Functional Data Analysis of Populations of Tree-structured Objects

by

Haonan Wang

CHAPTER 1

Introduction

1.1. General Introduction

A tree (see Section 2.1 for formal definition) is a simple graph such that there is a unique path (a set of edges) between every pair of nodes (vertices). In many applications, tree-valued complex objects are given, such as phylogenetic studies, clustering analysis, classification analysis and medical image analysis. For a population of tree-structured objects, many statistical notions, such as "center point" and "variation" are not clear. In this dissertation, a new method for understanding populations of tree-structured objects has been developed. This development includes, a new metric, a new "center point", and an analog of Principal Component Analysis (PCA) in tree space.

In phylogenetic studies (see for example, Li, et al, 2000 and Holmes, 1999), biologists build phylogenetic trees to illustrate the evolutionary relations among a group of organisms. Each node represents a taxonomic unit, such as a gene, or such as an individual represented by part of its genome, etc. The branching pattern (topology) represents the relationships between the taxonomic units. The lengths of the branches have meanings, such as the evolutionary time.

In cluster analysis (see Everitt, et al, 2001), a common practice is to obtain different cluster trees by using different algorithms, or by "bagging" or related methods (see Breiman, 1996), and then seek to do inference on the "central" tree. For cluster trees, the terminal nodes (external nodes, i.e., nodes at the tip of the tree) indicate

the objects to be grouped; while the interior nodes and the length of the paths bear no physical meaning.

In the classification and regression tree (CART) analysis (see Breiman, et al, 1984), researchers make a decision tree to categorize all of the data objects. First, all of the objects are in one big group, called the "root node". Then, according to a decision rule, each group of objects will be partitioned into two subgroups, called "nodes". For this type of classification tree, the branches indicate the responses to some decision rule. Each node represents a group of objects after applying a sequence of decision rules.

In medical image analysis, many organisms also have branching properties, such as blood vessel systems (see Bullitt and Aylward, 2002) and pulmonary airway systems (see Tschirren, et al, 2002). Each vessel (airway) system can be represented as a tree. For this vessel (airway) tree, each node represents a blood vessel (airway), and the branches only illustrate the connectedness property between two blood vessels (airways). For blood vessel trees (airway trees), both topological structure and geometric properties, such as the locations and orientations of the blood vessels (airways), are very important. Important geometric properties are numerically summarized using "attributes".

In statistical pattern recognition, a data vector is called a feature vector. Every data object is represented by a feature vector. Each entry in the feature vector is called a "feature". The term "attribute" has the same meaning as "feature" in this dissertation; while "attribute" is more specific in the field of graph and tree theory.

For general tree-structured objects, topological structures and nodal attributes are two important aspects of trees, with different importance for different examples. The attributes contain the full numerical summarization of the data objects. For the special case of cluster trees, numerical values (i.e., attributes) are not used; while, for the classification and regression trees, the attributes are the total numbers of objects

in each group, which is the numerical feature of each node. Those attributes will play a role in the analysis. For the blood vessel trees (airway trees), both topological structures and attributes (geometric properties) are very important.

In this dissertation, methods are developed for the study of populations of trees, not for individual trees. The "population" refers to the empirical population (or sample), not any notion of a theoretical population. Each tree has both topological structures and geometric properties.

A context where statistical analysis of populations of tree-structured objects is of interest is shape analysis in medical imaging.

Shape is an interesting and useful characteristic of objects. The problem of how to represent and classify shapes is very complicated. In medical research, various diseases, such as schizophrenia, have been associated with the shape of various brain parts (see Yushkevich, et al, 2001 for discussion and further references).

FIGURE 1.1. Example of a shape of interest.

For example, consider the shape in Figure 1.1 (from the work of Yushkevich, et al). It shows an example of one member of a population of shapes of corpora callosa. There are bendings at the two ends and one bump in the middle of the object.

A class of convenient and powerful shape representations is m-reps (see Pizer, et al, 1999). These are being developed by S. M. Pizer, and the Medical Image Display and Analysis Group (MIDAG) at UNC-Chapel Hill.[1] M-reps capture shape by dividing

[1]visit the MIDAG website at *http://midag.cs.unc.edu*

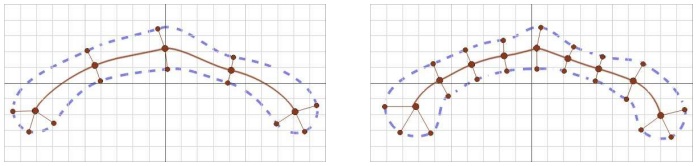

FIGURE 1.2. Coarse and fine scale m-reps.

it into parts coarsely or finely based on "medial" ideas. Figure 1.2 shows both coarse scale m-reps and fine scale m-reps of the shape shown in Figure 1.1. The m-rep parameters (location, radius and angles) are called "features" and are concatenated into a feature vector to provide a numerical summary of the shape.

The statistical analysis of populations of shapes represented by m-reps is straightforward when the general structures of the shapes are all the same because each member of the population is represented by a vector of the same length. But this is a rather restrictive assumption, and many medical imaging data sets need a more general representation. This can be done in the m-rep framework, but a more complicated tree-structured representation is needed.

M-reps provide a good shape representation for shapes that are not far from convex. If the shape is far from convex, a multi-figural representation is needed. A good example is the human hand. For a population of hands, the palm and each finger can be represented by a "figure", each of which is a collection of m-rep parameters. Therefore, each hand is a multi-figural object (see Figure 1.3).

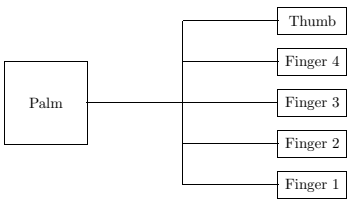

FIGURE 1.3. An example of multi-figural object — hand.

If every member of the population has five fingers, a simple approach is to put all of the features of one hand into a single long feature vector. Thus, the shape space is equivalent to Euclidean space, which is a linear vector space, and the addition and scalar multiplication operations are defined. So, statistical analysis is straightforward. For example, a useful notion of the population center point is the mean vector and the population variation is usually effectively analyzed by Principal Component Analysis, on the Euclidean space spanned by those feature vectors.

It is not straightforward to analyze population structure when some hands do not have five fingers or blood vessels do not have the same branching structures. In this case, the lengths of the feature vectors are not the same. Tree-structured objects are used to represent members of such a population. For example, to represent a hand (shown in Figure 1.3) using tree structure, the figures for the palm and fingers are the nodes of the tree. The palm is the root node, and each finger is a child node of the palm. Furthermore, the m-rep parameters of each figure, including figures for the palm and the fingers, are the attributes for that node.

The statistical analysis of tree-structured objects, such as population center point and population variation, is very complicated. Unlike Euclidean space where classical statistical methods are straightforward to implement, the space of tree-structured objects is non-Euclidean, in the sense that natural definitions of the fundamental linear operators of addition and scalar multiplication operations are unknown. Therefore, the population center point cannot be simply calculated as a mean vector and the variation cannot be analyzed by the regular PCA. Here, a careful axiomatic structure for understanding "center" and "variation" is developed, which avoids the need to define the linear operations.

A new method is required for the statistical analysis of a population of tree-structured objects. The approach is based on a new metric δ on tree space (see Section 3.1 and see Margush, 1982 for more discussion of metrics on trees). This

metric δ consists of two parts: the integer part d_I, which captures the topological aspects of the structure of the tree population (see Section 2.2 for more detail), and the fractional part f_δ, which captures features of the nodal attributes (see Section 3.1).

The metric δ provides a foundation for defining the notion of population center point. A new center point, the median-mean tree (see Section 3.3), is introduced as a combination of median and mean. It has properties similar to the median with respect to the integer part metric (see Section 2.3) and similar to the mean with respect to the fractional part metric (see Section 3.3).

Furthermore, it is of interest to quantify the variability of the population about the center point. Here, an analog of PCA, based on "treeline" which plays the role of "one-dimensional" subspace, is developed for tree space (see Sections 2.5 and 3.6). A key theoretical contribution was a fundamental theory of the variation decomposition in tree space, a tree version of the Pythagorean Theorem (see Sections 2.5 and 3.5), which allows ANOVA style decomposition of sums of squares.

This dissertation develops the statistical analysis of populations of tree-structured objects with (Chapter 3) or without (Chapter 2) attributes respectively.

1.2. Application to a Blood Vessel Data Set

In this section, the ideas of statistical analysis, such as "center point" and an analog of PCA, for a data set of tree-structured objects, are motivated by a data set of brain blood vessel trees. This illustrates the statistical analysis methods, which will be developed in this dissertation.

A good description of major blood vessels of the brain can be found at the website of *The Doctor's Lounge*:[2]

[2]From *http://www.thedoctorslounge.net/education/tutorials/cerebcirc/cerebcirc1.htm*

> The brain receives one fifth of the resting cardiac output. This blood supply is carried by the two internal carotid arteries (ICA) and the two vertebral arteries that anastomose at the base of the brain to form the circle of Willis. Carotid arteries and their branches (referred to as the anterior circulation) supply the anterior portion of the brain while the vertebrobasilar system (referred to as posterior circulation) supplies the posterior portion of the brain.

An example of brain blood vessels is shown in Figure 1.4, provided by Dr. E. Bullitt. This system has three important components: left carotid, right carotid and vertebrobasilar system.

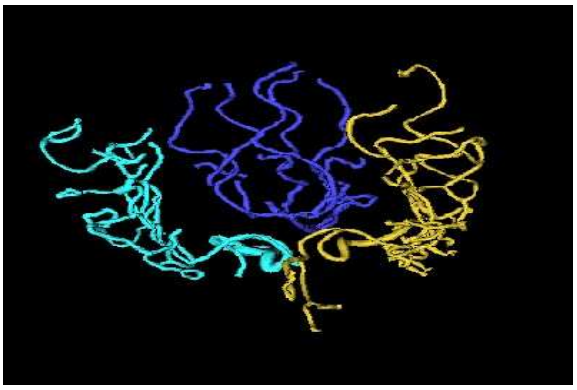

FIGURE 1.4. An example of brain blood vessels.

Because of the branching nature of blood vessel systems, a tree-structured data representation is very natural. This data set has 11 trees from 3 people. These are the left carotid, right carotid and vertebrobasilar system from each person, plus two smaller components from one person.

Each blood vessel branch is denoted as a node in the tree structure. For simplicity, here only a simple linear approximation of each branch is used. The attributes of the root node have the following form

$$[0, \text{three coordinates of the starting point}, \text{three coordinates of the ending point}];$$

while, the attributes of the non-root nodes are denoted as following

$$[p, 0, 0, 0, \text{three coordinates of the ending point}],$$

where p is the proportion parameter,

$$p = \frac{\text{Distance of starting point to point of attachment on its parent}}{\text{Distance of starting point to ending point on its parent}}.$$

The above data representation closely follows the form of the given data. However the segmentation algorithm makes some fairly arbitrary choices between "main vessel" and "branch". In later work, this issue will be explored by other representations of the data.

One approach is to embed the blood vessel data into a population of hierarchical tree structures by focussing on the blood vessel segments between the successive intersections. Take the blood vessel segment between the starting point (denoted by α_1) of the root blood vessel and the first point of intersection (denoted by β_1) as the first node, v_1. The attributes of the node v_1 are the coordinates of the starting point (i.e., α_1) and the ending point (i.e., β_1) of this first segment. The two vessel segments starting at the first intersection (i.e., the point β_1) will be nodes v_2 and v_3. The segment with larger radius at the starting point will be taken as the node v_2. The attributes of the nodes v_2 and v_3 are the coordinates of the ending points, denoted by β_2 and β_3 respectively. Note that the coordinates of the point β_1 is the "implicit attributes" of the nodes v_2 and v_3, and when the trees are reconstructed for visualization, the coordinates of β_1 are used. This assures that after operations, such as projection, the result is still a well defined tree. Iteratively, other vessel segments between successive intersections are assigned to nodes. This tree representation is much different from that above, and will be investigate in later work. In all examples studied here, the previous representation was used.

For computational speed (see the algorithm in Section 3.7 for more discussion), only a subtree (up to level 2 and three nodes) of each element among those 11 trees

is considered (see Figure 1.6). There are only two tree structures of these 11 trees, which are called Type I and Type II, shown in Figure 1.5. Among these 11 blood vessel trees, seven trees have Type I structure and four trees have Type II structure.

FIGURE 1.5. Two types of tree structures of the reduced blood vessel trees.

Figure 1.6 shows the reduced blood vessel trees for three people. The trees with thicker line are the median-mean trees (central tree with nodal attributes, see Section 3.3 for more discussion) in each figure. Note that the median-mean trees are "central" in terms of structure, size, and location, for each of the three people.

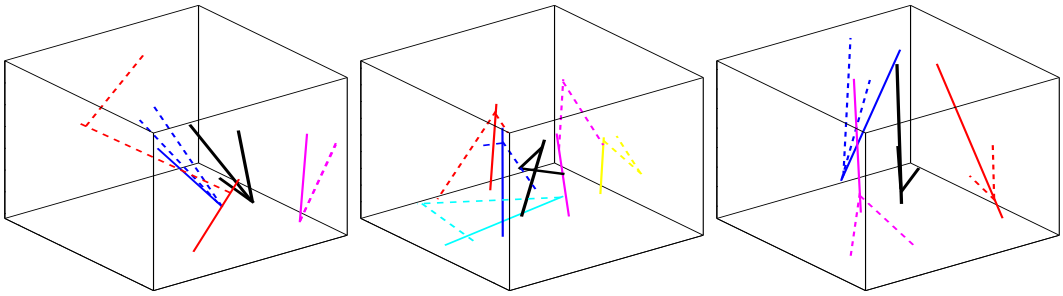

FIGURE 1.6. Reduced blood vessel trees (thin colored lines) and the median-mean trees (thicker black line) for each person. Root nodes are solid and children are dashed.

These trees are combined into a larger population in Figure 1.7. Again, the median-mean tree of the larger population is shown as a thicker black line. This time the median-mean tree is surprisingly small. This will be understood through careful analysis of the variation about the median-mean tree.

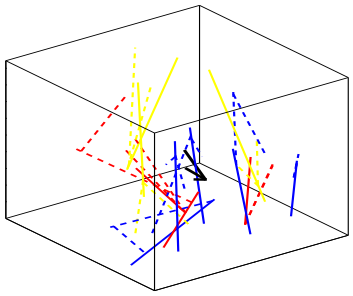

FIGURE 1.7. Combined population of reduced blood vessel trees and the median-mean tree.

Next, the tree version Principal Component Analysis (see Section 3.6) will be applied to the full blood vessel tree sample (denoted by T, shown in Figure 1.7).

An analog of one-dimensional representation, "treeline", is defined, which plays the role of "line" (a one-dimensional subspace in Euclidean space), in tree space (see Sections 2.4 and 3.5). Two useful treelines, the structure treeline (see Definition 3.5.1) and the attribute treeline (see Definition 3.5.3), are used in this dissertation.

The principal structure representation $l = \{u_0, u_1, u_2\}$ (i.e., structure treeline, see Definition 3.6.1) is shown in Figure 1.8 (structure only, without attributes) and Figure 1.9 (with attributes). On this treeline, the tree u_0 only has the root node and the right child. The trees u_1 and u_2 add one left child on u_0 and u_1 respectively. This shows that the dominant component of topological variation is towards branching in this direction.

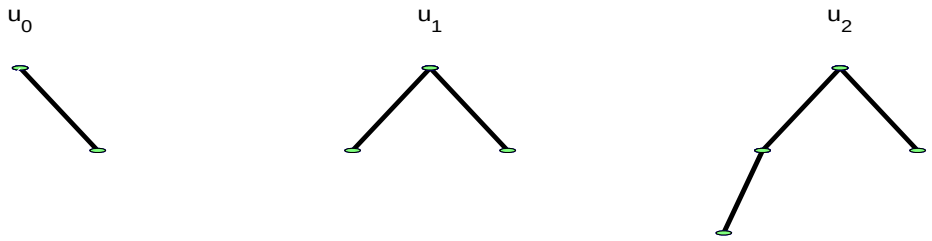

FIGURE 1.8. Principal structure treeline $l = \{u_0, u_1, u_2\}$ without nodal attributes.

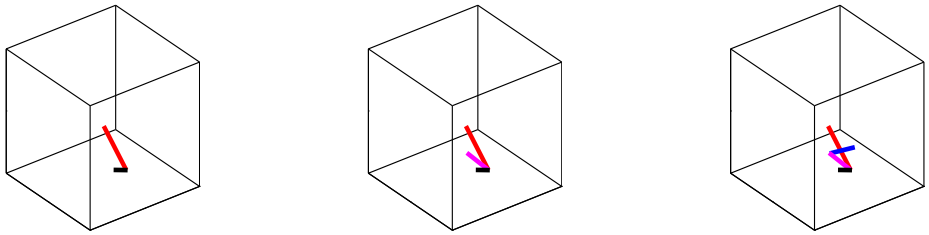

FIGURE 1.9. Principal structure treeline $l = \{u_0, u_1, u_2\}$ with nodal attributes.

Next, consider the principal attribute direction (see Definition 3.6.3). The induced attribute treelines passing through the median-mean tree and through the average support tree (see Definition 3.5.5) are shown in Figure 1.10 and Figure 1.12. There are six subplots in each figure. Each subplot depicts one location on the attribute treeline.

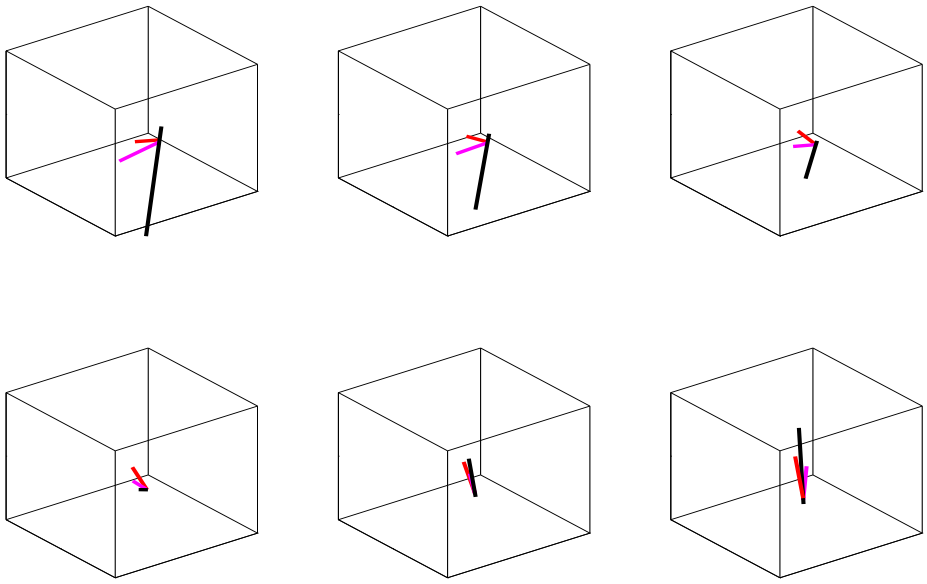

FIGURE 1.10. Induced attribute treeline passing through the median-mean tree.

In Figure 1.10, from the upper left subplot to the upper right one, it shows that the orientation of the main root (solid black line) is coming towards a horizontal line, and at the same time the length of the main root becomes shorter; while, from the

lower left to the lower right one, the main root (solid black line) flips in the opposite direction from the horizontal line and the length of the main root becomes longer. It shows that the main root flips over. This was a surprising feature of the population. Careful investigation showed that the given data set does not unanimously record the data according to the direction of blood flow. Some of them have the same direction; while, some of them have the inverse direction. Also, this can be verified from the projection coefficients of all 11 trees on the attribute treeline passing through the median-mean tree (shown in Figure 1.11). This shows that, there are two groups with a gap in the middle, six trees with negative projection coefficients and five with positive ones. This also shows that, no trees correspond to the fourth frame in Figure 1.10, with a very short root, as can be seen in the raw data in Figure 1.7.

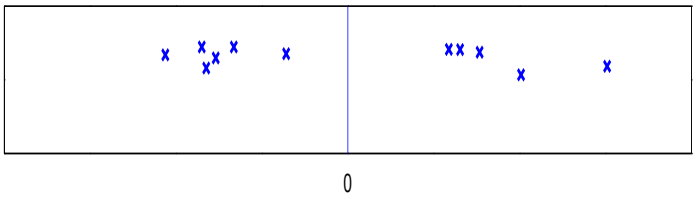

FIGURE 1.11. Projection coefficients of 11 trees on the attribute treeline passing through the median-mean tree.

Figure 1.12 shows the attribute treeline passing through the average support tree (see Section 3.5). Similar to Figure 1.10, the six frames show that the main root flips over and the length of the main root becomes shorter (three subplots on the top row) then becomes longer (three subplots on the bottom row). Similar to Figure 1.11, Figure 1.13 shows that these 11 trees are divided into two groups by projection on the attribute treeline passing through the average support tree with a gap in the middle.

In this example, the tree version PCA found a surprising characteristic of the population that there are two different orientations about the blood flow in the data

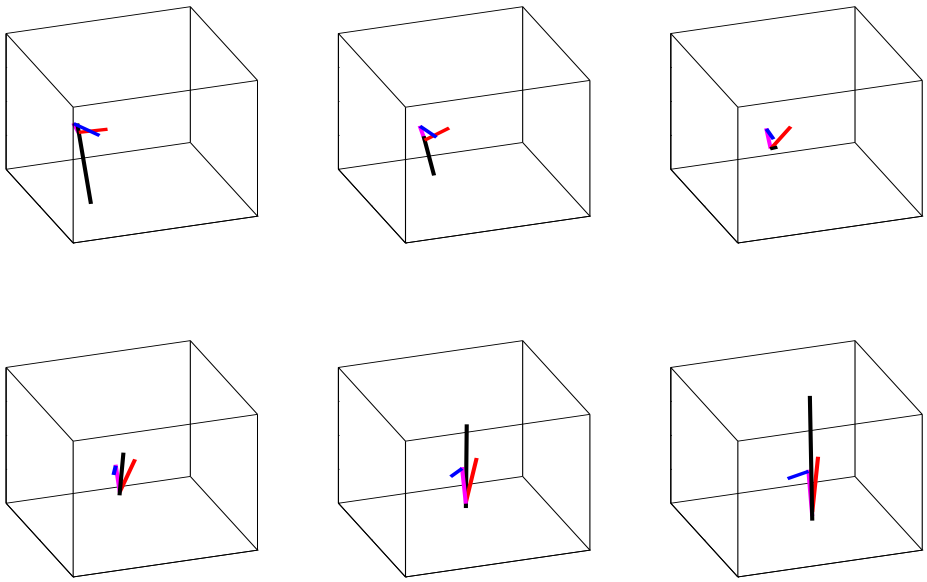

FIGURE 1.12. Induced attribute treeline passing through the average support tree.

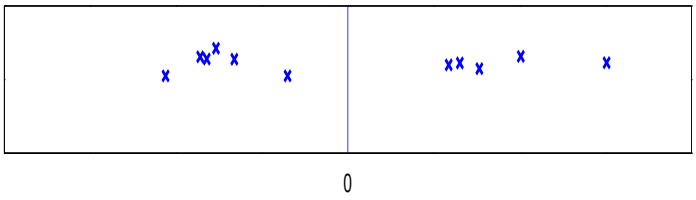

FIGURE 1.13. Projection coefficients of 11 trees on the attribute treeline passing through the average support tree.

set. This dominates the total variation, perhaps obscuring population features of more biological interest.

Also, the tree version PCA found interesting clusters in the data. The projections onto the dominant treeline provided a clear view of clustering. According to the projections on the two different types of treelines, the groupings may vary.

CHAPTER 2

Statistical Analysis on the Binary Tree Space without Nodal Attributes

2.1. Basic Definitions

In this research, a population of abstract complex multi-figural objects is considered. The single observation in this population is called a "tree". What is a "tree"?

DEFINITION 2.1.1. A **tree** is a simple graph such that there is a unique path (a set of edges) between every pair of nodes (vertices). The set of nodes and edges are denoted by V and E, respectively. Each edge can be denoted by an ordered pair of nodes.

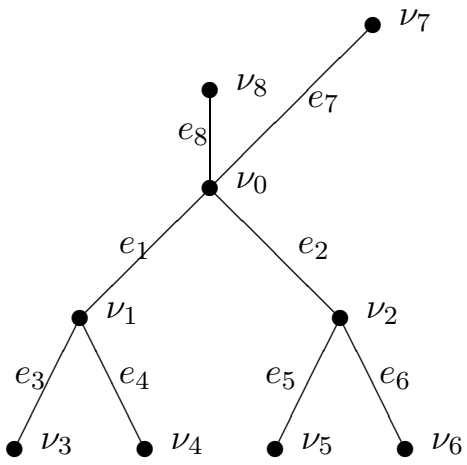

FIGURE 2.1. An example of tree.

DEFINITION 2.1.2. The **root** is one designated node. The **level of a node** is the length (number of edges) of the path to the root.

The only node with level 0 is the root. The maximum level of the nodes is called **level of the tree**. A tree with one node is called a trivial tree; otherwise, it is called the non-trivial tree.

EXAMPLE 2.1.1. The tree t in Figure 2.1 has 9 nodes and 8 edges.

$$V = \{\nu_0, \nu_1, \nu_2, \ldots, \nu_8\} \text{ and } E = \{e_1, e_2, \ldots, e_8\}.$$

Let ν_0 be the root of tree t. Note that $\{\nu_1, \nu_2, \nu_7, \nu_8\}$ have level 1, and $\{\nu_3, \nu_4, \nu_5, \nu_6\}$ have level 2. Thus, the level of the tree t is 2.

DEFINITION 2.1.3. A **binary tree** is a tree $t = (V, E)$, together with an edge labelling function $f : E \to \{0, 1\}$ such that every node has at most one edge incident from it labelled with 0 (called a left edge) and at most one edge incident from it labelled with 1 (called a right edge). For each left edge (ν, ω), ν is called the parent of ω and ω is called the left child of ν. Similarly, the right child is defined. A tree $t_1 = (V_1, E_1)$ is called a **subtree** of t, denoted by $t_1 \subseteq t$, if $V_1 \subseteq V$, $E_1 \subseteq E$ and the root of tree t is in the set V_1.

For simplicity, the **binary tree** will be considered first.

DEFINITION 2.1.4. Let t be a binary tree. Every node ω in t has a unique **level-order index** ($ind(\omega)$), defined as follows:

- If ω is the root, let $ind(\omega) = 1$;
- If ω is the left child of the node ν, let $ind(\omega) = 2 \times ind(\nu)$;
- Otherwise, if ω is the right child of the node ν, let $ind(\omega) = 2 \times ind(\nu) + 1$.

DEFINITION 2.1.5. A **complete binary tree** is a binary tree for which the level-order indices of the nodes form a complete interval $1, 2, \ldots, n$ of integers. Otherwise, it is called an **incomplete tree**.

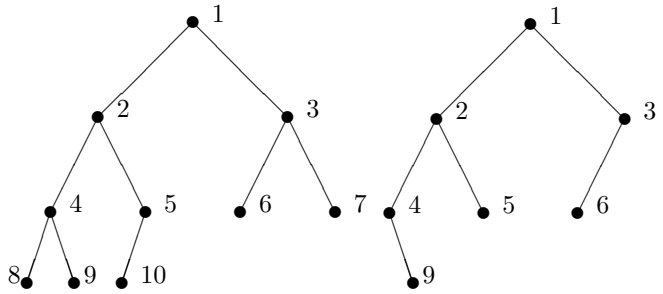

FIGURE 2.2. Examples of binary trees. The numbers are level-order indices.

EXAMPLE 2.1.2. In Figure 2.2, the tree on the left panel shows a complete binary tree and the tree on the right panel shows an incomplete binary tree.

DEFINITION 2.1.6. Let t be a binary tree. The set of all possible level-order indices of the i^{th} level is denoted by I_i and $I_i = \{2^i, 2^i + 1, \ldots, 2^{i+1} - 1\}$.

EXAMPLE 2.1.3. For any binary tree t, $I_0 = \{1\}$ and $I_2 = \{4, 5, 6, 7\}$.

REMARK 2.1.1. For a binary tree t, the set of level-order indices of the nodes is denoted by $Ind(t)$.

REMARK 2.1.2. For any binary tree t, the set of level-order indices of the nodes on the i^{th} level (denoted by $t(i)$) is a subset of I_i.

DEFINITION 2.1.7. Let t_1 and t_2 be two binary trees. A binary tree t is called the **union (intersection)** of binary trees t_1 and t_2 if the interval formed by the level-order indices of the nodes in tree t is a union (intersection) of those of binary trees t_1 and t_2. That is, $Ind(t) = Ind(t_1) \cup Ind(t_2)$ (or $Ind(t) = Ind(t_1) \cap Ind(t_2)$). It is denoted by $t = t_1 \cup t_2$ (or $t = t_1 \cap t_2$).

REMARK 2.1.3. The definitions of union and intersection of binary trees can be generalized to any tree population where a "level-order index" can be defined.

REMARK 2.1.4. All the definitions of the operations on the binary trees are based on the level-order indices of the nodes.

REMARK 2.1.5. The union tree provides a convenient framework for the development of the notion of "subspace" in tree space. To study a tree sample with n elements, we can consider the "subspace", in which every element is a subtree of the union tree (of those n elements).

2.2. Metric on Binary Trees without Nodal Attributes

In the previous section, some basic definitions were introduced. But, statistical analysis has not been developed for the binary tree population. There are two fundamental issues that will be addressed. The first issue for statistical analysis is, appropriate definition of a "center point" of the binary tree population.

A notion of "center point" of a population is the binary tree which is the "closest to all other trees". This requires a metric on the space of binary trees. Thus, the second fundamental issue is the definition of a distance between two trees.

The basic idea for a simple metric is illustrated by the two trees t_1 and t_2 shown in Figure 2.3. The tree t_2 can be obtained by adding two nodes and deleting one from the tree t_1; that is, the smallest number of addition and deletion of nodes from one tree to the other is 3. It will be shown that a tree metric can be defined, using this based on the total number of such deletions and additions.

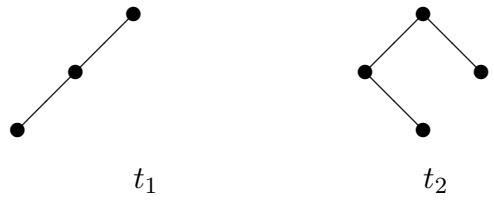

FIGURE 2.3. Binary trees t_1 and t_2.

For any two trees s and t, the difference of the i^{th} level will be studied, which will be a component of the binary tree metric.

DEFINITION 2.2.1. The total number of nodes which belong to $s(i) \triangle t(i)$ is called the difference of the i^{th} level (denoted by d_i), where $s(i) \triangle t(i) = (s(i) \cap \overline{t(i)}) \cup (t(i) \cap$

$\overline{s(i)})$ and $\overline{s(i)}$ is the complement of $s(i)$ in I_i. In other words,

$$d_i = d_i(s,t) = \sum_{k \in I_i} 1\{k \in s(i) \triangle t(i)\}.$$

Let L_s and L_t be the levels of tree s and t respectively. For any integer $n > \max(L_s, L_t)$, $d_n = 0$.

EXAMPLE 2.2.1. For the two binary trees t_1 and t_2 shown in Figure 2.3, $d_0(t_1,t_2) = 0$, $d_1(t_1,t_2) = 1$ and $d_2(t_1,t_2) = 2$.

EXAMPLE 2.2.2. In Section 1.2, there are two tree structures in the blood vessel data, Type I and Type II. Between those two tree structures, the difference of level 0 is 0, level 1 is 1 and level 2 is 1.

THEOREM 2.2.1. *d_i is a pseudo-metric on the binary trees.*

Proof. Suppose s, t and w are three binary trees.

(1) [Identity]

$$\begin{aligned} d_i(s,s) &= \sum_{k \in I_i} 1\{k \in s(i) \triangle s(i)\} \\ &= 0 \end{aligned}$$

(2) [Symmetry]

$$\begin{aligned} d_i(s,t) &= \sum_{k \in I_i} 1\{k \in s(i) \triangle t(i)\} \\ &= \sum_{k \in I_i} 1\{k \in t(i) \triangle s(i)\} \\ &= d_i(t,s) \end{aligned}$$

(3) [Triangle inequality]

Note that

$$\begin{aligned} d_i(s,w) &= \sum_{k\in I_i} 1\{k \in s(i) \triangle w(i)\} \\ &= \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{w(i)}\} + \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{s(i)}\} \\ &= \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{w(i)} \cap t(i)\} + \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{w(i)} \cap \overline{t(i)}\} \\ &+ \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{s(i)} \cap t(i)\} + \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{s(i)} \cap \overline{t(i)}\} \end{aligned}$$

Similarly,

$$\begin{aligned} d_i(w,t) &= \sum_{k\in I_i} 1\{k \in t(i) \triangle w(i)\} \\ &= \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{w(i)}\} + \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{t(i)}\} \\ &= \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{w(i)} \cap s(i)\} + \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{w(i)} \cap \overline{s(i)}\} \\ &+ \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{t(i)} \cap s(i)\} + \sum_{k\in I_i} 1\{k \in w(i) \cap \overline{t(i)} \cap \overline{s(i)}\} \end{aligned}$$

$$\begin{aligned} d_i(s,t) &= \sum_{k\in I(i)} 1\{k \in s(i) \triangle t(i)\} \\ &= \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{t(i)}\} + \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{s(i)}\} \\ &= \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{t(i)} \cap w(i)\} + \sum_{k\in I_i} 1\{k \in s(i) \cap \overline{t(i)} \cap \overline{w(i)}\} \\ &+ \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{s(i)} \cap w(i)\} + \sum_{k\in I_i} 1\{k \in t(i) \cap \overline{s(i)} \cap \overline{w(i)}\} \end{aligned}$$

Therefore,

$$
\begin{aligned}
& d_i(s,w) + d_i(w,t) - d_i(s,t) \\
& = 2\sum_{k\in I_i} 1\{k \in w(i) \cap \overline{s(i)} \cap \overline{t(i)}\} + 2\sum_{k\in I_i} 1\{k \in s(i) \cap t(i) \cap \overline{w(i)}\} \\
& \geq 0
\end{aligned}
$$

□

REMARK 2.2.1. From Example 2.2.1, $d_0(t_1, t_2) = 0$ where $t_1 \neq t_2$. Hence, d_0 is a pseudo-metric not a metric. Similarly, for $i = 1, 2, \ldots$, d_i is not a metric because there exist two different binary trees s and t such that $d_i(s,t) = 0$.

For any two binary trees s and t without nodal attributes, define the metric (Theorem 2.2.2 establishes that this is indeed a metric)

$$d_I(s,t) = \sum_{i=0}^{\infty} d_i(s,t), \tag{2.1}$$

where "I" means "integer" to contrast with a "fractional part" coming later. Then $d_I(s,t)$ is the total difference between two binary trees s and t.

The distance $d_I(s,t)$ is the sum of the differences of each level of two trees. Therefore, $d_I(s,t)$ counts the total number of nodes which show up only in either s or t, but not both of them. That is,

$$d_I(s,t) = \sum_{k=1}^{\infty} 1\{k \in Ind(s) \triangle Ind(t)\}. \tag{2.2}$$

REMARK 2.2.2. Since d_I is always an integer, it is called the integer tree metric.

EXAMPLE 2.2.3. Let t_1 and t_2 be the binary trees shown in Figure 2.3. $d_0 = 0$, $d_1 = 1$ and $d_2 = 2$. Therefore, the integer tree metric is

$$d_I(t_1, t_2) = \sum_{i=0}^{2} d_i(t_1, t_2) = 3.$$

Also,

$$Ind(t_1) = \{1, 2, 4\} \text{ and } Ind(t_2) = \{1, 2, 3, 5\}.$$

Therefore,

$$Ind(t_1) \triangle Ind(t_2) = \{3, 4, 5\}$$

and by Equation (2.2),

$$d_I(t_1, t_2) = 3.$$

EXAMPLE 2.2.4. For the blood vessel data in Section 1.2, the integer metric between two types of trees is 2.

THEOREM 2.2.2. *$d_I(s,t) = \sum_0^\infty d_i(s,t)$ is a metric on the binary tree space without nodal attributes.*

Proof. Suppose s, t and w are three binary trees without nodal attributes.

(1) [Identity]

It is easy to see that

$$d_I(s,s) = \sum_{i=0}^{\infty} d_i(s,s) = 0.$$

On the other hand, for two binary trees s and t, if $d_I(s,t) = 0$, then s and t must have the same tree structures because each item in the summation is zero. Hence, $s = t$.

(2) [Symmetry]

From Theorem 2.2.1, d_i is a pseudo-metric for all i; that is, $d_i(s,t) = d_i(t,s)$, $\forall i$. Therefore,

$$\begin{aligned} d_I(s,t) &= \sum_{i=0}^{\infty} d_i(s,t) \\ &= \sum_{i=0}^{\infty} d_i(t,s) \\ &= d_I(t,s) \end{aligned}$$

(3) [Triangle inequality]

By Theorem 2.2.1, $d_i(s,t) \leq d_i(s,w) + d_i(w,t)$ for all $i = 0, 1, \ldots$.

$$\begin{aligned} d_I(s,t) &= \sum_{i=0}^{\infty} d_i(s,t) \\ &\leq \sum_{i=0}^{\infty} (d_i(s,w) + d_i(w,t)) \\ &\leq d_I(s,w) + d_I(w,t) \end{aligned}$$

□

REMARK 2.2.3. There is an intuitive representation of the integer tree metric. It is the smallest total number of added and deleted nodes required to move from one binary tree to the other.

2.3. Finding the Median Tree on the Binary Tree Space without Nodal Attributes

In Section 2.2, an integer metric d_I was defined on the binary tree space without nodal attributes. Next, consider the question presented in the previous section, what is the "center point" of a sample of binary trees?

From now on, denote the set of all binary trees by $\mathcal{T}$ and the finite sample by $T = \{t_1, t_2, \ldots, t_n\}$.

DEFINITION 2.3.1. A tree is a **minimizer tree** according to the metric d_I if it minimizes $\sum_{i=1}^{n} d_I(t, t_i)$ over all binary trees $t \in \mathcal{T}$.

DEFINITION 2.3.2. A tree is called a **full (binary) tree** if it contains all the nodes in the binary tree sample T.

DEFINITION 2.3.3. The full tree with the minimum number of nodes is called **support (binary) tree**.

By the definitions of the metric d_I and the minimizer tree, the following property is given.

PROPOSITION 2.3.1. *A minimizer tree according to d_I cannot have a node which does not appear in the sample. That is, a minimizer tree is contained in the support binary tree.*

THEOREM 2.3.2. *If a tree s is a minimizer tree according to the metric d_I, then all the nodes of s must appear at least $\frac{n}{2}$ times in the binary tree sample T. Moreover, the minimizer tree s (according to d_I) must contain all the nodes, which appear more than $\frac{n}{2}$ times, and may contain any subset of nodes that appear exactly $\frac{n}{2}$ times.*

Proof. Let s be a minimizer tree according to the integer tree metric d_I. Suppose some of the nodes in s appear less than $\frac{n}{2}$ times and ν is the node with the largest level among all of those nodes. If a node appears less than $\frac{n}{2}$ times, so do its children. Thus, ν must be a terminal node of s.

For the binary tree $s' = s\backslash\{\nu\}$, the following equation is satisfied

$$\sum_{i=1}^{n} d_I(s', t_i) = \sum_{i=1}^{n} d_I(s, t_i) + n_\nu - (n - n_\nu), \tag{2.3}$$

where n_ν=#{appearance of the node ν in the sample T}. Since $n_\nu < \frac{n}{2}$,

$$\sum_{i=1}^{n} d_I(s', t_i) < \sum_{i=1}^{n} d_I(s, t_i),$$

which is a contradiction with the assumption that s is the minimizer tree.

From the proof above, if $n_\nu = \frac{n}{2}$, then $\sum_{i=1}^{n} d_I(s', t_i) = \sum_{i=1}^{n} d_I(s, t_i)$; that is, s' is also a minimizer tree. Therefore, the minimizer tree may contain any subset of the nodes that appear exactly $\frac{n}{2}$ times.

Finally, a proof is given of the fact that the minimizer binary tree s contains all the nodes which appear more than $\frac{n}{2}$ times.

Suppose the node ω appears more than $\frac{n}{2}$ times in the sample T and $\omega \notin s$. Without loss of generality, suppose that ω is a children of some node in the binary tree s. Otherwise, choose one of its ancestor nodes.

For the binary tree $s'' = s \cup \{\omega\}$, the following equation is satisfied

$$\sum_{i=1}^{n} d_I(s, t_i) = \sum_{i=1}^{n} d_I(s'', t_i) + n_\omega - (n - n_\omega), \tag{2.4}$$

where n_ω=#{appearance of the node ω in the sample T}. Since $n_\omega > \frac{n}{2}$,

$$\sum_{i=1}^{n} d_I(s'', t_i) < \sum_{i=1}^{n} d_I(s, t_i),$$

which is a contradiction with the assumption that s is the minimizer tree. □

COROLLARY 2.3.3. *If n is an odd number, then there is a unique minimizer tree (according to d_I), which consists of all the nodes with appearance more than $\frac{n}{2}$ times.*

REMARK 2.3.1. Banks and Constantine independently developed essentially the same notion of "central tree" and the algorithm of finding such tree, which is called the **majority rule** (see Banks and Constantine, 1998, page 204). In this dissertation, the algorithm derived in Theorem 2.3.2 is called the majority rule.

REMARK 2.3.2. Formulating this concept in statistical terms, the minimizer tree is called the **median tree** of the binary tree sample T.

REMARK 2.3.3. If n is an even number, then the median binary tree may be not unique because some nodes may have appearance number equal to $\frac{n}{2}$.

EXAMPLE 2.3.1. Among the 11 blood vessel trees in Section 1.2, seven trees have Type I structure and four trees have Type II structure. According to the majority rule, the median tree, the second tree in Figure 1.8, has the Type I tree structure. The tree structure of the median tree is the same as that of the median-mean tree (the second tree in Figure 1.9, see Definition 3.3.1 for more discussion).

DEFINITION 2.3.4. The median binary tree (according to the integer tree metric d_I) with the smallest number of nodes is called **minimal median binary tree**.

THEOREM 2.3.4. *The minimal median binary tree (according to the integer tree metric d_I) is unique.*

Proof. By the majority rule, the median binary tree contains all of the nodes with appearance number greater than $\frac{n}{2}$ and may contain any subset of the nodes with appearance number equal to $\frac{n}{2}$. Therefore, for any median binary tree, the unique minimal median binary tree can be obtained by deleting those nodes with $\frac{n}{2}$ appearance time. □

Since the integer tree metric d_I only counts the total number of nodes in the symmetric set of their level-order index sets. The following theorem provides a simple approach to easy calculations.

THEOREM 2.3.5. *T is a sample of binary trees with size n; that is,*

$$T = \{t_1, t_2, \ldots, t_n\}.$$

Suppose the full tree has level-order index set I and the corresponding numbers of appearance (of the nodes in I) are $n_i, i \in I$. Then,

$$\begin{aligned}\sum_{i=1}^{n} d_I(t_i, m) &= \sum_{i \in I} [n_i \cdot 1\{n_i \leq \frac{n}{2}\} + (n - n_i) \cdot 1\{n_i > \frac{n}{2}\}] \\ &= \sum_{i \in I} [\frac{n}{2} - \left|\frac{n}{2} - n_i\right|]\end{aligned}$$

where m is the median tree of this sample T.

Proof. For any node with level-order index j in the full tree, if $n_j > \frac{n}{2}$, then it will be included in the median binary tree by the majority rule. There are $n - n_j$ binary trees in T which do not have nodes with level-order index j. Hence, the contribution of the j^{th} node to the total sum $\sum_{i=1}^{n} d_I(t_i, m)$ would be $n - n_j$. If $n_j = \frac{n}{2}$, no matter

that j^{th} node is included in the median binary tree, the contribution to the total sum is $n_j = \frac{n}{2}$. Otherwise, this node will not be included in the median binary tree and its contribution to the sum would be n_j.

Furthermore, if $n_i \leq \frac{n}{2}$,

$$\frac{n}{2} - \left|\frac{n}{2} - n_i\right| = \frac{n}{2} - (\frac{n}{2} - n_i) = n_i.$$

Otherwise,

$$\frac{n}{2} - \left|\frac{n}{2} - n_i\right| = \frac{n}{2} + (\frac{n}{2} - n_i) = n - n_i.$$

□

EXAMPLE 2.3.2. T is a sample of binary trees with $n = 22$ members, $t_1, t_2, \ldots t_{22}$. There are four types of binary trees in T shown in Figure 2.4. Let $N_1 = 4, N_2 = 5, N_3 = 7, N_4 = 6$ be the numbers of trees of type I, II, III, IV respectively.

FIGURE 2.4. An example of a binary tree sample.

FIGURE 2.5. Support tree t_{sup} and median tree m of binary tree sample T.

The support binary tree of the sample T is shown in the left panel in Figure 2.5. The number of appearances of each node are $n_1 = 22, n_2 = 9, n_3 = 13, n_4 = 9, n_5 =$

$5, n_6 = 6, n_7 = 13$. According to the majority rule, the median binary tree is m shown in the right panel in Figure 2.5.

Then by Theorem 2.3.5, the total distance of binary trees in T to the median tree m is

$$\begin{aligned}\sum_{i=1}^{22} d_I(t_i, m) &= \sum_{i=1}^{7}(11 - |11 - n_i|) \\ &= 47.\end{aligned}$$

In Euclidean space, the total variation of a sample can be measured by the sum of squared distances to its sample mean. In the tree space without nodal attributes, the integer metric d_I can be written as a sum of zeros and ones (see Equation (2.2)). Therefore,

$$\sum_{k=1}^{\infty}(1\{k \in Ind(s) \triangle Ind(t)\})^2 = \sum_{k=1}^{\infty} 1\{k \in Ind(s) \triangle Ind(t)\} = d_I(s,t). \tag{2.5}$$

Thus, in the tree space without nodal attributes, the sum of distances to the median tree can be considered as the total variation of the sample. That is, the total variation is

$$\sum_{i=1}^{n} d_I(t_i, m).$$

2.4. Treeline and Projection in the Binary Tree Space without Nodal Attributes

In the binary tree space, each tree can be viewed as a point. Unlike Euclidean space, the binary tree space is a nonlinear space according to the previous metric d_I. Hence, the principal component analysis (PCA) in Euclidean space may not be applicable in the nonlinear binary tree space. So, the question is " how can an analogous way be developed to construct a manifold in binary tree space which consists of some binary trees that plays the role of a 'line', one-dimensional subspace in Euclidean space? "

First, consider the binary tree space without nodal attributes, $\mathcal{T}$. In this case, only the integer metric d_I will be used.

DEFINITION 2.4.1. Suppose $l = \{u_0, u_1, u_2, ...\}$ is a sequence of binary trees. l is called a **treeline** starting from u_0 if for $i = 1, 2, 3, \ldots$

(1) the tree u_{i-1} can be obtained by deleting a terminal node (denoted by ν_i) from u_i;

(2) the node ν_{i-1} is the parent of ν_i;

(3) there does not exist a subtree of u_0, denoted as u, such that u can be obtained by deleting some ancestor nodes of ν_1.

REMARK 2.4.1. From another point of view, the tree u_i is obtained by adding a node ν_i on the tree u_{i-1}.

DEFINITION 2.4.2. A treeline l is called **passing through** the tree u if the tree u is an element of the binary tree set l; i.e. $u \in l$.

EXAMPLE 2.4.1. In Figure 2.6, the tree u_1 is obtained by adding a node, ν_1, with level-order index 2 from the tree u_0. Similarly, the u_2 is obtained by adding a node, ν_2, with level-order index 4 from u_1. Therefore, there exists a treeline l passing through u_0, u_1 and u_2.

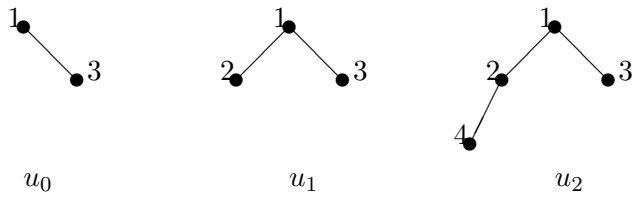

FIGURE 2.6. A tree sequence $l_0 = \{u_0, u_1, u_2, \ldots\}$ illustrating the idea of a treeline.

Recall that, from the blood vessel data in Section 1.2, Figure 1.8 shows the first three elements of a treeline without nodal attributes.

EXAMPLE 2.4.2. In Figure 2.7, the binary tree u_1 is obtained by adding a node with level-order index 2 from the binary tree u_0; while, the binary tree u_2 is obtained by adding a node with level-order index 3 from u_1. Those two adding nodes are on the same level of a binary tree. Therefore, there does not exist any treeline passing through u_0, u_1, u_2 and u_3.

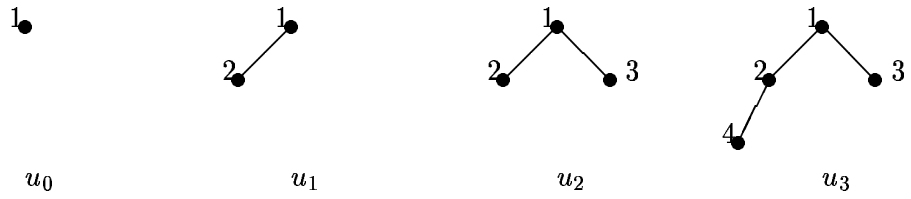

FIGURE 2.7. A tree sequence $l_1 = \{u_0, u_1, u_2, u_3 \ldots\}$ that is not a treeline.

DEFINITION 2.4.3. Suppose v is a tree and l is a treeline as in Definition 2.4.1. The tree $w \in l$ is called **the projection** of v on the treeline l, if w is the minimizer of $d_I(v, t)$ where t runs over all the binary trees on the treeline l.

PROPOSITION 2.4.1. *The projection of a tree on a treeline exists and is unique.*

Proof. Suppose $l = \{u_0, u_1, u_2, \ldots\}$ is a treeline. Let p be the index of the smallest d_I-closest, to the tree t, member of treeline l; i.e.,

$$p = inf\{i : d_I(u_i, t) \leq d_I(u_j, t), j = 0, 1, 2, \ldots\}.$$

First, consider the two elements u_p and u_{p+1} on the treeline l. By definition of the treeline, the tree u_p can be obtained by deleting a node ν_{p+1} from the tree u_{p+1}. It will now be shown that, $\nu_{p+1} \notin Ind(t)$. Otherwise,

$$d_I(u_{p+1}, t) = d_I(u_p, t) - 1,$$

which is a contradiction with the definition of p. Thus, $\nu_{p+1} \notin Ind(t)$, and

$$d_I(u_{p+1}, t) = d_I(u_p, t) + 1. \tag{2.6}$$

Iteratively, for $i = 1, 2, \ldots$, the tree u_{p+i} can be obtained by deleting a node ν_{p+i+1} from the tree u_{p+i+1}. The node ν_{p+i+1} is an offspring node of the node ν_{p+1}. Since $\nu_{p+1} \notin Ind(t)$, for $i = 1, 2, \ldots$, $\nu_{p+i+1} \notin Ind(t)$. Hence,

$$d_I(u_{p+i+1}, t) = d_I(u_{p+i}, t) + 1. \tag{2.7}$$

Next, consider the two trees u_{p-1} and u_p on the treeline l. The tree u_{p-1} can be obtained by deleting a node ν_p from the tree u_p. It will now be shown that, $\nu_p \in Ind(t)$. Otherwise,

$$d_I(u_{p-1}, t) = d_I(u_p, t) - 1,$$

which is a contradiction with the definition of p. Hence, $\nu_p \in Ind(t)$, and

$$d_I(u_{p-1}, t) = d_I(u_p, t) + 1. \tag{2.8}$$

Iteratively, for $i = 1, 2, \ldots, p-1$, the tree u_{p-i-1} can be obtained by deleting a node ν_{p-i} from the tree u_{p-i}. The node ν_{p-i} is an ancestor node of the node ν_p. Since $\nu_p \in Ind(t)$, for $i = 1, 2, \ldots, p-1$, $\nu_{p-i} \in Ind(t)$. Hence,

$$d_I(u_{p-i-1}, t) = d_I(u_{p-i}, t) + 1. \tag{2.9}$$

Hence, there is a unique tree u_p such that, for $i \neq p$

$$d_I(u_i, t) > d_I(u_p, t). \tag{2.10}$$

That is, the projection exists and is unique. □

From Proposition 2.4.1, it is straightforward to define the projection function

$$w = P_l(v),$$

where l, w and v are given in Definition 2.4.3.

2.5. Principal Component Analysis on Binary Tree Space without Nodal Attributes

In classical statistics, the principal component analysis (PCA) is a useful tool to capture the features of a data set by decomposing the total variation to the center point. In PCA analysis, the first principal component indicates the direction which captures the largest variation of the data. Furthermore, several other orthogonal directions, which often highlight additional interesting aspects of the data, can be obtained. Now consider the similar problem in the binary tree space, how can a method to analyze the variation of the data set be developed?

From the previous section, in binary tree space, the treeline plays the role of "line", i.e. one-dimensional representation in Euclidean space. Recall that, for any tree sample T, the median binary tree m plays the role of "center point". So, is it possible to define a treeline l, the one-dimensional representation in binary tree space, passing through the median tree m such that it maximizes the sum

$$\sum_{i=1}^{n} d_I(m, P_l(t_i))? \tag{2.11}$$

Recall from Section 2.3 that, if the population size n is odd, then the median tree is unique which is also the minimal median tree (see Definition 2.3.4). Otherwise, if n is an even number, those nodes with appearance $\frac{n}{2}$ can be included in, or deleted from the median tree. So, the median tree is not unique; while the minimal median tree is still unique.

Note that, for a sample T, the total variation does not depend on the choice of the median trees. It is convenient to use the minimal median tree because it is unique and it is a subtree of any other median trees.

The Pythagorean Theorem is a fundamental theorem for the decomposition of the variation in the PCA in Euclidean space. Now, an analogous theorem, which is

called a tree version of the Pythagorean Theorem, is developed in the binary tree space without nodal attributes.

THEOREM 2.5.1. *Let T be a sample of trees of size n and $T = \{t_1, t_2, \ldots, t_n\}$. P_l is a projection function where l is a treeline running through a tree u. Then, $\forall t \in T$,*

$$d_I(u, P_l(t)) + d_I(P_l(t), t) = d_I(u, t). \tag{2.12}$$

Proof. Suppose that the treeline $l = \{u_0, u_1, u_2, \ldots\}$. Without loss of generality, assume that $P_l(t) = u_k$.

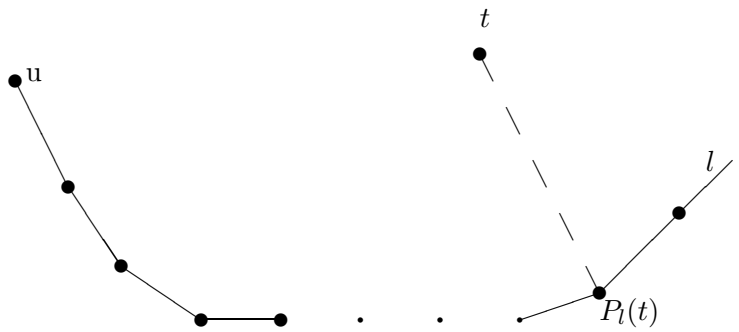

FIGURE 2.8. Projection of the tree t on the treeline l passing through u.

By the definition of treeline, there are two possible relations between u and u_k, either $u \subseteq u_k$ or $u_k \subseteq u$.

Case 1: $u \subseteq u_k$

Note that

$$Ind(u_k) \cap \overline{Ind(u)} \cap \overline{Ind(t)} = \emptyset. \tag{2.13}$$

In fact, if it is not empty, then there exists a terminal node of the tree u_k, ν, which is not included in the tree t and the tree u. Therefore, considering the binary tree $u_{k-1} = u_k \backslash \{\nu\}$,

$$d_I(u_{k-1}, u) = d_I(u_k, u) - 1,$$

which is a contradiction with the assumption that the tree u_k is the projection of the tree t.

Since the tree u is a subtree of the tree u_k, i.e. $Ind(u) \subseteq Ind(u_k)$, the following equations are established,

$$Ind(t) \cap \overline{Ind(u_k)} \cap Ind(u) = \emptyset \tag{2.14}$$

and

$$Ind(u_k) \cap \overline{Ind(t)} \cap Ind(u) = Ind(u) \cap \overline{Ind(t)}. \tag{2.15}$$

Using Equations (2.13), (2.14) and (2.15),

$$\begin{aligned}
& d_I(u, P_l(t)) + d_I(P_l(t), t) = d_I(u, u_k) + d_I(u_k, t) \\
= \;& \sum_j 1\{j \in Ind(u) \triangle Ind(u_k)\} + \sum_j 1\{j \in Ind(t) \triangle Ind(u_k)\} \\
= \;& \sum_j 1\{j \in Ind(u_k) \backslash Ind(u)\} + \sum_j 1\{j \in Ind(t) \triangle Ind(u_k)\} \\
= \;& \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(u)} \cap Ind(t)\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(u)} \cap \overline{Ind(t)}\} \\
& + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)}\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)}\} \\
= \;& \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(u)} \cap Ind(t)\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(u)} \cap \overline{Ind(t)}\} \\
& + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)} \cap Ind(u)\} + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)} \cap \overline{Ind(u)}\} \\
& + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)} \cap Ind(u)\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)} \cap \overline{Ind(u)}\} \\
= \;& \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u)}\} + \sum_j 1\{j \in Ind(u) \cap \overline{Ind(t)}\} \\
= \;& \sum_j 1\{j \in Ind(u) \triangle Ind(t)\} \\
= \;& d_I(u, t).
\end{aligned}$$

Case 2: $u_k \subseteq u$

Note that the tree u_k is the projection of the tree t, which implies

$$Ind(u) \cap \overline{Ind(u_k)} \cap Ind(t) = \emptyset, \tag{2.16}$$

by the same argument that was used to establish Equation (2.13).

Since the tree u_k is a subtree of the tree u, i.e. $Ind(u_k) \subseteq Ind(u)$, the following equations are established,

$$Ind(u_k) \cap \overline{Ind(t)} \cap \overline{Ind(u)} = \emptyset \tag{2.17}$$

and

$$\overline{Ind(u_k)} \cap \overline{Ind(u)} \cap Ind(t) = Ind(t) \cap \overline{Ind(u)}. \tag{2.18}$$

Using Equations (2.16), (2.17) and (2.18),

$$\begin{aligned}
& d_I(u, P_l(t)) + d_I(P_l(t), t) = d_I(u, u_k) + d_I(u_k, t) \\
= \;& \sum_j 1\{j \in Ind(u) \triangle Ind(u_k)\} + \sum_j 1\{j \in Ind(t) \triangle Ind(u_k)\} \\
= \;& \sum_j 1\{j \in Ind(u) \backslash Ind(u_k)\} + \sum_j 1\{j \in Ind(t) \triangle Ind(u_k)\} \\
= \;& \sum_j 1\{j \in Ind(u) \cap \overline{Ind(u_k)} \cap Ind(t)\} + \sum_j 1\{j \in Ind(u) \cap \overline{Ind(u_k)} \cap \overline{Ind(t)}\} \\
& + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)}\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)}\} \\
= \;& \sum_j 1\{j \in Ind(u) \cap \overline{Ind(u_k)} \cap Ind(t)\} + \sum_j 1\{j \in Ind(u) \cap \overline{Ind(u_k)} \cap \overline{Ind(t)}\} \\
& + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)} \cap Ind(u)\} + \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u_k)} \cap \overline{Ind(u)}\} \\
& + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)} \cap Ind(u)\} + \sum_j 1\{j \in Ind(u_k) \cap \overline{Ind(t)} \cap \overline{Ind(u)}\} \\
= \;& \sum_j 1\{j \in Ind(t) \cap \overline{Ind(u)}\} + \sum_j 1\{j \in Ind(u) \cap \overline{Ind(t)}\} \\
= \;& \sum_j 1\{j \in Ind(u) \triangle Ind(t)\} \\
= \;& d_I(u, t).
\end{aligned}$$

□

REMARK 2.5.1. The simplest form of the Pythagorean Theorem claims that in a right triangle with legs a, b and hypotenuse c, $c^2 = a^2 + b^2$.

In Euclidean space, a more general version of the Pythagorean Theorem provides the foundation of the ANOVA type of variation decomposition about the sample mean. For a set of numbers

$$\{x_1, x_2, \ldots, x_n\},$$

denote $\vec{x} = (x_1, x_2, \cdots, x_n)'$. Then

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} \overline{x}^2,$$

where $\overline{x}$ is the average of $x_1, x_2, \ldots, x_n$. That is,

$$\|(x_1, x_2, \ldots, x_n)'\|^2 = \|(x_1 - \overline{x}, \cdots, x_n - \overline{x})'\|^2 + \| \underbrace{(\overline{x}, \overline{x}, \ldots, \overline{x})'}_{n} \|^2, \tag{2.19}$$

where the vector

$$(\overline{x}, \overline{x}, \ldots, \overline{x})'$$

is the projection of the vector $\vec{x}$ onto the vector $(\vec{v})$ where all entries are equal, denoted by $P_{\vec{v}}(\vec{x})$. That is,

$$\|\vec{x} - \vec{0}\|^2 = \|\vec{x} - P_{\vec{v}}(\vec{x})\|^2 + \|P_{\vec{v}}(\vec{x}) - \vec{0}\|^2. \tag{2.20}$$

The tree version of Equation (2.20) is given in Theorem 2.5.1. In the tree space, the hypotenuse $(d_I(t, u))$ is the sum of the two legs, where d_I plays the role of squared Euclidean distance (see Theorem 2.5.1).

COROLLARY 2.5.2. *Let T be a sample of trees with median tree m. P_l is a projection function where l is a treeline running through m. Then, $\forall t \in T$,*

$$d_I(m, P_l(t)) + d_I(P_l(t), t) = d_I(m, t).$$

THEOREM 2.5.3. *Let $T = \{t_1, t_2, \ldots, t_n\}$ be a sample of trees. Maximizing the sum $\sum_{i=1}^{n} d_I(m, P_l(t_i))$ is equivalent to minimizing the sum $\sum_{i=1}^{n} d_I(P_l(t_i), t_i)$, where l runs over all treelines passing through the median tree m.*

Proof. From the tree version of the Pythagorean Theorem 2.5.1, for $i = 1, 2, \ldots, n$,

$$d_I(m, P_l(t_i)) + d_I(P_l(t_i), t_i) = d_I(m, t_i).$$

Therefore,

$$\sum_{i=1}^{n} d_I(m, P_l(t_i)) + \sum_{i=1}^{n} d_I(P_l(t_i), t_i) = \sum_{i=1}^{n} d_I(m, t_i).$$

□

DEFINITION 2.5.1. A treeline l_1, which maximizes the sum $\sum_{i=1}^{n} d_I(m, P_l(t_i))$ (or minimizes the sum $\sum_{i=1}^{n} d_I(P_l(t_i), t_i)$) over all treelines passing through the minimal median tree m, is called a **one-dimensional principal representation**, denoted by π_1.

REMARK 2.5.2. The one-dimensional principal representation, i.e. π_1, might not be unique, as shown in Example 2.5.1.

The following example also shows that the assumption of Definition 2.5.1 that only treelines passing through the median tree is very important. In particular, the sum of the distances to the projections on the treeline may be smaller for some other treeline.

EXAMPLE 2.5.1. Let $T = \{t_1, t_2, \ldots, t_9\}$ be a sample of binary trees as shown in Figure 2.9.

According to the majority rule, the median tree (m) has the same tree structure as that of tree t_5.

It is straightforward to see that, the one-dimensional principal representation is not unique. The one-dimensional principal representations, denoted by l_1 and l_2, are shown in Figures 2.10 and 2.11.

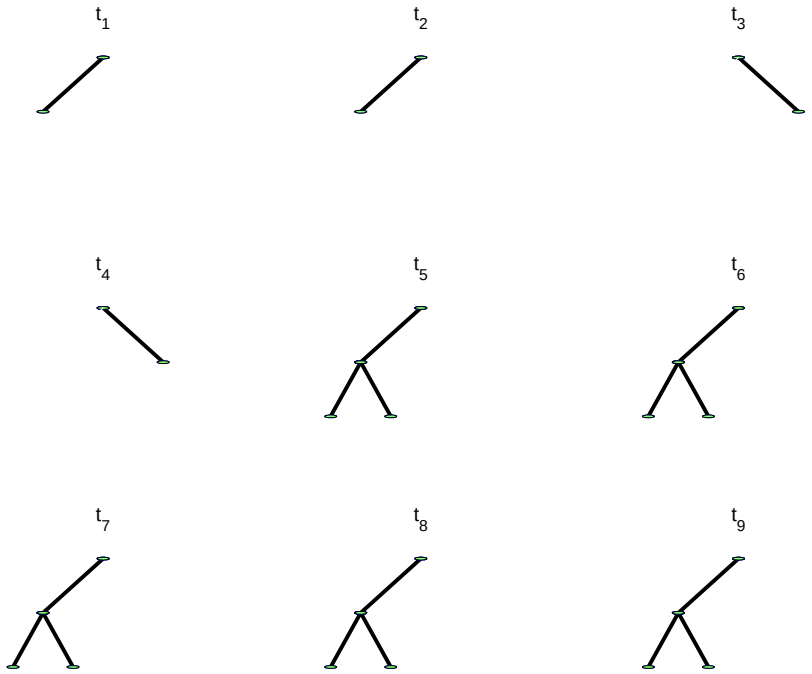

FIGURE 2.9. The tree sample T.

FIGURE 2.10. The one-dimensional principal representation, treeline l_1.

FIGURE 2.11. The one-dimensional principal representation, treeline l_2.

Therefore, the sums of the distances between the tree and its projection onto the treelines l_1 and l_2 are

$$\sum_{i=1}^{9} d_I(t_i, P_{l_1}(t_i)) = 8, \tag{2.21}$$

$$\sum_{i=1}^{9} d_I(t_i, P_{l_2}(t_i)) = 8. \tag{2.22}$$

Now consider another treeline l_3 (see Figure 2.12). The sum of distances to this treeline is

$$\sum_{i=1}^{9} d_I(t_i, P_{l_3}(t_i)) = 7. \tag{2.23}$$

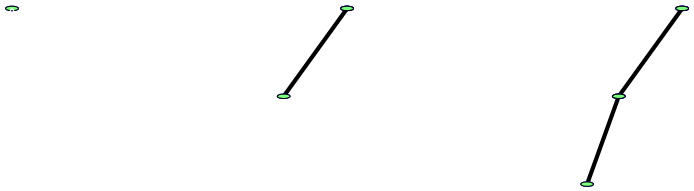

FIGURE 2.12. The treeline l_3.

REMARK 2.5.3. The assumption that the one-dimensional principal representation passes through the median tree shown in Example 2.5.1 to be a strong assumption. This choice is deliberately made as best reflecting the insight of the "variation about the center".

DEFINITION 2.5.2. For a tree sample $T = \{t_1, t_2, \ldots\}$, two treelines l_1 and l_2 are said to be **equivalent** if

$$P_{l_1}(t_i) = P_{l_2}(t_i), \forall i.$$

REMARK 2.5.4. This equivalence of two treelines is relative; that is, for different tree samples, their equivalence may be different.

REMARK 2.5.5. Let k be the maximum level of the trees of a tree sample T. If all the components with level no more than k are the same for two treelines l_1 and l_2, then l_1 and l_2 are equivalent for the sample T. Therefore, for simplicity, the treeline will be represented by the components with level no more than k.

2.6. Example

Some basic concepts and ideas were developed on trees (without nodal attributes). Now, these are illustrated with a toy example.

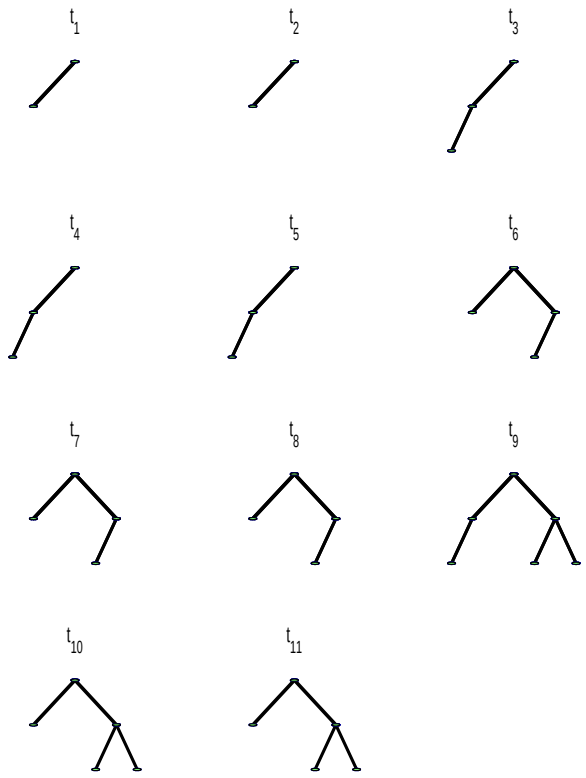

FIGURE 2.13. A sample of trees without nodal attributes.

T is a sample of trees with $n = 11$ members, $t_1, t_2, \ldots t_{11}$ shown in Figure 2.13. Based on the integer tree metric d_I, the support tree (t_{sup}) and median tree (m) are shown in Figure 2.14. Note that, $n = 11$ is an odd number. Therefore, the median tree is unique.

In the left panel of Figure 2.14, the level-order index set of the support tree is $\{1, 2, 3, 4, 6, 7\}$. And the numbers of appearance of each node are $n_1 = 11, n_2 = 11, n_3 = 6, n_4 = 4, n_5 = 0, n_6 = 6, n_7 = 3$, respectively.

According to the majority rule, the median tree consists of all nodes with appearance number more than $\frac{n}{2}$. The median tree is shown on the right panel in Figure 2.14.

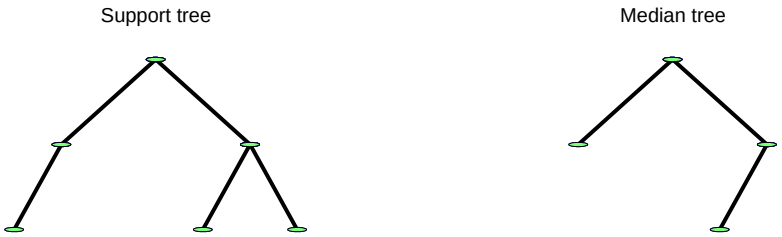

FIGURE 2.14. The support tree and the median tree.

The total variation of the sample T to its center, the sum of distances between each tree t_i and median tree m, is

$$\sum_{i=1}^{11} d_I(t_i, m) = 17.$$

Next, a treeline, one-dimensional representation in the tree space $\mathcal{T}$, which explains the greatest variability, will be found.

There are several treeline classes, which are different with respect to "equivalence" (see Definition 2.5.2), which pass through the median tree m. Three such treelines, l_1, l_2, and l_3, are shown in Figure 2.15, Figure 2.16 and Figure 2.17.

The projections of the tree sample T on the representative treeline l_1 are shown in Figure 2.18. The total distance for the median tree m to each of the projections of trees t_i on treeline l_1 is

$$\sum_{i=1}^{11} d_I(m, P_{l_1}(t_i)) = 4. \tag{2.24}$$

This can be seen from the fact that there are only four trees in Figure 2.18 that are different from the median tree in Figure 2.14 and each differs by only one branch. By the tree version of the Pythagorean Theorem (see Section 2.5), the total distance for

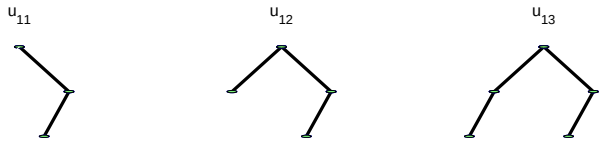

FIGURE 2.15. Representative treeline l_1.

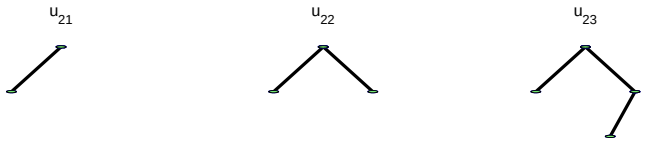

FIGURE 2.16. Representative treeline l_2.

FIGURE 2.17. Representative treeline l_3.

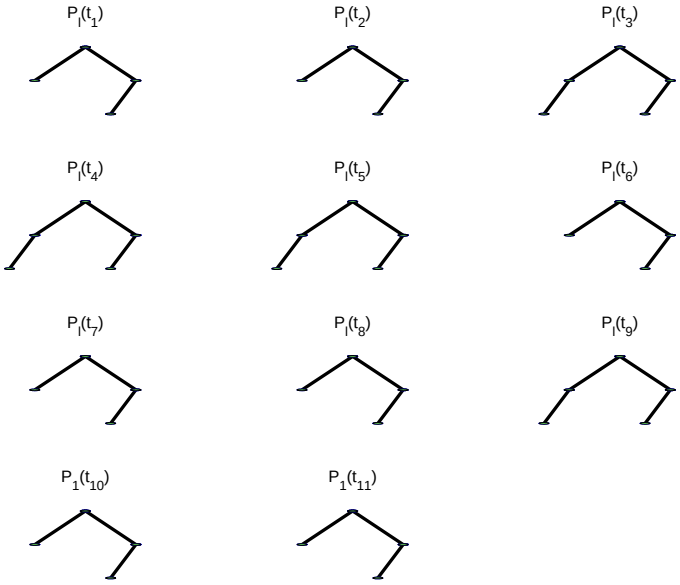

FIGURE 2.18. Projection of the tree sample on the treeline l_1.

the tree objects (t_i) to their projections on the treeline $(P_{l_1}(t_i))$ is

$$\sum_{i=1}^{11} d_I(t_i, P_{l_1}(t_i)) = \sum_{i=1}^{11} d_I(t_i, m) - \sum_{i=1}^{11} d_I(m, P_{l_1}(t_i)) = 13, \tag{2.25}$$

which is the variation unexplained by the treeline l_1.

Similarly, the projections of the tree sample T on treelines l_2 and l_3, as shown in Figure 2.19 and Figure 2.20, can be obtained.

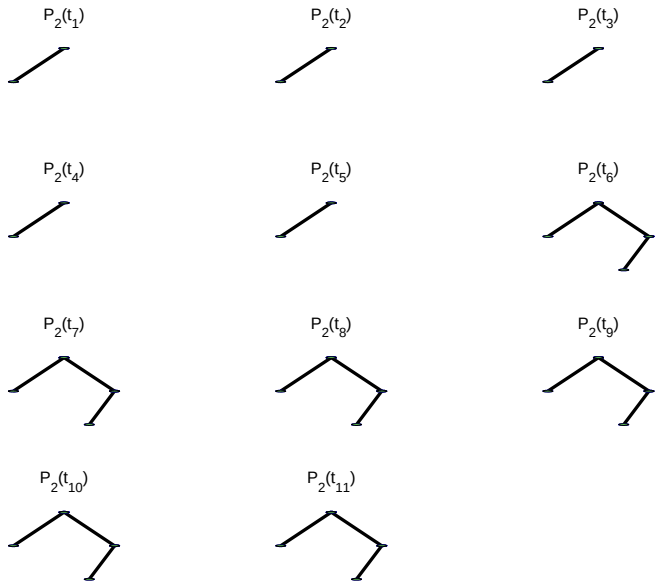

FIGURE 2.19. Projection of the tree sample on the treeline l_2.

The corresponding sums of distances for Figure 2.19 are

$$\sum_{i=1}^{11} d_I(m, P_{l_2}(t_i)) = 10, \tag{2.26}$$

$$\sum_{i=1}^{11} d_I(t_i, P_{l_2}(t_i)) = 7. \tag{2.27}$$

This can be seen that there are only six trees in Figure 2.19 that are different from the objects themselves among which five differs by one nodes and one by two nodes.

Also, the corresponding sums of distances for Figure 2.20 are

$$\sum_{i=1}^{11} d_I(m, P_{l_3}(t_i)) = 3, \tag{2.28}$$

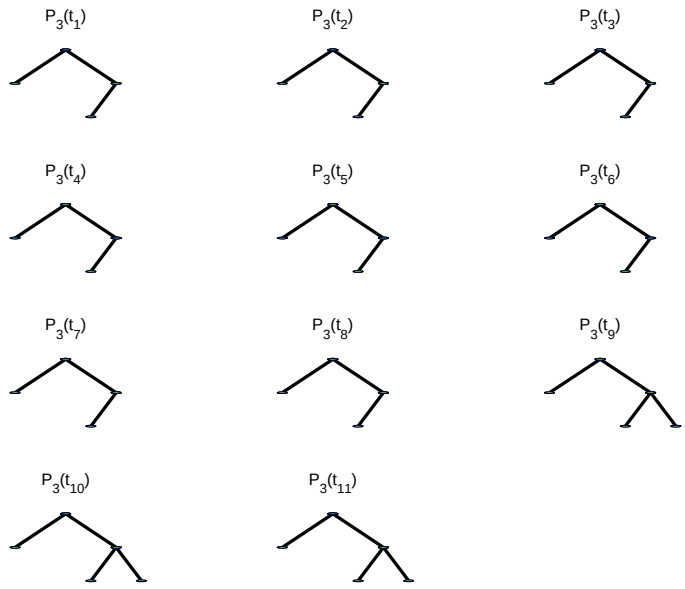

FIGURE 2.20. Projection of the tree sample on the treeline l_3.

$$\sum_{i=1}^{11} d_I(t_i, P_{l_3}(t_i)) = 14. \tag{2.29}$$

Comparing the results, it shows that, the total variations explained by the treeline l_1, (4, from Equation 2.24) and l_3 (3, from Equation 2.28) are very close. So, it is not surprising that it is difficult to see visually in Figure 2.18 and Figure 2.20 which captures more variability.

But, the total variation explained by the treeline l_2 (10, from Equation 2.26) is much further than those by l_1 and l_3. It can be seen visually that the projections on the treeline l_2 are closer to the objects themselves than those projections on the treelines l_1 or l_3. Therefore, tree line l_2 is the one-dimensional principal representation of the tree sample T.

This example shows that, a one-dimensional representation (i.e., treeline), even the one-dimensional principal representation, cannot explain all of the variation in the data. An approach to study additional population structure is to study analogs of higher dimensional subspaces. A more general 2-dimensional representation can

be generated by adding or deleting 1 or 2 terminal nodes starting from the median trees (see Chapter 4 for more discussion).

CHAPTER 3

Statistical Analysis on the Binary Tree Space with Nodal Attributes

In Chapter 2, the methodology was developed for statistical analysis on the binary tree space only considering the topological tree structure without nodal attributes. The integer metric d_I provided the foundation for finding the median tree and for quantifying the variation. Furthermore, an analogous variation analysis method was developed on the tree space without attributes. This chapter focuses on the statistical analysis on the binary tree space with nodal attributes, which is of primary interest in medical image analysis.

3.1. New Metric δ on Tree Space with Nodal Attributes

The integer tree metric d_I captures topological structure of the tree population. In many important cases, including image analysis, the nodes of the trees contain useful attributes (numerical values, see Section 1.1), which should also be used in the statistical analysis.

In this section, a metric will be defined on the trees with nodal attributes which extends the integer tree metric d_I. The attributes, contained in the node with level-order index k on the tree t, are denoted by $(x_{tk}, y_{tk}, \ldots)$. For simplicity, the case (x_{tk}, y_{tk}) will be treated explicitly. The general case is straightforward as discussed in Remark 3.1.1.

Generally, the values of the nodal attributes, x_{tk} and y_{tk}, have no restriction and can be any real value. But, after some appropriate transformation, the nodal attributes can be assumed to be bounded. This is important to control the attribute

component of the metric, with respect to the topological component. For example, for a mapping f,

$$f : x \mapsto \frac{1}{\pi\sqrt{2}} \arctan(x)$$

$f(x) \in [-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}]$.

From now on, assume that $x_{tk}, y_{tk} \in [-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}]$. The bound $\frac{\sqrt{2}}{4}$ is used because the Euclidean distance between two-dimensional vectors, all that are treated in these illustrative examples, whose entries satisfy this bound, is at most 1.

Recall from Equation (2.2) that, the integer metric d_I can be written as,

$$d_I(s,t) = \sum_{k=1}^{\infty} 1\{k \in Ind(s) \triangle Ind(t)\}.$$

For any trees s and t with nodal attributes, define the new metric (Theorem 3.1.2 establishes that this is indeed a metric)

$$\delta(s,t) = d_I(s,t) + f_\delta(s,t), \tag{3.1}$$

where

$$\begin{aligned} f_\delta(s,t) = \Bigg[& \sum_{k=1}^{\infty} \alpha_k((x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2) 1\{k \in Ind(s) \cap Ind(t)\} \\ & + \sum_{k=1}^{\infty} \alpha_k(x_{sk}^2 + y_{sk}^2) 1\{k \in Ind(s) \backslash Ind(t)\} \\ & + \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2 + y_{tk}^2) 1\{k \in Ind(t) \backslash Ind(s)\} \Bigg]^{\frac{1}{2}} \end{aligned} \tag{3.2}$$

and $\{\alpha_k\}$ is a non-negative weight series with $\sum_k \alpha_k = 1$. The last two summations in Equation (3.2) are included to avoid loss of information from those nodal attributes that are in one tree and not the other.

In Equation (3.1), the second term in the summation f_δ, where "f" means fractional part of the metric, is at most 1 (proof given in Proposition 3.1.1). Recall that, the first term in the summation is denoted as d_I where "I" means integer part of the metric (see Section 2.2).

Also, note that f_δ is a square root of a weighted sum of squares. When trees s and t have the same tree structure, $f_\delta(s,t)$ can be viewed as a weighted Euclidean distance. In particular, the nodal attributes can be combined into a single long vector. Then, $f_\delta(s,t)$ is a weighted Euclidean metric on these vectors.

When trees s and t have different tree structures, it is convenient to replace the nonexistent nodal attributes with $(0,0)$. Thus, f_δ can be rewritten as

$$\begin{aligned} f_\delta(s,t) = [&\sum_{k=1}^{\infty} \alpha_k((x_{sk}-x_{tk})^2+(y_{sk}-y_{tk})^2)1\{k \in Ind(s) \cap Ind(t)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((x_{sk}-0)^2+(y_{sk}-0)^2)1\{k \in Ind(s) \backslash Ind(t)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((0-x_{tk})^2+(0-y_{tk})^2)1\{k \in Ind(t) \backslash Ind(s)\}]^{\frac{1}{2}} \end{aligned} \tag{3.3}$$

This also allows the nodal attributes to be combined into a single long vector. Then, $f_\delta(s,t)$ is a weighted Euclidean metric on these vectors.

For another view of f_δ, rescale the entries of the vector by the square root of the weights α_k. Then, f_δ is the ordinary Euclidean metric on these rescaled vectors.

From now on, all the theorems are developed for general weight sequences. But, the power weight sequence, where the weight is $\{2^{-(2i+1)}\}$ for the node on the i^{th} level, $i = 0, 1, 2, \ldots$ in $\mathcal{T}$, will be used in the examples.

Insight into the metric δ comes from Example 3.1.1.

EXAMPLE 3.1.1. t_1 and t_2 are two trees with nodal attributes listed in Table 3.1.

level-order index	t_1	t_2
1	(0.3,0.1)	(0.2,0.1)
2	(0.15,0.25)	(0.3,0.2)

TABLE 3.1. Nodal attributes of the trees t_1 and t_2.

The following figure shows the graphical representation[1] of the two trees t_1 and t_2.

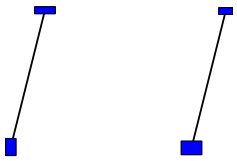

FIGURE 3.1. Graphical representation of trees t_1 and t_2 in Example 3.1.1.

Note that the trees t_1 and t_2 have the same tree structure which implies the integer part of the distance, $d_I(t_1, t_2) = 0$.

$$\begin{aligned}\delta(t_1, t_2) &= f_\delta(t_1, t_2) \\ &= \sqrt{\underbrace{\frac{1}{2}((0.3-0.2)^2 + (0.1-0.1)^2)}_{k=1} + \underbrace{\frac{1}{2^3}((0.15-0.3)^2 + (0.25-0.2)^2)}_{k=2}} \\ &= 0.0901\end{aligned}$$

where k is the level-order index, $\frac{1}{2}$ and $\frac{1}{2^3}$ are the weights of the two nodes, respectively.

As noted above, f_δ can be viewed as a weighted metric on the vectors (made up of combined nodal attributes) $[0.3, 0.1, 0.15, 0.25]'$ and $[0.2, 0.1, 0.3, 0.2]'$.

From the alternative point of view, $f_\delta(t_1, t_2)$ is the ordinary Euclidean distance between the two weighted vectors $\vec{v}_1$ and $\vec{v}_2$

$$\begin{aligned}\vec{v}_1 &= [\frac{0.3}{\sqrt{2}}, \frac{0.1}{\sqrt{2}}, \frac{0.15}{\sqrt{2^3}}, \frac{0.25}{\sqrt{2^3}}]'; \\ \vec{v}_2 &= [\frac{0.2}{\sqrt{2}}, \frac{0.1}{\sqrt{2}}, \frac{0.3}{\sqrt{2^3}}, \frac{0.2}{\sqrt{2^3}}]'.\end{aligned}$$

PROPOSITION 3.1.1. *For any two trees with nodal attributes, the fractional part is at most 1, i.e.,*

$$f_\delta(s, t) \leq 1.$$

[1]For every node with positive nodal attributes (x, y), x is taken as the length and y is taken as the width of the nodal box in the graphical representation.

Proof. Note that, for $k \in Ind(s) \cap Ind(t)$,

$$(x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2 \leq 1$$

because $x_{sk}, x_{tk}, y_{sk}, y_{tk} \in [-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}]$.

Similarly,

$$x_{sk}^2 + y_{sk}^2 \leq 1$$

for $k \in Ind(s) \backslash Ind(t)$, and

$$x_{tk}^2 + y_{tk}^2 \leq 1$$

for $k \in Ind(t) \backslash Ind(s)$. Therefore,

$$\begin{aligned} f_\delta^2(s,t) = & \sum_{k=1}^{\infty} \alpha_k \left[((x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2) 1\{k \in Ind(s) \cap Ind(t)\} \right. \\ & \left. + (x_{sk}^2 + y_{sk}^2) 1\{k \in Ind(s) \backslash Ind(t)\} + (x_{tk}^2 + y_{tk}^2) 1\{k \in Ind(t) \backslash Ind(s)\} \right] \\ \leq & \sum_{k=0}^{\infty} \alpha_k \\ = & 1. \end{aligned}$$

□

REMARK 3.1.1. In general, when the attribute vector contains more than two attributes, say N, $\frac{1}{2\sqrt{N}}$ can be taken as the bound.

Next, Theorem 3.1.2 shows that δ is a metric. This requires the following assumption.

ASSUMPTION 1. *The weight α_k is positive and $\sum \alpha_k = 1$.*

THEOREM 3.1.2. *Under Assumption 1, δ is a metric on the tree space with nodal attributes.*

Proof. Suppose s, t and u are any three trees with nodal attributes.

Note that

$$\delta(s,s) = d_I(s,s) + f_\delta(s,s) = 0.$$

On the other hand, for two binary trees s and t, if $\delta(s,t) = 0$, then $d_I(s,t) = 0$ and $f_\delta(s,t) = 0$ because both tree functions d_I and f_δ are non-negative. Therefore, the tree s and tree t have the same tree structure and nodal attributes. That is, $s = t$.

Also, the symmetry property is straightforward because d_I and f_δ are both symmetric functions on tree space.

Now, the triangle inequality will be proved; that is,

$$\delta(s,t) \leq \delta(s,u) + \delta(u,t). \tag{3.4}$$

Recall that d_I is a metric on the tree space without nodal attributes and pseudo-metric on the tree space with nodal attributes (see Theorem 2.2.2 in Section 2.2). Thus, the triangle inequality is satisfied; that is,

$$d_I(s,t) \leq d_I(s,u) + d_I(u,t).$$

Also, f_δ is the same as the weighted Euclidean distance between two attribute vectors. Therefore, the triangle inequality is satisfied.

Thus, in general, the triangle inequality (3.4) is satisfied. δ is a metric on the tree space with nodal attributes. □

REMARK 3.1.2. For a non-negative weight series $\{\alpha_k\}$, δ is a pseudo metric. In fact, if some weight is equal to zero, there exist two trees s and t ($s \neq t$), such that $\delta(s,t) = 0$.

When a general weight sequence is used, the fractional part can be very small. In some problems, the level of the trees in the population is finite. In that case, it might make sense to assign positive weights to all of those nodes only.

The set of all subtrees of a particular tree is an analog of the "subspace generated by the tree". In particular,

DEFINITION 3.1.1. For any tree w in $\mathcal{T}$, Let $\mathcal{T}_w$ denote the set of all subtrees (see Definition 2.1.3) of w. That is,

$$\mathcal{T}_w = \{t : Ind(t) \subseteq Ind(w)\}. \tag{3.5}$$

Note that the Definition 3.1.1 is not restrictive because w can be the union of any finite population of trees. Thus, $\mathcal{T}_w$ plays a role similar to the "subspace generated by a set of vectors".

In Definition 3.1.1, $\mathcal{T}_w$ consists of all of the subtrees of the tree w. Here, the term "subtree" refers to the topological relationship. That is, a tree t is a member of $\mathcal{T}_w$ (i.e., $t \in \mathcal{T}_w$), if $Ind(t) \subseteq Ind(w)$, without regard to the attributes. Hence, the name "topological subtree" is used for this relationship. For example, if $t \in \mathcal{T}_w$, then the tree t is a topological subtree of the tree w, denoted by $t \overset{T}{\subseteq} w$ or $w \overset{T}{\supseteq} t$. The following definition is for a different, but also useful, notion of subtree, the "attribute subtree".

DEFINITION 3.1.2. For two trees s and t, the tree s is called the **attribute subtree** of the tree t, denoted by $s \overset{A}{\subseteq} t$ or $t \overset{A}{\supseteq} s$, if

$$Ind(s) \subseteq Ind(t),$$

and for every node $k \in Ind(s)$, the two trees have the same nodal attributes.

REMARK 3.1.3. From Definition 3.1.2, if the tree s is an attribute subtree of the tree t, then s is also a topological subtree of the tree t.

DEFINITION 3.1.3. For two trees s and t ($s \overset{T}{\subseteq} t$), the tree s is called the **proper topological subtree** of the tree t, denoted by $s \overset{T}{\subset} t$ or $t \overset{T}{\supset} s$, if the set $Ind(s)$ is a proper subset of the set $Ind(t)$, i.e.,

$$Ind(s) \subset Ind(t).$$

DEFINITION 3.1.4. In Definition 3.1.2, the tree s is called the **proper attribute subtree** of the tree t, denoted by $s \overset{A}{\subset} t$ or $t \overset{A}{\supset} s$, if the set $Ind(s)$ is a proper subset of the set $Ind(t)$, i.e.,

$$Ind(s) \subset Ind(t).$$

PROPOSITION 3.1.3. *If two trees w_1 and w_2 have the same tree structures, i.e., $Ind(w_1) = Ind(w_2)$, then $\mathcal{T}_{w_1} = \mathcal{T}_{w_2}$.*

Using the notation $N(t)$ to denote the total number of nodes of tree t, $\forall t \in \mathcal{T}_w$, the following inequality holds

$$N(t) \leq N(w).$$

In the tree subspace $\mathcal{T}_w$, equal weight $\frac{1}{N(w)}$ is assigned to each node. That is, the weight α_k is

$$\alpha_k = \begin{cases} \frac{1}{N(w)}, & \text{if } k \in Ind(w) \\ \\ 0, & \text{if } k \notin Ind(w). \end{cases} \tag{3.6}$$

Thus, restrict the metric δ to the following metric ρ: for any two trees $s, t \in \mathcal{T}_w$,

$$\begin{aligned} \rho(s,t) = d_I(s,t) + \Bigg[& \frac{1}{N(w)} \sum_{k \in Ind(w)} ((x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2) 1\{k \in Ind(s) \cap Ind(t)\} \\ & + \frac{1}{N(w)} \sum_{k \in Ind(w)} ((x_{sk})^2 + (y_{sk})^2) 1\{k \in Ind(s) \backslash Ind(t)\} \\ & + \frac{1}{N(w)} \sum_{k \in Ind(w)} ((x_{tk})^2 + (y_{tk})^2) 1\{k \in Ind(t) \backslash Ind(s)\} \Bigg]^{\frac{1}{2}} \\ & = d_I(s,t) + f_\rho(s,t). \end{aligned} \tag{3.7}$$

From Remark 3.1.2, ρ is not a metric on the tree space $\mathcal{T}$ because the weight $\alpha_k = 0$ if $k \notin Ind(w)$. But, following the same proof as was used for Theorem 3.1.2, Proposition 3.1.4 holds.

PROPOSITION 3.1.4. *ρ is a metric on the tree subspace $\mathcal{T}_w$.*

EXAMPLE 3.1.2. Let t_1 and t_2 be the trees as given in Example 3.1.1. They are members of the tree subspace $\mathcal{T}_w$, where $Ind(w) = \{1, 2, 3\}$.

Note that

$$\begin{aligned}\rho(t_1, t_2) &= f_\rho(t_1, t_2) \\ &= \sqrt{\underbrace{\frac{1}{3}((0.3-0.2)^2 + (0.1-0.1)^2)}_{k=1} + \underbrace{\frac{1}{3}((0.15-0.3)^2 + (0.25-0.2)^2)}_{k=2}} \\ &= 0.1080.\end{aligned}$$

It shows that $\delta(t_1, t_2) < \rho(t_1, t_2)$. The reason is that, the metric δ puts much smaller weights on the nodes with larger level-order indices. Thus, the nodal attributes of the nodes with larger level-order indices have less impact on the distance. In Section 3.6, the metric ρ will be used for easy hand calculation; while in the other sections, the general metric δ is applied (the power sequence is used as the weights in δ for illustration).

3.2. Formulating the Nodal Attributes and Representing the Trees

In this section, the question of how to represent the trees in terms of their topologies and nodal attributes will be discussed.

In Example 3.1.1, a table is used to represent the trees by listing the level-order indices on the left column followed by the corresponding nodal attributes.

For example, t is a tree with level-order index set $Ind(t) = \{k_1, k_2, \ldots\}$, where $k_1 < k_2 < \cdots$. Then, the tree is represented as shown in Table 3.2.

Note that, for a node which does not appear in the tree t, its nodal attributes are recorded as "n/a" in the table.

level-order index	nodal attributes of t
$\vdots$	$\vdots$
k_1	(x_{tk_1}, y_{tk_1})
$\vdots$	$\vdots$
k_2	(x_{tk_2}, y_{tk_2})
$\vdots$	$\vdots$

TABLE 3.2. Representation of the tree t by using a table of level-order indices and corresponding nodal attributes.

As mentioned in the previous section, each tree is associated with a numerical data vector. The fractional part distance f_δ is the weighted Euclidean distance between those vectors. The following rule is used to formulate the nodal attribute vector.

Padding Rule for Attribute Vectors

For a tree t, its associated nodal attribute vector $\vec{v}$ is defined as

$$\vec{v} = [v_1, v_2, \ldots],$$

where for $k = 1, 2, \ldots,$

$$(v_{2k-1}, v_{2k}) = \begin{cases} (x_{tk}, y_{tk}), & \text{if } k \in Ind(t) \\ \\ (0,0), & \text{if } k \notin Ind(t). \end{cases} \tag{3.8}$$

If T is a sample of trees in the finite level tree subspace $\mathcal{T}_w$, then for every element in the sample T, $(v_{2k-1}, v_{2k}) = (0,0)$, when $k \notin Ind(w)$. Therefore, the nodal attributes can be simply recorded as a vector of length $2N(w)$, where $N(w)$ is the total number of nodes of the tree w.

Furthermore, the fractional part metric f_ρ on finite level trees is proportional to the ordinary Euclidean distance d. That is, for $t_1, t_2 \in \mathcal{T}_w$,

$$f_\rho(t_1, t_2) = \frac{1}{\sqrt{N(w)}} d(\vec{v}_1, \vec{v}_2),$$

where $\vec{v}_1$ and $\vec{v}_2$ are the nodal attribute vectors of the trees t_1 and t_2 respectively.

3.3. Median-mean Tree of the Tree Sample with Nodal Attributes

In Section 2.3, the problem of how to find the median tree for a tree sample $T = \{t_1, t_2, \ldots, t_n\}$ without nodal attributes has been solved. The solution for the metric d_I was the **majority rule** (see Theorem 2.3.2 for the algorithm). Now, a new metric δ, which also considers nodal attributes, is given. In this section, a new "center point" of the tree sample with nodal attributes, called the median-mean tree, will be developed. The name "median-mean" is used because it has properties of both a median with respect to d_I and a mean with respect to f_δ.

DEFINITION 3.3.1. A tree is called a **median-mean tree** for a tree sample $T = \{t_1, t_2, \ldots, t_n\}$, denoted by m_δ, if it minimizes

$$\sum_{i=1}^{n} d_I(t, t_i) \tag{3.9}$$

over all trees $t \in T$ and has nodal attributes, for the node $k \in Ind(m_\delta)$,

$$x_{m_\delta k} = \frac{\sum_{i=1}^{n} x_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{3.10}$$

$$y_{m_\delta k} = \frac{\sum_{i=1}^{n} y_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{3.11}$$

REMARK 3.3.1. The new "center point" m_δ is called **"median-mean"** because its tree structure complies with the majority rule with appearance number at least $\frac{n}{2}$ and because its nodal attributes can be calculated as a "sample mean".

The median-mean tree defined in Definition 3.3.1 may or may not be unique, as shown in Examples 3.3.1 and 3.3.2 below. A variation which is unique is given in Definition 3.3.2.

DEFINITION 3.3.2. The median-mean tree with the smallest number of nodes is called the **minimal median-mean** tree with nodal attributes (denoted by μ_δ).

The following example shows the lack of the uniqueness of median-mean tree. But the minimal median-mean tree is unique, as shown in Proposition 3.3.1.

EXAMPLE 3.3.1. For a tree sample $T = \{t_1, t_2\}$, the nodal attributes are listed in Table 3.3.

level-order index	t_1	t_2
1	(0.1,0.2)	(0.3,0.3)
2	(0.1,0.1)	N/A
3	N/A	(0.2,0.2)

TABLE 3.3. Nodal attributes of t_1 and t_2 in Example 3.3.1.

FIGURE 3.2. Graphical representation of the tree sample of Example 3.3.1.

There are four median-mean trees for this sample (see Figure 3.3). Because there are only two trees in the sample, all topological subtrees of the union tree satisfy the majority rule. There exists a median-mean tree for each topological subtree. The nodal attributes of the median-mean trees are shown in Table 3.4.

level-order index	attributes
1	(0.2,0.25)
2	N/A
3	N/A

level-order index	attributes
1	(0.2,0.25)
2	(0.1,0.1)
3	N/A

level-order index	attributes
1	(0.2,0.25)
2	N/A
3	(0.2,0.2)

level-order index	attributes
1	(0.2,0.25)
2	(0.1,0.1)
3	(0.2,0.2)

TABLE 3.4. Nodal attributes for the four median-mean trees.

FIGURE 3.3. Graphical representation of the four median-mean trees of the tree sample in Example 3.3.1.

The first one is the minimal median-mean tree μ_δ, which has the smallest number of nodes. Note that its structure is the same as that of the minimal median tree without nodal attributes, μ.

In the following Example 3.3.2, both the median-mean tree and the minimal median-mean tree are unique.

EXAMPLE 3.3.2. For a tree sample $T = \{t_1, t_2, t_3, t_4\}$ as shown in Figure 3.4, the nodal attributes are listed in Table 3.5.

In this example, by the majority rule, the median tree is unique without nodal attributes. Considering the nodal attributes, only one median-mean tree listed in Table 3.6, which is also the minimal median-mean tree, will be obtained.

The two examples above motivate the following proposition.

level-order index	t_1	t_2	t_3	t_4
1	(0.2,0.2)	(0.3,0.3)	(0.2,0.3)	(0.3,0.2)
2	(0.1,0.3)	(0.3,0.2)	(0.2,0.1)	N/A
3	(0.3,0.1)	(0.2,0.3)	N/A	(0.3,0.2)

TABLE 3.5. Nodal attributes of the tree sample in Example 3.3.2.

FIGURE 3.4. Graphical representation of the tree sample in Example 3.3.2.

level-order index	attributes
1	(0.25,0.25)
2	(0.20,0.20)
3	(0.267,0.20)

TABLE 3.6. Nodal attributes of the median-mean tree of Example 3.3.2.

PROPOSITION 3.3.1. *The minimal median-mean tree with nodal attributes is unique. Also, it has the same tree structure as that of the minimal median tree without nodal attributes.*

Proof. Since the median-mean tree minimizes Equation (3.9), the median-mean tree is also a median tree without nodal attributes. By Theorem 2.3.4, the minimal median tree is unique without nodal attributes. Hence, the minimal median-mean tree is also unique. □

The following Example 3.3.3 is used to show that the median-mean tree may not minimize the sum

$$\sum_{i=1}^{n} \delta(t_i, m_\delta).$$

EXAMPLE 3.3.3. For a tree sample $T = \{t_1, t_2, t_3\}$, the nodal attributes are listed in Table 3.7.

level-order index	t_1	t_2	t_3
1	(0.2,0.2)	(0.2,0.2)	(0.2,0.2)
2	(0,0.3)	(0.3,0)	(0,0)

TABLE 3.7. Nodal attributes of the tree sample T in Example 3.3.3.

In this example, there is a unique median-mean tree with nodal attributes m_δ, listed in Table 3.8.

level-order index	nodal attributes
1	(0.2,0.2)
2	(0.1,0.1)

TABLE 3.8. Nodal attributes of the median-mean tree of Example 3.3.3.

The total distance about the median-mean tree m_δ is

$$\sum_{i=1}^{3} \delta(t_i, m_\delta) = 0.2081.$$

Now, consider the tree s (see Table 3.9). The total distance from s to the other trees is

$$\sum_{i=1}^{3} \delta(t_i, s) = 0.2049.$$

level-order index	s
1	(0.2,0.2)
2	(0.06,0.06)

TABLE 3.9. Nodal attributes of the tree s.

Hence,

$$\sum_{i=1}^{3} \delta(t_i, m_\delta) > \sum_{i=1}^{3} \delta(t_i, s).$$

That is, the median-mean tree m_δ does not minimize the sum $\sum_i \delta(t_i, t)$ over all t.

REMARK 3.3.2. Recall from Section 2.3 that, the median tree without nodal attributes minimizes the sum $\sum_i d_I(t_i, t)$, over all t, while the median-mean tree with nodal attributes m_δ may not minimize the sum $\sum_i \delta(t_i, t)$. This is not surprising, because even in Euclidean space $\mathbb{R}^d$, the sample mean minimize the sum of **squared** distances to the data, **not** the sum of distances. The reason of making this choice of the median-mean tree is, that it fits best with the coming decomposition of variation, into topological and attribute components.

3.4. Quantifying the Variation in the Tree Space with Nodal Attributes

Now, for a tree sample T with nodal attributes, a metric δ was defined. The next question is how to quantify the variation of the sample to the "center point"—median-mean tree.

An important foundation of "variation" is the tree function:

$$V_\delta(s, t) = d_I(s, t) + f_\delta^2(s, t). \tag{3.12}$$

DEFINITION 3.4.1. Let T be a sample of trees with nodal attributes. The tree m_δ is a median-mean tree according to the metric δ. The **variation** of a tree t, in the sample, about the median-mean tree is defined as $V_\delta(t, m_\delta)$.

REMARK 3.4.1. $V_\delta(\cdot, m_\delta)$ is a function defined on a tree space, but it is not a metric because the triangle inequality is not satisfied, just as squared Euclidean distance is not a metric.

Recall that, the median-mean tree is not unique when the sample size n is an even number and some nodes appear $\frac{n}{2}$ times in the sample. Does the total variation depend on the choice of median-mean tree, when it is not unique? The following proposition answers this question.

THEOREM 3.4.1. *Let $T = \{t_1, t_2, \ldots, t_n\}$ be a finite sample of trees with nodal attributes. The sum of variation to the median-mean tree over elements in the sample*

$$\sum_{i=1}^{n} V_\delta(t_i, m_\delta) \tag{3.13}$$

is constant over all median-mean trees of the sample T.

Proof. Suppose $n = 2q$, where q is some positive integer, since otherwise the median-mean tree is unique. Let s be any median-mean tree, that is not the minimal median-mean tree μ_δ. A proof will be provided for the following equality

$$\sum_{i=1}^{n} V_\delta(t_i, s) = \sum_{i=1}^{n} V_\delta(t_i, \mu_\delta). \tag{3.14}$$

Since μ_δ is the minimal median-mean tree, μ_δ is an attribute subtree of s. Thus, there exists a sequence of median-mean trees $\{s_i\}$, such that

$$\mu_\delta = s_1 \overset{A}{\subseteq} s_2 \cdots \overset{A}{\subseteq} s_K = s,$$

where s_i is an attribute subtree of s_{i+1}, and s_{i+1} has one more node (denoted by k_{i+1}) than s_i, for $i = 1, \ldots, K-1$. It is straightforward that the node k_{i+1} appears exactly $q = \frac{n}{2}$ times in the sample.

For $1 \le p \le K-1$,

$$\begin{aligned}
&\sum_{i=1}^{n} V_\delta(t_i, s_{p+1}) \\
&= \sum_{i=1}^{n} V_\delta(t_i, s_p) - \sum_{i=1}^{n} \alpha_{k_{p+1}}(x^2_{t_i k_{p+1}} + y^2_{t_i k_{p+1}}) 1\{k_{p+1} \in Ind(t_i)\} \\
&+ q\alpha_{k_{p+1}}(x^2_{s_{p+1} k_{p+1}} + y^2_{s_{p+1} k_{p+1}}) \\
&+ \sum_{i=1}^{n} \alpha_{k_{p+1}}((x_{t_i k_{p+1}} - x_{s_{p+1} k_{p+1}})^2 + (y_{t_i k_{p+1}} - y_{s_{p+1} k_{p+1}})^2) 1\{k_{p+1} \in Ind(t_i)\}.
\end{aligned} \tag{3.15}$$

By the definition of the median-mean tree,

$$\begin{aligned}
&\sum_{i=1}^{n}((x_{t_i k_{p+1}} - x_{s_{p+1} k_{p+1}})^2 + (y_{t_i k_{p+1}} - y_{s_{p+1} k_{p+1}})^2) 1\{k_{p+1} \in Ind(t_i)\} \\
&= \sum_{i=1}^{n}(x^2_{t_i k_{p+1}} + y^2_{t_i k_{p+1}}) 1\{k_{p+1} \in Ind(t_i)\} \\
&- q(x^2_{s_{p+1} k_{p+1}} + y^2_{s_{p+1} k_{p+1}}).
\end{aligned} \tag{3.16}$$

Combining Equations (3.15) and (3.16),

$$\sum_{i=1}^{n} V_\delta(t_i, s_{p+1}) = \sum_{i=1}^{n} V_\delta(t_i, s_p).$$

Repeatly over $p = 1, 2, \ldots, K-1$,

$$\sum_{i=1}^{n} V_\delta(t_i, \mu_\delta) = \cdots = \sum_{i=1}^{n} V_\delta(t_i, s).$$

□

REMARK 3.4.2. This shows why the median-mean tree is a very natural notion of "center", as discussed in Section 3.3 .

3.5. Treeline and Projection in Finite Level Tree Subspace $\mathcal{T}_w$

In Sections 3.3 and 3.4, the center point of a sample of trees with nodal attributes and the total variation of the sample to its median-mean tree have been defined.

Also, according to Theorem 3.4.1, the total variation about the median-mean tree is constant over all choices of median-mean trees.

In Euclidean space, principal component analysis (PCA) provides a useful decomposition of complex data sets, in terms of simple one-dimensional representations of the data. Binary tree space is not a linear space, but useful one-dimensional representations are developed. There are two important types, defined below, both of which are called "treeline".

In this section, the treeline, which plays the role of line in Euclidean space, is defined in tree space. Hence, an analogy of PCA is developed to find treelines, which explain important features of the sample.

DEFINITION 3.5.1. Suppose $l = \{u_0, u_1, u_2, \ldots\}$ is a sequence of trees with nodal attributes in the subspace $\mathcal{T}_w$. The set l is called a **structure treeline** (s-treeline) starting from u_0 if for $i = 1, 2, 3, \ldots$,

(1) u_{i-1} can be obtained by deleting a terminal node (denoted by ν_i) from the tree u_i;
(2) The next node to be deleted, ν_{i-1} is the parent of ν_i;
(3) There does not exist an attribute subtree of u_0, denoted as u, such that u can be obtained by deleting some ancestor nodes of ν_1.

In Definition 3.5.1, the tree u_{i-1} is an attribute subtree of the trees u_i, u_{i+1}, etc. Therefore, the nodes with level-order index k in the s-treeline have the same nodal attributes. Since every element in the s-treeline is a topological subtree of w, the length of the s-treeline is finite and cannot exceed the number of levels of the tree w.

Figure 3.6 shows an example of an s-treeline in $\mathcal{T}_w$, where w has the tree structure shown in Figure 3.5.

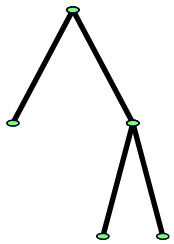

FIGURE 3.5. Tree structure of an example tree w.

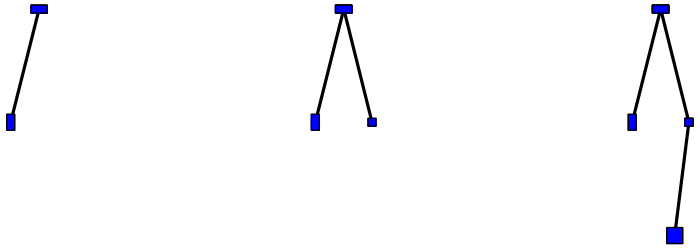

FIGURE 3.6. An example of an s-treeline in $\mathcal{T}_w$, for w defined in Figure 3.5.

DEFINITION 3.5.2. A structure treeline l is called **passing through** the tree u, if the tree u is an element of the tree set l, i.e., $u \in l$.

Recall from Section 1.2 that, in the blood vessel data, Figure 1.9 shows a structure treeline passing through the media-mean tree with nodal attributes.

An s-treeline indicates a direction of changing tree structures. The following definition will describe a quite different direction in which all trees have the same tree structure but changing nodal attributes.

DEFINITION 3.5.3. Suppose $l = \{u_\lambda : \lambda \in \mathbb{R}\}$ is a set of trees with nodal attributes in the subspace $\mathcal{T}_w$. The set l is called an **attribute treeline** (a-treeline) **passing through** a tree u_0 if

(1) every tree u_λ has the same tree structure as u_0;

(2) the nodal attribute vector is equal to $\vec{v}_0 + \lambda\vec{v}$, where $\vec{v}_0$ is the attribute vector of the tree u_0 and $\vec{v}$ is some fixed vector, $\vec{v} \neq \vec{0}$.

REMARK 3.5.1. An a-treeline is determined by the tree u_0 and vector $\vec{v}$. Also, it is a set of trees of the form

$$l = \{u : u \text{ has same structure as } u_0 \text{ with nodal attributes equal to } \vec{v}_0 + \lambda\vec{v}\}$$

Note that there are uncountably many elements on an a-treeline because it has the same cardinality as the real numbers. Figure 3.7 shows some elements with $\lambda = 0.5, 1.0, 1.2, 1.5$ and $\vec{v} = [0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.2, 0.2]'$ in an a-treeline in $\mathcal{T}_w$.

FIGURE 3.7. An example of an a-treeline in $\mathcal{T}_w$.

In Section 1.2, Figure 1.10 and Figure 1.12 illustrate two attribute treelines. There are six subplots in each figure, which depicts one location on the attribute treeline.

From now on, both s-treelines and a-treelines are called treelines. An analogy of the first principal component is the treeline which explains most of the data. Before finding this, the projection of a tree on a treeline is defined in the tree subspace $\mathcal{T}_w$.

DEFINITION 3.5.4. Let l be a treeline. For any tree t, a tree on the treeline is called a **projection** of the tree t if it minimizes $\delta(t, u)$ over all trees u on the treeline l.

Recall that, the projection of a point on a line is unique in Euclidean space. Is it still unique in the tree space with nodal attributes?

PROPOSITION 3.5.1. *Under Assumption 1, the projection of a tree t on a treeline l is unique.*

Proof. The proof will be provided for s-treelines and a-treelines separately.

Case 1: l is an s-treeline.

Suppose $l = \{u_0, u_1, u_2, \ldots\}$. First, the topological structure is considered. Let p be the index of the smallest d_I-closest, to the tree t, member of treeline l; i.e.,

$$p = inf\{i : d_I(u_i, t) \leq d_I(u_j, t), j = 1, 2, \ldots\}.$$

Consider the two elements u_p and u_{p+1} on the treeline l. By definition of the s-treeline, the tree u_p can be obtained by deleting a node ν_{p+1} from the tree u_{p+1}. It will now be shown that, $\nu_{p+1} \notin Ind(t)$. Otherwise,

$$d_I(u_{p+1}, t) = d_I(u_p, t) - 1,$$

which is a contradiction with the definition of p. Thus, $\nu_{p+1} \notin Ind(t)$, and

$$d_I(u_{p+1}, t) = d_I(u_p, t) + 1. \tag{3.17}$$

Iteratively, for $i = 1, 2, \ldots$, the tree u_{p+i} can be obtained by deleting a node ν_{p+i+1} from the tree u_{p+i+1}. The node ν_{p+i+1} is an offspring node of the node ν_{p+1}. Since $\nu_{p+1} \notin Ind(t)$, for $i = 1, 2, \ldots$, $\nu_{p+i+1} \notin Ind(t)$. Hence,

$$d_I(u_{p+i+1}, t) = d_I(u_{p+i}, t) + 1. \tag{3.18}$$

Next, consider the two trees u_{p-1} and u_p on the treeline l. The tree u_{p-1} can be obtained by deleting a node ν_p from the tree u_p. It will now be shown that, $\nu_p \in Ind(t)$. Otherwise,

$$d_I(u_{p-1}, t) = d_I(u_p, t) - 1,$$

which is a contradiction with the definition of p. Hence, $\nu_p \in Ind(t)$, and

$$d_I(u_{p-1}, t) = d_I(u_p, t) + 1. \tag{3.19}$$

Iteratively, for $i = 1, 2, \ldots, p-1$, the tree u_{p-i-1} can be obtained by deleting a node ν_{p-i} from the tree u_{p-i}. The node ν_{p-i} is an ancestor node of the node ν_p. Since

$\nu_p \in Ind(t)$, for $i = 1, 2, \ldots, p-1$, $\nu_{p-i} \in Ind(t)$. Thus,

$$d_I(u_{p-i-1}, t) = d_I(u_{p-i}, t) + 1. \tag{3.20}$$

Hence, there is a unique tree u_p such that, for $i \neq p$

$$d_I(u_i, t) > d_I(u_p, t). \tag{3.21}$$

Next, the attribute component of the metric is considered. It will be shown that the tree u_p is the unique projection of t on the s-treeline l by considering the fractional part f_δ as well. Recall that, for $i \neq p$,

$$\delta(u_i, t) - \delta(u_p, t) = (d_I(u_i, t) - d_I(u_p, t)) + (f_\delta(u_i, t) - f_\delta(u_p, t)).$$

Also, from Equation (3.21),

$$d_I(u_i, t) - d_I(u_p, t) \geq 1.$$

The proof will be finished by showing the following inequality

$$|f_\delta(u_i, t) - f_\delta(u_p, t)| < 1. \tag{3.22}$$

Since the fractional part of the distance is always no more than 1,

$$|f_\delta(u_i, t) - f_\delta(u_p, t)| \leq 1.$$

Note that, if

$$|f_\delta(u_i, t) - f_\delta(u_p, t)| = 1,$$

then

$$1 = |f_\delta(u_i, t) - f_\delta(u_p, t)| \leq |f_\delta(u_i, u_p)|,$$

because f_δ is the weighted Euclidean distance on the attribute vectors.

Since the fractional part metric is at most 1,

$$|f_\delta(u_i, u_p)| = 1.$$

In fact, for any two trees on the s-treeline, one of the two trees is an attribute subtree of the other one. Without loss of generality, assume that the tree u_i is an attribute subtree of the tree u_p, and

$$Ind(u_p) \backslash Ind(u_i) = K,$$

where the set K is some proper subset of the positive integers.

Furthermore,

$$1 = f_\delta^2(u_i, u_p) \leq \sum_{k \in K} \alpha_k < 1,$$

which is a contradiction.

Hence, the inequality (3.22) is satisfied. Thus,

$$\delta(u_i, t) - \delta(u_p, t) > 0$$

i.e., u_p is the unique projection.

Case 2: l is an a-treeline.

Suppose the a-treeline $l = \{u_\lambda; \lambda \in \mathbb{R}\}$ and all the elements have the same tree structure. In this case, the integer part metric $d_I(u_\lambda, t)$ is a constant over all λ. Also, the fractional part metric is the ordinary Euclidean distance between weighted attribute vectors. By the uniqueness of the projection in the Euclidean space, the projection of a tree t on an a-treeline is also unique. □

REMARK 3.5.2. From the proof above, the projection of a tree t on an s-treeline according to the metric δ has the same tree structure as that of the projection without nodal attributes (see Section 2.4).

REMARK 3.5.3. From the definition of the metric δ (see Equations 3.1 and 3.2 in Section 3.1), the Assumption 1 (see Section 3.1) is very critical for the uniqueness of the projection. If some weights are equal to zero or the sum is more than 1, then non-uniqueness may arise.

Since the projection of a tree t on a treeline l is unique, the projection is denoted by $P_l(t)$.

DEFINITION 3.5.5. A tree is called an **average support tree** (denoted by t_a) if it is a support tree and its nodal attributes are, for the node $k \in Ind(t_a)$,

$$x_{t_a k} = \frac{\sum_{i=1}^{n} x_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{3.23}$$

$$y_{t_a k} = \frac{\sum_{i=1}^{n} y_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}}. \tag{3.24}$$

PROPOSITION 3.5.2. *Let T be a sample of trees. The median-mean tree is an attribute subtree of the average support tree.*

In this paper, only the s-treelines, where every element is an attribute subtree of the average support tree, are considered, because this gives a tree version of the Pythagorean Theorem (Theorem 3.5.4).

The Pythagorean Theorem is critical to the decomposition of the sums of squares in classical analysis of variance (ANOVA). An analog of this is now developed for tree populations. Theorem 3.5.3 gives a Pythagorean Theorem for a-treeline.

THEOREM 3.5.3. *(Tree version of the Pythagorean Theorem: Part I) Let l be an a-treeline passing through a tree u in the tree space $\mathcal{T}$. Then, for any $t \in \mathcal{T}$,*

$$V_\delta(t, u) = V_\delta(t, P_l(t)) + V_\delta(P_l(t), u) \tag{3.25}$$

Proof. The projection tree $P_l(t)$ has the same tree structure as the tree u. Therefore,

$$d_I(P_l(t), u) = 0 \tag{3.26}$$

and

$$d_I(t, P_l(t)) = d_I(t, u).$$

Next, it needs to prove

$$f_\delta^2(t,u) = f_\delta^2(t,P_l(t)) + f_\delta^2(P_l(t),u). \tag{3.27}$$

for the a-treeline l.

Note that, for the nodes with level-order index $k \in Ind(t)\backslash Ind(u)$, the contribution of its nodal attributes to both sides of Equation (3.27) is the same. Thus, without loss of generality, assume that $Ind(t) \subseteq Ind(u)$. Its attribute vector has the same length as that of the tree u by adding zeroes on $Ind(u)\backslash Ind(t)$.

The metric δ is the same as the Euclidean distance of two weighted vectors. Thus, it is straightforward that Equation (3.27) follows from the ordinary Pythagorean Theorem. □

Theorem 3.5.4 gives a Pythagorean Theorem for s-treeline.

THEOREM 3.5.4. *(Tree version of the Pythagorean Theorem: Part II) Let $T = \{t_1, t_2, \ldots, t_n\}$ be a sample of finite level trees. Let l be an s-treeline where every element is an attribute subtree of the average support tree of T. Then, for any $u \in l$,*

$$\sum_{i=1}^{n} V_\delta(t_i,u) = \sum_{i=1}^{n} V_\delta(t_i,P_l(t_i)) + \sum_{i=1}^{n} V_\delta(P_l(t_i),u) \tag{3.28}$$

Proof. Theorem 2.5.1 showed that, for any i,

$$d_I(t_i,u) = d_I(t_i,P_l(t_i)) + d_I(P_l(t_i),u). \tag{3.29}$$

Therefore,

$$\sum_{i=1}^{n} d_I(t_i,u) = \sum_{i=1}^{n} d_I(t_i,P_l(t_i)) + \sum_{i=1}^{n} d_I(P_l(t_i),u). \tag{3.30}$$

Next, a proof will be provided for

$$\sum_{i=1}^{n} f_\delta^2(t_i,u) = \sum_{i=1}^{n} f_\delta^2(t_i,P_l(t_i)) + \sum_{i=1}^{n} f_\delta^2(P_l(t_i),u). \tag{3.31}$$

In fact, since l passes through the tree u, $P_l(t_i) \overset{A}{\subseteq} u$ or $u \overset{A}{\subseteq} P_l(t_i)$. Without loss of generality, assume that

$$P_l(t_1) \overset{A}{\subseteq} u, \ldots, P_l(t_K) \overset{A}{\subseteq} u, P_l(t_{K+1}) \overset{A}{\supseteq} u, \ldots, P_l(t_n) \overset{A}{\supseteq} u \tag{3.32}$$

for some $K \in \{0, 1, \ldots, n\}$. If $K = 0$, then the tree u is an attribute subtree of $P_l(t_i)$, for $i = 1, 2, \ldots, n$; while, if $K = n$, then $P_l(t_i)$ is an attribute subtree of the tree u, for $i = 1, 2, \ldots, n$.

First, for $i = 1, 2, \ldots, K$, $P_l(t_i)$ is an attribute subtree of u. A proof will be provided for

$$f_\delta^2(t_i, u) = f_\delta^2(t_i, P_l(t_i)) + f_\delta^2(P_l(t_i), u). \tag{3.33}$$

Suppose that t is a tree in the set $\{t_1, \ldots, t_K\}$, then $P_l(t) \overset{A}{\subseteq} u$, and for all $k \in Ind(P_l(t)) \cap Ind(u)$, two trees $P_l(t)$ and u have the same nodal attributes for node k. Therefore,

$$f_\delta^2(P_l(t), u) = \sum_{k=1}^{\infty} \alpha_k (x_{uk}^2 + y_{uk}^2) 1\{k \in Ind(u) \backslash Ind(P_l(t))\}. \tag{3.34}$$

Furthermore, the tree $P_l(t)$ is the projection of the tree t on the treeline l. The following equality will be demonstrated

$$Ind(t) \cap Ind(u) = Ind(t) \cap Ind(P_l(t)). \tag{3.35}$$

Since $P_l(t)$ is an attribute subtree of the tree u,

$$Ind(t) \cap Ind(u) \supseteq Ind(t) \cap Ind(P_l(t)).$$

Also, if there exists a node ν, such that

$$\nu \in Ind(t) \cap Ind(u), \text{but } \nu \notin Ind(t) \cap Ind(P_l(t)),$$

then, there exists a tree u^* on the treeline l, such that $Ind(u^*) \supseteq Ind(P_l(t)) \cup \{\nu\}$ and u^* is closer to the tree t than $P_l(t)$, which is a contradiction with the assumption that the tree $P_l(t)$ is the projection of the tree t. Therefore, Equation (3.35) holds.

Recall from Equation (3.2) that, the squared fractional part of the distance between the two trees t and u is,

$$\begin{aligned} f_\delta^2(t,u) = &\sum_{k=1}^{\infty} \alpha_k((x_{tk}-x_{uk})^2+(y_{tk}-y_{uk})^2)1\{k \in Ind(t) \cap Ind(u)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2+y_{tk}^2)1\{k \in Ind(t) \backslash Ind(u)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2+y_{uk}^2)1\{k \in Ind(u) \backslash Ind(t)\}. \end{aligned} \tag{3.36}$$

Similarly, the squared fractional part of the distance between the two trees t and $P_l(t)$ is,

$$\begin{aligned} f_\delta^2(t,P_l(t)) = &\sum_{k=1}^{\infty} \alpha_k((x_{tk}-x_{P_l(t)k})^2+(y_{tk}-y_{P_l(t)k})^2)1\{k \in Ind(t) \cap Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2+y_{tk}^2)1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{P_l(t)k}^2+y_{P_l(t)k}^2)1\{k \in Ind(P_l(t)) \backslash Ind(t)\}. \end{aligned} \tag{3.37}$$

By the assumption that, $P_l(t)$ is an attribute subtree of u, for any node $k \in Ind(P_l(t))$,

$$x_{P_l(t)k} = x_{uk} \text{ and } y_{P_l(t)k} = y_{uk}.$$

By Equation (3.35),

$$\begin{aligned} f_\delta^2(t,P_l(t)) = &\sum_{k=1}^{\infty} \alpha_k((x_{tk}-x_{uk})^2+(y_{tk}-y_{uk})^2)1\{k \in Ind(t) \cap Ind(u)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2+y_{tk}^2)1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2+y_{uk}^2)1\{k \in Ind(P_l(t)) \backslash Ind(t)\}. \end{aligned} \tag{3.38}$$

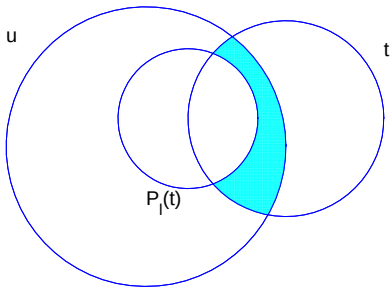

FIGURE 3.8. Venn diagram for the sets u, $P_l(t)$ and t when $P_l(t) \overset{A}{\subseteq} u$.

Thus, combining Equations (3.38) and (3.34),

$$\begin{aligned}
&f_\delta^2(t, P_l(t)) + f_\delta^2(P_l(t), u) \\
&= \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2) 1\{k \in Ind(t) \cap Ind(u)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2 + y_{tk}^2) 1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2 + y_{uk}^2) 1\{k \in Ind(P_l(t)) \backslash Ind(t)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2 + y_{uk}^2) 1\{k \in Ind(u) \backslash Ind(P_l(t))\}.
\end{aligned} \tag{3.39}$$

Using Equation (3.35),

$$Ind(t) \cap Ind(u) \cap \overline{Ind(P_l(t))} = Ind(t) \cap Ind(P_l(t)) \cap \overline{Ind(P_l(t))} = \emptyset. \tag{3.40}$$

Note that Equation (3.40) shows that the shaded area in Figure 3.8 is empty.

Now using the set relationship of the trees t, u, and $P_l(t)$ (see Figure 3.8) and Equation (3.40), the following equations are established,

$$Ind(t) \backslash Ind(P_l(t)) = Ind(t) \backslash Ind(u), \tag{3.41}$$

$$(Ind(P_l(t)) \backslash Ind(t)) \cup (Ind(u) \backslash Ind(P_l(t))) = Ind(u) \backslash Ind(t), \tag{3.42}$$

and

$$(Ind(P_l(t))\backslash Ind(t)) \cap (Ind(u)\backslash Ind(P_l(t))) = \emptyset. \tag{3.43}$$

Using Equations (3.36), (3.39), (3.41), (3.42) and (3.43),

$$\begin{aligned}
&f_\delta^2(t, P_l(t)) + f_\delta^2(P_l(t), u) \\
&= \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2) 1\{k \in Ind(t) \cap Ind(u)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2 + y_{tk}^2) 1\{k \in Ind(t)\backslash Ind(u)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2 + y_{uk}^2)(1\{k \in Ind(P_l(t))\backslash Ind(t)\} + 1\{k \in Ind(u)\backslash Ind(P_l(t))\}) \\
&= \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2) 1\{k \in Ind(t) \cap Ind(u)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{tk}^2 + y_{tk}^2) 1\{k \in Ind(t)\backslash Ind(u)\} \\
&+ \sum_{k=1}^{\infty} \alpha_k(x_{uk}^2 + y_{uk}^2) 1\{k \in Ind(u)\backslash Ind(t)\} \\
&= f_\delta^2(t, u).
\end{aligned}$$

By now, the single tree version of the Pythagorean Theorem is satisfied when the tree $P_l(t_i)$ is an attribute subtree of the tree u. That is, for $i \leq K$,

$$V_\delta(t_i, u) = V_\delta(t_i, P_l(t_i)) + V_\delta(P_l(t_i), u). \tag{3.44}$$

For $i > K$, $P_l(t_i) \overset{A}{\supseteq} u$. Note that the tree $P_l(t_i)$ is the projection of the tree t_i, which implies,

$$Ind(P_l(t_i)) \cap \overline{Ind(u)} \cap \overline{Ind(t_i)} = \emptyset, \tag{3.45}$$

by the same argument that was used to establish Equation (3.40). Equation (3.45) shows that the shaded region in Figure 3.9 is empty.

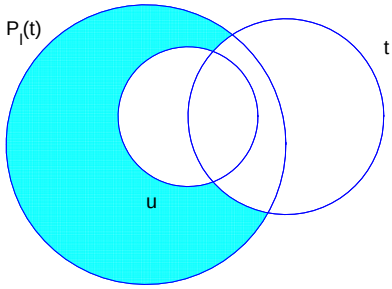

FIGURE 3.9. Venn diagram for the sets u, $P_l(t)$ and t when $P_l(t) \overset{A}{\supseteq} u$, where the tree t is a member of the set $\{t_{K+1}, \ldots, t_n\}$.

Thus, using the set relationship of the trees u, $P_l(t_i)$ and t_i (see Figure 3.9) and Equation (3.45), the following equations are established,

$$(Ind(t_i) \backslash Ind(P_l(t_i))) \cup (Ind(P_l(t_i)) \backslash Ind(u)) = Ind(t_i) \backslash Ind(u), \tag{3.46}$$

$$(Ind(t_i) \backslash Ind(P_l(t_i))) \cap (Ind(P_l(t_i)) \backslash Ind(u)) = \emptyset, \tag{3.47}$$

$$(Ind(P_l(t_i)) \backslash Ind(u)) \cup (Ind(t_i) \cap Ind(u)) = Ind(t_i) \cap Ind(P_l(t_i)), \tag{3.48}$$

and

$$Ind(P_l(t_i)) \backslash Ind(t) = Ind(u) \backslash Ind(t_i). \tag{3.49}$$

Hence, using Equations (3.36), (3.46) and (3.47),

$$\begin{aligned} f_\delta^2(t_i, u) = & \sum_{k \in Ind(t_i) \cap Ind(u)} \alpha_k ((x_{t_i k} - x_{uk})^2 + (y_{t_i k} - y_{uk})^2) \\ & + \sum_{k \in Ind(t_i) \backslash Ind(P_l(t_i))} \alpha_k (x_{t_i k}^2 + y_{t_i k}^2) + \sum_{k \in Ind(P_l(t_i)) \backslash Ind(u)} \alpha_k (x_{t_i k}^2 + y_{t_i k}^2) \\ & + \sum_{k \in Ind(u) \backslash Ind(t_i)} \alpha_k (x_{uk}^2 + y_{uk}^2) \end{aligned} \tag{3.50}$$

Using Equations (3.37), (3.48) and (3.49),

$$
\begin{aligned}
f_\delta^2(t_i, P_l(t_i)) = & \sum_{k \in Ind(P_l(t_i)) \setminus Ind(u)} \alpha_k((x_{t_ik} - x_{P_l(t_i)k})^2 + (y_{t_ik} - y_{P_l(t_i)k})^2) \\
& + \sum_{k \in Ind(t_i) \cap Ind(u)} \alpha_k((x_{t_ik} - x_{P_l(t_i)k})^2 + (y_{t_ik} - y_{P_l(t_i)k})^2) \\
& + \sum_{k \in Ind(t_i) \setminus Ind(P_l(t_i))} \alpha_k(x_{t_ik}^2 + y_{t_ik}^2) + \sum_{k \in Ind(u) \setminus Ind(t_i)} \alpha_k(x_{P_l(t_i)k}^2 + y_{P_l(t_i)k}^2).
\end{aligned} \tag{3.51}
$$

From the fact that, for $k \in Ind(t_i) \cap Ind(u)$ or $k \in Ind(u) \setminus Ind(t_i)$,

$$
x_{P_l(t_i)k} = x_{uk}, \text{ and } y_{P_l(t_i)k} = y_{uk}. \tag{3.52}
$$

Thus, Equation (3.51) can be rewritten as,

$$
\begin{aligned}
f_\delta^2(t_i, P_l(t_i)) = & \sum_{k \in Ind(P_l(t_i)) \setminus Ind(u)} \alpha_k((x_{t_ik} - x_{P_l(t_i)k})^2 + (y_{t_ik} - y_{P_l(t_i)k})^2) \\
& + \sum_{k \in Ind(t_i) \cap Ind(u)} \alpha_k((x_{t_ik} - x_{uk})^2 + (y_{t_ik} - y_{uk})^2) \\
& + \sum_{k \in Ind(t_i) \setminus Ind(P_l(t_i))} \alpha_k(x_{t_ik}^2 + y_{t_ik}^2) + \sum_{k \in Ind(u) \setminus Ind(t_i)} \alpha_k(x_{uk}^2 + y_{uk}^2).
\end{aligned} \tag{3.53}
$$

Also, because u is an attribute subtree of the tree $P_l(t_i)$,

$$
f_\delta^2(P_l(t_i), u) = \sum_{k \in Ind(P_l(t_i)) \setminus Ind(u)} \alpha_k(x_{P_l(t_i)k}^2 + y_{P_l(t_i)k}^2). \tag{3.54}
$$

By Equations (3.50), (3.53) and (3.54),

$$\begin{aligned}
&\sum_{i=K+1}^{n} (f_\delta^2(t_i,u) - f_\rho^2(t_i,P_l(t_i)) - f_\delta^2(P_l(t_i),u)) \\
&= \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} \alpha_k[(x_{t_ik}^2 + y_{t_ik}^2) - ((x_{t_ik} - x_{P_l(t_i)k})^2 \\
&+ (y_{t_ik} - y_{P_l(t_i)k})^2) - (x_{P_l(t_i)k}^2 + y_{P_l(t_i)k}^2)] \\
&= \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2 + y_{t_ik}y_{P_l(t_i)k} - y_{P_l(t_i)k}^2) \qquad (3.55) \\
&= \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2) \\
&+ \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} 2\alpha_k(y_{t_ik}y_{P_l(t_i)k} - y_{P_l(t_i)k}^2) \\
&= S_1 + S_2.
\end{aligned}$$

Next, it will be shown that $S_1 = 0$ and $S_2 = 0$. In fact,

$$\begin{aligned}
S_1 &= \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2) \\
&= \sum_{i=1}^{n} \sum_{k\in Ind(P_l(t_i))\backslash Ind(u)} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2)
\end{aligned} \qquad (3.56)$$

because for $i \le K$, $P_l(t_i) \overset{A}{\subseteq} u$, therefore

$$Ind(P_l(t_i))\backslash Ind(u) = \emptyset.$$

Furthermore, by Equation (3.56),

$$\begin{aligned}
S_1 &= \sum_{i=1}^{n}\sum_{k=1}^{\infty} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2)1\{k \in Ind(P_l(t_i))\backslash Ind(u)\} \\
&= \sum_{k=1}^{\infty}\sum_{i=1}^{n} 2\alpha_k(x_{t_ik}x_{P_l(t_i)k} - x_{P_l(t_i)k}^2)1\{k \in Ind(P_l(t_i))\backslash Ind(u)\}.
\end{aligned} \qquad (3.57)$$

Recall from Definition 3.5.1, two trees u_1 and u_2 on the s-treeline have one of the following two relations: $u_1 \overset{A}{\subseteq} u_2$ or $u_2 \overset{A}{\subseteq} u_1$. Therefore, without loss of generality,

assume

$$P_l(t_1) \overset{A}{\subseteq} P_l(t_2) \overset{A}{\subseteq} \cdots \overset{A}{\subseteq} P_l(t_n). \tag{3.58}$$

For simplicity, denote the tree $P_l(t_n)$ by U and the tree $P_l(t_1)$ by L. Since $P_l(t_i)$ is an attribute subtree of the tree U, for node $k \in Ind(P_l(t_i))$,

$$x_{P_l(t_i)k} = x_{Uk}, \text{ and } y_{P_l(t_i)k} = y_{Uk}. \tag{3.59}$$

Therefore,

$$\begin{aligned} S_1 &= \sum_{k=1}^{\infty} 2\alpha_k (\sum_{i=1}^{n} (x_{t_ik} x_{P_l(t_i)k} - x^2_{P_l(t_i)k}) 1\{k \in Ind(P_l(t_i)) \backslash Ind(u)\}) \\ &= \sum_{k=1}^{\infty} 2\alpha_k [\sum_{i=1}^{n} (x_{t_ik} x_{Uk} - x^2_{Uk}) 1\{k \in Ind(P_l(t_i)) \backslash Ind(u)\}] \\ &= \sum_{k=1}^{\infty} 2\alpha_k x_{Uk} [\sum_{i=1}^{n} (x_{t_ik} - x_{Uk}) 1\{k \in Ind(P_l(t_i)) \backslash Ind(u)\}] \\ &= \sum_{k=1}^{\infty} 2\alpha_k x_{Uk} S_{1,k} \end{aligned} \tag{3.60}$$

Next, it will be shown that $S_{1,k} = 0$.

If $k \notin \cup_{i=1}^{n} \{Ind(P_l(t_i)) \backslash Ind(u)\}$, then $S_{1,k} = 0$.

Otherwise, if $k \in \cup_{i=1}^{n} \{Ind(P_l(t_i)) \backslash Ind(u)\}$, by Equation (3.58), assume that $k \notin Ind(P_l(t_M)) \backslash Ind(u)$, but $k \in Ind(P_l(t_{M+1})) \backslash Ind(u)$, for some $M \geq 1$.

Figure 3.9 shows that $Ind(P_l(t_i)) \backslash Ind(u)$ is a subset of $Ind(t_i)$ because the shaded area is empty. Therefore, $k \in Ind(t_i)$, for $i > M$.

For $i \leq M$, if $k \in Ind(t_i)$, then there exists a tree u^{**} on the treeline, containing $Ind(P_l(t_i)) \cup \{k\}$ and closer to t_i which is a contradiction with the fact that $P_l(t_i)$ is the projection of t_i on l. Thus, $t_{M+1}, \ldots, t_n$ contain node k; while, $t_1, \ldots, t_M$, do not contain node k.

By the fact that $P_l(t_i)$ is an element on the s-treeline l, on which the nodal attributes can be calculated as a sample average (see Definitions 3.5.5 and 3.5.1) and

Equation (3.59),

$$x_{Uk} = \frac{\sum_{j=M+1}^{n} x_{t_j k}}{n - M}. \tag{3.61}$$

Moreover,

$$\begin{aligned} S_{1,k} &= \sum_{i=1}^{n} (x_{t_i k} - x_{Uk}) 1\{k \in Ind(P_l(t_i)) \backslash Ind(u)\} \\ &= \sum_{i=M+1}^{n} (x_{t_i k} - x_{Uk}) \\ &= 0. \end{aligned}$$

Therefore, $S_1 = 0$. Similarly, by the same argument, $S_2 = 0$. Finally, summing Equation (3.44) over $i = 1, 2, \ldots, K$ and combining Equation(3.55) and Equation (3.29), Equation (3.28) holds. □

REMARK 3.5.4. Note that, in Equation (3.56), while the summation of

$$\sum_{k \in Ind(P_l(t_i)) \backslash Ind(u)} 2\alpha_k (x_{t_i k} x_{P_l(t_i) k} - x^2_{P_l(t_i) k})$$

over i is zero, the individual entry need not be zero. Thus, the Pythagorean Theorem 3.5.4 is only true in the stated summation form, for some tree u.

3.6. Principal Component Analysis on Finite Level Trees with Nodal Attributes

In Section 3.1, a new metric δ was defined on the tree space with nodal attributes and a specific metric ρ was defined on the finite level tree space. Note that the metric δ (or, ρ) is the sum of the integer part metric d_I and the fractional part f_δ (f_ρ). Furthermore, the variation of a sample of trees about its "center point"—the median-mean tree, was defined. In this section, the problem of finding simple explanation of the variation of the sample will be discussed.

In standard statistics, principal component analysis (PCA) is a very useful tool to explain the variation in terms of a few orthogonal directions (i.e., one-dimensional

representations). But for tree space which is not a Euclidean space, can an analog of the PCA method be developed?

When all the trees in the sample T have the same tree structure, it is straightforward that the median-mean tree m_δ has the same tree structure as the other trees and the sum of the integer part distances

$$\sum_{i=1}^{n} d_I(t_i, m_\delta) = 0.$$

Also, the attribute vectors have the same length. The fractional part metric f_δ (or, f_ρ) is equivalent to the ordinary Euclidean space on the weighted attribute vectors. In particular, f_ρ is proportional to the ordinary Euclidean distance between two vectors (see Section 3.2). Therefore, the standard PCA can be applied in this case.

Next, a more difficult question is how to analyze the variation when not all the trees have the same tree structure in T. To analyze the variation, both the integer part metric and the fractional part need to be taken into account; that is, both tree structure and nodal attributes should be considered.

Recall that, in Section 2.4 and Section 2.5, the idea of treeline was developed as a one-dimensional representation of the data in the binary tree space. Also, the tree version PCA was developed on tree space without nodal attributes.

Now, on the binary tree space with nodal attributes, the tree version PCA without nodal attributes and the standard PCA on Euclidean space, will be combined to develop a new PCA on tree space with nodal attributes. The tree version PCA without nodal attributes will be used to capture interesting features of the tree structure and the idea of standard PCA will be used to analyze the nodal attributes.

DEFINITION 3.6.1. An s-treeline is called a **one-dimensional principal structure representation (treeline)** of the sample T if it minimizes the sum

$$\sum_{i=1}^{n} V_\delta(t_i, P_l(t_i)) \tag{3.62}$$

over all binary s-treelines l passing through the minimal median-mean tree μ_δ in the sample T.

According to the tree version of the Pythagorean Theorem (Theorem 3.5.4), minimizing the sum (3.62) is equivalent to maximizing the following sum

$$\sum_{i=1}^{n} V_\delta(\mu_\delta, P_l(t_i)). \tag{3.63}$$

Recall from the analysis of the blood vessel data in Section 1.2, Figure 1.9 shows the principal structure treeline $l = \{u_0, u_1, u_2\}$ with nodal attributes, where u_1 is the unique median-mean tree (also the minimal median-mean tree) of the sample. Figure 1.8 shows the topological tree structures of the principal structure treeline in Figure 1.9.

In Example 3.6.1 and 3.6.2, the metric ρ will be used to illustrate the new tree version PCA on the finite level trees with attributes.

EXAMPLE 3.6.1. Let $T = \{t_1, \ldots, t_5\}$ be a sample of finite level trees in $\mathcal{T}_w$ with sample size $n = 5$, where w is shown in Figure 3.10. The nodal attributes for those five trees are shown in Table 3.10.

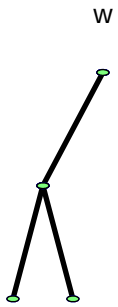

FIGURE 3.10. Binary tree w.

The support tree t_{sup} and average support tree t_a of the sample T are shown in Figure 3.12.

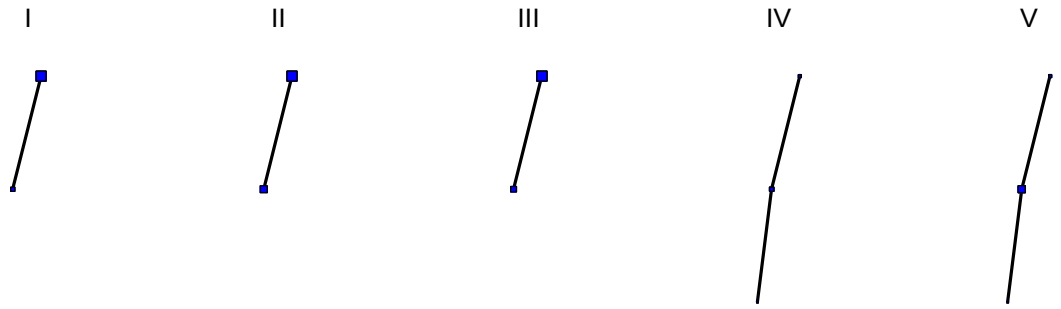

FIGURE 3.11. The binary tree sample T.

level-order index	I	II	III	IV	V
1	(0.35,0.35)	(0.35,0.35)	(0.35,0.35)	(0.10,0.10)	(0.10,0.10)
2	(0.15,0.15)	(0.25,0.25)	(0.20,0.20)	(0.15,0.15)	(0.25,0.25)
4	n/a	n/a	n/a	(0.05,0.05)	(0.05,0.05)

TABLE 3.10. Nodal attributes of the tree sample in Example 3.6.1.

FIGURE 3.12. Support tree and average support tree of the sample T.

The median-mean tree m_ρ, center point of the sample T, is shown in Figure 3.13. Note that there is a unique median-mean tree for the sample T. Therefore, the tree m_ρ is also the minimal median-mean tree. The nodal attributes of the average support tree t_a and the median-mean binary tree m_ρ are listed in Table 3.11.

Some calculation shows that the total variation to the center point is

$$\sum_{i=1}^{5} V_\rho(t_i, m_\rho) = 2.045 \tag{3.64}$$

where $V_\rho = d_I + f_\rho^2$ and $N(w) = 4$ in the definition of the factional part metric f_ρ.

level-order index	t_a	m_ρ
1	(0.25,0.25)	(0.25,0.25)
2	(0.2,0.2)	(0.2,0.2)
4	(0.05,0.05)	n/a

TABLE 3.11. Nodal attributes of the trees t_a and m_ρ.

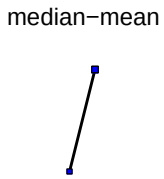

FIGURE 3.13. The median-mean tree m_ρ of the sample T.

Next, find the treeline to describe features of the data. Note that a reasonable s-treeline for this sample $l = \{u_1, u_2, u_3\}$ is shown in Figure 3.14.

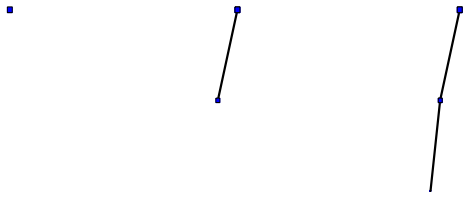

FIGURE 3.14. The s-treeline $l = \{u_1, u_2, u_3\}$ for the sample T. This is the one-dimensional principal structure representation.

Also, the projections of the five types of trees are u_2, u_2, u_2, u_3, u_3 respectively. Some calculation results in:

$$\sum_{i=1}^{5} V_\rho(P_l(t_i), m_\rho) = 2.0025$$

and

$$\sum_{i=1}^{5} V_\rho(P_l(t_i), t_i) = 0.0425,$$

which verifies the tree version of the Pythagorean Theorem, i.e.,

$$\sum_{i=1}^{5} V_\rho(t_i, m_\rho) = \sum_{i=1}^{5} V_\rho(P_l(t_i), m_\rho) + \sum_{i=1}^{5} V_\rho(P_l(t_i), t_i).$$

In Example 3.6.1, the proportion of variation that the one-dimensional structure representation explains is

$$\frac{\sum_{i=1}^{5} V_\rho(P_l(t_i), m_\rho)}{\sum_{i=1}^{5} V_\rho(t_i, m_\rho)} = \frac{2.0025}{2.045} = 97.92\%.$$

There is no other s-treeline to explain more about the total variation in Example 3.6.1. Now, the other type of treeline — a-treeline, will be applied.

Recall that, in Definition 3.5.3, an a-treeline is determined by a tree u_0 and an attribute vector $\vec{v}$.

DEFINITION 3.6.2. Let $\vec{c}$ be any vector of attributes. An a-treeline e, determined by u_0 and $\vec{v}$, is called a **$\vec{c}$-induced a-treeline** if $\vec{v}$ is a restriction of $\vec{c}$, in particular,

$$(v_{2k-1}, v_{2k}) = \begin{cases} (c_{2k-1}, c_{2k}), & \text{if } k \in Ind(u_0) \\ \\ (0,0), & \text{if } k \notin Ind(u_0). \end{cases} \tag{3.65}$$

Each tree t_j, it has a unique projection $P_l(t_j)$ on the s-treeline l, which is a one-dimensional structure representation. For any vector $\vec{c}$ and tree $P_l(t_j)$, there is a $\vec{c}$-induced a-treeline e_j. Now, find a vector (first principal direction $\vec{p}_1$) which minimizes

$$\sum_{j=1}^{n} V_\delta(t_j, P_{e_j}(t_j)). \tag{3.66}$$

over all vectors $\vec{c}$. The corresponding induced a-treelines are called principal attribute treelines.

DEFINITION 3.6.3. The vector $\vec{p}_1$, which minimizes Equation (3.66) over all vectors $\vec{c}$, is called the **principal attribute direction**.

Similar to the PCA in ordinary Euclidean space, other orthogonal (with respect to the inner product in the embedded Euclidean space defined at Equation (3.8)) vectors $\vec{p}_2, \vec{p}_3, \ldots$, can be defined. Furthermore, denote the induced a-treeline by the vector $\vec{p}_k$ passing through the tree $P_l(t_j)$ by e_{jk}.

The idea of the $\vec{c}$-induced a-treeline is now illustrated in the context of the previous Example 3.6.1. The first principal direction is

$$\vec{p_1} = [1, 1, 0, 0, 0, 0, 0, 0]',$$

because the dominant variation is in the direction of $x_1 = y_1$. The second principal direction is

$$\vec{p_2} = [0, 0, 1, 1, 0, 0, 0, 0]',$$

because the second orthogonal direction of variation is $x_2 = y_2$.

Thus,

$$\sum_{i=1}^{5} V_\rho(P_l(t_i), P_{e_{i1}}(t_i)) = 0.0375 \tag{3.67}$$

and

$$\sum_{i=1}^{5} V_\rho(P_l(t_i), P_{e_{i2}}(t_i)) = 0.005. \tag{3.68}$$

According to the Equations (3.67) and (3.68), the ANOVA

$$\sum_{i=1}^{5} V_\rho(P_l(t_i), P_{e_{i1}}(t_i)) + \sum_{i=1}^{5} V_\rho(P_l(t_i), P_{e_{i2}}(t_i)) = 0.0425 = \sum_{i=1}^{5} V_\rho(P_l(t_i), t_i)$$

is straightforward.

Note that, in Example 3.6.1, the total variation, 2.045, was decomposed into three parts. The first part, 2.0025, was explained by the first principal structure treeline. And, two attribute directions explained 0.0375 and 0.005 respectively.

EXAMPLE 3.6.2. Let w be a tree with level-order index set $Ind(w) = \{1, 2, 3, 7\}$. $T = \{t_1, t_2, \ldots, t_n\}$ is a sample of trees in the tree subspace $\mathcal{T}_w$. Also, there are four types of trees in the sample T, type I, II, III, IV (see Table 3.12 and Figure 3.15).

Type III and IV are replicated m times, so that the numbers of elements of each type are 1, 1, m and m ($m > 1$).

level-order index	I	II	III	IV
1	(0.3,0.3)	(0.2,0.2)	(0.3,0.3)	(0.2,0.2)
2	(0.2,0.2)	(0.2,0.2)	(0.2,0.2)	(0.2,0.2)
3	(0.1,0.1)	(0.3,0.3)	(0.1,0.1)	(0.3,0.3)
7	(0.1,0.1)	(0.1,0.1)	n/a	n/a

TABLE 3.12. Nodal attributes of the four basic trees.

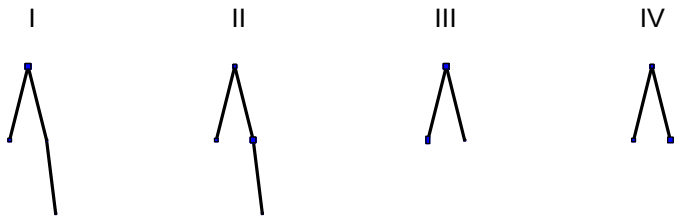

FIGURE 3.15. Four types of tree structures in Example 3.6.2.

The median-mean tree m_ρ (it is also the minimal median-mean tree, because it is unique when $m > 1$) and the average support tree are shown in Figure 3.16.

FIGURE 3.16. The median-mean tree and the average support tree in Example 3.6.2.

The total variation is

$$\sum_{i=1}^{2m+2} V_\rho(t_i, m_\rho) = 0.0125m + 2.0225. \tag{3.69}$$

Two reasonable s-treelines passing through the unique median-mean tree, l_1 and l_2, are shown in Figure 3.17 and Figure 3.18.

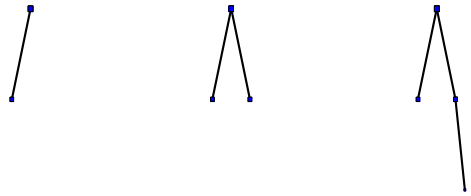

FIGURE 3.17. The structure treeline l_1 in Example 3.6.2.

FIGURE 3.18. The structure treeline l_2 in Example 3.6.2.

By calculation,

$$\sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), t_i) = 0.0125m + 0.0125,$$

and

$$\sum_{i=1}^{2m+2} V_\rho(P_{l_2}(t_i), t_i) = 0.0125m + 2.0225.$$

Hence, the s-treeline l_1 is the one-dimensional principal structure representation. By the tree version of the Pythagorean Theorem,

$$\sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), m_\rho) = \sum_{i=1}^{2m+2} V_\rho(t_i, m_\rho) - \sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), t_i) = 2.01.$$

Therefore, the proportion of the total variation that the one-dimensional principal structure representation l_1 explains is

$$\frac{\sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), m_\rho)}{\sum_{i=1}^{2m+2} V_\rho(t_i, m_\rho)} = \frac{2.01}{0.0125m + 2.0225}$$

Note that this proportion is arbitrarily small by taking m large. Thus, this is an example where the tree structure component of variability (in either the l_1 or l_2) is negligible. So, it is important to also analyze the attribute components.

Next, find the principal attribute direction to decompose the variation in the direction of the attributes.

By calculation, the first principal attribute direction is

$$\vec{p_1} = [-1, -1, 0, 0, 2, 2, 0, 0]';$$

Furthermore,

$$\sum_{i=1}^{2m+2} V_\rho(P_{e_{i1}}(t_i), t_i) = 0.$$

By the tree version of the Pythagorean Theorem, the proportion of variation explained by the first principal attribute direction is

$$\frac{0.0125m + 0.0125}{0.0125m + 2.0225},$$

which converges to 1 as $m \to \infty$.

3.7. Computation of the Principal Attribute Direction

In Section 3.6, a new tool of variation analysis on the tree space with nodal attributes, the tree version of the PCA, was developed. This new method decomposed the total variation by finding the one-dimensional principal structure representation (see Definition 3.6.1) and the principal attribute direction (see Definition 3.6.3).

For serious real data problems, computation of the variation analysis is challenging. In this section, an algorithm is proposed, which provides a simple approximation of the principal attribute direction.

Let $T = \{t_1, \ldots, t_n\}$ be a sample of trees. The treeline l is the one-dimensional principal structure representation. The projection of the tree t_i onto the treeline l is denoted by the projection function, $P_l(t_i)$, for $i = 1, \ldots, n$.

Next, consider the average support tree, w, of the tree sample T. Then, every tree t_i is a member of $\mathcal{T}_w$, i.e., $Ind(t_i) \subseteq Ind(w)$. Recall from Section 3.2 that, by the Padding Rule for Attribute Vectors, there exists an attribute vector, with length r (of the attribute vector of the tree w), associated with each tree t_i. The goal is to find a vector which minimizes

$$Z(\vec{v}) = \sum_{j=1}^{n} V_\delta(t_j, P_{e_j}(t_j)) \tag{3.70}$$

over all vectors $\vec{v}$, where e_j is the $\vec{v}$-induced attribute treeline passing through the tree $P_l(t_j)$. Figures 3.19 - 3.21 will illustrate the idea in two dimensional space.

Algorithm:

(1) For a starting direction $\vec{v}_0$ (shown as $\overrightarrow{OA_0}$ in Figure 3.19), calculate the sum, $Z(\vec{v}_0)$.

(2) Starting from the vector $\overrightarrow{OA_0}$, compare the sum Z for four directions $\overrightarrow{OB_i}$, $i = 1, 2, 3, 4$ (see Figure 3.20) and find the minimum value, without loss of generality, assume $\overrightarrow{OB_1}$ is the smallest.

(3) Normalize the vector $\overrightarrow{OA_1}$=$\overrightarrow{OB_1}/\|\overrightarrow{OB_1}\|$, and repeat step (2) by starting from the vector $\overrightarrow{OA_1}$ (see Figure 3.21).

Repeat this process, say k times, the approximate direction will be provided. For general nodal attribute vectors with length r, in step (2), Z will be calculated over $2r$, instead of 2^r, different directions. Those $2r$ directions are generated by starting from the direction $\overrightarrow{OA_0}$ choosing one coordinate each time, and adding or subtracting a constant from that coordinate.

The solution provided by this algorithm depends on the choice of the starting direction. Sometimes, when Z has multi-minima, this algorithm may provide a local minimum instead of the global minimum. It is of interest to compare the results by choosing different starting directions. For computational speed, all the results of the blood vessel data as shown in Sections 1.2 and 3.9 are based on a single starting

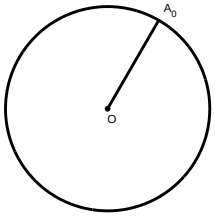

FIGURE 3.19. Step (1).

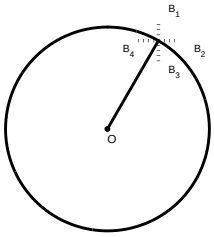

FIGURE 3.20. Step (2).

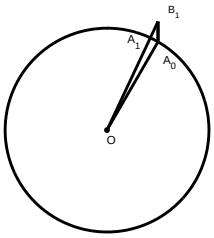

FIGURE 3.21. Step (3).

direction. Comparison of multiple choices of the starting direction has not been tried yet.

3.8. Comparison of the Tree Version PCA and Regular PCA

In this section, comparison between the tree version Principal Component Analysis and the regular Principal Component Analysis will be provided by using some toy examples.

EXAMPLE 3.8.1. The attributes of a toy sample of trees, all with same tree structure of size $n = 6$, are given in the Table 3.13.

Since all the trees have the same structure, for any two trees, the integer part metric is zero. Thus, the metric is determined by the fractional part metric. Recall that, in Section 3.1, the fractional part metric can be viewed as the regular Euclidean distance of two weighted attribute vectors.

level-order index	Attributes
1	(0.25,0.25)
2	(-0.25,-0.25)
3	(0.05,0.05)

level-order index	Attributes
1	(0.15,0.15)
2	(0.25,0.25)
3	(0.05,0.05)

level-order index	Attributes
1	(-0.20,-0.20)
2	(0.15,0.15)
3	(-0.05,-0.05)

level-order index	Attributes
1	(0.25,0.25)
2	(-0.15,-0.15)
3	(-0.05,-0.05)

level-order index	Attributes
1	(-0.20,-0.20)
2	(-0.35,-0.35)
3	(0.05,0.05)

level-order index	Attributes
1	(-0.25,-0.25)
2	(0.35,0.35)
3	(-0.05,-0.05)

TABLE 3.13. The attributes of six trees in the tree sample.

The corresponding weighted attribute vectors are

$$\begin{pmatrix} \frac{0.25}{\sqrt{2}} \\ \frac{0.25}{\sqrt{2}} \\ \frac{-0.25}{\sqrt{2^3}} \\ \frac{-0.25}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \end{pmatrix}, \begin{pmatrix} \frac{0.15}{\sqrt{2}} \\ \frac{0.15}{\sqrt{2}} \\ \frac{0.25}{\sqrt{2^3}} \\ \frac{0.25}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \end{pmatrix}, \begin{pmatrix} \frac{-0.20}{\sqrt{2}} \\ \frac{-0.20}{\sqrt{2}} \\ \frac{0.15}{\sqrt{2^3}} \\ \frac{0.15}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \end{pmatrix}, \begin{pmatrix} \frac{0.25}{\sqrt{2}} \\ \frac{0.25}{\sqrt{2}} \\ \frac{-0.15}{\sqrt{2^3}} \\ \frac{-0.15}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \end{pmatrix}, \begin{pmatrix} \frac{-0.20}{\sqrt{2}} \\ \frac{-0.20}{\sqrt{2}} \\ \frac{-0.35}{\sqrt{2^3}} \\ \frac{-0.35}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \\ \frac{0.05}{\sqrt{2^3}} \end{pmatrix}, \begin{pmatrix} \frac{-0.25}{\sqrt{2}} \\ \frac{-0.25}{\sqrt{2}} \\ \frac{0.35}{\sqrt{2^3}} \\ \frac{0.35}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \\ \frac{-0.05}{\sqrt{2^3}} \end{pmatrix}.$$

The total variation, in the regular PCA sense, is 0.3975. Also, three non-zero eigenvalues are

$$\lambda_1 = 0.0611, \lambda_2 = 0.0178, \lambda_3 = 0.0006.$$

The first eigenvector is

$$[-0.6814, -0.6814, 0.1870, 0.1870, -0.0280, -0.0280]'.$$

Recall that, this eigenvector has been computed for the transformed data. Applying the inverse-transformation, the eigenvector is

$$\vec{v}_1 = [-0.6183, -0.6183, 0.3393, 0.3393, -0.0508, -0.0508]'.$$

Note that the first principal component explains

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 76.86\%$$

of the total variation.

Now applying the tree version PCA, the total variation is 0.3975 and the *approximate* first principal (attribute) direction

$$\vec{v}_2 = [-0.6180, -0.6182, 0.3396, 0.3397, -0.0512, -0.0504]'.$$

Note that there is some difference between $\vec{v}_1$ and $\vec{v}_2$ due to the error in the approximation of $\vec{v}_2$. The variation explained by the first principal (attribute) direction $\vec{v}_2$ is 0.3055. Hence, the proportion explained by the first principal (attribute) direction $\vec{v}_2$ is

$$\frac{0.3055}{0.3975} = 0.7686.$$

Example 3.8.1 verifies that, for the sample of trees with the same structure, the tree version PCA and the regular PCA obtain the same result essentially. In the following Example 3.8.2, the members in a sample of trees have different structures. And, the tree version PCA finds a more appropriate mode of variation than the one given by regular PCA.

EXAMPLE 3.8.2. Let $T = \{t_1, t_2, \ldots, t_{13}\}$ be a sample of trees with size $n = 13$. Each member in T has one of the two structures shown in Figure 3.22. The attributes have the form shown in Table 3.14, where x and y are real values in $[0, \frac{\sqrt{2}}{4}]$. Note

that, for the trees without node 3, the corresponding nodal attributes are denoted as $(\star, \star)$ (see Table 3.15). Thus, trees $t_1, t_2, \ldots, t_7$ have three nodes, while the others have two nodes.

FIGURE 3.22. Two types of tree structures in T.

Level-order index	Attributes
1	(0.1,0.1)
2	(x,x)
3	(y,y)

TABLE 3.14. Attribute form of trees in T.

	1	2	3	4	5	6	7
x	0.267	0.280	0.250	0.241	0.242	0.251	0.252
y	0.220	0.230	0.200	0.180	0.180	0.190	0.190
	8	9	10	11	12	13	
x	0.276	0.285	0.266	0.210	0.220	0.200	
y	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	

TABLE 3.15. Values of x's and y's for all trees in T.

To apply the regular PCA, a sequence of equal-dimensional vectors are needed. A natural approach is to substitute the non-existent nodal attributes $\star$ by the sample average of all the others.

The corresponding weighted attributes are

$$(\frac{x}{\sqrt{2^3}}, \frac{x}{\sqrt{2^3}}) \text{ and } (\frac{y}{\sqrt{2^3}}, \frac{y}{\sqrt{2^3}})$$

for node 2 and node 3 respectively. Hence, the weighted attribute vector can be written as

$$[\frac{0.1}{\sqrt{2}}, \frac{0.1}{\sqrt{2}}, \frac{x}{\sqrt{2^3}}, \frac{x}{\sqrt{2^3}}, \frac{y}{\sqrt{2^3}}, \frac{y}{\sqrt{2^3}}].$$

It shows that $x/\sqrt{8}$ and $y/\sqrt{8}$ are the two important components of the attribute vector. For simple visualization, in the later study, the principal components will be represented in two-dimensional space of x and y, instead of six-dimensional space.

The scatter plot of the attributes, x and y, is shown in Figure 3.23. It shows that, the attributes of the Type II trees (seven, shown with "+") forms a pattern from lower left to upper right. The attributes of the Type I trees (six, shown with "×") have been divided into two groups with a gap in the middle.

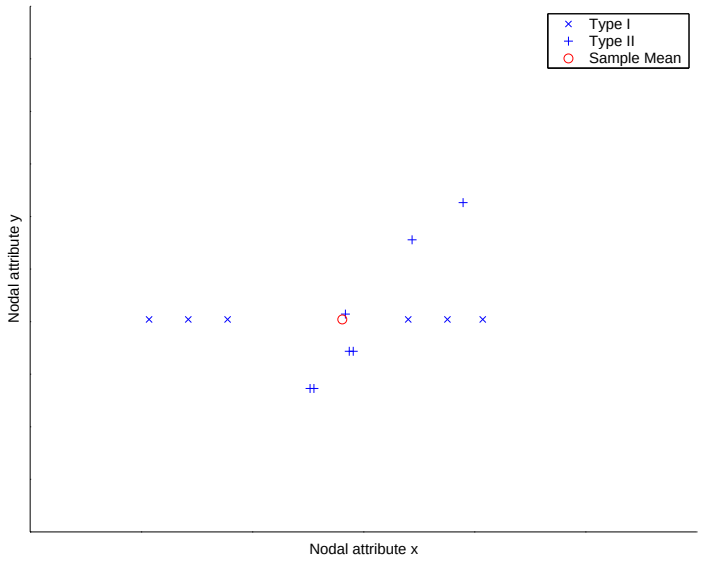

FIGURE 3.23. Scatter plot of the nodal attributes.

Applying the regular PCA to the weighted attribute vectors, gives the first principal direction (first eigenvector, the solid line in Figure 3.24). It shows that the trees

with the Type I structure have a strong effect on the attribute direction, pulling it towards a horizontal line.

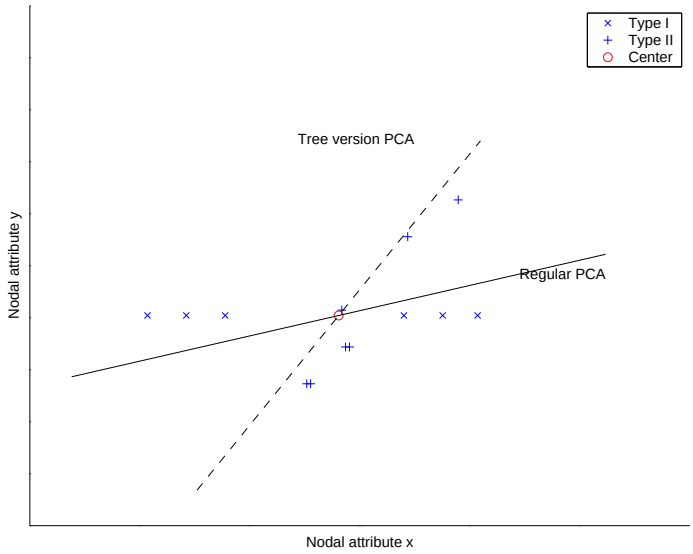

FIGURE 3.24. Principal attribute directions given by Regular PCA and Tree version PCA.

Next, the tree version PCA will be applied to the tree sample T. The tree version PCA has two steps, finding the principal structure direction and finding the principal attribute direction.

The first two elements (denoted as u_0 and u_1) on the principal structure treeline l is shown in Figure 3.25. Note that u_1 is the median-mean tree of the sample T. Moreover, the elements in T can be categorized by projection on the treeline l. The trees with Type I structure have projection u_0 on the treeline l; while, the trees with Type II structure have projection u_1 instead.

FIGURE 3.25. Principal structure treeline $l = \{u_0, u_1\}$.

Based on the principal structure treeline, the principal attribute direction is calculated and shown as the dashed line in Figure 3.24. Comparing with the direction given by regular PCA, it is more appropriate for the reason that it represents the relation of the (weighted) attributes. The Type I elements should not influence the direction because they contain no information about the relationship between the attributes x and y.

Next, the attributes of the six trees with Type I structure will be studied. All these six trees have a common projection on the principal structure treeline u_0. The projection coefficients of these trees on the attributes treeline passing through u_0 are shown in Figure 3.26. Note that there is a big jump from the negative coefficients to the positive ones.

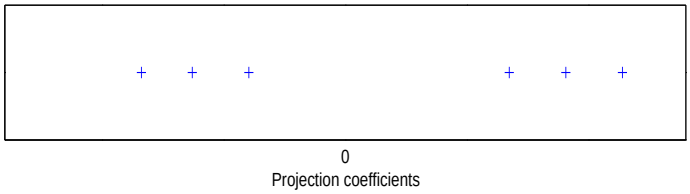

FIGURE 3.26. Projection coefficients of the trees with Type I structure on the principal attribute direction.

This section shows that the tree version PCA is a generalization of the regular PCA. When all the trees in the sample have the same structure, the principal attribute direction is the same as the first eigenvector given by the regular PCA with inverse-transformation. When the structures are not all the same, the tree version PCA will give a more appropriate attribute direction.

3.9. More Data Analysis on the Blood Vessel Data

In Section 1.2, an exploratory data analysis based on the eleven blood vessel trees from three people was discussed. In that analysis, the "central trees", median-mean trees, are given for the reduced linear trees of each patient (see Figure 1.6) and the

combined population (see Figure 1.7). Furthermore, the tree version PCA found a surprising characteristic of the population, that there are two different orientations about the blood flow in the data set. This dominated the total variation, perhaps obscuring population features of more biological interest.

In this section, the brain blood vessel data collected from 10 people has been provided. An example of a brain blood vessel system is shown in Figure 1.4. There are three important components for the brain blood vessel system: left carotid, right carotid and vertebrobasilar system. Each component will be represented as a tree-structured object. This data set has 30 trees from 10 people, that is, three components for each person.

Each component consists of one root vessel and many offspring vessels. Each blood vessel is denoted as a node in the tree structure. For each blood vessel, information such as parent's ID, the point of attachment on the parent, the coordinates of a sequence of points along this blood vessel are recorded.

To avoid the difference between patients, the shape will be rescaled by subtracting the minimum and dividing the range, for each coordinate. Those rescaled coordinates are used in the following.

Like the analysis in Section 1.2, here only a simple linear approximation of each blood vessel is used. The nodal attributes for the root node have the following form

$$[0, \text{three coordinates of the starting point}, \text{three coordinates of the ending point}];$$

while, the following attributes of the non-root nodes are used

$$[p, 0, 0, 0, \text{three coordinates of the ending point}],$$

where p is the proportion parameter,

$$p = \frac{\text{Distance of starting point to point of attachment on its parent}}{\text{Distance of starting point to ending point on its parent}}.$$

In this application, each node has seven attributes and each attribute is between 0 and 1. Recall from Section 3.1 that, to make the fractional part metric f_δ no more than 1, the assumption that each node consists of two attributes and each attribute is bounded by $[-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}]$ was made. Here, the attributes are divided by $2\sqrt{7}$. Hence, in this new analysis, the attributes are between 0 and $\frac{1}{2\sqrt{7}}$.

By the definition of the metric δ (see Equations (3.1), (3.2) and (3.3)), the nodal attributes, which only appeared in one tree, are treated as $(0, 0)$. Here, the value 0 is an extreme of the nodal attributes. Therefore, for each node, those nodal attributes need to be centralized by subtracting the average nodal attributes of this node (the sum of nodal attributes divided by the number of appearances of this node in this population).

Again, for reasons of computational tractability (see Section 3.7 for the current algorithm), like the analysis in Section 1.2, an attribute subtree of each component is considered. Next, take the root node and the first two nodes closest to the starting point of the root node by comparing the total number of voxels between the starting point of the root and the vessels. Therefore, there are two types of tree structures among those 30 trees (see Figure 3.27). Note that, in Figure 3.27, the edge between two nodes only shows that the corresponding two vessels are connected, and its length and orientation are not meaningful here.

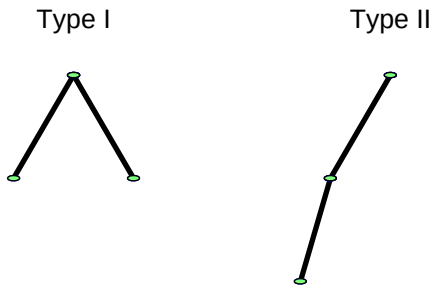

FIGURE 3.27. Two types of tree structures of the reduced blood vessel trees.

Next, the features of left carotid and right carotid systems will be explored.

First, look at the population of all 10 left carotid vessel trees (see Figure 3.28). Among those 10 left carotid trees, six of them have Type I tree structure and the other four trees have Type II structures. By the majority rule (see Theorem 2.3.2 for the algorithm), the median-mean tree must have Type I structure.

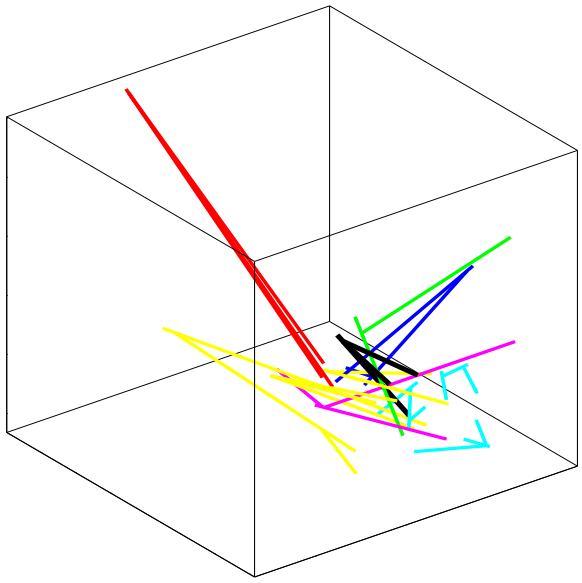

FIGURE 3.28. The population of 10 left carotid vessel trees. The black thicker tree is the median-mean tree.

Figure 3.29 illustrates the principal structure treeline without nodal attributes; while Figure 3.30 shows the same structure treeline with nodal attributes. Figure 3.29 shows the dominate direction of the topological structure changing, u_1 and u_2 adding one left node on u_0 and u_1. It is hard to see differences in the parts of Figure 3.30, because the added branches are very close to each other.

Note that no one in the left carotid population has projection of u_0 onto this structure treeline. Also, the Type I tree has projection u_1 on the treeline and Type II tree has projection u_2.

Next, find the principal attribute direction. Recall from Section 3.5 that, there are uncountably many elements on the attribute treeline. Figure 3.31 and Figure 3.33 illustrate the principal attribute treelines passing through the median-mean tree

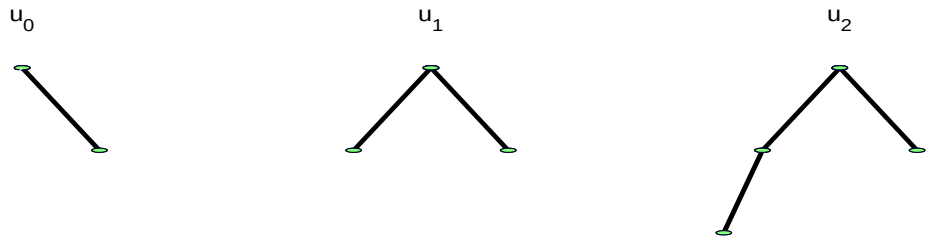

FIGURE 3.29. Principal structure treeline $l = \{u_0, u_1, u_2\}$ without nodal attributes.

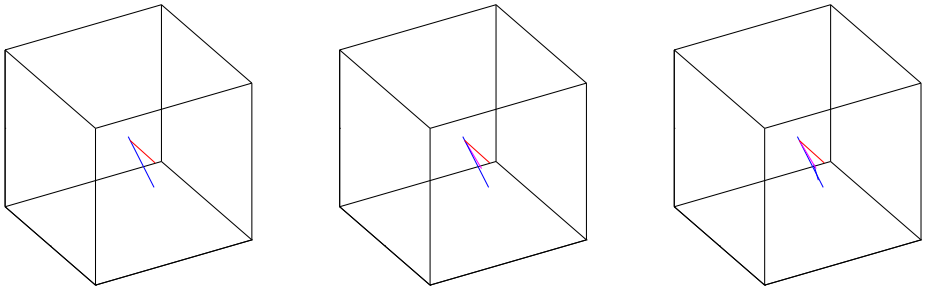

FIGURE 3.30. Principal structure treeline $l = \{u_0, u_1, u_2\}$ with nodal attributes, with 2, 3 and 4 nodes respectively.

and the average support tree.[2] There are 9 subplots in each figure and each subplot depicts one location on the attribute treeline.

In Figure 3.31, from the first 4 subplots, it shows that the vessel tree becomes smaller, but the orientation of the main root does not change too much. In the next 5 subplots, the vessel tree becomes larger and the orientation of the main root is changing to a different direction at the same time. This is not a surprising feature of the population, because in Figure 3.28, the trees in red color and yellow color form a pattern of becoming smaller and shifting down (which is illustrated in the first 4 subplots in Figure 3.31); while, for the green-colored and cyan-colored trees, the orientation of the main root goes in a much different direction. Also, Figure 3.32 shows the projection coefficients of the 10 trees on the attribute treeline passing

[2]In this section, the principal attribute direction is calculated for the centralized data. But, the central tree will be "added" when plot those blood vessel trees.

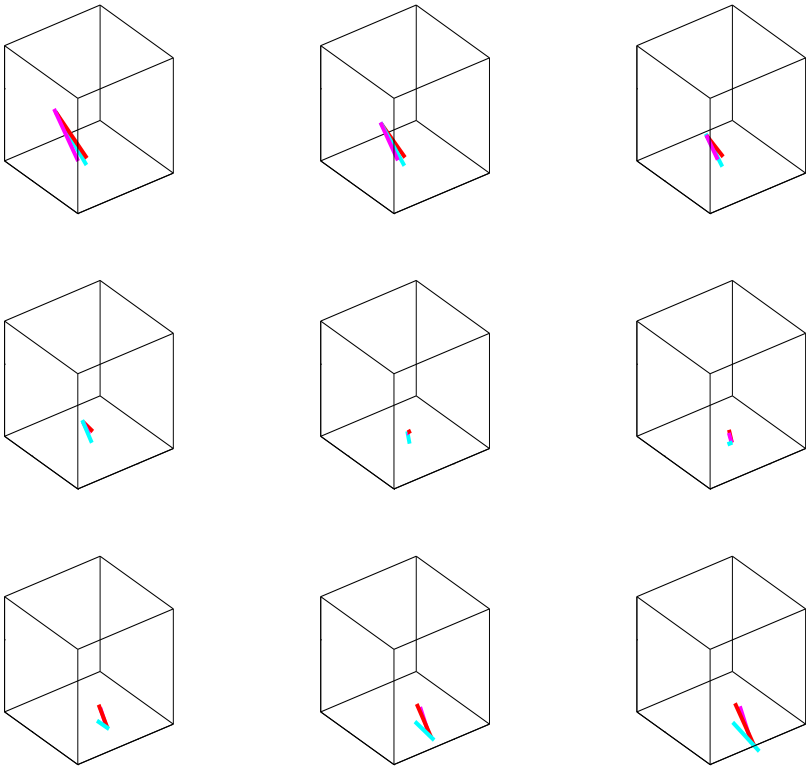

FIGURE 3.31. Principal attribute treeline passing through the median-mean tree for the population of left carotid trees. The cyan-colored branch is the root.

through the median-mean tree. There are two small clusters, indicated by yellow and cyan colors. The red-colored tree is far from the center, which can be seen from the population plot (Figure 3.28).

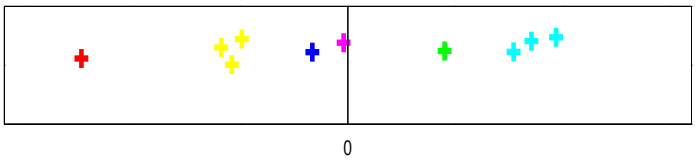

FIGURE 3.32. Projection coefficients of 10 trees on the attribute treeline passing through the median-mean tree for the population of left carotid trees, colored as in Figure 3.28.

Figure 3.33 shows the attribute treeline passing through the average support tree. Similar to Figure 3.31, it shows the same features that the vessel trees become smaller and then the orientation of the main root is jumping out. The projections shown in Figure 3.34 are very similar to those in Figure 3.32. This shows that for this data set, there is little difference between projection on the treelines passing through median-mean tree and the average support tree.

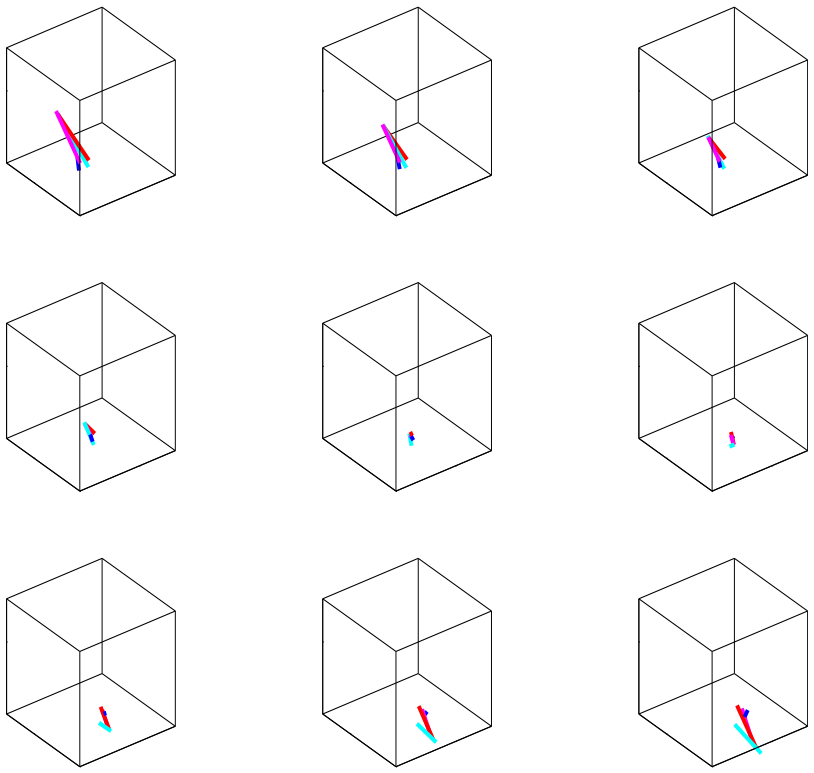

FIGURE 3.33. Principal attribute treeline passing through the average support tree for the population of left carotid trees. The cyan-colored branch is the root.

Next, the population of right carotid vessel trees will be analyzed as shown in Figure 3.35. Among those 10 trees, there are the same two possible tree structures as shown in Figure 3.27, and 5 trees each.

By the majority rule (Theorem 2.3.2), the median-mean tree is not unique. Figure 3.36 shows all four median-mean trees (structure only, without nodal attributes).

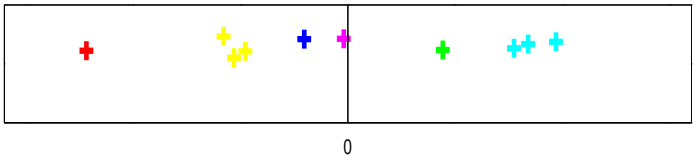

FIGURE 3.34. Projection coefficients of 10 trees on the attribute tree-line passing through the average support tree for the population of left carotid trees, colored as in Figure 3.28.

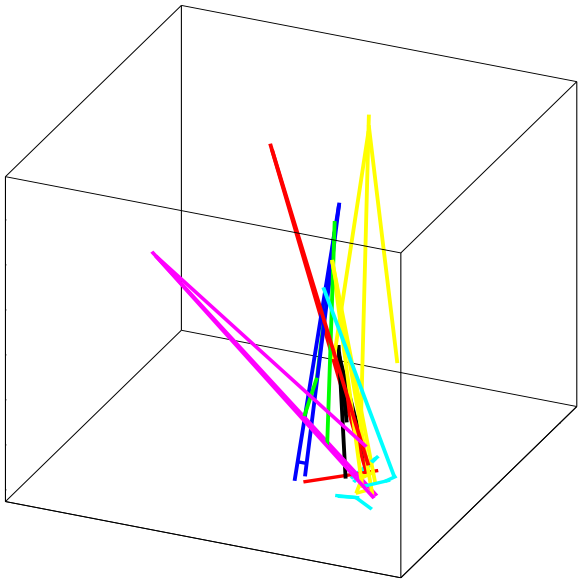

FIGURE 3.35. A population of 10 right carotid vessel trees, colored as in Figure 3.28. The black thicker tree is one of the four median-mean trees, and its topological structure is the second shown in Figure 3.36.

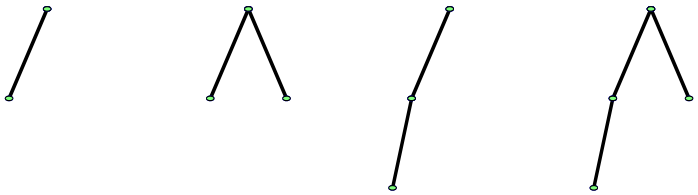

FIGURE 3.36. The topological structures of four median-mean trees of the population of 10 right carotid vessel trees.

Recall from Section 3.4 that, the total variation about the median-mean tree is constant and does not depend on how the tie is broken between the median-mean trees. In this paper, the general recommendation is to use the minimal median-mean tree (denoted by m_1, the first one shown in Figure 3.36) because it is unique. But, because 5 elements in the population have the same tree structure as the second median-mean tree (denoted by m_2) and none have the m_1 structure, the following calculations are based on the second median-mean tree, m_2.

There are two structure treelines that are important to this population l_1 and l_2, shown in Figure 3.37 and Figure 3.38.

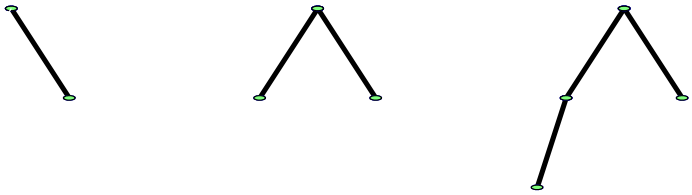

FIGURE 3.37. Structure treeline l_1.

FIGURE 3.38. Structure treeline l_2.

By calculation, both l_1 and l_2 are one-dimension principal structure representation. In the process of finding the principal attribute direction, it may depend on the choice of the structure treeline. Here, the treeline l_1 is used for the calculation purpose of the principal attribute direction.

Like Figure 3.31, Figure 3.39 illustrates the principal attribute treeline passing though the second median-mean tree by depicting 9 locations on this attribute treeline. The first four subplots indicate the attribute changing in the direction of the

orientation of the root node; while the next five subplots indicate the changing in terms of the size of the vessel trees. Because of human symmetry, similarities are expected in Figure 3.31. A clear difference is visible in the ordering, which is the same phenomenon as the arbitrary directions of eigenvector (positive or negative) in PCA. When the ordering in 3.39 is reversed, then a common structure can be seen.

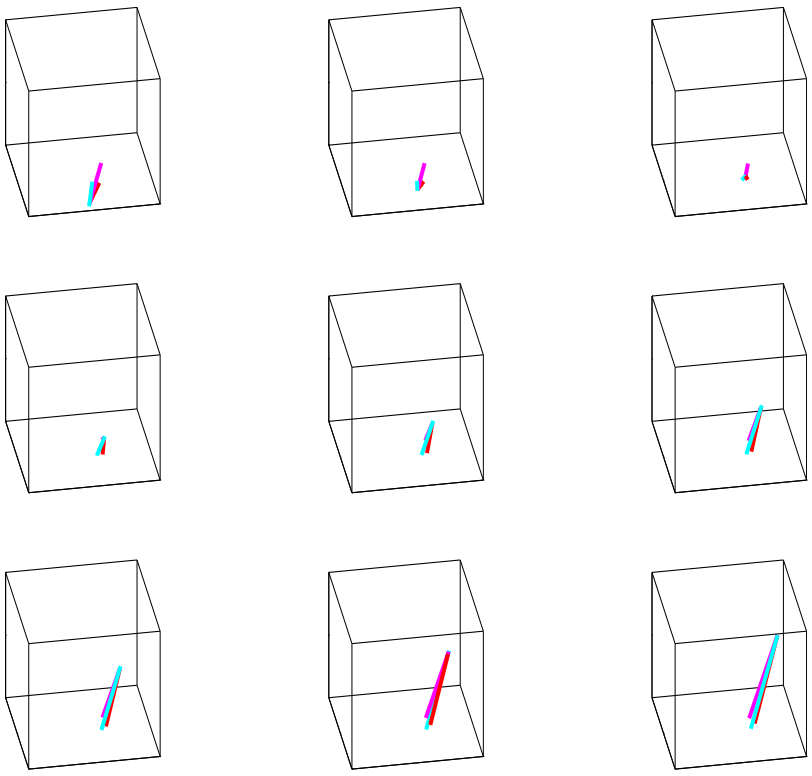

FIGURE 3.39. Principal attribute treeline passing through the median-mean tree for the population of right carotid trees. The cyan-colored branch is the root.

Figure 3.40 shows the projection coefficients on this attribute treeline. Comparing with the projection coefficients shown in Figures 3.32 and 3.34, these 10 right carotid trees are divided into two groups by the projection with a gap in the middle, denoted as "$\bigtriangledown$" and "$\bigtriangleup$". A further check of symmetry comes from comparison with Figure 3.32, so the same colors have been used. The groupings from Figure 3.32 still hold

roughly although one yellow point has changed group. Thus, approximate symmetry holds, but that case exhibits clear asymmetry.

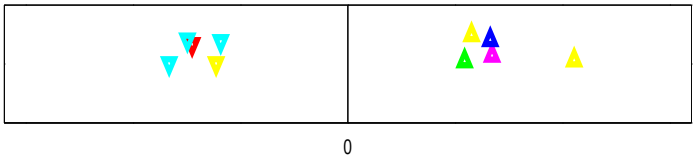

FIGURE 3.40. Projection coefficients of 10 trees on the attribute treeline passing through the median-mean tree for the population of right carotid trees. Colors are the same as Figure 3.32.

Figure 3.41 shows the principal attribute treeline passing through the average support tree. And, Figure 3.42 shows the projections on this treeline. The main ideas are quite similar as for Figure 3.39, with again similar lessons. As in Figure 3.40, the population of right carotid trees are divided into two groups.

The analysis of the vertebrobasilar vessel trees is not shown here, because it was quite similar to the left and right carotid trees.

In this example, the tree version PCA gave an analysis of the characteristics of two populations (left carotid and right carotid). The dominant component of variation of the attributes consists of the orientation changing of the root node and the enlargement of the size of the vessel trees.

In this data analysis, the projection coefficient plots illustrated interesting clusters of the data. The yellow and cyan colored groupings for the left carotid trees roughly hold for the right carotid trees, except that one yellow point changed group. This indicated approximate symmetry, but asymmetry is also visible in the case of that yellow point which changed group.

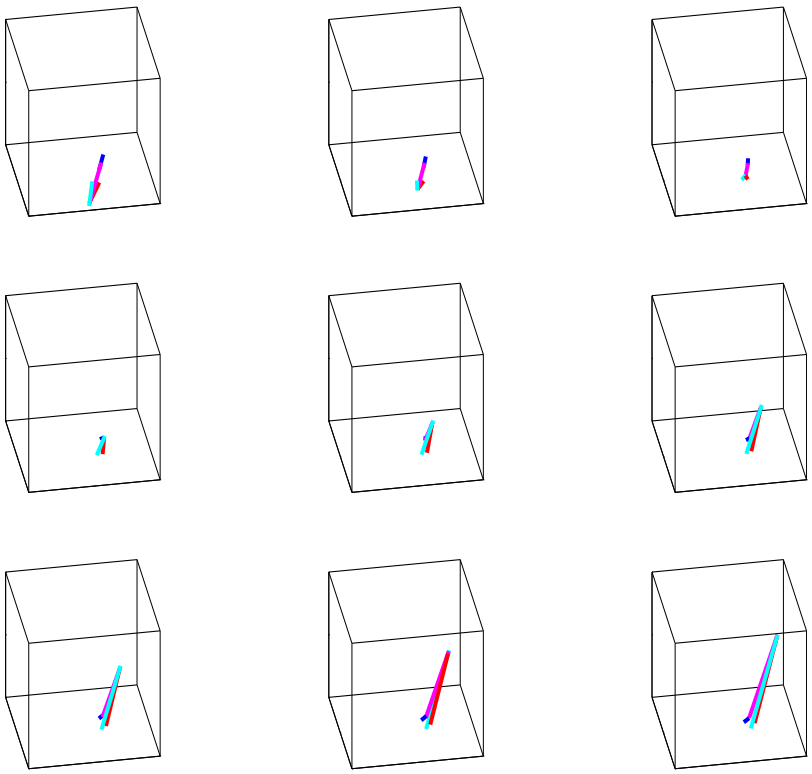

FIGURE 3.41. Principal attribute treeline passing through the average support tree for the population of right carotid trees. The cyan-colored branch is the root.

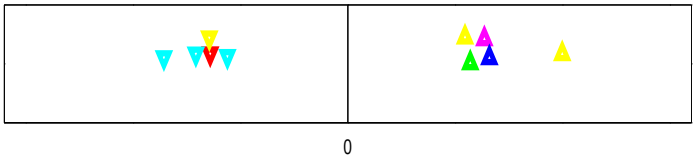

FIGURE 3.42. Projection coefficients of 10 trees on the attribute treeline passing through the average support tree for the population of right carotid trees. Colors are the same as Figure 3.34.

CHAPTER 4

Conclusions and Discussion

In this dissertation, a new method for understanding the structure of populations of tree-structured objects was developed. This method was based on the new metric δ, which measures the difference of topological structure and nodal attributes. For tree-structured objects with this non-linear metric δ, many standard notions of classical statistics, such as population center point, have been developed here.

A "central tree", the median-mean tree, was introduced as a combination of the idea of "sample median" with respect to the topological properties and "sample mean" with respect to the geometric properties (i.e., attribute properties). Quick and easy computation of the topological structure of the median-mean tree was developed through the majority rule. The structure of the median-mean tree (which has the same topological structure as the median tree) was determined by the nodal appearance number, which must be at least $\frac{n}{2}$. Also, the nodal attributes of the median-mean tree can be calculated as the sample average.

Furthermore, a new tool of variation analysis, a tree version PCA was developed. This was based on the notion of "treelines" which played the role of one-dimensional representation in tree space. A key theoretical contribution was the tree version of the Pythagorean Theorem that provided the foundation of the ANOVA type of variation decomposition. This tree version PCA analysis provided a useful tool for finding the characteristics of the topological structure and of the nodal attributes by pointing out the dominant directions of change in structure and of change in nodal attributes.

For both real data examples in Sections 1.2 and 3.9, only the trees up to three levels are considered. For more complex situations, the median-mean tree is still

computationally tractable by using the majority rule. But, the variation analysis is computationally intensive (see Section 3.7 for the current algorithm), and a more efficient algorithm is needed, in order to address problems of the scale needed in medical image analysis.

The example with blood vessel data described in Section 1.2 showed how the method can find interesting clusters in the data. The projections onto the dominant treeline provided a clear view of clustering. According to the projections on the two different types of treeline, the groupings into clusters may vary.

In this example, the 11 trees formed two different "bimodal distributions" from the two viewpoints of the structure and of the attributes. The bimodality of the projections on the principal attribute direction was caused by the use of two different arbitrary directions of blood flow, with no biological interest. Moreover, in this example, the median-mean tree was not close to any of those 11 trees. Therefore, this notion of "center" is not representative of any individual, as is common with bimodal populations.

In most cases, the two one-dimensional representations of the principal structure representation and the principal attribute direction, are not enough. A simple approach is to consider more one-dimensional attribute directions (other than the principal attribute direction). A more complicated approach is to study analogs of higher dimensional subspaces.

The simple approach works in the attribute directions by methods similar to additional principal directions in regular PCA. For these additional directions, it is conjectured that the uniqueness of the projection (Proposition 3.5.1) and the tree version of the Pythagorean Theorem (Part I, Theorem 3.5.3) will still hold.

For the more complicated structure representation, a more general 2-dimensional structure representation can be generated by adding or deleting 1 or 2 terminal nodes

starting from the median-mean trees. For the higher dimensional structure representations, the projection may not be unique; but it is conjectured the tree version of the Pythagorean Theorem (Part II, Theorem 3.5.4) still holds.

These will be considered in future research.

The example in Section 3.9 showed another application of the new methodology. The tree version PCA gave an analysis of the characteristics of the left and right carotid trees. The principal attribute directions for both left and right carotid trees consist of the orientation changing of the root node and the enlargement of the size of the vessel trees. Also, the projection coefficient plots showed that the yellow and cyan colored groupings for the left carotid trees roughly hold for the right carotid trees, except that one yellow point changed group. This indicated the approximate symmetry between the left and right carotid trees, except for the yellow case.

In the two examples described in Sections 1.2 and 3.9, the attributes used for the statistical analysis include the coordinates of the starting point and ending point for the root node, and the coordinates of the ending point plus the proportion parameter for the non-root nodes. A linear approximation of each blood vessel was considered. Many other attributes, such as radii and arc length of the blood vessels and the additional location information along each blood vessel, have been ignored. In future research, those attributes could be included.

Next, since the median-mean trees were calculated and the variations were quantified for the left and right carotid samples, statistical inference about the left and right carotid tree populations is an interesting problem. For example, a maximum likelihood estimator and confidence "region" for the population (not the empirical population) central tree could be developed. Hypotheses testing is also of interest. This motivates the definition of a probability measure which provides the foundation for statistical inference in tree space.

In this dissertation, each node was assumed to contain some attributes, and the weight assigned to that node is positive. In some conceivable applications, some nodes may not contain any attributes. A similar mathematical approach can be taken by putting positive weights onto only the set of nodes with nodal attributes.

In this research, only binary tree-structured objects are considered. This is straightforward to generalize to general tree-structured objects. Great care in this generalization will be needed in defining the labels of the nodes, such as the level-order index. A much more challenging generalization will be to graph-structured objects. Here a new metric is needed to take the possibilities of isolated nodes and loops into account. These will be considered in future research.

REFERENCES

[1] Banks, D. and Constantine, G. M. (1998), "Metric Models for Random Graphs", Journal of Classification 15:199-223.

[2] Breiman, L., Friedman, J. H., Olshen, J. A., Stone, C. J. (1984), "Classification and Regression Trees". Belmont, CA: Wadsworth.

[3] Breiman, L., (1996), "Bagging Predictors", Machine Learning, vol 24, Number 2, 123-140.

[4] Bullitt, E. and Aylward, S. (2002), "Volume rendering of segmented image objects". IEEE-TMI 21:998-1002.

[5] Everitt, B. S., Landau, S., Leese, M. (2001), "Cluster Analysis" (4th edition). Oxford University Press, New York.

[6] Holmes, S. (1999). "Phylogenies: An Overview", IMA series, vol 112, on Statistics and Genetics, (ed. Halloran and Geisser), 81-119 Springer Verlag, New York.

[7] Li, S., Pearl, D. K., Doss, H. (2000) "Phylogenetic Tree Constructure Using Markov Chain Monte Carlo", Journal of the American Statistical Association 95:493-508.

[8] Margush, T. (1982), "Distances Between Trees", Discrete Applied Mathematics 4:281-290.

[9] Mott, J. L., Kandel, A., Baker, T. P., (1986), "Discrete Mathematics for Computer Scientists and Mathematicians" (2nd edition), Prentice-Hall, New Jersey.

[10] Pizer, S. M., Thall, A., Chen, D. (1999), "M-Reps: A New Object Representation for Graphics", Submitted to ACM TOG.

[11] Ramsay, J. O. and Silverman, B. W. (1997), "Functional Data Analysis". Springer Verlag, New York.

[12] Ramsay, J. O. and Silverman, B. W. (2002), "Applied Functional Data Analysis". Springer Verlag, New York.

[13] Shannon, W. and Banks, D. (1999), "Combining Classification Trees Using MLE", Statistics in Medicine 18:727-740.

[14] Tschirren, J., Palágyi, K., Reinhardt, J. M., Hoffman, E. A., Sonka, M. (2002). "Segmentation, Skeletonization, and Branchpoint Matching — A Fully Automated Quantitative Evaluation of Human Intrathoracic Airway Trees", Proc. 5th Int. Conf. Medical Image Computing and Computer-Assisted Intervention, MICCAI, Part II, 12-19, Tokyo, Japan.

[15] Yushkevich, P., Pizer, S. M., Joshi, S., Marron, J. S. (2001), "Intuitive, Localized Analysis of Shape Variablity", in Information Processing in Medical Imaging (IPMI), 402-408.