Functional Data Analysis of Populations of Tree-structured Objects

Haonan Wang
Department of Statistics
University of North Carolina at Chapel Hill

October 28, 2000

1 Introduction

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

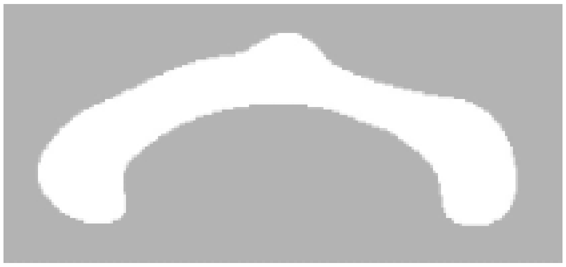

Figure 1.1: Example of shapes of interest

For example, consider the shape in Figure 1.1. It shows an example of one member of a population of shapes of interest. There are bendings at the two ends and one bump in the middle of the object.

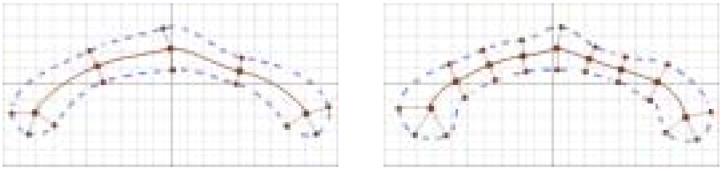

Figure 1.2: Coarse and fine scale M-reps

A class of convenient and powerful shape representations is M-reps (see Pizer, S.M., 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 the shape into parts coarsely or finely. Figure 1.2 shows both a coarse scale M-rep and a fine scale M-rep of the shape shown in Figure 1.1.

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 imagining 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.

For example, for a population of hands, the palm and each finger can be represented by a figure, which is a collection of M-rep parameters. Each hand is a multi-figural object (see Figure 1.3).

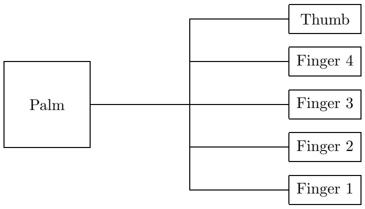

Figure 1.3: An example of multi-figural object — hand

If every member in the population has five fingers, we can simply put all of the features of one hand into a feature vector. Thus, the shape space is equivalent the Euclidean space. And, we can do statistical analysis, such as finding center point and quantifying the variation, on the Euclidean space spanned by those feature vectors.

[1] visit the MIDAG web site at *http://www.cs.unc.edu/Research/Image/MIDAG*

It is not straight forward to analyze population structures when some hands do not have five fingers. In this case, we can not get feature vectors of the same length. We use tree structure to represent members of such a population.

To do a statistical analysis on a population of tree-structured objects, we want to find the "center point" which is the closest objects to all the others. Furthermore, we want to quantify the variability of the population based on the center point. A careful axiomatic structure is developed here because it is a priori unclear which ideas from linear vector spaces apply in this nonlinear spaces.

2 Basic definitions

In this research, we will deal with a population of multi-dimensional objects. The single observation in this population is called a "tree". What is a "tree"?

Definition 2.1. 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.

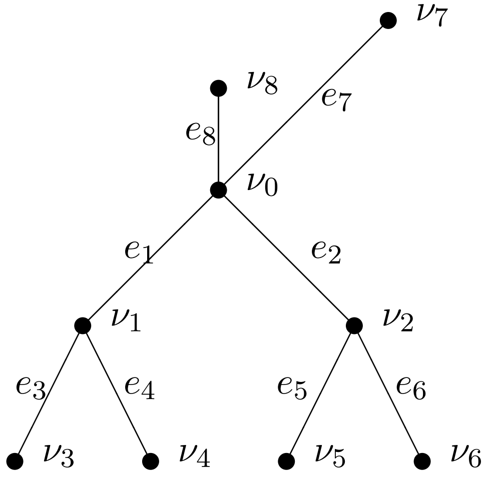

Figure 2.1: tree

Definition 2.2. 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. We call the maximum level of the nodes **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. The tree t in Figure 2.1 has 9 nodes and 8 edges. $V = \{\nu_0, \nu_1, \nu_2, \dots, \nu_8\}$, and $E = \{e_1, e_2, \dots, e_8\}$. 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 2.3. A **binary tree** is a tree $t = (V, E)$, together with an edge labeling 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, we can define the right child. A tree $t_1 = (V_1, E_1)$ is called a **subtree** of t if $V_1 \subseteq V$, $E_1 \subseteq E$ and the root of tree t is in the set V_1.

For simplicity, we will deal with the **binary tree** first.

Definition 2.4. 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 v is the right child of the node ν, let $ind(\omega) = 2 \times ind(\nu) + 1$.

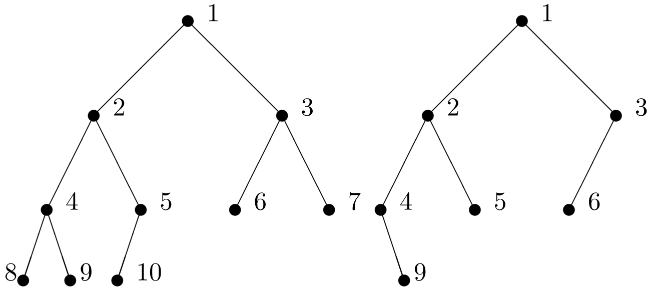

Figure 2.2: Examples of binary trees. The numbers are level-order indices.

Definition 2.5. A **complete binary tree** is a binary tree for which the level-order indices of the nodes form a complete interval $1, 2, \dots, n$ of integers. Otherwise, it is called an **incomplete tree**.

Example 2.2. 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 2.6. 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 2.3. For any binary tree t, $I_0 = \{1\}$ and $I_2 = \{4, 5, 6, 7\}$.

Remark 2.1. For a binary tree t, we denote the set of level-order indices of the nodes by $Ind(t)$.

Remark 2.2. 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 2.7. 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 $Ind(t) = Ind(t_1) \cap Ind(t_2)$). We denote it by $t = t_1 \cup t_2$ (or $t = t_1 \cap t_2$).

Remark 2.3. The definitions of union and intersection of binary trees can be generalized to any tree population where we can define a "level-order index".

Remark 2.4. All the definitions of the operations on the binary trees are based on the level-order indices of the nodes.

3 Metric on binary trees without nodal information

In the previous section, we introduced some basic definitions. But we still can not do statistical analysis on the binary tree population. A first question for statistical analysis is, what is the "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. So, how can we measure the distance between two trees?

Suppose we have two trees t_1 and t_2 shown in Figure 3.1. We can obtain t_2 from t_1 by adding two nodes and deleting one from t_1; that is, the smallest number of addition and deletion of nodes from one tree to the other is 3. So, can we define a tree metric based on the total number of such deletions and additions?

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

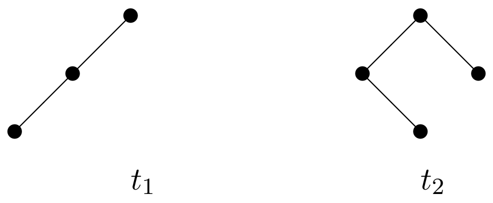

Figure 3.1: Binary trees t_1 and t_2.

Definition 3.1. 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 $n > \max(L_s, L_t)$, $d_n = 0$.

Theorem 3.1. *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, we have

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

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

   Therefore,

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

□

Example 3.1. For the two binary trees t_1 and t_2 shown in figure 3.1, $d_0(t_1, t_2) = 0$, $d_1(t_1, t_2) = 1$ and $d_2(t_1, t_2) = 2$.

Remark 3.1. From the example above, $d_0(t_1, t_2) = 0$ 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 we can find two different binary trees s and t such that $d_i(s, t) = 0$.

For any two binary trees s and t without nodal information, denote

$$d_I(s,t) = \sum_{i=0}^{\infty} d_i(s,t), \tag{3.1}$$

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.

Theorem 3.2. *$d_I(s,t) = \sum_0^\infty d_i(s,t)$ is a metric on the binary tree space without nodal information.*

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

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 3.1, 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 3.1, we have $d_i(s,t) \leq d_i(s,w) + d_i(w,t)$ 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 3.2. Since d_I is always an integer, we called it the integer tree metric.

Remark 3.3. There is an intuitive representation of the integer part metric. It is the smallest total number of added and deleted nodes required to move from one binary tree to the other.

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,

$$d_I(s,t) = \sum_{k=1}^{\infty} 1\{k \in Ind(s) \triangle Ind(t)\}. \tag{3.2}$$

Example 3.2. Let t_1 and t_2 be the binary trees shown in Figure 3.1. $d_0 = 0$, $d_1 = 1$ 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 (3.2), we have $d_I(s,t) = 3$.

4 Finding the median tree on the binary tree space without nodal information

Now we have an integer metric d_I on the binary tree space without nodal information. We can try to answer 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 we are interested by $T = \{t_1, t_2, \ldots, t_n\}$.

Definition 4.1. A tree is a minimizer tree according to the metric λ if it minimizes $\sum_{i=1}^{n} \lambda(t, t_i)$ over all binary trees $t \in \mathcal{T}$.

Definition 4.2. A tree is called a **full binary tree** if it contains all the nodes the binary tree sample T.

Definition 4.3. The full tree with the minimum number of nodes is called **support binary tree**.

By the definition of the metric d_I and minimizer, we have the following property.

Proposition 4.1. *A minimizer tree according to d_I can not have a node which does not appear in the sample. That is, a minimizer tree is contained in the support binary tree.*

Theorem 4.2. *If a tree s is a minimizer 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 according to the tree integer measure 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. We have that ν must be a terminal node of s.

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

$$\sum_{i=1}^{n} d_I(s', t_i) = \sum_{i=1}^{n} d_I(s, t_i) + n_\nu - (n - n_\nu), \tag{4.1}$$

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

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

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

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 still a minimizer. Therefore, we have the minimizer may contain any subset of the nodes that appear exactly $\frac{n}{2}$ times.

Finally, we will prove 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, we suppose ω is a children of some node in the binary tree s. Otherwise, we can choose one of its ancestor nodes.

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

$$\sum_{i=1}^{n} d_I(s, t_i) = \sum_{i=1}^{n} d_I(s'', t_i) + n_\omega - (n - n_\omega), \tag{4.2}$$

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

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

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

Corollary 4.3. *If n is an odd number, then there is a unique minimizer (according to d_I), which consists of all the nodes with appearance more than $\frac{n}{2}$ times.*

Remark 4.1. The theorem 4.2 is also called the **majority rule** (See David Banks, 1998, page 204).

Remark 4.2. Formulating this concept in statistical terms, we call the minimizer the **median** of the binary tree sample T.

Remark 4.3. If n is an even number, then the median binary tree may be not unique because some nodes may have appearance number equal to $\frac{n}{2}$.

Definition 4.4. The median binary tree (according to the binary tree metric d_I) with the smallest number of nodes is called **minimal median binary tree**.

Theorem 4.4. *The minimal median binary tree (according to the integer binary 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, we delete those nodes with $\frac{n}{2}$ appearance time to obtain the unique minimal median binary tree. □

Since the integer tree metric d_I only counts the total number of nodes in the symmetric set of their level-order index sets. We have the following theorem that allows easy calculations.

Theorem 4.5. *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 order index set $I \subset \{1, 2, \ldots, k\}$ and the corresponding numbers of appearance are $n_i, i = 1, 2, \ldots, k$. Then,

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

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 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 4.1. 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 4.1. 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.

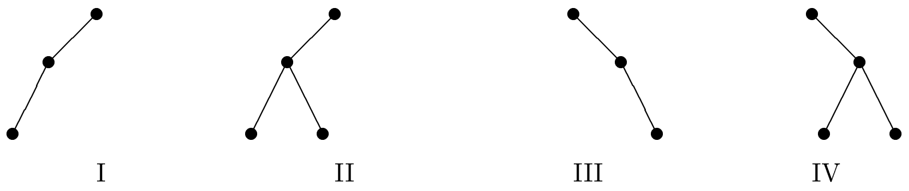

Figure 4.1: An example of a binary tree sample

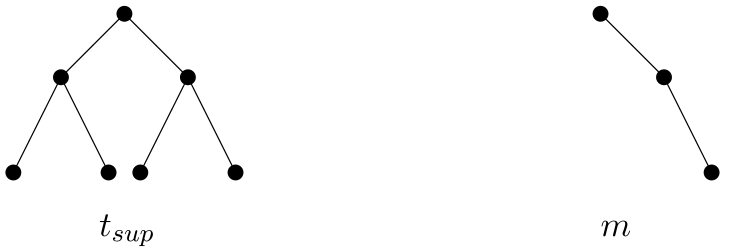

Figure 4.2: Support tree t_{sup} and median tree m of binary tree sample T

The support binary tree of the sample T is shown in the left panel in Figure 4.2. 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 4.2.

Then by Theorem 4.5, 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 the tree space without nodal information, we can treat the sum of distances to the median tree as the total variation of the sample.

5 Line and Projection in the binary tree space without nodal information

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 principle component analysis (PCA) in Euclidean space may not be applicable in the nonlinear binary tree space. So, the question is " Can we find an analogy way to construct a manifold in binary tree space which consists of some binary trees plays the role of a 'line', one-dimensional subspace in Euclidean space? "

First, let's consider the binary tree space without nodal information, $\mathcal{T}$. In this case, we will only consider the integer metric d_I.

Definition 5.1. 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 5.1. From another point of view, the tree u_i is obtained by adding a node ν_i on the tree u_{i-1}.

Example 5.1. In Figure 5.1, 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 tree line l passing through u_0, u_1 and u_2.

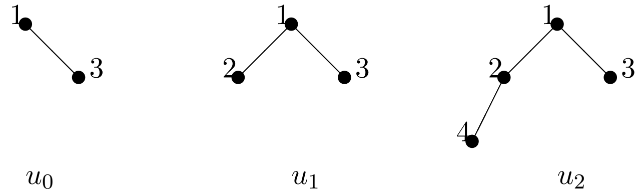

Figure 5.1: A tree sequence $l_0 = \{u_0, u_1, u_2, \ldots\}$ illustrating the idea of a treeline

Example 5.2. In Figure 5.2, 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 tree line passing through u_0, u_1, u_2 and u_3.

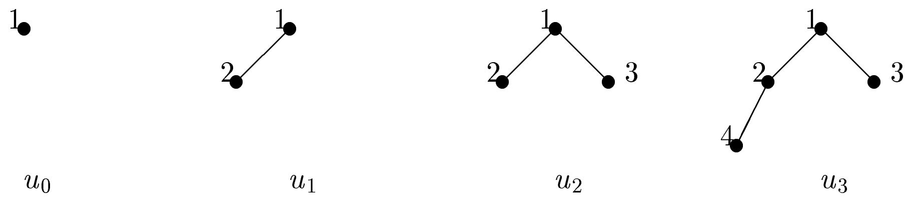

Figure 5.2: A tree sequence $l_1 = \{u_0, u_1, u_2, u_3 \ldots\}$ that is not be a treeline

Definition 5.2. 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., $u \in l$.

Definition 5.3. Suppose v is a tree and l is a treeline as in Definition 5.1. 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 5.1. *The projection of a tree on a treeline exists and is unique.*

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

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

Consider the two elements u_p and u_{p+1} in the treeline l. By definition of the treeline, u_p can be obtained by deleting a node ν_{p+1} from the tree u_{p+1}. Therefore, $\nu_{p+1} \notin Ind(t)$. Otherwise,

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

which is a contradiction with the definition of p. Thus,

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

Repeatly, for $i \geq p$, we have

$$d_I(u_{i+1}, t) = d_I(u_i, t) + 1.$$

Similarly, we have, for $i \leq p$

$$d_I(u_{i-1}, t) = d_I(u_i, t) + 1.$$

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

$$d_I(u_i, t) > d_I(u_p, t).$$

That is, the projection is unique. □

From Proposition 5.1, it makes easy to define projection function

$$w = P_l(v),$$

where l, w and v are given above.

6 Principal Component Analysis on Binary Tree Space without Nodal Information

In classical statistics, the principal component analysis (PCA) is a useful tool to capture the feature 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, we can find several other orthogonal directions which often highlight additional interesting aspects of the data. Now consider the similar problem in the binary tree space, can we develop a method to analyze the variation of the data set?

As we know from the previous section, the treeline plays the role of “line”, i.e. one-dimensional representation, in binary tree space. Recall that, for any tree sample T, the median binary tree m plays the role of “center point”. So, can we find a treeline l, one-dimensional representation in binary tree space, passing through the median tree m such that it maximizes the sum

$$\sum_{i=1}^{n} d_I(m, P_l(t_i))? \tag{6.1}$$

Recall that, if the population size n is odd, then the median tree is unique which is also a minimal median tree. 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.

Pythagorean theorem is a fundamental theorem for the decomposition of the variation in the PCA in Euclidean space. Now, we will develop an analog theorem, which is called tree version Pythagorean theorem in the binary tree space without nodal information.

Theorem 6.1. *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 m. Then, $\forall t \in T$,*

$$d_I(m, P_l(t)) + d_I(P_l(t), t) = d_I(m, t). \tag{6.2}$$

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

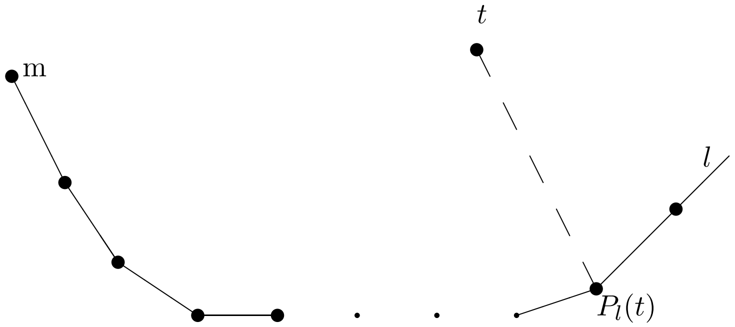

Figure 6.1: Projection of the tree t on the tree line l passing through m

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

1. $m \subset u_k$

   Note that,

   $$u_k \cap \overline{m} \cap \overline{t} = \emptyset. \tag{6.3}$$

   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 m. Therefore, considering the binary tree $u_{k-1} = u_k \backslash \{\nu\}$,

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

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

Furthermore,

$$\begin{aligned}
& d_I(P_l(t), m) + d_I(P_l(t), t) \\
= \; & d_I(u_k, m) + d_I(u_k, t) \\
= \; & \sum_j 1\{j \in m \triangle u_k\} + \sum_j 1\{j \in t \triangle u_k\} \\
= \; & \sum_j 1\{j \in u_k \backslash m\} + \sum_j 1\{j \in t \triangle u_k\} \\
= \; & \sum_j 1\{j \in m \triangle t\} + 2 \sum_j 1\{j \in u_k \cap \overline{m} \cap \overline{t}\} \\
= \; & d_I(t, m)
\end{aligned}$$

because u_k is a projection of t, hence $u_k \cap \overline{m} \cap \overline{t} = \emptyset$.

2. $u_k \subset m$,

   Similarly as Equation 6.3, we have

$$m \cap \overline{u_k} \cap t = \emptyset. \tag{6.4}$$

Also,

$$\begin{aligned}
& d_I(P_l(t), m) + d_I(P_l(t), t) \\
= \; & d_I(u_k, m) + d_I(u_k, t) \\
= \; & \sum_j 1\{j \in m \triangle u_k\} + \sum_j 1\{j \in t \triangle u_k\} \\
= \; & \sum_j 1\{j \in m \backslash u_k\} + \sum_j 1\{j \in t_i \triangle u_k\} \\
= \; & \sum_j 1\{j \in m \triangle t\} + 2 \sum_j 1\{j \in m \cap \overline{u_k} \cap t\} \\
= \; & d_I(t, m)
\end{aligned}$$

because u_k is a projection of t, hence $m \cap \overline{u_k} \cap t = \emptyset$.

□

Remark 6.1. In Euclidean space, Pythagorean theorem claims that in a right triangle with legs a, b and hypotenuse c, $c^2 = a^2 + b^2$. In the tree space, the hypotenuse ($d_I(t, m)$) is the sum of two legs.

Corollary 6.2. *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, $\forall t \in T$,*

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

Theorem 6.3. *Let $T = \{t_1, t_2, \ldots, t_n\}$ be a sample. Maximizing the sum $\sum_{i=1}^{n} d_I(m, P_l(t_i))$ is equivalent to minimizing the sum $\sum_{i=1}^{n} d_I(t_i, P_l(t_i))$ where l runs over all treelines passing median tree m.*

Proof. From the tree version Pythagorean theorem 6.1, we have, for $i = 1, 2, \ldots, n$,

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

Therefore,

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

□

Definition 6.1. The treeline l_1 above is called **principal one-dimensional representation**, denoted by π_1.

Remark 6.2. The principal one-dimensional representation, i.e. π_1, might not be unique.

Definition 6.2. 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 6.3. This equivalence of two treelines is relative; that is, for different tree samples, their equivalence may be different.

Remark 6.4. Let k be the maximum level of a tree sample T. If all the components with level no more than k are the same for two tree lines l_1 and l_2, then l_1 and l_2 are equivalent. Therefore, for simplicity, we only represent the tree line by the components with level no more than k.

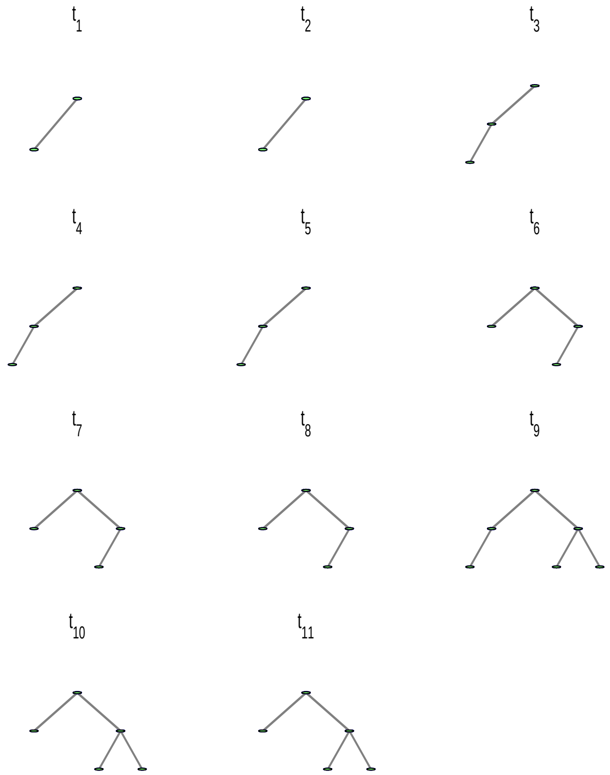

Figure 7.1: A sample of trees without nodal information.

7 Example

We have developed some basic concepts and ideas on trees (without nodal information). Now, let's look at a toy example.

T is a sample of trees with $n = 11$ members, $t_1, t_2, \dots t_{11}$. Based on the integer tree metric d_I, the support tree (t_{sup}) and median tree (m) are shown in Figure 7.1. Note that, $n = 11$ is an odd number. Therefore, the median tree is unique.

In the left panel, 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 was shown on the right panel in figure 7.2.

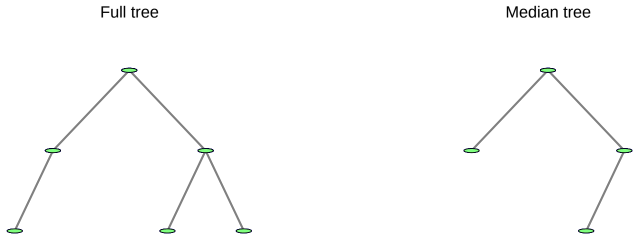

Figure 7.2: Support tree and median tree

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

There are three different equivalent treeline classes passing through the median tree m. We have three representative treelines shown in Figure 7.3, l_1, l_2, and l_3.

The projections of tree sample T on representative treeline l_1 are shown in Figure 7.6. The total distance of the median tree m and the projection of tree t_i on tree line l_1 is

$$\sum_{i=1}^{11} d_I(m, P_1(t_i)) = 4$$

Similarly, we can obtain the projection of the tree sample T on treelines l_2 and l_3.

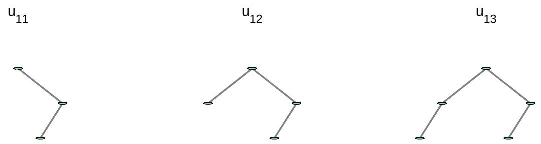

Figure 7.3: Representative treeline l_1

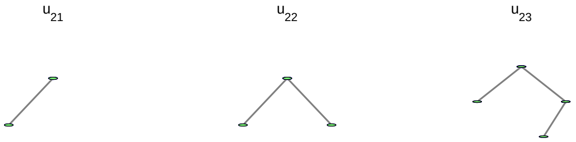

Figure 7.4: Representative treeline l_2

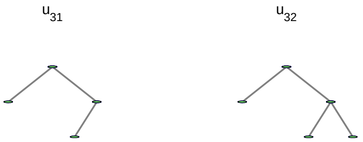

Figure 7.5: Representative treeline l_3

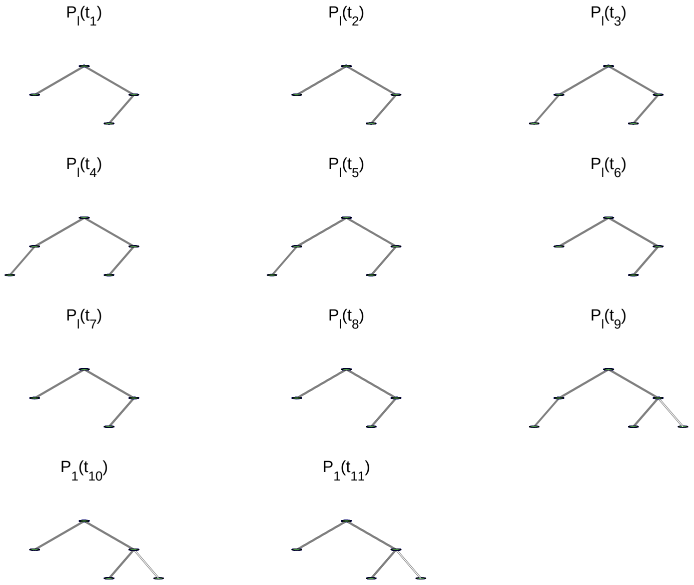

Figure 7.6: Projection of the tree sample on the treeline l_1

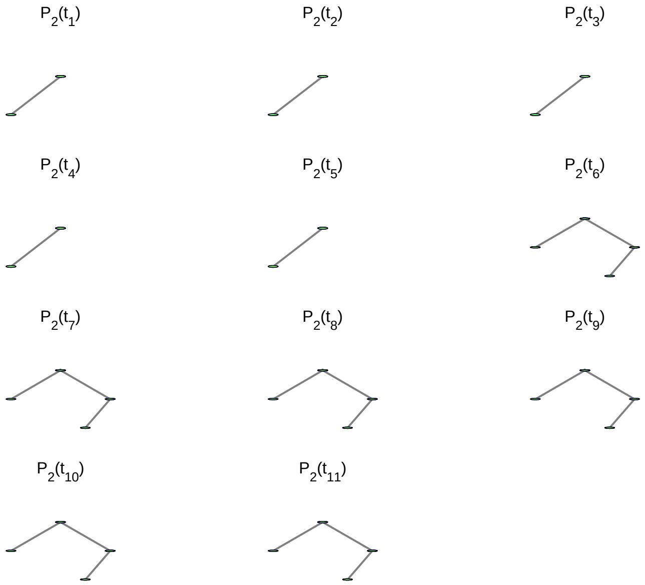

Figure 7.7: Projection of the tree sample on the treeline l_2

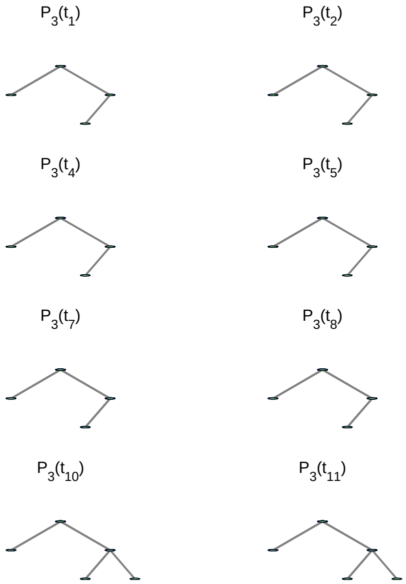

Figure 7.8: Projection of the tree sample on the treeline l_3

$$\sum_{i=1}^{11} d_I(m, P_2(t_i)) = 10$$

$$\sum_{i=1}^{11} d_I(m, P_3(t_i)) = 3$$

Therefore, tree line l_2 is the one-dimensional representation of the tree sample T.

8 New metric δ on tree space with nodal information

The integer tree metric d_I captures some structure of the tree population. Sometimes, the nodes of the trees contain some useful information, which should also be used in the statistical analysis.

Now, we will define a metric on the trees with nodal information which extends the integer tree metric. Denote the information contained in the node with level-order index k on the tree t by $(x_{tk}, y_{tk}, \ldots)$. For simplicity, we explicitly treat the case (x_{tk}, y_{tk}).

Generally, the values of the nodal information, x_{tk} and y_{tk}, have no restriction and can be any real value. But, after some appropriate transformation, the nodal information can be assumed to be bounded. For example, for a mapping f,

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

$f(x) \in [0, \frac{\sqrt{2}}{2}]$. From now on, we assume that $x_{tk}, y_{tk} \in [0, \frac{\sqrt{2}}{2}]$. We take $\frac{\sqrt{2}}{2}$ as the bound because the Euclidean distance between two-dimensional vectors, whose entries satisfy this bound, is at most 1.

For any trees s and t, define the new metric (proof given in Theorem 8.3)

$$\begin{aligned}\delta(s,t) = d_I(s,t) + \Bigg[&\sum_{k=0}^{\infty} \alpha_k((x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2)1\{k \in Ind(s) \cap Ind(t)\} \\ &+ \sum_{k=0}^{\infty} \alpha_k(x_{sk}^2 + y_{sk}^2)1\{k \in Ind(s) \backslash Ind(t)\} \\ &+ \sum_{k=0}^{\infty} \alpha_k(x_{tk}^2 + y_{tk}^2)1\{k \in Ind(t) \backslash Ind(s)\}\Bigg]^{\frac{1}{2}}\end{aligned} \tag{8.1}$$

where $\{\alpha_k\}$ is a positive weight series with $\sum_k \alpha_k = 1$.

In equation (8.1), the second term in the summation is at most 1 (proof given in proposition 8.1). We denote the second term as f_δ where "f" means fractional part of the metric. Recall that, the first term in the summation is denoted as d_I where "I" means integer part of the metric (see section 3). Therefore, we can rewrite δ as

$$\delta = d_I + f_\delta \tag{8.2}$$

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 information 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 missing nodal information with $(0,0)$. Thus, we can rewrite f_δ as

$$\begin{aligned} f_\delta(s,t) = [&\sum_{k=0}^{\infty} \alpha_k((x_{sk}-x_{tk})^2 + (y_{sk}-y_{tk})^2)1\{k \in Ind(s) \cap Ind(t)\} \\ &+ \sum_{k=0}^{\infty} \alpha_k((x_{sk}-0)^2 + (y_{sk}-0)^2)1\{k \in Ind(s) \backslash Ind(t)\} \\ &+ \sum_{k=0}^{\infty} \alpha_k((0-x_{tk})^2 + (0-y_{tk})^2)1\{k \in Ind(t) \backslash Ind(s)\}]^{\frac{1}{2}} \end{aligned} \tag{8.3}$$

This also allows the nodal information 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 is to 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, we will develop all the theorems for general weight sequences. But, we will use the power weight sequence, where the weight is $\{2^{-(2i+1)}\}$ for the node on the i^{th} level, $i = 0, 1, 2, \dots$ in $\mathcal{T}$ in the examples.

Insight into the metric δ comes from Example 8.1.

Example 8.1. t_1 and t_2 are two trees with nodal information listed below.

level-order index	t_1	t_2
1	(0.5,0.5)	(0.2,0.5)
2	(0,0.1)	(0.7,0.1)

The following figure shows the graphical representation [2]of two trees t_1 and t_2.

[2] For every node with nodal information (x,y), we take x as the length and y as the width of the nodal box in the graphical representation.

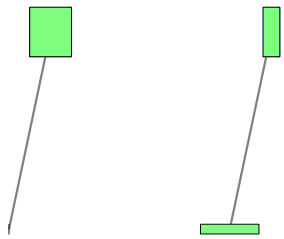

Figure 8.1: Graphical representation of trees t_1 and t_2 in Example 8.1

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

$$\begin{aligned}\delta(t_1, t_2) &= f_\delta(t_1, t_2) \\ &= \sqrt{\underbrace{\frac{1}{2}((0.5-0.2)^2 + (0.5-0.5)^2)}_{k=1} + \underbrace{\frac{1}{2^3}((0-0.7)^2 + (0.1-0.1)^2)}_{k=2}} \\ &= 0.3260\end{aligned}$$

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 information) $[0.5, 0.5, 0, 0.1]'$ and $[0.2, 0.5, 0.7, 0.1]'$.

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 $\vec{v}_2$

$$\begin{aligned}\vec{v}_1 &= [\frac{0.5}{\sqrt{2}}, \frac{0.5}{\sqrt{2}}, \frac{0}{\sqrt{2^3}}, \frac{0.1}{\sqrt{2^3}}]'; \\ \vec{v}_2 &= [\frac{0.2}{\sqrt{2}}, \frac{0.5}{\sqrt{2}}, \frac{0.7}{\sqrt{2^3}}, \frac{0.1}{\sqrt{2^3}}]'.\end{aligned}$$

Proposition 8.1. *For any two trees with nodal information, the fractional part is at most 1, i.e.*

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

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

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

because $x_{sk}, x_{tk}, y_{sk}, y_{tk} \in [0, \frac{\sqrt{2}}{2}]$.

Similarly, we have

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

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,

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

□

When we use a general weight sequence, the fractional part can be very small. In some problems, the level of the trees in the population is finite. We can assign equal weight on those nodes.

A convenient finite population of trees are all subtrees of a particular tree. In particular,

Definition 8.1. Let w be any tree in $\mathcal{T}$. A tree t is said to be a member of $\mathcal{T}_w$ if

$$Ind(t) \subset Ind(w) \tag{8.4}$$

Note that, the previous definition 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".

Proposition 8.2. *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, $\forall t \in \mathcal{T}_w$, we have

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

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

$$\alpha_k = \begin{cases} \frac{1}{N(w)}, & \text{if } k \in Ind(w) \\ \\ 0, & \text{if } k \notin Ind(w). \end{cases} \tag{8.5}$$

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

$$\begin{aligned}\rho(s,t) = d_I(s,t) + \Bigg[&\frac{1}{N(w)}\sum_{k=1}^{N(w)}((x_{sk}-x_{tk})^2 + (y_{sk}-y_{tk})^2)1\{k \in Ind(s) \cap Ind(t)\} \\ &+ \frac{1}{N(w)}\sum_{k=1}^{N(w)}((x_{sk})^2 + (y_{sk})^2)1\{k \in Ind(s) \backslash Ind(t)\} \\ &+ \frac{1}{N(w)}\sum_{k=1}^{N(w)}((x_{tk})^2 + (y_{tk})^2)1\{k \in Ind(t) \backslash Ind(s)\}\Bigg]^{\frac{1}{2}} \\ = d_I(s,t) + f_\rho(s,t)&\end{aligned} \tag{8.6}$$

Example 8.2. Let t_1 and t_2 be the trees as given in Example 8.1. They are members in the tree subspace $\mathcal{T}_w$, where $Ind(w) = \{1, 2, 3\}$.

Note that,

$$\begin{aligned}\rho(t_1,t_2) &= f_\rho(t_1,t_2) \\ &= \sqrt{\underbrace{\frac{1}{3}((0.5-0.2)^2 + (0.5-0.5)^2)}_{k=0} + \underbrace{\frac{1}{3}((0-0.7)^2 + (0.1-0.1)^2)}_{k=1}} \\ &= 0.4397\end{aligned}$$

We can see that $\delta(t_1, t_2) < \rho(t_1, t_2)$. The reason is that, the metric δ puts much smaller weights on higher levels. Thus, the nodal information of the nodes on higher level has small impact on the distance. Because this property is unappealing, for finite level tree spaces, we will use the metric ρ instead of δ.

Now, we will show that δ is a metric.

Theorem 8.3. *δ is a metric on the tree space with nodal information.*

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

Note that

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

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

Now, we will prove the triangle inequality; that is,

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

Recall that d_I is a metric on the tree space without nodal information and pseudometric on the tree space with nodal information (see Theorem 3.2 in section 3). 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 information vectors. Therefore, the triangle inequality is satisfied.

Thus, in general, the triangle inequality is satisfied. δ is a metric on the tree space with nodal information. □

9 Formulating the nodal information and representing the tree

In this section, we will discuss how to represent the tree of different tree structures and nodal inforamtion.

In Example 8.1, we use a table to represent the trees by listing the level-order indices on the left column followed by the corresponding nodal information.

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, we can represent the tree in the following table.

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

Note that, for a node which does not appear in the tree t, we will record its nodal information as “n/a” in the table above.

As we 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. We use the following rule to formulate the nodal information vector.

Rule of nodal information vector

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

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

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

$$(v_{2k-1}, v_{2k}) = \begin{cases} (x_{tk}, y_{tk}), & \text{if } k \in Ind(t) \\ \\ (0, 0), & \text{if } k \notin Ind(t). \end{cases} \tag{9.1}$$

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, we can simply record the nodal information as a vector of length $2N(w)$.

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 information vectors of the trees t_1 and t_2 respectively.

10 Median-mean tree of the tree sample with nodal information

In section 4, we have already discussed "how to find the median tree for a tree sample $T = \{t_1, t_2, \ldots, t_n\}$ without nodal information? ". The final solution to this problem for the metric d_I is the **majority rule**. Now, we have a new metric δ which also considers nodal information. We will develop a new "center point" of the tree sample with nodal information called the median-mean tree. 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 10.1. A tree is called a **median-mean tree** for a sample T, denoted by m_δ, if it minimizes

$$\sum_{i=1}^{n} d_I(t, t_i) \tag{10.1}$$

over all trees $t \in T$ and has nodal information

$$x_{m_\delta k} = \frac{\sum_{i=1}^{n} x_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{10.2}$$

$$y_{m_\delta k} = \frac{\sum_{i=1}^{n} y_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{10.3}$$

Remark 10.1. The new "center point" m_δ is called **"median-mean"** because its tree structure complies with the majority rule with appearance number at least $\frac{n}{2}$ times and its nodal information can be calculated as a "sample mean".

The median-mean tree defined in Definition 10.1 may or may not be unique, as shown in Examples 10.1 and 10.2 below. A variation which is unique is given in Definition 10.2.

Definition 10.2. The median-mean tree with the smallest number of nodes is called the **minimal median-mean** tree with nodal information (denoted by μ_δ).

The following example shows the lack of the uniqueness of median-mean tree. But the minimal median-mean is unique.

Example 10.1. For a tree sample $T = \{t_1, t_2\}$, the nodal information is listed below.

level-order index	t_1	t_2
1	(0.3,0.4)	(0.5,0.5)
2	(0.1,0.1)	N/A
3	N/A	(0.2,0.2)

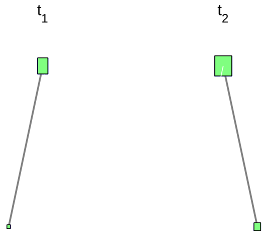

Figure 10.1: Graphical representation of the tree sample of Example 10.1

There are four median-mean trees for this sample. Their nodal information is:

level-order index	information
1	(0.4,0.45)
2	N/A
3	N/A

level-order index	information
1	(0.4,0.45)
2	(0.1,0.1)
3	N/A

level-order index	information
1	(0.4,0.45)
2	N/A
3	(0.2,0.2)

level-order index	information
1	(0.4,0.45)
2	(0.1,0.1)
3	(0.2,0.2)

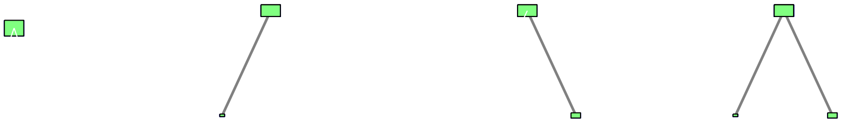

Figure 10.2: Graphical representation of the four median-mean trees in Example 10.1

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 information, μ.

In the following Example 10.2, both the median-mean and the minimal median mean tree are unique.

Example 10.2. For a tree sample $T = \{t_1, t_2, t_3, t_4\}$, the nodal information is listed below.

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)	(0.3,0.5)	(0.2,0.1)	N/A
3	(0.3,0.1)	(0.2,0.3)	N/A	(0.3,0.4)

In this example, the median tree is unique without nodal information. If we consider the nodal information, we will get only one median-mean tree listed below which is also the minimal median-mean tree.

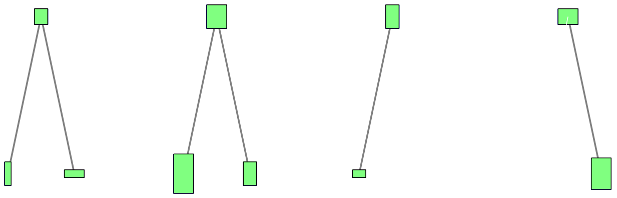

Figure 10.3: Graphical representation of the tree sample in Example 10.2

level-order index	information
1	(0.25,0.25)
2	(0.15,0.2)
3	(0.2,0.2)

The two examples above motivate the following proposition.

Proposition 10.1. *The minimal median-mean tree with nodal information is unique. Also, it has the same tree structure as that of the minimal median tree without nodal information.*

Proof. Since the median-mean tree minimizes equation (10.1), the median-mean tree is also a median without nodal information. By Theorem 4.4, the minimal median is unique without nodal information. Hence, the minimal median-mean tree is also unique. □

We will use the following Example 10.3 to show that the median-mean tree may not minimize the sum

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

Example 10.3. For a tree sample $T = \{t_1, t_2, t_3\}$, the nodal information is listed below.

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)

In this example, there is a unique median-mean tree with nodal information m_δ, listed below.

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

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

Now we consider a subtree s of this median-mean tree.

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

The total distance from s to the other trees is:

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

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 10.2. The median tree without nodal information minimizes the sum $\sum_i d_I(t_i, t)$, overall t, while the median-mean tree with nodal information 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.

Now, for a tree sample T, we have a metric δ. Our 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:

$$V_\delta(s, t) = d_I(s, t) + f_\delta^2(s, t). \tag{10.4}$$

Definition 10.3. Let T be a sample of trees with nodal information. m_δ is a median-mean according to the metric δ. The **variation** of a tree t to the median-mean is defined as $V_\delta(t, m_\delta)$.

Remark 10.3. $V_\delta(\cdot, m_\delta)$ 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? The following proposition will answer this question.

Theorem 10.2. *T is a finite sample of trees with nodal information. The sum of variation to the median-mean of each element in the sample*

$$\sum_{i=1}^{n} V_\delta(t_i, m_\delta) \tag{10.5}$$

is a 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. We will prove that

$$\sum_{i=1}^{n} V_\delta(t_i, s) = \sum_{i=1}^{n} V_\delta(t_i, \mu_\delta). \tag{10.6}$$

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

$$\mu_\delta = s_1 \subset s_2 \ldots \subset s_K = s.$$

where s_{i+1} has one more node (denoted by k_{i+1}) than s_i. It is straight forward that the node k_{i+1} appears exactly $q = \frac{n}{2}$ times in the sample.

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

$$\begin{aligned}
&\sum_{i=1}^{n} V_\delta(t_i, s_{p+1}) \\
&= \sum_{i=1}^{n} V_\delta(t_i, s_p) - \sum_{i=1}^{n} \alpha_{k_{p+1}}(x^2_{t_i k_{p+1}} + y^2_{t_i k_{p+1}}) 1\{k_{p+1} \in Ind(t_i)\} \\
&+ q\alpha_{k_{p+1}}(x^2_{s_{p+1}k_{p+1}} + y^2_{s_{p+1}k_{p+1}}) \\
&+ \sum_{i=1}^{n} \alpha_{k_{p+1}}((x_{t_i k_{p+1}} - x_{s_{p+1}k_{p+1}})^2 + (y_{t_i k_{p+1}} - y_{s_{p+1}k_{p+1}})^2) 1\{k_{p+1} \in Ind(t_i)\}
\end{aligned} \tag{10.7}$$

By the definition of the median-mean tree, we have

$$\begin{aligned}
&\sum_{i=1}^{n}((x_{t_i k_{p+1}} - x_{s_{p+1}k_{p+1}})^2 + (y_{t_i k_{p+1}} - y_{s_{p+1}k_{p+1}})^2) 1\{k_{p+1} \in Ind(t_i)\} \\
&= \sum_{i=1}^{n}(x^2_{t_i k_{p+1}} + y^2_{t_i k_{p+1}}) 1\{k_{p+1} \in Ind(t_i)\} \\
&- q(x^2_{s_{p+1}k_{p+1}} + y^2_{s_{p+1}k_{p+1}})
\end{aligned} \tag{10.8}$$

Combining equations (10.7) and (10.8), we have

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

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

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

□

Remark 10.4. This shows why the median-mean tree is a very natural notion of "center".

11 Treeline and Projection in finite level tree subspace $\mathcal{T}_w$

In section 10, we have defined the center point of a sample of trees with nodal information and the total variation of the sample to its median-mean tree. Also, according theorem 10.2, the total variation 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 representation. Binary tree space is not a linear space, but we still seek useful one-dimensional representations. There are two important types, defined below, and called "treeline".

In this section, we will define the treeline which plays the role of line in Euclidean space. Hence, we will develop an analogy of PCA to find treelines, which explain important features of the sample.

Definition 11.1. Suppose $l = \{u_0, u_1, u_2, \ldots\}$ is a sequence of trees with nodal information 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 a subtree of u_0, denoted as u, such that u can be obtained by deleting some ancestor nodes of ν_1.

In this definition, the nodes in the s-treeline with level-order index k have the same nodal information. Since every element in the s-treeline is a subtree of w, the length of the s-treeline is finite and can not exceed the level of the tree w.

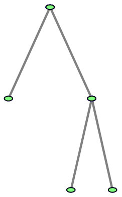

Figure 11.1: Tree structure of an example tree w.

Figure 11.2 shows an example of an s-treeline in $\mathcal{T}_w$, where w has the tree structure shown in figure 1

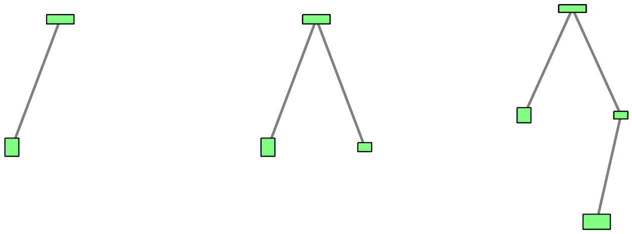

Figure 11.2: An example of an s-treeline in $\mathcal{T}_w$, for w defined in Figure 11.1.

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 information.

Definition 11.2. Suppose $l = \{u_\lambda : \lambda \in \mathbb{R}\}$ is a set of trees with nodal information in the subspace $\mathcal{T}_w$. The set l is called an **information treeline** (i-treeline) passing through a tree u_0 if

1. every tree u_λ has the same tree structure as u_0;
2. the nodal information vector is equal to $\vec{v}_0 + \lambda\vec{v}$, where $\vec{v}_0$ is the information vector of the tree u_0 and $\vec{v}$ is some fixed vector, $\vec{v} \neq \vec{0}$.

Remark 11.1. An i-treeline is determined by the tree u_0 and vector $\vec{v}$. Also, it is a set of trees of the form

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

Note that, there are uncountably many elements in an i-treeline because it has the same cardinality as the real numbers. Figure 11.3 shows some elements with $\lambda = 1.0, 1.75, 3.0, 3.5$ and $\vec{v} = [0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.2, 0.2]'$ in an i-treeline in $\mathcal{T}_w$.

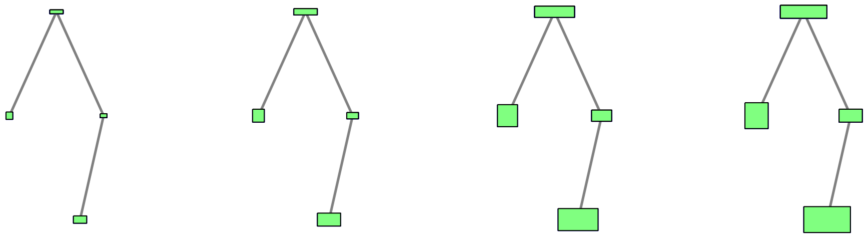

Figure 11.3: An example of an i-treeline in $\mathcal{T}_w$

From now on, both s-treelines and i-treelines are called treelines. An analogy of the first principal component is the treeline which explains most of the data. Before finding this, let's define the projection of a tree on a treeline in the tree subspace $\mathcal{T}_w$.

Definition 11.3. 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 $\rho(t, u)$ 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 information?

Proposition 11.1. *The projection of a tree t on a treeline l is unique.*

Proof. We will prove it for s-treelines and i-treelines separately.

1. l is an s-treeline.

   Suppose $l = \{u_0, u_1, u_2, \ldots\}$. Let p be the index of the smallest d_I closest member of treeline l; i.e.,

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

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

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

which is a contradiction with the definition of p. Thus,

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

Repeatly, for $i \geq p$, we have

$$d_I(u_{i+1}, t) = d_I(u_i, t) + 1.$$

Similarly, we have, for $i \leq p$

$$d_I(u_{i-1}, t) = d_I(u_i, t) + 1.$$

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

$$d_I(u_i, t) > d_I(u_p, t).$$

Now, we will prove that the tree u_p is the unique projection of t on the s-treeline l by considering the fractional part f_ρ as well. Recall that, for $i \neq p$,

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

Also,

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

Furthermore, we will prove that

$$|f_\rho(u_i, t) - f_\rho(u_p, t)| < 1.$$

Note that, since the fraction part of the distance is always no more than 1,

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

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

$$Ind(u_p) \backslash Ind(u_i) = \{k_1, k_2, \dots, k_q\}.$$

Note that, if

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

then, $f_\rho(u_i,t)=0$ and $f_\rho(u_p,t)=1$. Therefore,

$$1=f_\rho^2(u_p,t)-f_\rho^2(u_i,t)\le\sum_{i=1}^{q}\alpha_{k_i}<1.$$

Hence, the inequality is satisfied. Thus,

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

i.e., u_p is the unique projection.

2. l is an i-treeline. Suppose the i-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 information vectors. By the uniqueness of the projection in the Euclidean space, the projection of a tree t on an i-treeline is also unique.

□

Remark 11.2. From the proof above, the projection of a tree t on an s-treeline according to the metric ρ has the same tree structure as that of the projection without nodal information.

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

Definition 11.4. A tree is called an **average support tree** (denoted by t_a) if it is a support tree and its nodal information is

$$x_{t_ak}=\frac{\sum_{i=1}^{n}x_{t_ik}1\{k\in Ind(t_i)\}}{\sum_{i=1}^{n}1\{k\in Ind(t_i)\}}\tag{11.1}$$

$$y_{t_ak}=\frac{\sum_{i=1}^{n}y_{t_ik}1\{k\in Ind(t_i)\}}{\sum_{i=1}^{n}1\{k\in Ind(t_i)\}}.\tag{11.2}$$

Proposition 11.2. *Let T be a sample of trees. The median-mean tree is a subtree of the average support tree.*

In this paper, we focus on s-treelines where every element is a subtree of the average support tree which is an important assumption in the tree version Pythagorean Theorem 11.4.

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 population.

Theorem 11.3. *(Tree version Pythagorean Theorem: Part I) Let l be an i-treeline passing through a tree u in the tree space $\mathcal{T}$. Then, for any $t \in \mathcal{T}$,*

$$V_\rho(t,u) = V_\rho(t, P_l(t)) + V_\rho(P_l(t), u) \tag{11.3}$$

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

$$d_I(P_l(t), u) = 0 \tag{11.4}$$

and

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

We also need to prove

$$f_\rho^2(t,u) = f_\rho^2(t, P_l(t)) + f_\rho^2(P_l(t), u). \tag{11.5}$$

for the i-treeline l.

Note that, for the nodes with level-order index $k \in Ind(t) \backslash Ind(u)$, the contribution of its nodal information to both sides of equation (11.5) is the same. Thus, without loss of generality, we assume $Ind(t) \subset Ind(u)$. Its information vector has the same length as that of the tree u by adding zeroes on $Ind(u) \backslash Ind(t)$.

The metric ρ is the same as the Euclidean distance of two weighted vectors. Thus, it is straight forward that equation (11.5) follows from the ordinary Pythagorean theorem. □

Theorem 11.4. *(Tree version 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 a subtree of the average support tree of T. Then, for any $u \in l$,*

$$\sum_{i=1}^{n} V_\rho(t_i, u) = \sum_{i=1}^{n} V_\rho(t, P_l(t_i)) + \sum_{i=1}^{n} V_\rho(P_l(t_i), u) \tag{11.6}$$

Proof. In Theorem 6.1, we have proved that, for any i,

$$d_I(t_i, u) = d_I(t, P_l(t_i)) + d_I(P_l(t_i), u). \tag{11.7}$$

Therefore,

$$\sum_{i=1}^{n} d_I(t_i, u) = \sum_{i=1}^{n} d_I(t_i, P_l(t_i)) + \sum_{i=1}^{n} d_I(P_l(t_i), u). \tag{11.8}$$

We need to prove that

$$\sum_{i=1}^{n} f_\rho^2(t_i, u) = \sum_{i=1}^{n} f_\rho^2(t_i, P_l(t_i)) + \sum_{i=1}^{n} f_\rho^2(P_l(t_i), u). \tag{11.9}$$

In fact, since l passes through the tree u, we have $P_l(t_i) \subset u$ or $u \subset P_l(t_i)$. Without loss of generality, we assume that

$$P_l(t_1) \subset u, \ldots, P_l(t_K) \subset u, P_l(t_{K+1}) \supset u, \ldots, P_l(t_n) \supset u \tag{11.10}$$

for some $K = 0, 1, \ldots, n$.

If $P_l(t) \subset u$, for $k \in Ind(P_l(t)) \cap Ind(u)$, two trees $P_l(t)$ and u have the same nodal information, therefore,

$$f_\rho^2(P_l(t), u) = \sum_{k=1}^{\infty} \alpha_k((x_{uk})^2 + (y_{uk})^2) 1\{k \in Ind(u) \backslash Ind(P_l(t))\}. \tag{11.11}$$

Furthermore, the tree $P_l(t)$ is the projection of the tree t on the treeline l. Therefore,

$$Ind(t) \cap Ind(u) = Ind(t) \cap Ind(P_l(t)) \tag{11.12}$$

since $P_l(t)$ is a subtree of the tree u,

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

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 we can find a tree u^*, such that $Ind(u^*) \supset Ind(u) \cup \{\nu\}$, which is a contradiction with the assumption that the tree $P_l(t)$ is the projection of the tree t. Therefore, we have Equation 11.12.

Recall that, the squared fractional part distance between the two trees t and u is,

$$\begin{aligned} f_\rho^2(t, u) = & \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2) 1\{k \in Ind(t) \cap Ind(u)\} \\ & + \sum_{k=1}^{\infty} \alpha_k((x_{tk} - 0)^2 + (y_{tk} - 0)^2) 1\{k \in Ind(t) \backslash Ind(u)\} \\ & + \sum_{k=1}^{\infty} \alpha_k((0 - x_{uk})^2 + (0 - y_{uk})^2) 1\{k \in Ind(u) \backslash Ind(t)\} \end{aligned} \tag{11.13}$$

Thus, by the definition of f_ρ^2, and by Equation (11.12),

$$\begin{aligned} f_\rho^2(t, P_l(t)) &= \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{P_l(t)k})^2 + (y_{tk} - y_{P_l(t)k})^2)1\{k \in Ind(t) \cap Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((x_{tk} - 0)^2 + (y_{tk} - 0)^2)1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((0 - x_{P_l(t)k})^2 + (0 - y_{P_l(t)k})^2)1\{k \in Ind(P_l(t)) \backslash Ind(t)\} \\ &= \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2)1\{k \in Ind(t) \cap Ind(u)\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((x_{tk} - 0)^2 + (y_{tk} - 0)^2)1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\ &+ \sum_{k=1}^{\infty} \alpha_k((0 - x_{uk})^2 + (0 - y_{uk})^2)1\{k \in Ind(P_l(t)) \backslash Ind(t)\} \end{aligned} \tag{11.14}$$

Note that, the tree $P_l(t)$ is a subtree of the tree u; that is,

$$Ind(P_l(t)) \subset Ind(u).$$

Also, by Equation 11.12,

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

Thus, combining Equations (11.14) and (11.11), and using Equation (11.13), we have

$$
\begin{aligned}
& f_\rho^2(t, P_l(t)) + f_\rho^2(u, P_l(t)) \\
= & \sum_{k=1}^{\infty} \alpha_k((x_{tk} - x_{uk})^2 + (y_{tk} - y_{uk})^2) 1\{k \in Ind(t) \cap Ind(u)\} \\
& + \sum_{k=1}^{\infty} \alpha_k((x_{tk} - 0)^2 + (y_{tk} - 0)^2) 1\{k \in Ind(t) \backslash Ind(P_l(t))\} \\
& + \sum_{k=1}^{\infty} \alpha_k((0 - x_{uk})^2 + (0 - y_{uk})^2) 1\{k \in Ind(P_l(t)) \backslash Ind(t)\} \\
& + \sum_{k=1}^{\infty} \alpha_k((x_{uk})^2 + (y_{uk})^2) 1\{k \in Ind(u) \backslash Ind(P_l(t))\} \\
= & f_\rho^2(u, t) + \sum_{k=1}^{\infty} \alpha_k((x_{tk} - 0)^2 + (y_{tk} - 0)^2) 1\{k \in Ind(t) \cap Ind(u) \cap \overline{Ind(P_l(t))}\} \\
& + \sum_{k=1}^{\infty} \alpha_k((x_{uk} - 0)^2 + (y_{uk} - 0)^2) 1\{k \in Ind(t) \cap Ind(u) \cap \overline{Ind(P_l(t))}\} \\
= & f_\rho^2(u, t)
\end{aligned}
$$

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

By now, the single tree version Pythagorean theorem is satisfied when the tree $P_l(t_i)$ is a subtree of the tree u. That is, for $i < K$,

$$V_\rho(t_i, u) = V_\rho(t_i, P_l(t_i)) + V_\rho(P_l(t_i), u). \tag{11.15}$$

For $i > K$, $P_l(t_i) \supset u$. Note that, the tree $P_l(t_i)$ is the projection of the tree t, which implies,

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

Thus,

$$(Ind(t_i) \backslash Ind(P_l(t_i))) \cup (Ind(P_l(t_i)) \backslash Ind(u)) = Ind(t_i) \backslash Ind(u), \tag{11.16}$$

$$(Ind(P_l(t_i)) \backslash Ind(u)) \cup (Ind(t_i) \cap Ind(u)) = Ind(t) \cap Ind)(P_l(t)) \tag{11.17}$$

and

$$Ind(P_l(t_i)) \backslash Ind(t) = Ind(u) \backslash Ind(t). \tag{11.18}$$

Hence, using Equation (11.13) and Equation (11.16),

$$\begin{aligned} f_\rho^2(t_i,u) = & \sum_{k\in Ind(t_i)\cap Ind(u)} \alpha_k((x_{t_ik}-x_{uk})^2+(y_{t_ik}-y_{uk})^2) \\ & + \sum_{k\in Ind(t_i)\setminus Ind(P_l(t_i))} \alpha_k(x_{t_ik}^2+y_{t_ik}^2) + \sum_{k\in Ind(P_l(t_i))\setminus Ind(u)} \alpha_k(x_{t_ik}^2+y_{t_ik}^2) \\ & + \sum_{k\in Ind(u)\setminus Ind(t_i)} \alpha_k(x_{uk}^2+y_{uk}^2) \end{aligned} \tag{11.19}$$

Using Equation (11.14), Equation (11.17) and Equation (11.18),

$$\begin{aligned} f_\rho^2(t_i,P_l(t_i)) = & \sum_{k\in Ind(P_l(t_i))\setminus Ind(u)} \alpha_k((x_{t_ik}-x_{uk})^2+(y_{t_ik}-y_{uk})^2) \\ & + \sum_{k\in Ind(t_i)\cap Ind(u)} \alpha_k((x_{t_ik}-x_{uk})^2+(y_{t_ik}-y_{uk})^2) \\ & + \sum_{k\in Ind(t_i)\setminus Ind(P_l(t_i))} \alpha_k(x_{t_ik}^2+y_{t_ik}^2) + \sum_{k\in Ind(u)\setminus Ind(t_i)} \alpha_k(x_{uk}^2+y_{uk}^2) \end{aligned} \tag{11.20}$$

Also,

$$f_\rho^2(P_l(t_i),u) = \sum_{k\in Ind(P_l(t_i))\setminus Ind(u)} \alpha_k(x_{uk}^2+y_{uk}^2) \tag{11.21}$$

By equations (11.19), (11.20) and (11.21), we have

$$\begin{aligned} & \sum_{i=K+1}^{n} (f_\rho^2(t_i,u) - f_\rho^2(t_i,P_l(t_i)) - f_\rho^2(P_l(t_i),u)) \\ & = \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\setminus Ind(u)} \alpha_k[(x_{t_ik}^2+y_{t_ik}^2) - ((x_{t_ik}-x_{uk})^2+(y_{t_ik}-y_{uk})^2) - (x_{uk}^2+y_{uk}^2)] \\ & = \sum_{i=K+1}^{n} \sum_{k\in Ind(P_l(t_i))\setminus Ind(u)} 2\alpha_k(x_{t_ik}x_{uk} - x_{uk}^2 + y_{t_ik}y_{uk} - y_{uk}^2) \\ & = 0 \end{aligned} \tag{11.22}$$

Note that, each entry in the summation is not equal to zero.

Finally, summarizing Equation (11.15) over $i = 1, 2, \ldots, K$ and combining Equation(11.22) and Equation (11.7), the Pythagorean theorem without nodal information, we have Equation (11.9).

□

12 Principal Component Analysis on finite level trees with nodal information

In section 8, we have defined a new metric δ on the tree space with nodal information and a specific metric ρ on the finite level tree space. Note that, δ (ρ) is the sum of the integer part metric d_I and the fractional part f_δ (f_ρ). Furthermore, we have defined the variation of a sample of trees about its "center point"— median-mean tree. In this section, we will mainly discuss the problem of finding simple explanation the variation of the sample.

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 we develop an analog of the PCA method?

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

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

Also, f_ρ is proportional to the Euclidean distance between two vectors. Therefore, we can apply standard PCA in this case.

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

Recall that, in section 5 and section 6, we have developed the idea of tree line as a one-dimensional representation of the data in the binary tree space. Also, we developed the tree version PCA on tree space without nodal information.

Now, on the binary tree space with nodal information, we will combine the tree version PCA and the standard PCA on Euclidean space to develop a new PCA on tree space with nodal information. We will use the tree version PCA to capture interesting features of the tree structure and use standard PCA to analyze the nodal information.

Definition 12.1. An s-treeline is called a **one-dimensional principal structure representation** of the sample T if it minimizes the sum

$$\sum_{i=1}^{n} V_\rho(t_i, P_l(t_i)) \tag{12.1}$$

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

According to the tree version Pythagorean theorem, (Theorem 11.4) minimizing the sum (12.1) is equivalent to maximizing the following sum

$$\sum_{i=1}^{n} V_\rho(\mu_\rho, P_l(t_i)) \tag{12.2}$$

Example 12.1. Let T be a sample of finite level trees in $\mathcal{T}_w$ with sample size $n = 5m$, where w is shown in figure 12.1. There are five types of trees and each has m elements in T. Suppose trees $t_{(i-1)m+1}, \ldots, t_{im}$ have type i, for $i = 1, 2, 3, 4, 5$.

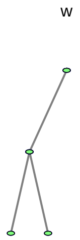

Figure 12.1: Binary tree w

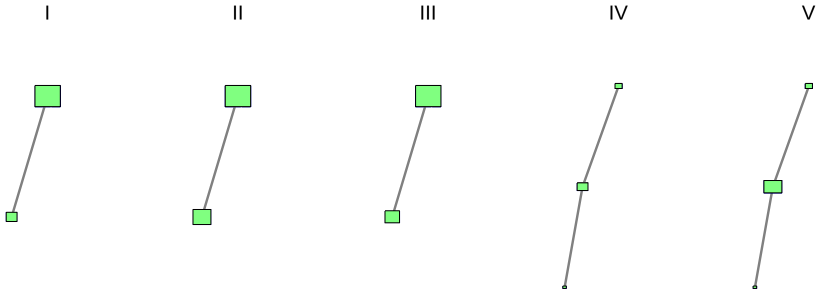

Figure 12.2: Representative of the binary tree sample T. There are m elements of each type.

level-order index	I	II	III	IV	V
1	(0.7,0.7)	(0.7,0.7)	(0.7,0.7)	(0.2,0.2)	(0.2,0.2)
2	(0.3,0.3)	(0.5,0.5)	(0.4,0.4)	(0.3,0.3)	(0.5,0.5)
4	n/a	n/a	n/a	(0.1,0.1)	(0.1,0.1)

The support tree t_{sup} and average support tree t_a of the sample T are shown in the figure 12.3.

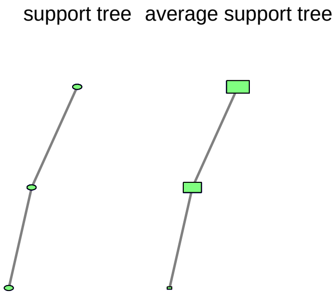

Figure 12.3: Support tree and average support tree of the sample T

The median-mean tree m_ρ, center point of the sample T, is shown in Figure 12.4. Note that, there is a unique median-mean tree of the sample T. The nodal information of the average support tree t_a and the median-mean binary tree m_ρ is listed in the table below.

level-order index	t_a	m_ρ
1	(0.5,0.5)	(0.5,0.5)
2	(0.4,0.4)	(0.4,0.4)
4	(0.1,0.1)	n/a

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

$$\sum_{i=1}^{5m} V_\rho(t_i, m_\rho) = 2.18m \tag{12.3}$$

where $N(w) = 4$ in the definition of the metric ρ.

Now, we will find the treeline to describe features of the data. Note that, there is a unique s-treeline $l = \{u_1, u_2, u_3\}$ in this example.

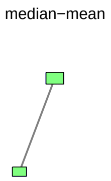

Figure 12.4: The median-mean tree m_ρ of the sample T

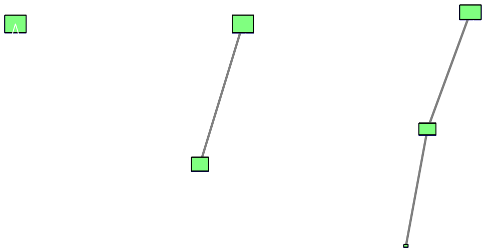

Figure 12.5: The unique s-treeline l for the sample T.

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:

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), m_\rho) = 2.01m$$

and

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), t_i) = 0.17m$$

which verifies the tree version Pythagorean theorem, i.e.,

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

In the example 12.1, the proportion of variation that the one-dimensional structure representation explains is

$$\frac{\sum_{i=1}^{5m} V_\rho(P_l(t_i), m_\rho)}{\sum_{i=1}^{5m} V_\rho(t_i, m_\rho)} = \frac{2.01}{2.18} = 92.2\%.$$

We can see that, there is no other s-treeline to explain more about the total variation in the example 12.1. Now we will use the other type of treeline — i-treeline.

Recall that, in Definition 11.2, an i-treeline is determined by a tree u_0 and an information vector $\vec{v}$.

Definition 12.2. Let $\vec{c}$ be any vector of information. An i-treeline e, determined by u_0 and $\vec{v}$, is called a **$\vec{c}$-induced i-treeline** if $\vec{v}$ is a restriction of $\vec{c}$, in particular,

$$(v_{2k-1}, v_{2k}) = \begin{cases} (c_{2k-1}, c_{2k}), & \text{if } k \in Ind(u_0) \\ \\ (0,0), & \text{if } k \notin Ind(u_0). \end{cases} \tag{12.4}$$

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 i-treeline e_j. Now, we will find a vector (first principal component $\vec{p_1}$) which minimizes

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

over all vectors $\vec{v}$.

Similar to the PCA in ordinary Euclidean space, we will find the orthogonal vectors $p2, p3, \ldots$. Furthermore, denote the induced i-treeline by the vector $\vec{p_k}$ passing through the tree $P_l(t_j)$ by e_{jk}.

The idea of $\vec{c}$-induced i-treeline is now illustrated in the context of the previous example 12.1,

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

and

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

Thus,

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), P_{e_{i1}}) = 0.15m \tag{12.5}$$

and

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), P_{e_{i2}}) = 0.02m \tag{12.6}$$

According to the equations (12.5) and (12.6), it is straight forward that

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), P_{e_{i1}}) + \sum_{i=1}^{5m} V_\rho(P_l(t_i), P_{e_{i2}}) = 0.17m = \sum_{i=1}^{5m} V_\rho(P_l(t_i), t_i).$$

Note that, in Example 12.1, the total variation, $2.18m$, was decomposed into three parts. The first part, $2.01m$, was explained by the first principal structure treeline. And, two information treelines explain $0.15m$ and $0.02m$ respectively.

Example 12.2. w is 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 the table below). The numbers of elements of each type are 1, 1, m and m ($m > 1$).

level-order index	I	II	III	IV
1	(0.5,0.5)	(0.3,0.3)	(0.5,0.5)	(0.3,0.3)
2	(0.2,0.2)	(0.2,0.2)	(0.2,0.2)	(0.2,0.2)
3	(0.3,0.3)	(0.7,0.7)	(0.3,0.3)	(0.7,0.7)
7	(0.1,0.1)	(0.1,0.1)	n/a	n/a

The total variation is

$$\sum_{i=1}^{2m+2} V_\rho(t_i, m_\rho) = 0.05m + 2.06. \tag{12.7}$$

There are two s-treelines passing through the unique median-mean tree, l_1 and l_2. By calculation,

$$\sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), t_i) = 0.05m + 0.05.$$

and

$$\sum_{i=1}^{2m+2} V_\rho(P_{l_2}(t_i), t_i) = 0.05m + 2.06.$$

Hence, the s-treeline l_1 is the one-dimensional structure representation. By the tree version 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 structure representation l_1 explains is

$$\frac{\sum_{i=1}^{2m+2} V_\rho(P_{l_1}(t_i), m_\rho)}{\sum_{i=1}^{2m+2} V_\rho(t_i, m_\rho)} = \frac{2.01}{0.05m + 2.06}$$

Note that, this proportion is arbitrary 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 information structure.

Next, we will find the principal i-treeline to decompose the variation in the direction of information.

By calculation, the first principal component 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 Pythagorean theorem, the proportion of variation explained by the i-treeline is

$$\frac{0.05m + 0.05}{0.05m + 2.06},$$

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

References

[1] Banks, D. and Constantine, G. M. (1998), "Metric Models for Random Graphs", Journal of Classification 15:199-223.

[2] Shannon, William and Banks, David (1999), "Combining Classification Trees Using MLE", Statistics in Medicine 18:727-740.

[3] Margush, T. (1982), "Distances Between Trees", Discrete Applied Mathematics 4:281-290.

[4] Yushkevich, Paul, et al (2001), "Intuitive, Localized Analysis of Shape Variablity".

[5] Pizer, S.M., A Thall, Chen, D.. (1999). "M-Reps: A New Object Representation for Graphics".