ of a tree in $\mathcal{O}$ concatenated to produce edge $e_i \in T$ sum precisely to w_i. So solving the system of equations in (3.1) finds vectors of possible edge lengths in tree topologies which project to T.

Definition 3.3.3. Given an orthant $\mathcal{O} \in S_T^N \in \mathcal{T}^{n+\ell}$, which, alternatively, has splits corresponding to a clique in G_T^N and a topology in C_T^N, we call the

set of $\mathbf{x}^{\mathcal{O}}$ satisfying (3.1) the **extension space of T in $\mathcal{O}$**, denoted $E_T^{\mathcal{O}}$. The **extension space of T in $\mathcal{T}^N$** is defined to be the union of extension spaces over all orthants in the connection space:

$$E_T^N := \bigcup_{\mathcal{O} \in S_T^N} E_T^{\mathcal{O}}.$$

Note that the image of $\mathbf{Q} = \{Q_1, \ldots, Q_{2N-3}\}$ under tree dimensionality reduction to $\mathcal{L}(T)$ gives a partition of the set into precisely $2n-3$ components, because $\Psi_{\mathcal{L}}(\mathbf{Q})$ is well-defined and surjective on P_i's. Because it is a partition and $w_i > 0$, we are guaranteed a solution of dimension $\sum_j m_{ij} - 1$ to (3.1), and a total solution space of dimension

$$\sum_{i=1}^{2n-3} \left(\left(\sum_{j=1}^{2N-3} m_{ij} \right) - 1 \right) = \sum_{j=1}^{2N-3} \sum_{i=1}^{2n-3} m_{ij} - (2n-3) = (2N-3) - (2n-3) = 2\ell.$$

The extension space E_T^N generalizes the single leaf extension case in that, after the equations are solved for all orthants, the result is the direct product of a piecewise-linear connected ℓ-manifold (intersecting a strict subset of orthants each in an ℓ-dimensional linear subspace), with $(\mathbb{R}_{\geq 0})^{\ell}$. Connectivity follows from the consideration that if two orthants share a k-dimensional face, then that face is represented as a k-clique in the connection graph, and the metric extension space meets the face in a set of equations of precisely the same sort on each side.

Proposition 3.3.7. *For leaf set $\mathcal{L} \subset [N]$, let $T \in \mathcal{T}^{\mathcal{L}}$ be a binary tree. The extension space of T, E_T^N, is connected. Furthermore, for adjacent orthants $\mathcal{O}_1, \mathcal{O}_2 \subset S_T^N$, $E_T^{\mathcal{O}_1 \cap \mathcal{O}_2} = E_T^{\mathcal{O}_1} \cap \mathcal{O}_2 = \mathcal{O}_1 \cap E_T^{\mathcal{O}_2}$.*

Proof. For each orthant $\mathcal{O} \subset S_T^N$, the extension space $E_T^{\mathcal{O}}$ is connected, since it is the solution of a linear system of equations, restricted to the non-negative orthant. Any two adjacent orthants $\mathcal{O}_1, \mathcal{O}_2 \subset S_T^N$ share some k-dimensional boundary orthant, which corresponds to a k-clique in the connection graph. Suppose the k splits in the clique are $Q_{j_1}, Q_{j_2}, \ldots, Q_{j_k}$. Then any solutions $\mathbf{x}^{\mathcal{O}_1}$, $\mathbf{x}^{\mathcal{O}_2}$ on the boundary only have non-zero weights for the splits $Q_{j_1}, Q_{j_2}, \ldots, Q_{j_k}$. Furthermore, since the projection of each Q_j onto a unique split P_i in $S(T)$ does not depend on the orthant, when we remove the 0 weights from each system of equations ($M_T^{\mathcal{O}_1} \cdot \mathbf{x}^{\mathcal{O}_1} = \mathbf{w}$ and $M_T^{\mathcal{O}_2} \cdot \mathbf{x}^{\mathcal{O}_2} = \mathbf{w}$), the two systems of equations will now be identical. Therefore the intersection of $E_T^{\mathcal{O}_1}$ and $E_T^{\mathcal{O}_2}$ is precisely each of their intersections with the boundary orthant $\mathcal{O}_1 \cap \mathcal{O}_2$. □

Example 3.3.8. *Returning to the tree T from Examples 3.3.2 and 3.3.3, based on the projection $\Psi_{\bar{4}}(Q_j)$ which deletes the label "4", we set up the following linear system.*

$$\begin{cases} x_{25|134} + x_{13|245} = 0.2 = w_{13|25} \\ x_{24|135} + x_{2|1345} = 0.3 = w_{2|135} \\ x_{45|123} + x_{5|1234} = 0.25 = w_{5|123} \\ x_{14|235} + x_{1|2345} = 0.15 = w_{1|235} \\ x_{34|125} + x_{3|1245} = 0.2 = w_{3|125} \\ x_j \geq 0 \quad \forall j \end{cases}$$

Without the leaf dimensions, the portion of the extension space pictured in Example 3.4 is specified by the first equation and the non-negative constraints.

We now show that the extension space E_T^N defined in Definition 3.3.3 is indeed the pre-image of the tree dimensionality reduction map $\Psi_{\mathcal{L}} : \mathcal{T}^N \to \mathcal{T}^{\mathcal{L}}$.

Theorem 3.3.1. *Let $\mathcal{L} \subset [N]$ and $T \in \mathcal{T}^{\mathcal{L}}$. Then $E_T^N = \Psi_{\mathcal{L}}^{-1}(T) \subset \mathcal{T}^N$.*

Proof. By construction and Proposition 3.3.4, $E_T^N \subset S_T^N$, so $\Psi_{\mathcal{L}}(S(\bar{T})) = S(T)$ for each $\bar{T} \in E_T^N$, i.e. E_T^N and $\Psi_{\mathcal{L}}^{-1}(T)$ intersect the same orthant set, given by S_T^N. Furthermore, the procedure of dimension reduction as given in Definition 3.2.1 guarantees that each edge $e_i \in \mathcal{E}(\Psi_{\mathcal{L}}(\bar{T}))$ will be obtained by concatenating edges $\bar{e_j}$ projecting to e_i. Thus, to satisfy $T = \Psi_{\mathcal{L}}(\bar{T})$, for a fixed orthant $\mathcal{O} \in S_T^N$, there is a fixed procedure of dimensionality reduction, and a fixed set of splits $\{Q_j\}$, each with weight $\bar{w_j}$, projecting to some $P_i \in S(T)$. Therefore $\Psi_{\mathcal{L}}(\bar{T}) = T$ is equivalent to having $\sum_{j:\Psi_{\mathcal{L}}(Q_j)=P_i} \bar{w_j} = w_i$ for each $e_i \in \mathcal{E}(T)$ with weight w_i, which is precisely the condition specified by the equations of $E_T^{\mathcal{O}}$. Since E_T^N and $\Psi_{\mathcal{L}}^{-1}(T)$ agree in each orthant, we have the result. □

Complexity of the Extension Algorithm

If we restrict our computation to a single orthant, the matrix $M_T^{\mathcal{O}}$ can be computed by calculating each $\Psi_{\mathcal{L}}(Q_j)$ and matching with P_i, which is $O(N)$. Each such computation determines a column of $M_T^{\mathcal{O}}$ (with unique non-zero entry in i-th position), so $M_T^{\mathcal{O}}$ is computed in $O(N^2)$.

The barrier to a polynomial time algorithm is the size of C_T^N, which by [46] is

$$\frac{(2(n+\ell)-5)!!}{(2n-5)!!} \in O(N^{\ell}).$$

These two estimates imply that computing all extension matrices is less than quadratic in the support size of the space.

Proposition 3.3.9. *The computation of the collection of matrices $M_T^{\mathcal{O}}$ is $O(N^{\ell+2})$, which dominates the complexity of the previous steps in the extension algorithm. Thus, the complexity of the extension algorithm is $O(N^{\ell+2})$.*

Proof. The complexity of $M_T^{\mathcal{O}}$ follows from the above observations. Combined with Proposition 3.3.6, the complete extension algorithm will be dominated by $N^{\ell+2} + 2^{3\ell}n^3$, and so we have the complexity bound given in the statement. For $\ell << n$ fixed, this is polynomial of degree $\ell + 2$. □

The actual space of solutions, a convex affine polytope, can be presented by its boundary vertices in each orthant; interior points can then be expressed as convex combinations of boundary vertices. These convex combinations can be computed, but there are a lot of them: since M is rank n, we expect around $\binom{N}{n}$ basic feasible solutions, which gives an estimate for boundary vertices. In low dimensions, enumeration might be reasonable; there exist algorithms to do this. In general, we will operate on this space in indirect ways.

Lemma 3.3.10. *Let binary tree $T \in \mathcal{T}^{\mathcal{L}}$ with $\mathcal{L} \subset [N]$, $|\mathcal{L}| = n$, and $|N \backslash \mathcal{L}| = \ell$. To test whether a point $\bar{x} \in \mathcal{T}^N$ is in E_T^N, it is sufficient to check whether $\Psi_{\mathcal{L}}(\bar{x}) = T$, which is $O(N)$.*

Proof. The first part is obvious from Theorem 3.2.1. For the complexity, we note that in order to check the latter condition, we must perform dimensionality reduction on $\bar{x}$, which can be done in $O(\ell)$ from the tree representation of

$\bar{x}$: each successive leaf removal results in at most one concatenation (see Definition 3.2.1). Then we must compare $\Psi_{\mathcal{L}}(\bar{x})$ to T. Since they are both binary trees in $\mathfrak{T}^{\mathcal{L}}$, they each have $2n-3$ splits and, as graphs, $2n-4$ vertices. We can therefore determine isometry by traversing the two trees simultaneously, starting at the same leaf, which is $O(n)$. Since $N > n, \ell$, we have the result, which is not tight. □

For the more general statement of Lemma 3.3.10, see Proposition 3.4.5.

Remark 3.3.1. To find a point $\bar{x}$ in $E_T^{\mathcal{O}}$ which optimizes a linear function $f(\bar{x})$ in orthant $\mathcal{O}$, standard linear programming methods will find a global solution in polynomial time, with an average runtime $\sim N^3 B$ using the simplex method. To estimate B, we note that matrices $M_T^{\mathcal{O}}$ will always be $(2n-3) \times (2N-3)$ (binary) matrices, with $2n-3$ edge lengths in floating point numbers, requiring a total of $O(Nn)$ bits, for a total average run time on the order of $N^4 n$.

3.3.4 Comparing extension spaces

One might hope that, as we have $d_{\mathcal{T}^{\mathcal{L}}}(\cdot,\cdot)$ which gives a well-defined metric on $\mathfrak{T}^{\mathcal{L}}$, we can use this metric to define a meaningful distance between $E_{T_1}^N$ and $E_{T_2}^N$ as sets. Though this calculation is possible, distances between the sets E_1 and E_2 in $\mathfrak{T}^N$ do not produce a metric on extension spaces.

Remark 3.3.2. The distance function $d_{E^N} : (E_T^N, E_{T'}^N) \mapsto \inf_{\bar{T}\in E_T^N, \bar{T}'\in E_{T'}^N} d_{\mathcal{T}^N}(\bar{T}, \bar{T}')$ is not a pseudometric. To see this, take two distinct points $\mathfrak{T}_1, \mathfrak{T}_2$ in a non-trivial extension space E; they are each trivial extensions of themselves, so they

are in the domain of the distance function, and there is a positive tree space distance $d_{\mathcal{T}^N}(\mathfrak{T}_1, \mathfrak{T}_2) = d_{E^N}(\mathfrak{T}_1, \mathfrak{T}_2)$. However, each have $\inf_{\bar{T} \in E}(\mathfrak{T}_i, \bar{T}) = 0$, $i = 1, 2$, so $\inf_{\bar{T} \in E}(\mathfrak{T}_1, \bar{T}) + \inf_{\bar{T} \in E}(\mathfrak{T}_2, \bar{T}) = 0$, which violates the triangle inequality. Furthermore, $d_{E^N}(E_1, E_2) = 0$ and $d_{E^N}(E_2, E_3) = 0$ do not imply $d_{E^N}(E_1, E_3) = 0$.

However, the vanishing of this quantity is meaningful, and corresponds to a "compatibility" of trees:

Lemma 3.3.11. *Let E_T^N and $E_{T'}^N$ be extension spaces of $T \in \mathcal{T}^{\mathcal{L}}$ and $T' \in \mathcal{T}^{\mathcal{L}'}$, respectively, where $\mathcal{L}, \mathcal{L}' \subseteq [N]$. Then $d_{E^N}(E_T^N, E_{T'}^N) = 0$ if and only if there exists a tree $\mathfrak{T} \in \mathcal{T}^N$ which contains all the splits of T and all the splits of T', with lengths as in T and T'.*

Proof. If distance is zero then they intersect, since extension spaces are locally affine. If they intersect, their intersection is non-empty, and we can choose a tree $\mathfrak{T}$ in this intersection. Then by Proposition 3.3.1, $\mathfrak{T}$ projects to each of T and T' under $\Psi_{\mathcal{L}(T)}$ and $\Psi_{\mathcal{L}(T')}$, and so $\mathfrak{T}$ contains a preimage of each split $P \in T, P' \in T'$, which separates the same leaves that P and P' do. Furthermore by previous results we know that the pairwise distances between leaves are preserved between T and $\mathfrak{T}$ (and T' and $\mathfrak{T}$). □

Then $\mathfrak{T}$ can be seen as combining the information of T and T', as in the case that T and T' are samples of a larger tree on different taxa subsets, and this $d_{E^N}(E_T^N, E_{T'}^N) = 0$ case (and later, $d_{E^N}(E_T^N, E_{T'}^N) < \epsilon$) is what we will explore in the next section.

3.4 Extension of tree sets

By Theorem 3.3.1, an intersection point of two extension spaces is an intersection of the preimages. In particular, if $\bar{T} \in \mathcal{T}^N$ is contained in $\Psi^{-1}_{\mathcal{L}(T)}(T)$ and $\Psi^{-1}_{\mathcal{L}(T')}(T')$, then by definition, $\Psi_{\mathcal{L}(T)}(\bar{T}) = T$ and $\Psi_{\mathcal{L}(T')}(\bar{T}) = T'$. Thus $\bar{T}$ can be seen as "combining" the information of two "compatible" trees, with different leaf sets $\mathcal{L}(T)$ and $\mathcal{L}(T')$.

Example 3.4.1. *Building on Example 3.3.2, suppose we have a second tree T' with labels $\mathcal{L}(T') = \{1, 2, 3, 4\}$, leaf edge lengths $(0.15, 0.35, 0.2, 0.35)$ respectively, and interior edge $13|24$ with length 0.15, pictured on the left in Figure 3.5. Then the preimage of T', shown in the center of Figure 3.5, under pruning of the 5th leaf is also a T'-shaped subspace of $\mathcal{T}^5$, and it intersects $\Psi^{-1}_{\bar{5}}(T)$ in a single point (circled), $(0.05, 0.15)$ in the $(13) - (25)$ plane (green), representing the tree pictured on the right in Figure 3.5, with leaf edges $(0.15, 0.3, 0.2, 0.35, 0.25)$, respectively. The combined information of these two trees can also be realized as the pairwise path distance matrix of $\bar{T}$, which contains the distance matrices for T and T' as distinct submatrices.*

$$A_{\bar{T}} = \begin{pmatrix} 0 & .65 & .35 & .65 & .6 \\ .65 & 0 & .7 & .7 & .55 \\ .35 & .7 & 0 & .7 & .65 \\ .65 & .7 & .7 & 0 & .65 \\ .6 & .55 & .65 & .65 & 0 \end{pmatrix}$$

In this section, we are interested in characterizing non-empty intersection points, and quickly computing the equations which define the complete

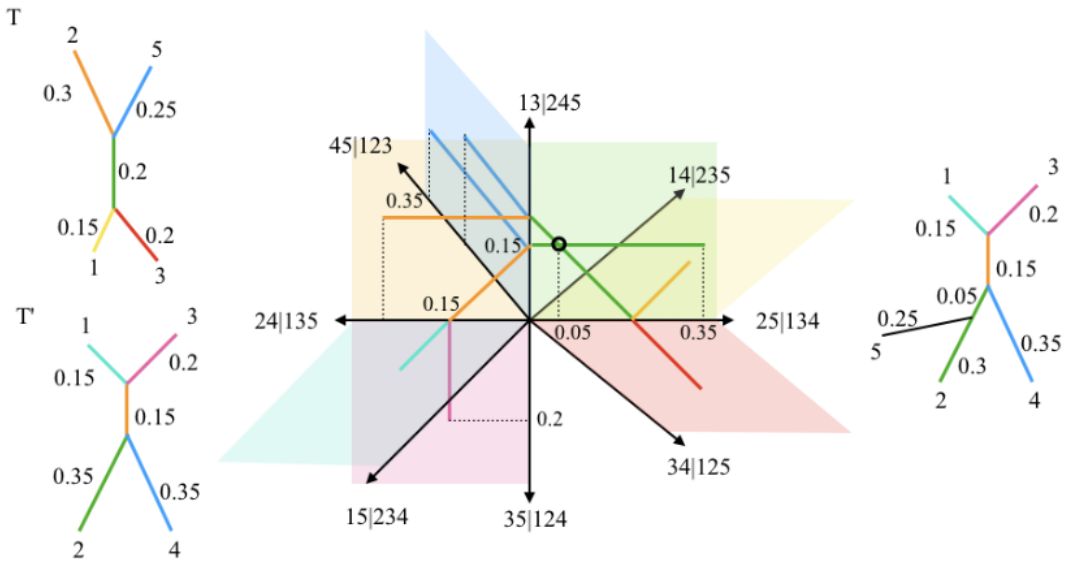


Figure 3.5: Left, tree T (repeated from Figure 3.2) and a second tree T' with leaves $\{1, 2, 3, 4\}$. Center, the T-shaped subspace of $\Psi_{\bar{5}}^{-1}(T)$ and the T'-shaped subspace of $\Psi_{\bar{5}}^{-1}(T')$, with their unique intersection circled. Right, the tree at the intersection point of the two subspaces.

set. More generally, consider a collection of trees $\mathbf{T} = \{T_1, \ldots, T_k\}$ with any leaf sets $\mathcal{L}_r$, where $|\mathcal{L}_r| = n_r$. By fixing $\ell_r = N - n_r$ we consider their tree dimensionality reduction preimages $\Psi_{\mathcal{L}_r}^{-1}(T_r)$ collectively in $\mathcal{T}^N$. We can now define generalizations of the

- connection cluster $C_{\mathbf{T}}^N := \cap_r C_{T_r}^N$,
- connection space $S_{\mathbf{T}}^N := \cap_r S_{T_r}^N$, and
- connection graph $G_{\mathbf{T}}^N := \cap_r G_{T_r}^N$.

These generalizations, $C_{\mathbf{T}}^N$, $S_{\mathbf{T}}^N$, and $G_{\mathbf{T}}^N$, correspond to the topologies in $\mathcal{T}^N$ which simultaneously extend $S(T_r)$ for all $T_r \in \mathbf{T}$.

As in Section 3, where $\mathbf{T} = \{T\}$, we will first find $C_{\mathbf{T}}^N$, and then find solutions to a system of metric constraints, which gives the **intersection extension space** $E_{\mathbf{T}}^N := \cap_r E_{T_i}^N$. However, due to the high codimension of $E_{T_r}^N$, the extension space of $\mathbf{T}$ can be unstable under small treespace perturbations of the T_r. In the next section, we will present two relaxations which will allow for bounded independent perturbations of $T_1, \ldots, T_k$, which produces a neighborhood of each $E_{T_r}^N$ for transverse intersection. These relaxations also give rise to two "measures of compatibility", $\alpha_{\mathbf{T}}$ and $p_{\mathbf{T}}$, the minimum parameter under two relaxation regimes giving a non-empty extension intersection. In the final section, we will discuss methods for consolidating more diverse tree topologies, which will choose orthants of highest likelihood for analysis.

We first give a few remarks on N. We are assuming that the data has consistently labeled trees - i.e. that label j represents the same sample across trees in $\mathbf{T}$. If the labels are numbers, we could take N equal to the maximum label, to represent missing taxa, but it might also make sense to take N equal to the number of different labels, which would simplify the solution space and decrease computation time, and add degrees of freedom later. Whatever N is chosen, we will assume that the label set $\mathcal{L}_r$ of T_r is a subset of $[N]$, and we will denote by $\Psi_{\mathcal{L}_r}$ the TDR projection map from $\mathcal{T}^N$ to $\mathcal{T}^{\mathcal{L}_r}$.

3.4.1 Combinatorial intersection

Given $T_1, \ldots, T_k$ binary trees with leaves $\mathcal{L}_r$ such that $L_r \subset [N]$ for each r, we can construct $G^N_{T_r}$ for each r, and take the intersection, to find tree topologies which project under $\Psi_{\mathcal{L}_r}$ to $S(T_r)$ for each r. However, if we are starting from the split sets $S(T_r)$, it is much more efficient to construct the intersection itself, since it can be much smaller than the largest $G^N_{T_r}$. The algorithm is as follows.

Algorithm 2 Computation of the combinatorial intersection

1: Reindex the trees so that T_1 has the greatest number of leaves n_1, and therefore the smallest ℓ_1. This step will ensure that we begin with the smallest connection graph.
2: Generate $G = G^N_{T_1}$.
3: For each $Q \in V(G)$, check if $\Psi_{\mathcal{L}_r}(Q) \in S(T_r)$ for all $r = 2, \ldots, k$. If not, remove Q from G, as well as all of its incident edges.
4: Find $(2N-3)$-cliques in G, output this set as $C^N_{\mathbf{T}}$.

Proposition 3.4.2. *Given* $\mathbf{T} = \{T_r\}$ *a finite set of binary trees, and* N *such*

that $\mathcal{L}_r \subset [N]$ *for each* r*, then* $G = \bigcap_r G^N_{T_r}$*, and therefore topology* $C \in C^N_{\mathbf{T}}$ *if and only if* $\Psi_{\mathcal{L}_r}(S(C)) = S(T_r)$ *for each* T_r*.*

Proof. By construction of the final graph G in Algorithm 2, $V(G)$ consists of splits Q_j such that $\Psi_{\mathcal{L}_r}(Q_j) \in S(T_r)$ for each r. This is the vertex set of $\cap_r G^N_{T_r}$, by construction. The edges of G, formed in Step 2 of Algorithm 2, come from pairwise compatibility, which is independent of the original tree set. We know also that compatibility determines adjacency equally for each $G^N_{T_r}$, so that the intersection of connection graphs is the full subgraph of the intersection of the vertex set in L^1_N, and any edge which is present in $G^N_{T_1}$ is present in all G_{T_r} containing both endpoints. Therefore all edges of $\cap_r G^N_{T_r}$ are present in Step 2, and none are deleted, since their endpoints remain. So $G = \cap_r G^N_{T_r}$.

We can also note that if K is a maximal $(2N-3)$-clique in G, then K is also a maximal clique in each $G^N_{T_r}$, and conversely, so that $C^N_{\mathbf{T}} = \cap_r C^N_{T_r}$.

Next, we note that by Proposition 3.3.4, topology $C \in C^N_{T_r}$ if and only if $\Psi_{\mathcal{L}_r}(S(C)) = S(T_r)$. Then since $C^N_{\mathbf{T}} = \cap_r C^N_{T_r}$, it follows that $C \in C^N_{\mathbf{T}}$ if and only if $\Psi_{\mathcal{L}_r}(N) = S(T_r)$ for each r. □

Definition 3.4.1. We call a set $\mathbf{T} = \{T_r\}$ of binary trees **combinatorially compatible** if $C^N_{\mathbf{T}} \neq \emptyset$.

Definition 3.4.1 relates to edge compatibility (Definition 1.1.2), but edge compatibility is not a special case of it. The requirement that the inputs be binary trees would need to be generalized.

Proposition 3.4.3. *If $N \cdot k < 2^{2\ell_1}$, then the complexity of Algorithm 2 is $O(2^{3\ell_1} n_1^3)$. If $N \cdot k > 2^{2\ell_1}$, then it is $O(2^{\ell_1} n_1^4 k^2)$. Either way, it is $O(2^{3\ell_1} n_1^4 k^2)$.*

Proof. Reindexing the trees to put the tree with the most leaves first is $O(k)$. By Proposition 3.3.6, we have that Step 2 is $O(2^{2\ell_1} n_1^3)$. For Step 3, we iterate through each of $\sim 2^{\ell_1} n_1$ vertices, and for each, delete leaves to get down to L_r (order N) and compare with the $2n_r - 3$ splits of T_r (order $n_r(2n_r - 3) \sim 2n_r^2 \lessapprox 2n_1^2$). In total, then, Step 3 is $O(2^{\ell_1} n_1^3 N k)$, and we can simplify to $O(2^{\ell_1} n_1^4 k^2)$ by noting that $N < k \cdot n_1$. For Step 4, in the worst case, the size of G is comparable to $G_{T_1}^N$, so by Proposition 3.3.6, Step 4 is $O(2^{3\ell_1} n_1^3)$. If $N \cdot k < 2^{2\ell_1}$, then Step 4 dominates. If not, Step 3 does. □

3.4.2 Metric intersection

Given a binary topology $C \in C_{\mathbf{T}}^N$with splits $Q_1, \ldots, Q_{N-3}$, plus leaf splits $Q_{N-2}, \ldots, Q_{2N-3}$, we have an $2\ell_r$-dimensional solution space for each T_r, cut out by a set of equations

$$x_{m_1} + x_{m_2} + \cdots + x_{m_{a_j}} = w_i$$

for each $P_i \in S(T_r)$, $i = 1, \ldots, 2n_r - 3$. The collection of equations from all T_r defines a solution space: either it is empty, or there is some linear subspace of solutions, with dimension at most $\min_r \ell_r$, which simultaneously satisfies the collection of metric constraints. Unlike the single-tree extension case, this system can be overdetermined, and have no solution in an orthant $\mathcal{O} \in S_{\mathbf{T}}^N$.

Definition 3.4.2. Let $\mathcal{O} \in S_{\mathbf{T}}^N$ be an orthant in the intersection cluster, with split lengths parametrized by respective coordinates $(x_1, \ldots, x_{2N-3})$. Let $M_{T_r}^{\mathcal{O}}$ be the $(2n_r - 3) \times (2N - 3)$ projection matrix of S to $\mathcal{T}^{\mathcal{L}_r}$. We then write

$$\begin{pmatrix} \frac{M_{T_1}^{\mathcal{O}}}{M_{T_2}^{\mathcal{O}}} \\ \hline \vdots \\ M_{T_k}^{\mathcal{O}} \end{pmatrix} \mathbf{x}^{\mathcal{O}} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_k \end{pmatrix}, \quad \mathbf{x}^{\mathcal{O}} \geq 0 \tag{3.2}$$

Then the solution space of $\mathbf{x}^{\mathcal{O}}$ satisfying (3.2) is denoted $E_{\mathbf{T}}^{\mathcal{O}}$. In (3.2), the matrix on the left is denoted $M_{\mathbf{T}}^{\mathcal{O}}$ for brevity, and the vector on the right hand side $\mathbf{w}_{\mathbf{T}}$, so expressing the equation more compactly, $M_{\mathbf{T}}^{\mathcal{O}} \mathbf{x}_{\mathcal{O}} = \mathbf{w}_{\mathbf{T}}$. The **intersection extension space** of a collection $\mathbf{T}$ of trees is defined to be

$$E_{\mathbf{T}}^N := \bigcup_{\mathcal{O} \in S_{\mathbf{T}}} E_{\mathbf{T}}^{\mathcal{O}},$$

where as before, N is taken to be the size of the total leaf set $L(\mathbf{T})$ and $\ell_r = N - n_r$ for $T_r \in \mathbf{T}$ of size n_r.

Note that when $\mathbf{T} = \{T\}$, $E_{\mathbf{T}}^N = T$, since N is set to $\mathcal{L}(T)$, unless we set a larger extension space, in which case $E_{\mathbf{T}}^N = E_T^N$, and so the results of Section 3 are a special case of Definition 3.4.2 and the algorithm for computing the intersection extension space.

Definition 3.4.3. Given a finite set of binary trees $\mathbf{T}$, we call the set **compatible** if $E_{\mathbf{T}} \neq \emptyset$.

Trivially, for $T \in \mathcal{T}^N$, $\Psi_{\mathcal{L}}(T)$ and $\Psi_{\mathcal{L}'}(T)$ are compatible for $\mathcal{L}, \mathcal{L}' \subset [N]$.

Proposition 3.4.4. *For a collection* $\mathbf{T}$ *of trees with total leaf set of size* N*, the intersection extension region of* $\mathbf{T}$ *is the intersection of the extension regions of* $T \in \mathbf{T}$*. That is,* $E^{\mathcal{O}}_{\mathbf{T}} = \bigcap_{T\in\mathbf{T}} E^{\mathcal{O}}_T$*,* $E^{N}_{\mathbf{T}} = \bigcap_{T\in\mathbf{T}} E^{N}_T$*.*

Proof. From Proposition 3.4.2, we know that the orthant support of the intersection is the intersection of the orthant supports. Thus

$$\bigcap_T E^N_T = \bigcap_T \bigcup_{\mathcal{O}} E^{\mathcal{O}}_T = \bigcup_{\mathcal{O}} \bigcap_T E^{\mathcal{O}}_T = \bigcup_{\mathcal{O}} E^{\mathcal{O}}_{\mathbf{T}},$$

where the first equality is by definition of the intersection extension space, the middle from finiteness of this union and intersection, and the last equality follows from the fact that the intersection of real linear varieties is the vanishing set of the collection of generating equations. □

Complexity of computing the intersection extension space

As in Section 3, we can quickly do the operations that size allows. For $\mathcal{C} = \max\{\sum_{T_r\in\mathbf{T}} 2n_r - 3, N\}$, equation (3.2) is a $\mathcal{C}$-dimensional system of equations which can be set up in $O(kN^2)$ time. As before, this solution space is cumbersome to describe enumeratively and quick to search.

Proposition 3.4.5. *Given* $\mathbf{T} = \{T_r\}_{r=1,\dots,k}$*,* $[N] = \cup L_r$*, and a tree* $T \in \mathfrak{T}^N$*, the decision problem "Is* T *in* $E^N_{\mathbf{T}}$*?" can be solved in* $O(kN)$ *time.*

Proof. To answer the decision problem, it suffices to check, for each $T_r \in \mathbf{T}$, if $\Psi_{L_r}(T) = T_r$. By Lemma 3.3.10, each can be done in $O(N)$ time, so the problem is $O(kN)$ time. □

$C^N_{\mathbf{T}}$ may be substantially smaller than $C^N_{T_1}$ (which is on the order of N^{ℓ_1}), so a complete description may be possible. A starting point is linear feasibility, i.e. determining if the system (3.2) has a solution, which, in contrast to the single-tree case, is not automatically true. To solve, we introduce $\mathcal{C}$ slack variables $\mathbf{y}_P$ and a ℓ_∞-norm variable α, and we minimize α subject to

$$\begin{aligned} \left(\begin{array}{cc} M^{\mathcal{O}}_{\mathbf{T}} & I \end{array} \right) \left(\frac{\mathbf{x}^{\mathcal{O}}}{\mathbf{y}_P} \right) &= \left(\begin{array}{c} \mathbf{w_T} \end{array} \right) \\ x_{m_a} &\geq 0 \\ \alpha \geq y_P &\geq 0 \end{aligned} \tag{3.3}$$

This LP has an initial feasible solution: $\mathbf{x}^{\mathcal{O}} = \mathbf{0}$, $\mathbf{y}_P = \mathbf{w_T}$, and $\min \alpha = 0$ if and only if there is an $\mathbf{x}^{\mathcal{O}}$ satisfying (3.2). This step takes as long as your favorite LP solver, for example the simplex method, which will have an average runtime of $O(\mathcal{C}^5)$. In the next section we will investigate the case $\min \alpha > 0$. For the LP formulation, skip to Section 3.5.1.1.

3.5 Relaxation

Since each $E^N_{T_r}$ (for collection $\{T_r\}$ as in Section 3.4 with fixed $N = n_r + \ell_r$) is locally a submanifold of codimension $2n_r - 3$ in each orthant, for $n_r + n_{r'} > N + 1$, two extension manifolds will not intersect stably. Thus, a small perturbation in two different projections of an N-tree may give the impression of subtree incompatibility, as illustrated in Example 3.5.1 below. In the language of our linear optimization problem (3.3), given a small amount of sampling error in compatible trees, we may obtain an approximate solution with small, but positive, objective value. To ensure stability of inter-

section, we find a minimum amount of error $\alpha_{\mathbf{T}}$, and find intersections of $\alpha_{\mathbf{T}}$-neighborhoods of the $E^N_{T_r}$ in each orthant.

Example 3.5.1. *Suppose tree T is as shown in Figure 3.5 and previous examples, and let T'' be tree T' in Figure 3.5 with the weight of the leaf 2 edge, $w_{2|134}$, being 0.3 instead of 0.35. Consider the 3-dimensional orthant $\mathcal{O}$ corresponding to splits* $13|245$, $25|134$, *and* $2|1345$. *Then the intersection of E^N_T with $\mathcal{O}$ is the solution to*

$$\begin{aligned} x_{13|245} &= 0.15 \\ x_{2|1345} + x_{25|134} &= 0.3 \end{aligned} \tag{3.4}$$

and the intersection of $E^N_{T''}$ with $\mathcal{O}$ is the solution to

$$\begin{aligned} x_{2|1345} &= 0.3 \\ x_{13|245} + x_{25|134} &= 0.2 \end{aligned} \tag{3.5}$$

However, there is no common solution to both (3.4) and (3.5), as shown in Figure 3.6. Thus the perturbation of the leaf 2 edge weight by 0.05 (or any other small amount) in tree T'' means the extension spaces E^N_T and $E^N_{T''}$ no longer intersect.

3.5.1 Uniform α-relaxation

We can uniformly expand a single orthant of extension region $E^{\mathcal{O}}_T$ by replacing each equation of the form

$$x_{m_1} + x_{m_2} + \cdots + x_{m_{a_j}} = w_i$$

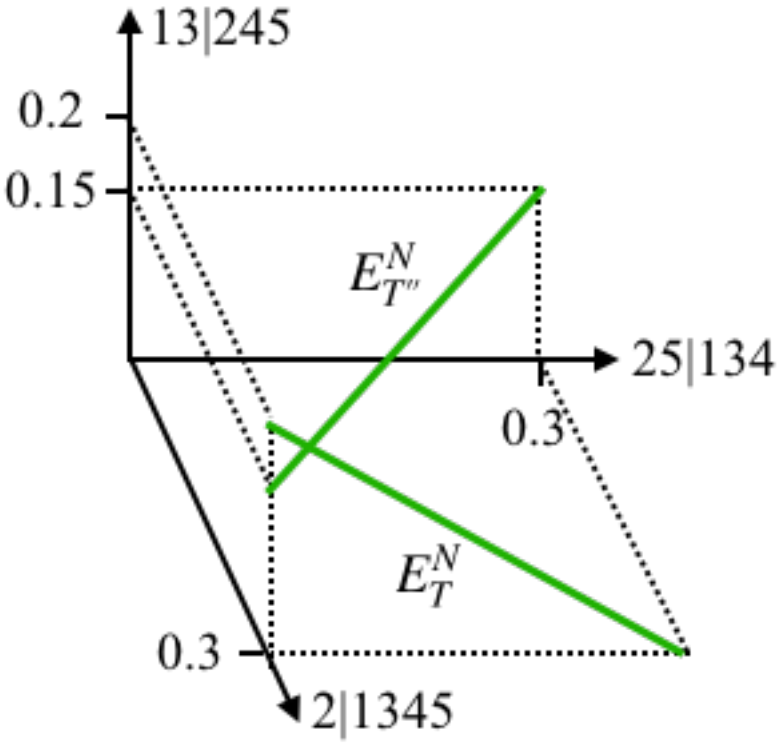


Figure 3.6: The extension spaces E^N_T and $E^N_{T''}$ from Example 3.5.1 intersected with the orthant corresponding to splits 13|245, 25|134, and 2|1345. Note that if the extension spaces are projected onto the 2-dimensional orthant corresponding to splits 13|245 and 25|134 they appear to intersect.

with a pair of equations of the form

$$x_{m_1} + x_{m_2} + \cdots + x_{m_{a_j}} \geq w_i - \alpha$$

$$x_{m_1} + x_{m_2} + \cdots + x_{m_{a_j}} \leq w_i + \alpha$$

Formally, we expand the equation (3.2) to the set of inequalities

$$\left(\begin{array}{c} \frac{M^{\mathcal{O}}_{T_1}}{M^{\mathcal{O}}_{T_2}} \\ \hline \vdots \\ M^{\mathcal{O}}_{T_k} \end{array}\right) \mathbf{x}^{\mathcal{O}} \geq \left(\begin{array}{c} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_k \end{array}\right) - \alpha \cdot \mathbf{1}, \quad \left(\begin{array}{c} \frac{M^{\mathcal{O}}_{T_1}}{M^{\mathcal{O}}_{T_2}} \\ \hline \vdots \\ M^{\mathcal{O}}_{T_k} \end{array}\right) \mathbf{x}^{\mathcal{O}} \leq \left(\begin{array}{c} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_k \end{array}\right) + \alpha \cdot \mathbf{1}, \quad \mathbf{x}^{\mathcal{O}} \geq 0 \tag{3.6}$$

For a single tree T_r, the solution space in a fixed orthant $\mathcal{O}$ is the extension space of a rectangular α-neighborhood of T_r in $\mathcal{T}^{\mathcal{L}_r}$, and we will see that it contains a neighborhood of the $2\ell_r$-plane $E^{\mathcal{O}}_T$ in $\mathcal{T}^N$. When $\alpha < w_i$ for all $P_i \in S(T)$, the solution space does not contain the cone point. The orthant solution space for $\mathbf{T}$ then becomes a (bounded or unbounded, empty or non-empty) polytope $E^{\mathcal{O}}_{\mathbf{T}}(\alpha)$. We choose α uniformly across orthants to ensure that the extension polytope is closed for small α.

Definition 3.5.1. For a given tree $T \in \mathcal{T}^{\mathcal{L}}$, define $E^N_T(\alpha) := \bigcup_{\mathcal{O} \in S^N_T} E^{\mathcal{O}}_T(\alpha)$ as the **α-extension region of T in $\mathcal{T}^N$.**

Example 3.5.2. *Let $\alpha = 0.05$, then the α-extension region of our first example is shown in Figure 3.7.*

Definition 3.5.2. For a finite collection $\mathbf{T} = \{T_r\}$ of binary trees and orthant $\mathcal{O} \in S^N_{\mathbf{T}}$, the α-relaxation of the equations (3.2) gives a (possibly empty)

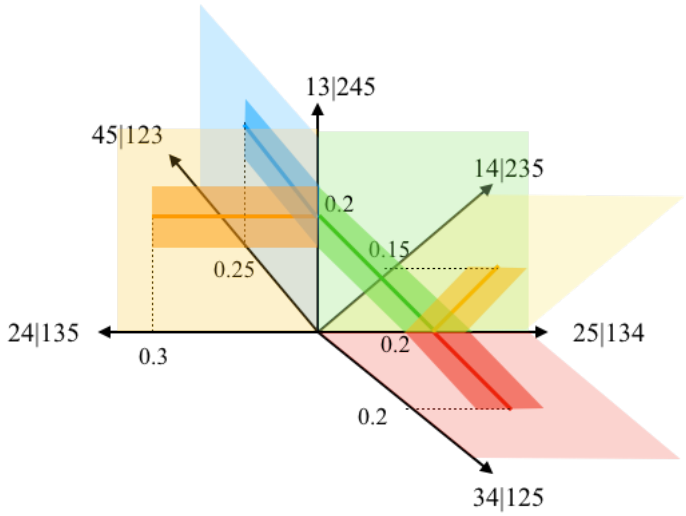


Figure 3.7: The α-extension region of tree T from Example 3.3.2 is the darker shaded region within the 5 orthants. Here $\alpha = 0.05$.

polytope in $\mathcal{O}$, denoted $E^{\mathcal{O}}_{\mathbf{T}}(\alpha)$. The **$\alpha$-intersection region** $E^{N}_{\mathbf{T}}(\alpha)$ of $\mathbf{T}$ is defined to be

$$E^{N}_{\mathbf{T}}(\alpha) := \bigcup_{\mathcal{O} \in S^{N}_{\mathbf{T}}} E^{\mathcal{O}}_{\mathbf{T}}(\alpha),$$

where as before, N is taken to be the size of the total leaf set $\cup \mathcal{L}_r$ and $\ell_r = N - n_r$.

Proposition 3.5.3. *Let binary tree $T \in \mathcal{T}^{\mathcal{L}}$ have leaf set $\mathcal{L} \subset [N]$. If tree $\mathfrak{T} \in E^{\mathcal{O}}_{\mathbf{T}}(\alpha)$, then $d_{\mathcal{T}^{\mathcal{L}}}(\Psi_{\mathcal{L}}(\mathfrak{T}), T) < c\alpha$ for all $T \in \mathbf{T}$, where c is a constant depending on $\mathcal{O}$ and $\mathcal{L}(T)$.*

Proof. If $\mathfrak{T} \in E^{\mathcal{O}}_{\mathbf{T}}(\alpha)$, then there is some $\mathfrak{T}' \in E^{\mathcal{O}}_{\mathbf{T}}$ such that $d(\mathfrak{T}, \mathfrak{T}') < \alpha$. Since $\mathfrak{T}' \in E^{\mathcal{O}}_{\mathbf{T}}$, we have that $\Psi_{\mathcal{L}}(\mathfrak{T}') = T$ for all $T \in \mathbf{T}$. So by Section 4.3 in [62], we can take $c = \log_2(N)$ to be the max number of edges concatenated in $\Psi_{\mathcal{L}}$

acting on $S(\mathcal{T}^N)$. $\square$

Note that $E_T^N(\alpha)$ is not defined as an α-neighborhood of E_T^N, but its restriction to each orthant in S_T^N is an α-neighborhood in that orthant. Furthermore, for small α, $E_T^N(\alpha)$ is closely related to the neighborhood.

Proposition 3.5.4. *Let $T \in \mathcal{T}^{\mathcal{L}}$ be a binary tree with leaf set $\mathcal{L} \subset [N]$. For $\alpha < \log_2(N)^{-1} \min_{e \in \mathcal{E}(T)} w_e$, $E_T^N(\alpha)$ contains the α-neighborhood of E_T^N in $\mathcal{T}^N$.*

Proof. The α-neighborhood $N_\alpha := N_\alpha(E_T^N) \subset \mathcal{T}^N$ is path-connected. Suppose $\mathfrak{T} \in N_\alpha \backslash E_T^N(\alpha)$. Since $N_\alpha \cap \mathcal{O} = E_T^N(\alpha)$ for $\mathcal{O} \subset S_T^N$, we conclude that $\mathfrak{T} \notin \mathcal{O}$ for any orthant of the connection space, so the orthant $\mathcal{O}'$ containing $\mathfrak{T}$ does not contain a preimage of some edge $e' \in \mathcal{E}(T)$, i.e. $e' \notin \Psi_{\mathcal{L}}(S(\mathcal{O}))$. Since the neighborhood is path-connected though, between $\mathfrak{T}$ and E_T^N there is some geodesic path γ contained in N_α, corresponding to a deformation of $\mathfrak{T}$ to some tree $\bar{T} \in E_T^N$.

Consider the image of γ under $\Psi_{\mathcal{L}}$. $\Psi_{\mathcal{L}}(\mathfrak{T})$ does not have edge e', so $\Psi_{\mathcal{L}}(\gamma)$ must have length at least the length of the projection to e'. Therefore the length of $\Psi_{\mathcal{L}}(\gamma)$ must be greater than α since the e' component of the path has length at least $w_{e'} \geq \min_e w_e > \log_2(N)\alpha$. By [61] geodesic lengths grow by at most $\log_2(N)$ under $\Psi_{\mathcal{L}}$, which implies $\mathfrak{T} \notin N_\alpha$, a contradiction. $\square$

Lemma 3.5.5. *Let $\mathbf{T} = \{T_r\}$ be a finite set of binary trees in $\mathcal{T}^N$, each with leaf set $\mathcal{L}(T_r) \subset [N]$. If $\alpha_1 < \alpha_2$, then $E_{\mathbf{T}}^N(\alpha_1) \subset E_{\mathbf{T}}^N(\alpha_2)$. For $T, T' \in \mathcal{T}^{\mathcal{L}}$ with $d_{\mathcal{T}^{\mathcal{L}}}(T, T') < \min\{\alpha, \log_2(N)^{-1} \min_{e_j \in T} w_j\}$, we have the inclusion $E_{T'}^N \subset E_T^N(\alpha)$.*

Proof. The first statement is clear from construction. For the second, if $d_{\mathcal{TL}}(T, T') < \min w_j / \log(N)$, then we have that T' has the same split set as T, with w'_j the corresponding lengths. Each $w'_j < w_j + d_{\mathcal{TL}}(T, T') < w_j + \alpha$, and similarly $w'_j > w_j - d_{\mathcal{TL}}(T, T') > w_j - \alpha$, so solutions to $x_{m_1} + x_{m_2} + \cdots + x_{m_{a_j}} = w'_j$ satisfy both inequalities. □

Definition 3.5.3. For a finite, combinatorially compatible collection $\mathbf{T}$ of trees $\{T_r\}$, each with leaf set $\mathcal{L}(T_r) \subset [N]$, and a given orthant $\mathcal{O} \in S^N_{\mathbf{T}}$, we denote by $\alpha^{\mathcal{O}}_{\mathbf{T}}$ the infimum of α such that $E^{\mathcal{O}}_{\mathbf{T}}(\alpha)$ is non-empty. Then the **intersection parameter** $\alpha_{\mathbf{T}} := \min_{\mathcal{O}} \alpha^{\mathcal{O}}_{\mathbf{T}}$.

If $\mathbf{T}$ can be obtained from a single N-tree by deleting subsets of the leaves, then $\alpha_{\mathbf{T}} = 0$. We also have a natural upper bound on $\alpha^{\mathcal{O}}_{\mathbf{T}}$ given by the length of the longest edge in $\mathbf{T}$ (so that $E^N_{\mathbf{T}}(\alpha)$ contains all $E^N_T(\alpha)$), so $\alpha^{\mathcal{O}}_{\mathbf{T}}$ is guaranteed to be finite. The parameter $\alpha_{\mathbf{T}}$ represents minimum amount the preimages of the trees T_r must be perturbed to have a metric solution, assuming combinatorial compatibility.

3.5.1.1 Computing $\alpha_{\mathbf{T}}$

When the system of equations (3.3) has a non-zero optimal solution, we conclude that (3.2) had no solutions in that orthant, but we also obtain a valuable by-product: a measure of the degree to which the extension spaces $E^N_{T_r}$ miss each other. For a solution $\mathbf{x}^{\mathcal{O}}, \mathbf{y}_P$ to (3.3), for each $r = 1, \ldots, k$ we have a unique subset $(\mathbf{y}_P)_r \subset \mathbf{y}_P$, satisfying only the system of equations

corresponding to the $M^{\mathcal{O}}_{T_r}$ rows of $M^{\mathcal{O}}_{\mathbf{T}}$. Rearranging those rows,

$$(\mathbf{y}_P)_r = \mathbf{w}_r - M^{\mathcal{O}}_{T_r}\mathbf{x}^{\mathcal{O}} \tag{3.7}$$

Thus the $(\mathbf{y}_P)_r$ can be viewed as representing the edge lengths of a positive "error tree" in orthant $\mathcal{O}$ of $\mathcal{T}^{\mathcal{L}_r}$, and the maximum entry in $(\mathbf{y}_P)_r$ is the minimum amount of ℓ_∞ error between T_r and a tree satisfying the T_r rows of equation (3.2). Then a global solution is the minimum ℓ_∞ error which must be tolerated to include all $T_r \in \mathbf{T}$.

To make this argument precise, we must add another relaxation variable to stretch $E^N_{T_i}$ to include larger trees as well as smaller ones.

Proposition 3.5.6. *The uniform relaxation parameter $\alpha^{\mathcal{O}}_{\mathbf{T}}$ of a tree set $\mathbf{T}$ in orthant $\mathcal{O} \in S^N_{\mathbf{T}}$ is equal to the objective value of the linear program*

$$\begin{array}{ll} \textit{minimize} & \alpha \\ \textit{s.t.} & \left(\begin{array}{ccc} M^{\mathcal{O}}_{\mathbf{T}} & I & -I \end{array} \right) \left(\begin{array}{c} \mathbf{x}^{\mathcal{O}} \\ \hline \mathbf{y}_P \\ \hline \mathbf{y}_N \end{array} \right) = \left(\begin{array}{c} \mathbf{w_T} \end{array} \right) \\ & 0 \leq x_{m_a} \\ & 0 \leq y_{P,m}, y_{N,m} \leq \alpha \end{array} \tag{3.8}$$

To use the intrinsic BHV metric, which is piecewise ℓ_2, we could use the objective function $\min \sum y^2_{P,m} + \sum y^2_{N,m}$, or in order to preserve linearity of the objective function, we can use the ℓ_1 metric in tree space, minimizing $\sum y_{P,m} + \sum y_{N,m}$.

Regarding the complexity, if $\mathcal{C} = \max\{\sum_{T_r \in \mathbf{T}} 2n_r - 3, N\}$, then each matrix has $\sim \mathcal{C}^2$ entries, so the simplex algorithm will run in $(\mathcal{C}^5)$ time on

average, although this is emphatically not a worst-case estimate. This step will solve $\alpha_{\mathbf{T}}$, but again we may not want to enumerate the boundary points.

3.5.1.2 Computing $E^N_{\mathbf{T}}(\alpha)$

Using $M^{\mathcal{O}}_{\mathbf{T}}$, $\mathbf{w_T}$, $\mathbf{x}^{\mathcal{O}}$, $\mathbf{y}_P$, $\mathbf{y}_N$ as defined previously, $\mathcal{O} \in S^N_{\mathbf{T}}$ and $\alpha \geq \alpha^S_{\mathbf{T}}$, the α-relaxed extension space of $\mathbf{T}$ is defined by the equation

$$\left(\begin{array}{ccc} M^{\mathcal{O}}_{\mathbf{T}} & I & 0 \\ M^{\mathcal{O}}_{\mathbf{T}} & 0 & -I \end{array} \right) \left(\begin{array}{c} \mathbf{x}^{\mathcal{O}} \\ \hline \mathbf{y}_P \\ \hline \mathbf{y}_N \end{array} \right) = \left(\begin{array}{c} \mathbf{w_T} + \alpha \\ \mathbf{w_T} - \alpha \end{array} \right) . \qquad (3.9)$$
$$x_{m_a}, y_{m,P}, y_{m,N} \geq 0$$

We can use this description to search $E^N_{\mathbf{T}}(\alpha)$ for optimal solutions to a linear function (i.e. a function on $\mathcal{T}^N$ whose restriction to orthants is linear, or a linear function supported in a limited number of orthants).

3.5.2 Proportional relaxation

The α-extension region, which is closely related to the α neighborhood of $E_{\mathbf{T}}$ for small α (Proposition 3.5.4), is a natural choice for relaxation, but we can also choose a neighborhood proportional to the extension region by solving the inequalities

$$M^{\mathcal{O}}_{\mathbf{T}}\mathbf{x}^{\mathcal{O}} \geq (1 - p_\alpha)\mathbf{w_T}, \quad M^{\mathcal{O}}_{\mathbf{T}}\mathbf{x}^{\mathcal{O}} \leq (1 + p_\alpha)\mathbf{w_T}, \quad \mathbf{x}^{\mathcal{O}} \geq 0 \qquad (3.10)$$

Definition 3.5.4. Let $\mathbf{T} = \{T_r\}$ be a finite set of binary trees, $\mathcal{L}_r \subset [N]$, $C^N_{\mathbf{T}}$ nonempty, and let $\mathcal{O} \in S^N_{\mathbf{T}}$. Then for a fixed $p_\alpha \in [0, 1]$, the non-negative

solutions to (3.10) in $\mathbb{R}^N_{\geq 0}$ give a $(2N-3)$-dimensional solution space in $\mathcal{O}$; the polytope generated with such a p_α is denoted $E^{\mathcal{O}}_{\mathbf{T}}(p_\alpha)_p$, with corresponding (p_α)**-proportional extension region**

$$E^N_{\mathbf{T}}(p_\alpha)_p = \bigcup_{\mathcal{O}\in S^N_{\mathbf{T}}} E^{\mathcal{O}}_{\mathbf{T}}(p_\alpha)_p.$$

Then define the **proportional intersection parameter**

$$p_{\mathbf{T}} = \inf_{E^N_{\mathbf{T}}(p_\alpha)_p \neq \emptyset} p_\alpha$$

Proposition 3.5.7. *The proportional intersection parameter* $p_{\mathbf{T}} \in [0,1]$. *For each* $\mathcal{O} \in S^N_{\mathbf{T}}$, *set*

$$p^{\mathcal{O}}_{\mathbf{T}} := \inf_{E^{\mathcal{O}}_{\mathbf{T}}(p_\alpha)_p \neq \emptyset} p_\alpha.$$

Then for $p_\alpha < 1$, $p_{\mathbf{T}} = \min_{\mathcal{O}} p^{\mathcal{O}}_{\mathbf{T}}$.

Proof. For $p_\alpha < 0$, $1 - p_\alpha > 1 + p_\alpha$, so the system (3.10) has no solutions. Thus $E^{\mathcal{O}}_{\mathbf{T}}(p_\alpha)_p = \emptyset$ for all $\mathcal{O}$, which implies $p^{\mathcal{O}}_{\mathbf{T}} \geq 0$.

For $p_\alpha > 1$, $1 - p_\alpha < 0$, so $\mathbf{x}^{\mathcal{O}} = \mathbf{0}$ is a solution to (3.10). Since $\mathbf{0}$ is identified in each orthant, $E^N_{\mathbf{T}}(p_\alpha)_p$ is formally non-empty. Thus $p_{\mathbf{T}} \leq 1$, and for $p_{\mathbf{T}} < 1$, the cone point is not in $E^N_{\mathbf{T}}(p_\alpha)_p$. In this case, since $E^N_{\mathbf{T}}(\cdot)_p = \cup_{\mathcal{O}} E^{\mathcal{O}}_{\mathbf{T}}(\cdot)_p$, $E^N_{\mathbf{T}}(\cdot)_p$ is nonempty precisely when one of $E^{\mathcal{O}}_{\mathbf{T}}(\cdot)_p$ is non-empty, which occurs at $\min_{\mathcal{O}} p^{\mathcal{O}}_\alpha$, showing equality with $p_{\mathbf{T}}$. □

Note that as with the uniform parameter α, the p_α case gives the original extension regions, but unlike the α case, p_α has a maximum, 1, which includes boundaries of each orthant, including the cone point. Thus we are guar-

anteed non-empty relaxed intersection extension region for some value of p_α. Also, for $\alpha < \frac{p_\alpha}{\log_2(N)} \cdot \min_{e \in \mathbf{T}} w_e$, by Proposition 3.5.4 $N_\alpha \subset E^N_{\mathbf{T}}(\alpha) \subset E^N_{\mathbf{T}}(p_\alpha)_p$.

We are also led to a slightly different notion of stability, or alternately, the condition on the following lemma can be strengthened to $d_{\mathfrak{T}^{\mathcal{L}}}(T, T') < \min_{e \in \mathcal{E}(T)} p_\alpha \cdot w_e$ to obtain the same inclusion.

Lemma 3.5.8. *For any $N \in \mathbb{N}$ with leaf set $\mathcal{L} \subset N$, let $T, T' \in \mathcal{O} \in \mathfrak{T}^{\mathcal{L}}$, and let $p_\alpha \in [0, 1)$. If $|w_e - w'_e| < p_\alpha w_e$ for each $e \in \mathcal{E}(T)$, then $E^N_{T'} \subset E^N_T(p_\alpha)_p$ for any extension codomain $\mathfrak{T}^N$.*

Proof. Similar to the proof of Lemma 3.5.5, we can easily see that solutions to equations for $E^N_{T'}$ satisfy the inequalities defining $E^N_T(p_\alpha)$. □

Proposition 3.5.9. *The proportional relaxation parameter $(p_\alpha)^{\mathcal{O}}_{\mathbf{T}}$ of a tree set $\mathbf{T}$ in orthant $\mathcal{O} \in S^N_{\mathbf{T}}$ is equal to the objective value of the linear program*

$$\begin{array}{ll} \textit{minimize} & p_\alpha \\ \textit{s.t.} & \left(\begin{array}{ccc} M^{\mathcal{O}}_{\mathbf{T}} & I & -I \end{array} \right) \left(\begin{array}{c} \mathbf{x}^{\mathcal{O}} \\ \hline \mathbf{y}_P \\ \hline \mathbf{y}_N \end{array} \right) = \left(\begin{array}{c} \mathbf{w_T} \end{array} \right) \\ & 0 \leq x_{m_a}, y_{P,m}, y_{N,m} \\ & 0 \leq p_\alpha \cdot w_m - y_{P,m} \\ & 0 \leq p_\alpha \cdot w_m - y_{N,m} \end{array} \tag{3.11}$$

Chapter 4

Manifold Learning and Dimensionality Reduction for Non-trivial Topology

In this chapter we give exposition of some techniques in manifold learning, and outline three new heuristic methods currently in development for preserving various topological features in the process. The main output will be a set of non-linear projections for the manifold depending on the local distributions of data - after fitting a mixture of locally flat models, we group the local subspaces based on topological data and align each in low-dimensional Euclidean space or on a sphere.

4.1 Introduction

Given a set of sample points $\mathbf{Y} = (y_1, \ldots, y_N) \subset \mathbb{R}^n$, and a suspicion that they may lie on or near some lower-dimensional embedded manifold $\mathcal{M} \subset \mathbb{R}^n$, manifold learning either attempts to construct a non-linear dimensionality reduction map (NLDR), or to provide a description of the best-fit manifold for the data, freely or within a parametrized family. Manifold learning is a very active and applied area, but most techniques assume that the manifold in question is contractible or relatively flat, often trying to find the

best representation in $\mathbb{R}^2$ or $\mathbb{R}^3$ independent of structure.

There are some notable exceptions. Riemannian manifold learning [37], for example, embeds a base tangent plane isometrically and extends iteratively outward, minimizing distortion to angle and geodesic length. This can be done locally at points of interest or globally at the centroid.

By fitting not just the manifold, but the tangent bundle, we move toward additional geometric structure. If $\mathcal{M}$ is *flat*, meaning with zero curvature at every point, then a map from the tangent bundle to $\mathbb{R}^d$ gives the unique flat connection, meaning parallel transport along curves is path-independent and given by translation of the vector in $\mathbb{R}^d$. Tangent space alignment [63] uses local PCA and a technique similar to the least squares method of Section 4.5.1 to align local frames in $\mathbb{R}^d$. This also works to preserve local geometry and global structure, although [63], like manifold charting [13], assembles a single flat chart. This works best if $\mathcal{M}$ is close to a compact subset of a linear affine subspace in $\mathbb{R}^n$.

Recently, Scoccola and Perea have developed a technique of approximating Euclidean vector bundles using nearest-neighbors PCA and the orthogonal Procrustes alignment between pairs of approximate tangent spaces [49]. This allows them to specify an orthogonal structure group, and to go on to define approximate cocycle conditions, estimates of characteristic classes, and a reconstruction theorem that allows for precise guarantees on homotopy equivalence.

In all of these techniques there is a tradeoff between topological fidelity and speed of computation[1], and we aim toward bridging the gap.

Our approach is to extend the least squares alignment of [13] to the case where $\mathcal{M}$ is

- a sphere
- non-contractible, with high reach and bounded curvature
- a union of contractible manifolds, not necessarily disjoint, of possibly mixed dimension, intersecting transversely.

Using the same flat tangent space alignment optimization, we patch together the local linear subspace arrangements resulting from the symmetric block decomposition of Kileel and Pereira [33]. Following their GPCA algorithm, we decompose the $2n$-th data moment, where $n \geq 2$, to robustly approximate the local structure, which may be a transverse intersection of tangent spaces of various dimension.

We also generalize the alignment algorithm to optimize connections on a sphere of dimension d. A theorem of Kobayashi states that the Levi-Civita connection on a smooth surface is the pullback under the Gauss map of the Levi-Civita connection on the sphere. In projecting to the sphere with

[1]For example, the Niyogi-Smale-Weinberger result guaranteeing homotopy equivalence of the Čech complex requires quite high amounts of samples, which then create an intractible load on the already computationally intensive persistent-homology algorithms.[43]

minimal distortion, we construct a discrete dimension-reduced approximation of the Gauss map, which will give us a dimension-reduced approximation of the Levi-Civita connection on an unknown manifold, which may very well have torsion/holonomy.

Our procedure assumes that Y lies on some unknown compact manifold $\mathcal{M} \subset \mathbb{R}^n$, with bounded reach and curvature. We will construct a (probabilistic) open cover consisting of ellipsoidal distributions in $\mathbb{R}^n$, together with projection maps to local coordinates $U_j \subset \mathbb{R}^d$, and find a small set of relatively flat charts to cover $\mathcal{M}$, which can be quickly aligned in $\mathbb{R}^d$.

1. (Section 4.3.3) Estimate the intrinsic dimension[s] d of the data locally.

2. Estimate the embedded tangent bundle structure with a Gaussian mixture model $\{\pi_j, \mathcal{N}(\mu_j, \Sigma_j)\}_{j=1,\dots,k}$. Sections 4.2 and 4.3 give different methods, for manifolds and stratified/mixed manifolds, respectively, or any GMM approximation will suffice. In either case, we take the d principal components of the local model to represent the tangent plane $T_{\mu_j}\mathcal{M}$.

3. Instead of set inclusion determining membership $y_i \in U_j$, we compute stochastic membership weights from the density functions of our cover. This is a significant relaxation of the notion of an open set which accounts for both off-manifold and on-manifold noise. To each point $y \in \mathbf{Y}$, calculate p_{ky} recording the relative likelihood that y belongs to chart k. This is either the normalization of a vector of density functions $f_k(y)$ for

each Gaussian, or we may use projected distances to various spaces in the subspace arrangement.

4. (Section 4.4) We use the point-chart probabilities p_{ky} to approximate intersections of charts, and an approximate *nerve* of the cover by Gaussians. The nerve reflects topological information about $\mathcal{M}$, and we can perform topology-preserving operations that drastically reduce the number of charts needed to represent the data. Using the link condition of [20], Algorithm 3 clusters the charts into contractible homogeneous components of low curvature variation.

5. (Section 4.5.1) For each component, let U_j be the local coordinates of $y \in \mathbf{Y}$ projected to $T_{\mu_j}(\mathcal{M})$. We choose a set G of affine maps in $\mathbb{R}^d$ to assemble the local projections U_j into a single neighborhood of $0 \in \mathbb{R}^d$, via a least squares minimization of weighted point-to-point errors. Alternately, solve a constrained optimization problem to arrange data on $S^d \subset R^{d+1}$ (Section 4.5.3).

4.2 Gaussian mixture model fitting

A Gaussian mixture model (GMM) is a collection (μ_i, Σ_i) of multivariate Gaussians in $\mathbb{R}^n$, with respective weight vector $\{w_i\}$. These Gaussians will represent the tangent plane locally, and we will use their associated density

functions to assign points to charts.[2]

We will optimize the choice of Gaussian mixture model by two heuristics:

1. Maximize the likelihood of the points $\mathbf{Y}$.

2. Minimize the curvature and complexity of the resulting manifold $\mathcal{M}$.

(1) is presented via the standard likelihood function

$$P(y_i|\mu, \Sigma) := \sum_j f(y_i|\mu_j, \Sigma_j)p_j \tag{4.1}$$

where f_j is the density function of $\mathcal{N}(\mu_j, \Sigma_j)$:

$$f_j(x) = \frac{1}{(\sqrt{2\pi})^k \det(\Sigma_j)^{1/2}} \exp(-(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)) \tag{4.2}$$

Multivariate normal distributions (MVN) are affine transformations of the product of standard normal random variables: if A is a matrix such that $AA^T = \Sigma_j$, then $\mathcal{N}_j = AZ + \mu_j$, where Z is the random vector $(Z_1, Z_2, \ldots, Z_d)$ for $Z_i \sim N(0,1)$ independent and identically distributed. Conversely, Σ_j must be symmetric, $n \times n$, and positive semi-definite.

Remark 4.2.1. $\{N_j\}$ represent local concentrations of points in an open manifold in $\mathbb{R}^n$. Where $d < n$, this open manifold is a neighborhood of $\mathcal{M}$.

[2]For the purposes of the following sections, any method can be used to estimate the best-fit GMM. This is one suggestion for bounded-curvature manifolds with high error, proposed in [13] to prevent over-fitting. We might also prefer a requirement that Gaussians be equal volume or roughly equal weight, as in [50].

The operation (2) will be to set a prior distribution:

$$p(\mu, \Sigma) := e^{-\sum_{i \neq j} m_i(\mu_j) KL(\mathcal{N}_i || \mathcal{N}_j)} \tag{4.3}$$

where $m_i(\mu_j)$ is a function that increases in distance between the centers μ_i and μ_j, and KL is the Kullback-Liebler divergence, or cross-entropy, of the two distributions $\mathcal{N}_i$ and $\mathcal{N}_j$. Effectively, this ensures that the dominant axes of the charts are penalized for differing substantially over a small distance, smoothing the charts to prevent over-fitting and ensure a good approximation of continuity of derivative along paths. We use these curvature weights $m * KL$ in Section 4.4 as well.

The Kullback-Liebler divergence of two multivariate normal distributions is given by

$$D(\mathcal{N}_1 || \mathcal{N}_2) = (\log |\Sigma_1^{-1} \Sigma_2| + \operatorname{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1} (\mu_2 - \mu_1) - n)/2$$

Together these two equations give a posterior distribution

$$\arg\max_{\mu, \Sigma} P(\mu, \Sigma | \mathbf{Y}) = \arg\max_{\mu, \Sigma} \left\{ \left(\sum_{y_i \in \mathbf{Y}} P(y_i | \mu, \Sigma) \right) P(\mu, \Sigma) \right\}$$

For the functions m_i we have an assortment of reasonable choices - we can take it to be uniform and depending on the injectivity radius r, or we can use an approximation of local curvature to de-emphasize linearity of neighboring components in high-curvature areas. In [13], the function m_i is the probability $\mathcal{N}(\mu_j; \mu_i, (r/2)^2)$, which concentrates weight largely within the injectivity radius.

4.3 Tensor Decomposition

For spaces $\mathcal{M}$ that have closed, measure zero subsets which are not locally diffeomorphic to an open subset of Euclidean space, such as intersections and singularities, a multivariate Gaussian will not approximate the tangent space well. Instead we will use principal components of the higher-order moments of the k-nearest neighbors at singular points to produce a mixed-dimension collection of planes, with no curvature prior, and cluster them by rank and component for alignment (see Section 4.5).

4.3.1 Data Moments

In Principal Component Analysis (PCA), the data covariance matrix

$$\Sigma = YY^T = \sum_y yy^T$$

is decomposed into its principal components, given by the eigenvectors of Σ with highest eigenvalue. This set of eigenvectors, based at μ_Y, can also be seen as an optimal rank d linear approximation of the data, or a low-compression tangent plane to Y at μ_Y.

Instead of decomposing the second cumulant of the data (covariance), we can take higher order moments, expressed:

$$M_i = \sum_{y \in Y} (y - \mu_Y)^{\otimes i}$$

$$M_i' = \sum_{y \in Y} y^{\otimes i}$$

for the centralized moment and the moment about the origin, respectively.[3] This i-moment is a real symmetric tensor of order i.

Summarizing the data via its principal components works most accurately for Gaussian random variables, for which the principal components are the orthogonal directions with highest subsequent variance, and the distribution is independent along each, so that it can be completely specified by a mean and covariance matrix. In general, higher order moments (and more directly, cumulants) can be seen as some measure of *non-Gaussian* behavior - cumulants of a multivariate normal distribution vanish after the first and second.

4.3.2 GPCA using symmetric block decomposition

In [33], Kileel and Pereira define a symmetric block decomposition algorithm using Sylvester's catalecticant method, which factors a symmetric tensor T into a sum of *real symmetric Tucker products*:

$$T = \sum_{i=1}^{R} (A_i; A_i; \ldots; A_i) \cdot \Lambda_i$$

for a collection of *core tensors* $\Lambda_i \in \mathrm{Sym}\mathfrak{T}^m_{\ell_i}$, and *factor matrices* $A_i \in M_{n\times\ell_i}$. This is called an $(A_1, \ldots, A_R)$-*symmetric block term tensor decomposition*. This is similar to other block term decompositions (see, e.g. [34][35][36]), except the decomposition itself is symmetric.

[3] We note that this can be computationally intensive. Recent results of Sherman and Kolda allow for implicit computation of low-rank symmetric approximations to the higher-order moment tensors[51].

Suppose $Y \in \mathbb{R}^n$ is a random variable supported on a subspace arrangement $\mathcal{S} = \cup_{i=1}^{R} S_i \subset \mathbb{R}^n$, where each S_i is a linear subspace of respective dimension d_i. Then for $A_i \in M_{n \times d_i}$ such that $S_i = \text{colspan}(A_i)$, for each m, the moment tensor $\mathbb{E}[Y^{\otimes m}] \in \text{Sym}\mathcal{T}_n^m$ admits a symmetric block term decomposition as above, with A_i as factor matrix coefficients. This is Lemma 6.1 in [33]; we replicate the proof here to give demonstrate the particular significance of the decomposition.

Proof. First, we decompose Y. Let x be the discrete random variable over $[R]$ with probabilities w_i corresponding to the measure of Y restricted to the subspace S_i. For a choice of basis $b_1, \ldots, b_{d_i}$ of S_i, let B_i be the $n \times d_i$ matrix $(b_1 b_2 \ldots b_{d_i})^T$. Let y_i be the random variable in $\mathbb{R}^{d_i}$ induced by the projection $B_i : S_i \to \mathbb{R}^{d_i}$. Then $Y = \{B_i y_i\}_x$, and

$$\mathbb{E}[Y^{\otimes m}] = \sum_{i=1}^{R} w_i \mathbb{E}[(B_i y_i)^{\otimes m}] \tag{4.4}$$

Multilinearity of the m-way tensor product and linearity of expectation give

$$= \sum_{i=1}^{R} w_i (B_i; \ldots; B_i) \cdot \mathbb{E}[y_i^{\otimes m}]. \tag{4.5}$$

Setting $\Lambda_i = w_i \mathbb{E}[y_i^{\otimes m}]$, we see that this decomposes the m-moment of Y from an m-tensor of length n to a sum of R m-tensors of respective length d_i, each corresponding to the n-moment of the restriction of Y to S_i using a particular choice of basis. □

Of course, the choice of basis for a subspace S_i is only unique up to action of $GL_d(\mathbb{R})$. As for the converse - if we decompose the moment tensor into

symmetric blocks, is Y supported minimally on that subspace arrangement? - computational convergence may depend on properties of the arrangement, such as the dimension of pairwise intersections.

4.3.3 Local rank estimation

The naive approach to adapting [63], [49] to the stratified or transversely intersecting setting would be to apply GPCA to k-nearest neighbors at each point. This is possible, but since GPCA detects linear subspace arrangements, and not affine subspace arrangements, the results we get a small distance from an intersection locus will not reflect the local structure accurately.

A better method would be to detect points x at which the tangent space is a union of linear subspaces based at x in $\mathbb{R}^n$. To accomplish this, we assume we have a uniform sampling density ρ, and examine the growth of neighborhoods based at x.

Let $\beta_x(r)$ be the number of points $y \in Y$ such that $||y - x|| \leq r$. Then for a d-dimensional locally linear neighborhood, under ideal circumstances,

$$\beta_x(r) = \rho A_d r^d$$

where A_d is the volume of a unit ball in $\mathbb{R}^d$, and

$$(\log(\beta_x(r)))' = \frac{d}{r},$$

so that the dimension is approximately the slope of the plot $\log(\beta_x(r))$. In practice, these values are computed with $\beta_x(r)$ as the independent variable:

ordering nearest neighbors by distance, $\beta_x(r)$ takes on discrete values $1, 2, \ldots, k$ (if we are using the k nearest neighbors), and the radius of each new point is recorded. We can average the slope by computing

$$\frac{1}{\log(k)} \sum_{i=1}^{k} \frac{\log(i) - \log(i-1)}{r_i - r_{i-1}} \cdot r_i$$

However, at small scales, noise will cause $\beta_x(r)$ to grow as $\rho A_n r^n$, and as r exceeds the injectivity radius, reach, or nears the radius of curvature in any direction, $\beta_x(r)$ will grow in excess of d again. If we have bounds on curvature, reach, injectivity, and noise, then we can take the sum

$$\frac{1}{\log(\max\{i : r_i < \kappa\} / \min\{i : r_i > \epsilon\})} \sum_{i : \epsilon < r_i < \kappa} r_i \frac{\log(i) - \log(i-1)}{r_i - r_{i-1}}$$

for only those neighbors in the annulus $\epsilon < r < \kappa$, for ϵ the upper bound on noise and κ the lower bound on reach, injectivity, and curvature in any direction, to increase accuracy.

If x is on the singular locus of $\mathcal{M}$, contained in the closure of a d-dimensional stratum, then the growth will look similar:

$$\beta_x(r) = m_x \rho A_d r^d$$

where m_x can be an integer, if $T_x(\mathcal{M})$ is a union of m_x linear subspaces; or a multiple of $1/2$, if x lies on the boundary of a halfspace; or another real number if x is a cone-type singularity.

Given d, to estimate m_x we compute the values

$$\frac{\beta_x(r)}{\rho A_d r^d}$$

for all r. We suspect this can be used to give a rank estimate for the local moment tensor $\sum_{y \in kNN(x)} y^{\otimes i}$, given some restrictions on the singularity type.

If x is a singular point lying in the closure of a number of strata of different dimensions d_i, then

$$r \log(\beta_x(r))' = \frac{\sum d_i m_{x,i} A_{d_i} r^{d_i}}{\sum m_{x,i} A_{d_i} r^{d_i}}$$

will not be linear, but it will be continuous - this distinguishes it from the case where T_x can't be estimated accurately by tensor decomposition; if x is near a singular point, or the neighborhood radius exceeds the reach, then $\log(\beta_x(r))'$ will be discontinuous, and we should not use this neighborhood for tangent plane inference.

4.4 Multiple charts

Let $\{\mu_i, \Sigma_i\}$ be a Gaussian mixture fit to the data Y. Associated to this mixture are the density functions $f_i(y)$ (See 4.2). For $(i_1, \ldots, i_\ell) \in [k]^\ell$, define the probability vector

$$q_{(i_1,\ldots,i_\ell)}(y) = \min(f_{i_1}(y), f_{i_2}(y), \ldots, f_{i_\ell}(y))$$

We choose a threshold value t for the nerve complex. A reasonable value of t will depend on the ambient dimension and the size of $\mathcal{M}$, such as $t \sim (2\pi)^{-n/2}|Y|^{-1/2}$ where $|Y|$ is the volume of the convex hull of points Y.

Definition 4.4.1. We define the *nerve complex* $\Delta_t(Y, \{\mu_i, \Sigma_i\})$ for $t \in [0, 1]$:

- $\Delta^0 = [k]$, for k the number of charts

- For $\ell \geq 2$, $(i_1, \ldots, i_\ell) \in \Delta(Y)^\ell$ when $||q_{(i_1,\ldots,i_\ell)}||_\infty > t$.

For $\mathcal{M}$ a stratified manifold or union of manifolds, we may want to add the pairwise Kullback-Leibler divergence of $(i_1, \ldots, i_\ell)$ into $q_{(i_1,\ldots,i_\ell)}$, so that transverse tangent spaces are less likely to be highly connected. For $\mathcal{M}$ a Riemannian manifold, we can relax restrictions on curvature by using intersection for adjacency.

Definition 4.4.2. $\Delta(Y, \{\mu_i, \Sigma_i\})$ is called *flag* if for every $v_1, \ldots, v_\ell, v_{\ell+1} \in V(\Delta(Y))$ such that $(v_i, v_j) \in E(\Delta(Y))$ for all $i \neq j \in [\ell{+}1]$, then $(v_1 \ldots v_{\ell+1}) \in \Delta^\ell$ is an ℓ-simplex in Δ (see 2.2.1).

If $\Delta(Y)$ is flag, then by definition, it can be stored by its graph adjacency matrix.

Lemma 4.4.1. *$\Delta(Y)$ is a flag simplicial complex, i.e. $\sigma' \subset \sigma$ implies $\sigma' \in \Delta(Y)$ for every $\sigma \in \Delta(Y)$, and if $\sigma \in \Delta(Y)$ for every face σ of σ', then $\sigma' \in \Delta(Y)$.*

Proof. That $\Delta(Y)$ is a simplicial complex follows from the fact that

$$\min(f_{i_1}, \ldots, f_{i_\ell}, f_{i_{\ell+1}}) \leq \min(f_{i_1}, \ldots, f_{i_\ell}),$$

so that if the former is greater than t for some y, and therefore a simplex in $\Delta(Y)$, using the same y, its faces are as well.

To show that $\Delta(Y)$ is flag, suppose $\max_y(q_{(j,k)}(y)) > t$ for all $j, k \in (i_1, \ldots, i_\ell)$. Then

$$\max_y(q_{(i_1,\ldots,i_\ell)}(y)) = \max_y(\min_i f_i(y)) = \max_y(\min_{i,j} q_{(i,j)}(y)) > t$$

shows that $(i_1, \ldots, i_\ell)$ is the basis of a simplex in $\Delta(Y)$. $\square$

The nerve complex represents the nerve of the open cover $(U_i, [V_i]_d)$ of Y (the projection to the principal components of Σ_i, see Section 4.5.1), a discrete approximation of $\mathcal{M}$, where $\mathcal{M}$ is compact. We are inspired by the Nerve Theorem:

Theorem 4.4.1. *(Nerve Theorem, see Hatcher [29]) If X is a paracompact space, and U is an open cover of X such that the intersection of any finite subfamily of U is either empty or contractible, then $|\Delta(U)| \simeq X$, i.e. the geometric realization of $\Delta(U)$ is homotopy equivalent to X.*

Assuming that U is a Čech cover of $\mathcal{M}$, the nerve preserves the homotopy type of $\mathcal{M}$. Using operations which preserve the topology of $|\Delta(U)|$, we construct a simpler complex which contains instructions for the combination of multiple tangent planes into charts.

The main technique we will use is edge contraction. In [20], it is proven that if the edge ab satisfies a link condition in the complex Δ, then the contraction $\Delta/C(a, b) \simeq \Delta$. Regarding the charts, contraction will mean combining the charts U_a and U_b, or if a and b already represent index sets A and B, then contraction will result in a vertex label $A \cup B$, so that all charts U_i for $i \in A \cup B$ are aligned using Section 4.5.1.

Definition 4.4.3. (See 1.1.3 for comparison; this is slightly more general) The *star* of a set $X \subset \Delta$ denoted $St(X)$, is the set of cofaces of all $\sigma \in X$, that

is, all simplices containing σ as a face. For a subset S of Δ, the closure of S, denoted $\bar{S}$, is the set of simplices in S and all of their faces. Then the *link* of X, denoted $Lk(X)$, is the set of simplices in $\overline{St(X)} \setminus St(\bar{X})$.

The *link condition* for an edge ab is satisfied if $Lk(ab) = Lk(a) \cap Lk(b)$. To check this, we must be able to compute the link: find all simplices σ containing a (resp. b, ab), list all faces, and use set operations to compare $Lk(ab)$ with $Lk(a)$ and $Lk(b)$.

Lemma 4.4.2. *If Δ is a flag complex, the link condition can be checked using the adjacency matrix, without constructing higher simplices.*

Proof. We first show that if v is a 0-simplex, then $Lk(v) = Lk(\bar{v})$ can be computed using adjacencies. Let $w_1, \ldots, w_m$ be the set of neighbors of v, and find $\{e \in \Delta : e = (w_i, w_j)\}$. Then the link of v is given by the flag complex over $w_1, \ldots, w_m$, $\{e = (w_i, w_j)\}$: since v is adjacent to all w_i, if a set of $w_{i_1}, w_{i_2}, \ldots, w_{i_k}$ are pairwise adjacent, then since Δ is a flag complex, $< w_{i_1}, w_{i_2}, \ldots, w_{i_k} >$ is a face of the simplex $< v, w_{i_1}, w_{i_2}, \ldots, w_{i_k} >$ in δ. For a 1-simplex $e = (v, w)$, $Lk(e)$ is given by the flag complex over the induced subgraph on $N(v) \cap N(w)$. So the link condition can be checked by computing the neighbor sets $N(v)$ and $N(w)$, taking the intersection $N(v) \cap N(w)$, finding the induced subgraph $\Delta_{N(v)}, \Delta_{N(w)}, \Delta_{N(v,w)}$ for each, and comparing the intersection $\Delta_{N(v)} \cap \Delta_{N(w)}$ with $\Delta_{N(v,w)}$. □

4.4.1 Procedure

Once we have our nerve complex Δ, we search for a cover of Δ via contractible subcomplexes, favoring neighbors which are closer in mean and tangent space spanned.

1. For each $e = (v, w) \in E(\Delta)$, let $F_{vw} = e^{-|m_v(\mu_w)+m_w(\mu_v)|*KL(\mathcal{N}_v||\mathcal{N}_w)}$, for m as in Section 4.2 and KL the Kullback-Liebler divergence of Gaussian distributions $\mathcal{N}_b$ and $\mathcal{N}_v$.

2. Begin with a random basepoint $b \in V(\Delta)$.

3. For all edges (b, v) incident to b, check the link condition. Denote by E_b the set of edges satisfying the link condition.

4. Choose $v = \arg\min_v F_{bv}$.

5. Contract edge (b, v): for each simplex containing v, map $\sigma = (\dots, v, \dots) \mapsto (\dots, b, \dots)$. When a simplex contains both b and v, it collapses down one dimension. Relabel b as $b \cup v$. If σ is a 1-simplex (edge), it retains its value F_{vw}, except when σ is produced by the contraction of a 2-simplex to a 1-simplex; in that case, $F_{(b\cup v)w} := \min(F_{bw}, F_{vw})$.

6. At the i-th iteration, basepoint b_I now has labels $b, v_1, \dots, v_{i-1}$. Again, we check the link condition for all neighbors, choose the neighbor $v_i = \arg\min F_{b_I v}$, and contract.

7. Stop when $|I| \geq$ `maxsize`, or when no incident edges satisfy the link condition.

8. Repeat the process, choosing a new basepoint when necessary, until no edges satisfy the link condition. The number of vertices in the final complex is the *chart number* C, and the set of vertex labels is the *nerve cover* $I_1, \ldots, I_C$.

9. Once we have the nerve cover $\{I_1, \ldots, I_C\}$, we pass each index set I to the flat alignment algorithm of Section 4.5.1: create the submatrix of QQ^T (as in 4.5.1) with rows and columns indexed by I, take the trailing eigenvectors of $Q_I Q_I^T + \mathbf{1}$ to get G_I, which maps the sets U_j for $j \in I$ to a common chart in $\mathbb{R}^d$.

10. The result is C charts, with transition maps as defined in Section 4.4.2.

Algorithm 3 Nerve Decomposition Algorithm

for $(v, w) \in \Delta^1$, compute F_{vw} from $\{\mu_i, \Sigma_i\}$ values.
`points` = random ordering of Δ^0
for $b \in$ `points`:
    set $I[b] = \{b\}$.
    while($|I[b]| \leq$`maxsize`):
        Find `neighb` $= \{(b, v) \in \Delta^1\}$
        for $(b, v) \in$ `neighb`:
            If `link_condition`(b, v) = true and v not in I already:
                add (b, v) to E_b
        if $E_b =$ `NULL`: `break`
        else:
            find argument (b, v^*) of $\min\{F_{bv} : (b, v) \in E_b\}$.
        $(\Delta, F) =$ `contract`$(\Delta, F, (b, v^*)))$
        add v^* to $I[b]$.
        remove v^* from `points`
$C = \text{length}(I)$
return (Δ, I)

Since we are adding vertices by adjacency, U_I always remains connected. Similarly, U_I is contractible, since by results of [20], I is produced by topology-preserving contraction of the nerve complex. The homotopy type of the tangent space cover $\{U_i\}$ is given by the type of the resulting contracted nerve. The number of charts C is bounded below by the topological complexity of the cover U_i, which approximates $TC(\mathcal{M})$.

For $k < \sqrt{N}$, the alignment step dominates runtime, but efficiencies can be obtained in reducing the storage of Δ.

4.4.2 Transition Maps

There are a couple distinct natural ways to define the transition maps $\phi'_{ij} : U'_i \to U'_j$.

By linear alignment: for each pair U'_i and U'_j of new charts, if their intersection on Y is non-empty (with respect to the threshold), there is a subset $v_i \in U'_i$ and $w_j \in U'_j$ such that v_i is contained in a simplex that intersects U'_j, and similarly with w_j. Then the *transition maps* are defined on connected components of $v_i \cup w_j$ by the linear alignment $G_{C(v_i \cup w_j)}$.

By interpolation of data: $U'_i \to U'_j$ for $U'_i \cap U'_j \neq \emptyset$ are given on Y by the image of y in each - if $p_{yi'} > t, p_{yj'} > t$, then $\phi_{ij}(G_{U'_i} y) = G_{U'_j}(y)$. This map will not be linear, continuous, or well-defined on points not in Y, but it will provide the best preservation of paths in Y.

4.4.3 Intersection Spaces

Suppose we have a local decomposition of tensors as given in Section 4.3.2, i.e. a collection of R weights w_i, projection matrices B_i, and moment tensors Λ_i.

If $\sum(B_i; B_i; B_i; B_i)\Lambda_i$ is a 4^{th} moment, and if a neighborhood of the singular point x at which GPCA has been performed looks like a mixture of Gaussians based at x and supported on the subspaces generated by B_i, then Wick's Theorem implies that on each subspace,

$$E[(y_i - \mu_i)^{\otimes 4}]_{ijk\ell} = \Sigma_{ij}\Sigma_{k\ell} + \Sigma_{ik}\Sigma_{j\ell} + \Sigma_{i\ell}\Sigma_{jk}$$

so that the covariance matrix entries generate the fourth moment, and with enough information, can be recovered. In [23], a technique is described to give a maximum likelihood mixture of mean-zero Gaussians using tensor decomposition of the 3rd, 4th, and 6th moments. We propose either an analogous technique, or to use a direct tangent space alignment such as [63] which does not depend on a maximum variance basis for the tangent plane.

Once we have a collection of transverse Gaussians centered at μ, we can alternately depend on the high Kullback-Liebler divergence between different subspaces to prevent adjacency in the nerve complex, adjust $m_\mu(\mu)$ to be quite large, or manually enforce that Gaussians based at the same mean are an independent set in the complex. This will allow Algorithm 3 to separate the charts into different components.

4.4.4 Nerve Conjectures

Conjecture 1. *Let $\mathcal{M} \subset \mathbb{R}^n$ be a smooth manifold, with reach ρ and curvature bounded by κ. Let $\epsilon < \rho/2$ be given. Suppose Y is a random uniform sample of sufficiently high density. If the nerve $\Delta(Y, \{\mu_i, \Sigma_i\})$ is contractible and k sufficiently large, then $\mathbb{P}(\mathcal{M} \simeq \cdot) \to 1$.*

Conjecture 2. *Let $\mathcal{M}$ be a manifold in $\mathbb{R}^n$, and let $\{U_i\} \subset \mathcal{M}$ be a Čech cover of open balls of radius r, with r less than the reach and injectivity radius. Replace each U_i with a Gaussian distribution centered at p with axes in the tangent plane to p of length r, and normal axes of length ϵ. Let $Y \sim Unif(\mathcal{M}_\epsilon)$ be a sample of size N. Then $\mathbb{P}[\Delta(Y, \{\mu_i, \Sigma_i\}) \simeq \mathcal{M}] \to 1$ as $N \to \infty$.*

4.5 The alignment G

Once we have the model best fitting the data, we can take advantage of the intrinsic dimension d of the data to compute a dimensionality reduction map which reflects the local geometry. If $\mathcal{M}$, or a suitable subset of $\mathcal{M}$, is contractible and close to flat, then we will be able to assemble the local charts linearly into a best-fit map to $\mathbb{R}^d$.

4.5.1 Flat alignment of Gaussians

Here we follow a technique similar to [13] or [63], with some modifications as noted.

N	number of data points in Y
n	original dimension, $y \in \mathbb{R}^n$
d	intrinsic dimension
D	ambient dimension of desired embedding, $D \geq d$

Let $D \geq d$ be the chosen ambient embedding dimension for our alignment. A smaller D produces more data compression; $D = d$ produces a classic tangent space alignment. An ambient codimension of 1 or 2 may be desired to preserve intrinsic features of $\mathcal{M}$, for example if $\mathcal{M}$ is not contractible or has high curvature, keeping in mind that in some cases, $\mathcal{M}$ might not isometrically embed without an ambient dimension over $2d$.

Per Section 4.2, we have a set $\{(\mu_k, \Sigma_k)\}$ of multivariate Gaussians with global weight vector w_k. Using the corresponding density functions f, this gives rise to pointwise assignment weights $w_{ky} = f_{\mu_k,\Sigma_k}(y) * w_k$ of each

data point y to each chart. Denote by P the $k \times N$ matrix of w_{ky} values, normalized by column so that P is a stochastic matrix. Then $p_{iy} \in P$ gives the likelihood that y is generated by Gaussian $\{\mu_i, \Sigma_i\}$. Each row $P_{i\cdot}$ gives the *membership vector* for chart i.

If $\Sigma_k = V_k \Lambda_k V_k^T$, with Λ_k a diagonal matrix of decreasing eigenvalues, then we take the first d rows of V_k, or the first d columns of V_k^T. For the Gaussian distribution, this is equivalent to performing Principal Component Analysis on the distribution and taking the first d components.[4] We define the projection matrix

$$U_k := \begin{pmatrix} [V_k]_d (Y - \mu_k) \\ 1 \dots 1 \end{pmatrix}; \quad (U_k)_y = \begin{pmatrix} u_{ky} \\ 1 \end{pmatrix}$$

U_k is a $(d+1) \times N$ matrix of local coordinates centered at μ_k, with an additional row of 1's. This will allow us to define affine transformations of U.

An important property to note about P is that the normalization of $w_{ky} = f_{\mu_k, \Sigma_k}(y) * w_k$ is a continuous partition of unity on $\mathbb{R}^n$, practical to compute on a neighborhood of $\mathcal{M}$. This allows for linear interpolation of sheaf-theoretic local data on $\mathcal{M}$: if I have local sections (e.g. defined on the local tangent plane approximations U_k), then I can use the weights to extend this data to a global section.

We will denote by G_k the affine transformation mapping $U_k \subset \mathbb{R}^d$ neighborhood of 0 into the connection space $\mathbb{R}^D$. Our goal in choosing G is to

[4] We note that this is different from taking PCA of the data itself, because of the addition of the prior.

minimize

$$\sum_y \sum_{i \geq j} \left\| \left[G_i \begin{pmatrix} [V_i]_d(y - \mu_i) \\ 1 \end{pmatrix} - G_j \begin{pmatrix} [V_j]_d(y - \mu_j) \\ 1 \end{pmatrix} \right] p_{iy} p_{jy} \right\|^2$$

$$= \sum_{i \geq j} ||[G_i U_i - G_j U_j)] P_i P_j||_F^2 \tag{4.6}$$

where P_k is the $N \times N$ diagonal matrix of p_{ky} values, and $||\cdot||_F$ is the Frobenius norm. This is the distance between the image of y according to chart j and chart k, weighted by the probability that y associates to both of them. This records the error in the transition maps - since we are relating the charts linearly, we will not be able to entirely eliminate error arising from curvature.

Each G_k is a $D \times (d+1)$ matrix $(v_1 v_2 \dots v_a | a_k)$. We stack them for computation:

$$G = \begin{pmatrix} G_1 & G_2 & \dots & G_k \end{pmatrix}$$

Then we find an expression equivalent to (4.6). Let Q^{ij}, for $i \leq j$, be the block matrix

$$Q^{ij} := \begin{pmatrix} 0 \\ \vdots \\ U_i P_i P_j \\ 0 \\ \vdots \\ -U_j P_i P_j \\ 0 \\ \vdots \end{pmatrix},$$

and let Q be $(Q^{12} Q^{13} \dots Q^{1k} Q^{23} \dots)$ with the standard lexicographic ordering of $\binom{k}{2}$. Then

$$GQ = ((G_1 U_1 - G_2 U_2) P_1 P_2 \ \ (G_1 U_1 - G_3 U_3) P_1 P_3 \ \ \dots (G_i U_i - G_j U_j) P_i P_j \dots)$$

so that the sum of squared error (Equation 4.6) is given by the Frobenius norm of GQ:

$$||GQ||_F = \mathrm{Tr}(GQQ^TG^T). \tag{4.7}$$

This definition of Q departs from the technique of Brand. It increases complexity, but also avoids degeneracy. By Lemma 4.5.1, QQ^T can be computed directly in blocks; then G is minimized by choosing as columns of G^T the D trailing eigenvectors of QQ^T.

This ensures that the norm of Equation 4.7, which records a sum of point-to-point errors, is as close to 0 as possible.

We note, however, that this technique guarantees independence of the *rows* of G, not the columns. To see that this may produce degenerate solutions, consider the connection matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

for $D = 3$ and any $k > 1$, which sends all charts except the first to 0.

To help alleviate this problem, we condition 4.7 by eigendecomposing $QQ^T + \mathbf{1}$ instead. This minimizes 4.7 and also $||G\mathbf{1}||_F$, which counts row sums. Favoring rows which sum to 0 helps prevent solutions like G above, and balances the charts somewhat. Degenerate solutions (in the sense of an individual G_k having rank less than d) are still possible.

Remark 4.5.1. QQ^T can be computed directly as a block matrix given by the

	dimension	array	contents
Y	$n \times N$	$\begin{pmatrix} y_1 & y_2 & \dots & y_N \end{pmatrix}$	cols are data points
$\mathbf{M}$	$n \times k$	$\begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_k \end{pmatrix}$	cols are chart centers
$\mathbf{\Sigma}$	$n \times n \times k$	$\begin{pmatrix} \Sigma_1 & \Sigma_2 & \dots & \Sigma_k \end{pmatrix}$	K cov. matrices
w	$k \times 1$	$(w_1, w_2, \dots, w_k)^T$	mixture weights
V	$n \times d \times k$	$\begin{pmatrix} v_1^j & v_2^j & \dots & v_d^j \end{pmatrix}_{j=1,\dots,k}$	1st d eigenvecs of Σ_k
G_k	$D \times (D+1)$	$\begin{pmatrix} c_{11}^k & c_{12}^k & \dots & c_{1D}^k & a_1^k \\ c_{21}^k & c_{22}^k & \dots & c_{2D}^k & a_2^k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{D1}^k & c_{D2}^k & \dots & c_{DD}^k & a_D^k \end{pmatrix}$	affine transformation
G	$D \times (k(D+1))$	$\begin{pmatrix} G_1 & G_2 & \dots & G_k \end{pmatrix}$	all G_k
P	$k \times N$	$\left(p_{iy} = \frac{w_{iy}}{\sum_{j=1}^k w_{jy}}\right)_{y \in Y, i \in [k]}$	stochastic matrix
P_i	$N \times N$	$\begin{pmatrix} p_{iy_1} & 0 & \dots & 0 \\ 0 & p_{iy_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{iy_N} \end{pmatrix}$	i-th chart probabilities
U_i	$(d+1) \times N$	$\begin{pmatrix} u_{iy_1} & \dots & u_{iy_N} \\ 1 & \dots & 1 \end{pmatrix}$	local coordinates + 1
Q	$k(d+1) \times \binom{k}{2} N$	$\begin{pmatrix} 0 \\ \vdots \\ U_i P_i P_j \\ 0 \\ \vdots \\ -U_j P_i P_j \\ 0 \\ \vdots \end{pmatrix}_{i,j \in [k]}$	Q^{ij} in lex. order
QQ^T	$k(d+1) \times k(d+1)$		See Remark 4.5.1

Figure 4.1: Array reference

U_j and P_j:

$$\begin{pmatrix} U_1(P_1^2(\sum_{i=2}^{k} P_i^2))U_1^T & U_1(P_1^2P_2^2)U_2^T & \dots & U_1(P_1^2P_k^2)U_k^T \\ U_2(P_1^2P_2^2)U_1^T & U_2(P_2^2(\sum_{i\neq 2} P_i^2))U_2^T & \dots & U_2(P_2^2P_k^2)U_k^T \\ \vdots & \vdots & \ddots & \vdots \\ U_k(P_1^2P_k^2)U_1^T & U_k(P_2^2P_k^2)U_2^T & \dots & U_k(P_k^2(\sum_{i=1}^{k-1} P_i^2))U_k^T \end{pmatrix} \tag{4.8}$$

I.e. (i,i) diagonal blocks are $U_iU_i^TP_i^2(\sum_{j\neq i} P_j^2)$, and (i,j) off-diagonal blocks are $U_iU_j^TP_i^2P_j^2$. This bypasses the need to construct Q, which is much larger.

Remark 4.5.2. Because some of the probabilities p_{ky} will be quite small, there may be some variation in the result based on numerical imprecision. We avoid this danger by thresholding P_k at a reasonable uncertainty level α. This also increases sparsity of P_k, and QQ^T; if we know which charts have $P_i^2P_j^2 = 0$, which for a large number k of charts should be quite common, those blocks need not be computed.

With G in hand, we can finally construct the NLDR map $\sum G_kU_kP_k$, a $D \times N$ matrix whose columns represent image of y, computed as a weighted average in $\mathbb{R}^a$.

$$y \mapsto \left(\sum_k G_k(U_k)P_k\right)_{\cdot y} \tag{4.9}$$

The objective value (4.7) gives a measurement of the degree of distortion induced by the map G. To compare these distortions, we calculate the mean squared error

$$\text{MSE}(G,Y) := \frac{1}{N}||GQ||_F \tag{4.10}$$

If we have multiple charts, the mean squared error is given by

$$\text{MSE}(G, Y) := \frac{1}{CN} \left(\sum_{j=1}^{C} ||G_j Q_j||_F \right)$$

If there exists an affine subspace A with projection map $P_A : \mathbb{R}^n \to A$ such that $||y - P_A(y)|| < \epsilon$ for all $y \in Y$, then we call Y *ϵ-flat.*

Conjecture 3. *Let $\epsilon > 0$ be given. Let $\delta < \sin(\epsilon/2)$. Let Y be an δ-flat random sample of $\mathcal{M}$ (i.e. with normal noise bounded by δ), where $\mathcal{M}$ is a contractible open subset of a d-dimensional affine subspace A of $\mathbb{R}^n$, and such that $\delta < var(P_L(Y))$ for P_L the projection in $\mathbb{R}^n$ to any affine line $L \subset A$ contained in A. Let $k = 1$, and let μ, Σ be the result of maximum a posteriori approximation as described in Section 4.2. Let G be the least squares embedding in $\mathbb{R}^d$ as given in Equation 4.7. Then the map $GU : \mathcal{M} \to \mathbb{R}^d$, a composition of a linear and an affine map, is Lipschitz with constant bounded in ϵ, as is its reverse map to the principal eigenspace of Σ, $\phi^{-1} : \mathbb{R}^d \to \mathbb{R}^n$.*

Conjecture 4. *Suppose $\mathcal{M}$ is a contractible manifold in $\mathbb{R}^n$, Y a random sample in $\mathcal{M}_\epsilon$, with k and N sufficiently large, $P_{\mathcal{M}}(Y)$ sufficiently dense, ϵ sufficiently small, that Conj. 3 is satisfied for any ellipsoidal neighborhood contained in a ball of radius $(\epsilon/2, \epsilon)$. Then for $x, y \in Y$, $||x - y|| < \delta$, GP_{T_i} is a Lipschitz map for all μ_i, Σ_i such that $p_{ix}, p_{iy} > 0$.*

4.5.2 Example

An ellipsoidal gaussian mixture model was fit to 1000 points on a unit sphere using Mclust [50], and the chart groupings computed by nerve contrac-

tion as in Section 4.4. The output contracted nerve is the boundary of a 3-simplex (homeomorphic to the sphere), with basis $\{(4), (2, 6, 8, 10), (7), (1, 3, 5, 9)\}$. The grouped charts were then aligned using Section 4.5.1, and plotted according to Equation 4.9, with size of point given by the probability it belongs to that chart collection. The resulting visualization in $\mathbb{R}^2$ is given in Figure 4.2.

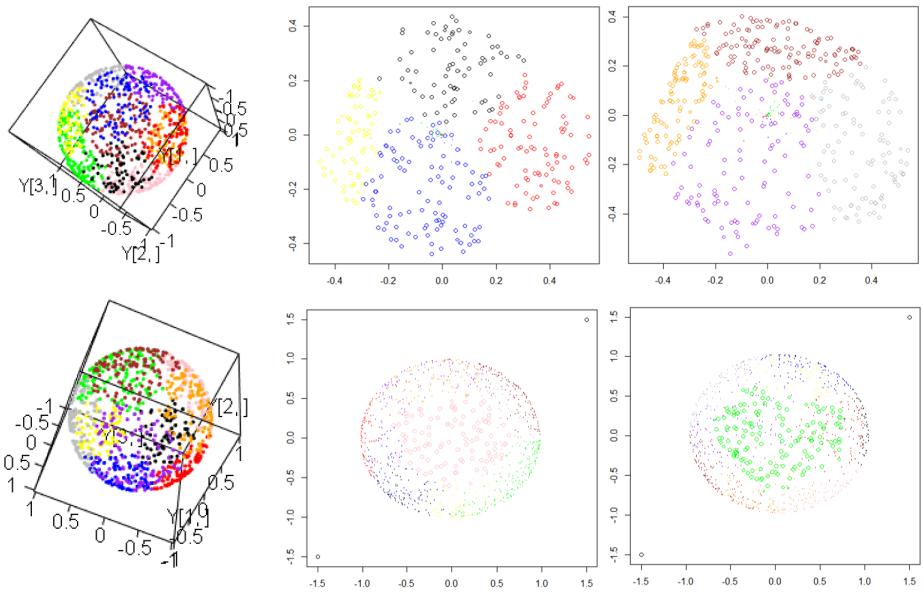


Figure 4.2: Left, 1000 points on a sphere in $\mathbb{R}^3$. Right, the visualized charts.

4.5.3 Spherical Alignment

If $\mathcal{M}$ is not contractible, then it will not embed diffeomorphically in $\mathbb{R}^d$; however, we may have a reasonable embedding in $\mathbb{R}^{d+1}$ or $\mathbb{R}^{d+2}$.

Here we restrict to the special case where $D = d+1$, and we would like to fit the data to the unit sphere S^d.

We modify the technique of the previous section, adding constraints to the optimization problem (4.7).

$$||a_i|| = 1, \langle c_j^i, a_i \rangle = 0 \tag{4.11}$$

for c_j^i columns of G_i. This ensures that the center of the tangent plane is translated to a point on S^d, and that $\text{span}(c_1^i, c_2^i, \ldots, c_d^i)$ lies in $T_{S^d}(a_i) \subset \mathbb{R}^{d+1}$.

Let λ be a vector of Lagrange multipliers

$$(\lambda_1^1, \lambda_2^1, \ldots, \lambda_d^1, \lambda_{d+1}^1, \lambda_1^2, \lambda_2^2, \ldots, \lambda_d^2, \lambda_{d+1}^2, \ldots, \ldots, \lambda_1^k, \lambda_2^k, \ldots, \lambda_d^k, \lambda_{d+1}^k)^T$$

where λ_j^i corresponds to the j-th column vector of G_i via the equations

$$\mathcal{L}(\lambda, G) = \text{Tr}(GQQ^TG^T) - \sum \lambda_j^i \langle c_j^i, a_i \rangle$$

$$0 = \nabla_G \text{Tr}(GQQ^TG^T) - \sum \lambda_j^i \nabla_G \langle c_j^i, a_i \rangle$$

$$0 = 2GQQ^T - \sum_{j=1}^{d} \sum_i \lambda_j^i \begin{pmatrix} \ldots 0 & a_i & \ldots & c_j^i & 0 & \ldots \end{pmatrix}$$

$$+ \sum_i \lambda_{d+1}^i \begin{pmatrix} 0 & \ldots & 0 & 2a_i & \ldots & 0 \end{pmatrix}$$

$$0 = 2GQQ^T - G\Lambda,$$

where Λ is the matrix

$$\Lambda = \begin{pmatrix} B_1 & 0 & \ldots & 0 \\ 0 & B_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & B_k \end{pmatrix}; \qquad B_i = \begin{pmatrix} 0 & \ldots & 0 & \lambda_1^i \\ 0 & \ldots & 0 & \lambda_2^i \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & \lambda_d^i \\ \lambda_1^i & \ldots & \lambda_d^i & 2\lambda_{d+1}^i \end{pmatrix}$$

So we have $G(2QQ^T - \Lambda) = 0$, which together with the constraints $\langle c_j^i, a_i \rangle = 0, \langle a_i, a_i \rangle = 1$, makes $(d+1)(d+2)k$ equations in $(d+1)(d+2)k$ variables.

Index

Bibliography

[1] Alex Abreu and Marco Pacini. The automorphism group of $M_{0,n}^{\mathrm{trop}}$ and $\bar{M}_{0,n}^{\mathrm{trop}}$. Journal of Combinatorial Theory, Series A, 154:583–597, 2018.

[2] W. A. Akanni, M. Wilkinson, C. J. Creevey, P. G. Foster, and D. Pisani. Implementing and testing bayesian and maximum-likelihood supertree methods in phylogenetics. Royal Society Open Science, 2(8), 08.

[3] David Ayala, John Francis, and Hiro Lee Tanaka. Local structures on stratified spaces. Advances in Mathematics, 307:903–1028, 2017.

[4] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 2139–2147, Red Hook, NY, USA, 2013. Curran Associates Inc.

[5] Dennis Barden and Huiling Le. The logarithm map, its limits and fréchet means in orthant spaces. Proceedings of the London Mathematical Society, 117(4):751–789, jun 2018.

[6] Dennis Barden, Huiling Le, and Megan Owen. Central limit theorems for Fréchet means in the space of phylogenetic trees. Electronic Journal of Probability, 18(none):1 – 25, 2013.

[7] M. Bačák. Computing medians and means in Hadamard spaces. SIAM Journal on Optimization, 24:1542–1566, 09 2014.

[8] P. Benner, M. Bačák, and P. Y. Bourguignon. Point estimates in phylogenetic reconstructions. Bioinformatics, 30:i534–i540, 08 2014.

[9] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics, 27(4):733 – 767, 2001.

[10] O. R. Bininda-Emonds, editor. Phylogenetic supertrees: combining information to reveal the tree of life, volume 4 of Computational Biology. Springer Netherlands, 2004.

[11] Andrew J. Blumberg, Prithwish Bhaumik, and Stephen G. Walker. Testing to distinguish measures on metric spaces, 2018.

[12] Debra Boutin. Identifying graph automorphisms using determining sets. Electr. J. Comb., 13, 09 2006.

[13] M. Brand. Charting a manifold. In NIPS, 2002.

[14] Corey Bregman. Isometry groups of cat(0) cube complexes, 2017.

[15] Daniel G. Brown and Megan Owen. Mean and Variance of Phylogenetic Trees. Systematic Biology, 69(1):139–154, 06 2019.

[16] Peter Buneman. The recovery of trees from measures of dissimilarity. In Mathematics the the Archeological and Historical Sciences, pages 387–395, United Kingdom, 1971. Edinburgh University Press.

[17] Dmitri Burago, Yuri Burago, and Sergei Ivanov. A Course in Metric Geometry, volume 33 of Graduate Studies in Mathematics. American Mathematical Society, 2001.

[18] José Cáceres, Delia Garijo, Antonio Gonzalez, Alberto Márquez, and María Puertas. The determining number of kneser graphs. Discrete Mathematics and Theoretical Computer Science. DMTCS [electronic only], 15, 01 2013.

[19] Damien M. de Vienne, Sébastien Ollier, and Gabriela Aguileta. Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution, 29(6):1587–1598, 01 2012.

[20] Tamal K. Dey, Herbert Edelsbrunner, Sumanta Guha, and Dmitry V. Nekhayev. Topology preserving edge contraction. Publications de l' Institut Mathematique, 60:23–45, 1999.

[21] A.J. Drummond and A. Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7(214), 2007.

[22] P. Erdös, Chao Ko, and R. Rado. Intersection theorems for systems of finite sets. The Quarterly Journal of Mathematics, 12(1):313–320, 01

1961.

[23] Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC '15, page 761–770, New York, NY, USA, 2015. Association for Computing Machinery.

[24] Chris Godsil and Gordon Royle. Algebraic Graph Theory, volume 207 of Graduate Texts in Mathematics. Springer-Verlag New York, 2001.

[25] Mark Goresky and Robert MacPherson. Stratified Morse Theory. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, 1988.

[26] K. Gori, T. Suchan, N. Alvarez, N. Goldman, and C. Dessimoz. Clustering genes of common evolutionary history. Molecular biology and evolution, 33:1590–1605, 2016.

[27] Gillian Grindstaff. The isometry group of phylogenetic tree space is S_n. Proceedings of the American Mathematical Society, 2020.

[28] Gillian Grindstaff and Megan Owen. Representations of partial leaf sets in phylogenetic tree space. SIAM Journal on Applied Algebra and Geometry, 3:691–720, 2019.

[29] Allen Hatcher. Algebraic Topology. Cambridge University Press, December 2001.

[30] J. Heled and A. J. Drummond. Bayesian inference of species trees from multilocus data. Molecular biology and evolution, 27:570–580, 2009.

[31] Susan Holmes. Statistical approach to tests involving phylogenies. Mathematics of Evolution and Phylogeny, pages 91–120, 2005.

[32] J.P. Huelsenbeck and F. Ronquist. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics, 17:754–755, 2001.

[33] Joe Kileel and João M. Pereira. Subspace power method for symmetric tensor decomposition and generalized pca, 2020.

[34] L. Lathauwer. Decompositions of a higher-order tensor in block terms - part i: Lemmas for partitioned matrices. SIAM J. Matrix Anal. Appl., 30:1022–1032, 2008.

[35] L. Lathauwer. Decompositions of a higher-order tensor in block terms - part ii: Definitions and uniqueness. SIAM J. Matrix Anal. Appl., 30:1033–1066, 2008.

[36] Lieven Lathauwer and Dimitri Nion. Decompositions of a higher-order tensor in block terms—part iii: Alternating least squares algorithms. SIAM J. Matrix Analysis Applications, 30:1067–1083, 01 2008.

[37] Tong Lin and Hongbin Zha. Riemannian manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(5):796–809, 2008.

[38] L. Liu. Best: Bayesian estimation of species trees under the coalescent model. Bioinformatics, 24:2542–2543, 2008.

[39] Wayne P. Maddison. Gene trees in species trees. Systematic Biology, 46(3):523–536, 1997.

[40] Ezra Miller, Megan Owen, and J. Scott Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. Advances in Applied Mathematics, 68:51 – 91, 2015.

[41] Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics, 31(12):i44–i52, 06 2015.

[42] Anthea Monod, Bo Lin, Ruriko Yoshida, and Qiwen Kang. Tropical geometry of phylogenetic tree space: A statistical perspective, 2020.

[43] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. Discrete and Computational Geometry, 39:419–441, 2008.

[44] Megan Owen. Computing geodesic distances in tree space. SIAM Journal on Discrete Mathematics, 25:1506–1529, 2011.

[45] Megan Owen and Scott Provan. A fast algorithm for computing geodesic distances in tree space. IEEE/ACM Trans. Computational Biology and Bioinformatics, 8:2–13, 2011.

[46] Y. Ren, S. Zha, J. Bi, J.A. Sanchez, C. Monical, M. Delcourt, R. Guzman, and R. Davidson. A combinatorial method for connecting bhv spaces representing different numbers of taxa. 2017.

[47] J. A. Rhodes. Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Transactions in Computational Biology and Bioinformatics, 17(6):2107–2118, 2020.

[48] Michah Sageev. CAT (0) cube complexes and groups, volume 21 of IAS/Park City Mathematics Series, pages 7–54]. American Mathematical Society, 2014.

[49] Luis Scoccola and Jose A. Perea. Approximate and discrete euclidean vector bundles. 2021.

[50] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal, 8(1):289–317, 2016.

[51] Samantha Sherman and Tamara G. Kolda. Estimating higher-order moments using symmetric tensor decomposition. SIAM Journal on Matrix Analysis and Applications, 41(3):1369–1387, 2020.

[52] Cuong Than, Derek Ruths, and Luay Nakhleh. Phylonet: A software package for analyzing and reconstructing reticulate evolutionary relationships. BMC bioinformatics, 9:322, 02 2008.

[53] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 61(3):611–622, 1999.

[54] Shuji Tsukiyama, Mikio Ide, Hiromu Ariyoshi, and I. Shirakawa. A new algorithm for generating all the maximal independent sets. SIAM J. Comput., 6:505–517, 09 1977.

[55] Tandy Warnow. Supertree construction: Opportunities and challenges, 2018.

[56] Stephen Watson. The classification of metrics and multivariate statistical analysis. Topology and its Applications, 99(2):237–261, 1999.

[57] Grady Weyenberg, Peter Huggins, Christopher Schardl, Daniel Howe, and Ruriko Yoshida. Kdetrees: Non-parametric estimation of phylogenetic tree distributions. Bioinformatics (Oxford, England), 30, 04 2014.

[58] Mark Wilkinson, James A. Cotton, Chris Creevey, Oliver Eulenstein, Simon R. Harris, Francois-Joseph Lapointe, Claudine Levasseur, James O. Mcinerney, Davide Pisani, and Joseph L. Thorley. The Shape of Supertrees to Come: Tree Shape Related Properties of Fourteen Supertree Methods. Systematic Biology, 54(3):419–431, 06 2005.

[59] Amy Willis. Confidence sets for phylogenetic trees. Journal of the American Statistical Association, 114(525):235–244, 2019.

[60] Niko Yasui, Chrysafis Vogiatzis, Ruriko Yoshida, and Kenji Fukumizu. imphy: Imputing phylogenetic trees with missing information using mathematical programming. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(4):1222–1230, 2020.

[61] Sakellarios Zairis, Hossein Khiabanian, Andrew J. Blumberg, and Raul Rabadan. Moduli spaces of phylogenetic trees describing tumor evolutionary patterns. In Dominik Ślęzak, Ah-Hwee Tan, James F. Peters, and Lars Schwabe, editors, Brain Informatics and Health, pages 528–539, Cham, 2014. Springer International Publishing.

[62] Sakellarios Zairis, Hossein Khiabanian, Andrew J. Blumberg, and Raul Rabadan. Genomic data analysis in tree spaces, 2016.

[63] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM JOURNAL ON SCIENTIFIC COMPUTING, pages 313–338, 2004.

Vita

Gillian Roxanne Grindstaff was born in Long Beach, California on May 6, 1992, the daughter of Charles C. Grindstaff and Randi M. Summer. In 2010, she graduated from Highland Park High School in Dallas, Texas, and moved to Claremont, California for a liberal arts education at Pomona College, including a semester abroad in Budapest. She spent summers at the Claremont Colleges and Oregon State University on undergraduate research projects. She received a Bachelor of Arts degree from Pomona College in 2014, majoring in Mathematics. After graduation she worked remotely designing curriculum for Minerva Schools at KGI, and attended the Math in Moscow program. During her time in Russia, she was accepted to the University of Texas at Austin mathematics program. She began her graduate studies here in 2015.

Permanent address: 1156 Kenilworth Ave.
Kenwood, CA 95452

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.