An Introduction to Functional Data Analysis of Populations of Tree-structured Objects

Haonan Wang[1] and J. S. Marron[2]

[1] Department of Statistics
University of North Carolina at Chapel Hill
wanghn@email.unc.edu
[2] Department of Statistics
University of North Carolina at Chapel Hill
marron@stat.unc.edu

Abstract. This paper proposes a new method for understanding the structure of populations of complex objects in the area of medical image analysis. The new methods require invention of approaches to the statistical analysis of a population of tree-structured objects. The approach is based on a metric in tree space. The metric provides a foundation for defining a notion of population center. In Functional Data Analysis, variation about the center is usually analyzed by Principal Component Analysis. Here an analog of PCA is developed for tree space.

1 Introduction

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, [1] for discussion and further references).

Fig. 1. Example of shapes of interest

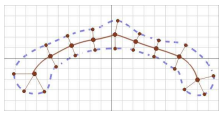

Fig. 2. Representation by M-reps

For example, consider the shape in Figure 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.

A class of convenient and powerful shape representations is M-reps (see Pizer, S.M., et al, [2]). These are being developed by the Medical Image Display and Analysis Group at UNC[3]. The M-rep parameters (location, radius, angles) are concatenated into a vector to provide a numerical summary of the shape.

[3] visit the MIDAG web site at *http://midag.cs.unc.edu*

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 3).

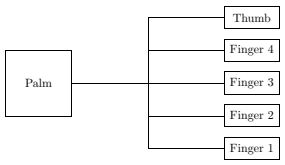

Fig. 3. An example of multi-figural object — hand

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 to 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.

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.

In section 2, a brief overview [4] of the methodology on tree space is given, e.g., a new metric, a population "center point" and an analog of Principal Component Analysis. In section 3, an application of the tree version PCA is discussed.

2 Development of the Method On the Tree Space

In this research, a population of abstract complex multi-figural objects is considered. The single observation in this population is called a "tree". For simplicity, a special case, the binary tree, is studied.

A binary tree is a tree such that every node has at most two children (left child and right child). Also, every node is uniquely labelled by a natural number, *level-order index*.

Tree structure represents the topological aspects of the data. Sometimes, the nodes of the trees contain attributes, numerical values such as M-rep parameters,

[4] A draft of my dissertation proposal can be viewed at *http://www.cs.unc.edu/Research/MIDAG/pubs/papers*

which should also be used in the statistical analysis. In later research, we will study both aspects of the population, structure and attributes.

A first question for statistical analysis is, what is the "center point" of a tree population? A notion of "center point" of a population is the tree which is the "closest to all other trees". This requires a metric on the tree space.

A new metric (δ) is defined on the tree space as the summation of two parts: an integer part (d_I) and a fractional part (f_δ); that is, $\delta = d_I + f_\delta$. The integer part metric measures the topological difference between two trees by counting the number of different nodes; while, the fractional part metric measures the attribute difference, equivalent to the weighted Euclidean distance on the attribute vectors. (See Margush, [3] for more discussion of metrics on trees.)

Next, we define the "center point" of a tree population with structure only. Considering a finite tree sample $T = \{t_1, t_2, \ldots, t_n\}$, the "center point", the median tree, is defined as the minimizer of the summation $\sum_{i=1}^{n} d_I(t, t_i)$ over all trees t. The minimizer tree must follow the *majority rule* (Banks and Constantine, [4], page 204), a node is in the median tree if and only if it is present in more than half of the trees (non-uniqueness may arise when n is even).

Furthermore, a new "center point", the median-mean tree, is introduced as a combination of median and mean for a tree sample with nodal attributes. Its tree structure complies with the majority rule and its nodal attributes can be calculated as a "sample mean".

For a tree sample, we can quantify the variation as

$$\sum_{i=1}^{n} V_\delta(t_i, m_\delta) = \sum_{i=1}^{n} d_I(t_i, m_\delta) + \sum_{i=1}^{n} f_\delta^2(t_i, m_\delta)$$

where m_δ is a "central" tree and $V_\delta = d_I + f_\delta^2$ is the variation function.

In Euclidean space, Principal Component Analysis (PCA) provides a useful decomposition of complex data sets, in terms of simple one-dimensional representations. In tree space, a challenging non-linear space, the "treeline" plays the role of the line in Euclidean space, i.e. a one-dimensional subset. There are two important types, the structure treeline and the attribute treeline (shown in Figure 4 and Figure 5), which indicate the variation of tree structure and attributes respectively. (See Wang, [7]). Note that, the size of the box associated with each node visually indicates the magnitude of the corresponding attributes.

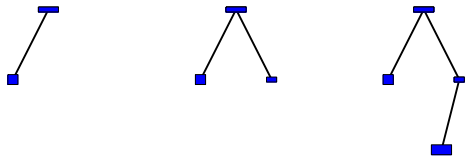

Fig. 4. An example of a structure treeline with nodal attributes

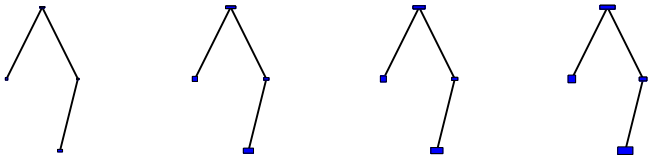

Fig. 5. An example of an attribute treeline

For any tree t and treeline l, the projection of t onto l is the tree which minimizes the distance $\delta(t, \cdot)$ over all trees on the treeline l and is denoted by $P_l(t)$.

Next, a tree version Principal Component Analysis is developed on the tree space. The fundamental theorem is a tree version of the Pythagorean theorem. (See Wang, [7]). The total variation $\sum V_\delta(t_i, m_\delta)$ can be decomposed as

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

where l is a structure treeline passing through the central tree m_δ or any attribute treeline.

The tree version PCA is a two-step variation analysis. First, it finds the principal structure treeline as the minimizer of the sum $\sum V_\delta(P_l(t_i), t_i)$ over all structure treelines passing through the central tree m_δ, which quantifies the topological aspects of the structure of the tree population. Next, it finds the principal attribute direction, which quantifies features of the nodal attributes.

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 of a weighted PCA. When the structures are not all the same, the tree version PCA will give a more appropriate attribute direction. We illustrate this idea using the following toy example.

Example 1. Let T be a sample of trees with size $n = 13$. Each member in T has one of the two structures shown in Figure 6. The tree attributes have the form shown in Table 1, where x_i and y_i are real values, $i = 1, 2, \ldots, 13$.

Fig. 6. Two types of tree structures in T

Table 1. Attribute form of trees in T

Level-order index	Attributes
1	(0.1,0.1)
2	(x_i, x_i)
3	(y_i, y_i)

Six of the trees have just two nodes, so the y_i is non-existent. The x_i and y_i values are shown in Figure 7, with the missing y_i replaced by the average of the non-missings. Note that, the red circle indicates the center point, which is the sample mean in the regular PCA and the nodal attributes of the median-mean in the tree version PCA.

Applying a weighted PCA to the attribute vectors, gives the first principal component (solid line in Figure 7). It shows that the trees with the Type I structure have a strong effect on the attribute direction, pulling it towards a horizontal line.

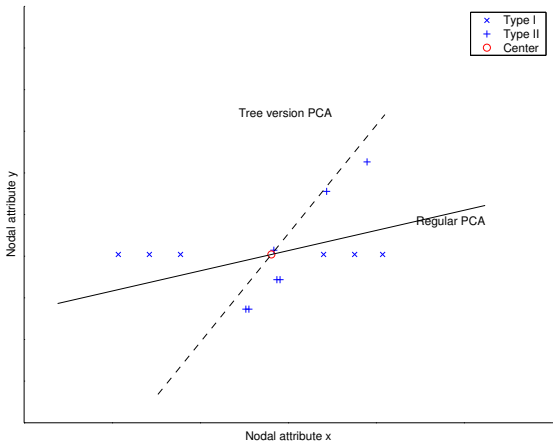

Fig. 7. Scatter plot of attributes and principal attribute directions given by Regular PCA and Tree version PCA

Next, we will apply the tree version PCA 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 8. 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.

Fig. 8. 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 7. Comparing with the direction given by regular PCA, it is more appropriate because it correctly represents the variation in the attributes. The Type I elements should not influence the direction because they contain no information about the relationship between the attributes.

3 Application of Tree Version PCA

In section 2, the PCA on tree space was developed. Now, we will apply the approach to a sample of blood vessel trees.

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. [5]

An example of brain blood vessels is shown in Figure 9, provided by Dr. E. Bullitt, UNC Department of Surgery [6]. This system has three important components: left carotid, right carotid and Vertebrobasilar system, shown in different colors.

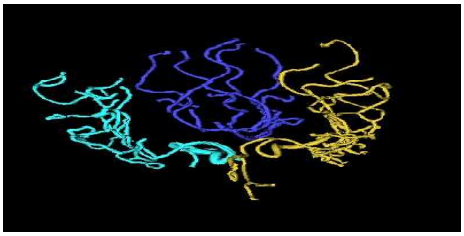

Fig. 9. An example of brain blood vessels

[5] From *http://www.thedoctorslounge.net/education/tutorials/cerebcirc/cerebcirc1.htm.*
[6] Dr. Bullitt's webpage can be viewed at *http://casilab.med.unc.edu/Bullitt_Home.htm.*

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. The attributes of the root node are, the three coordinates of the starting point and the three coordinates of the ending point. The attributes of the non-root nodes are the three coordinates of the ending point and the proportion parameter,

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

which determines the location of the starting point by interpolation of the parent's starting and ending points.

For simplicity and computational speed, we only consider a subtree (up to level 2 and three nodes) of each element among those 11 trees (See Figure 10). The trees with thicker black lines are the median-mean trees in each figure. Note that, the median-mean trees are "central" in terms of structure, size, and location, for each of the three people.

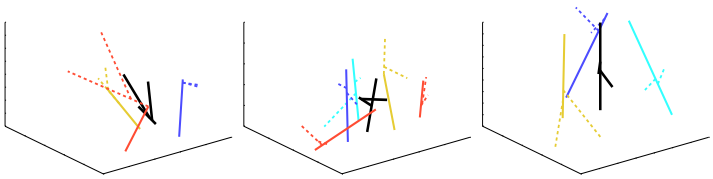

Fig. 10. Reduced blood vessel trees (thin colored lines) and the median-mean trees (thicker black line) for each person. Root nodes are solid and children are dashed.

These trees are combined into a larger population in Figure 11. Again, the median-mean of the larger population is showed as a thick 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.

Next, we will apply the tree version PCA to the blood vessel tree sample (denoted by T). There are only two types of tree structures of these 11 trees, shown in Figure 12.

The principal structure treeline $l = \{u_0, u_1, u_2\}$ is shown in Figure 13 (structure only, without attributes) and Figure 14 (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.

Next, consider the principal attribute direction. The attribute treelines passing through the median-mean and through the full support tree are shown in

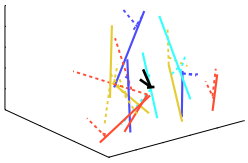

Fig. 11. Combined population of reduced blood vessel trees and the median-mean tree

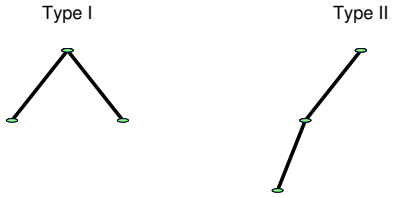

Fig. 12. Two types of tree structures in T

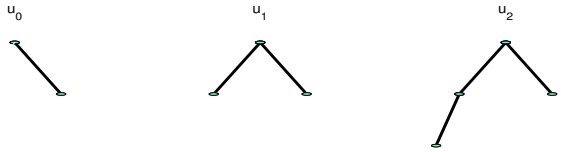

Fig. 13. Principal structure treeline $l = \{u_0, u_1, u_2\}$ without nodal attributes

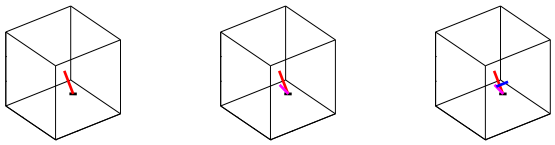

Fig. 14. Principal structure treeline $l = \{u_0, u_1, u_2\}$ with nodal attributes

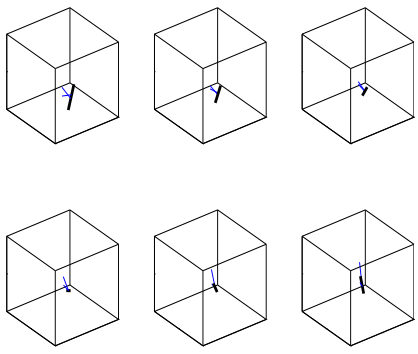

Fig. 15. Attribute treeline passing through the median-mean tree.

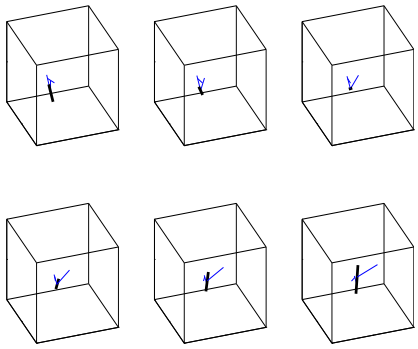

Fig. 16. Attribute treeline passing through the full support tree.

Figure 15 and Figure 16. There are six subplots in each figure. Each subplot depicts one location on the attribute treeline.

In Figure 15, from the upper left subplot to the lower right one, we can see that the orientation of the main root (solid black line) changes very substantially, in fact "flipping", with the top becoming the bottom. This was a surprising feature of the population. Careful investigation showed that the given data set used two orientations to the direction of blood flow. Some of them have the same direction; while, some of them have the inverse direction. Also, we can verify these two clusters in the data from the projection coefficients of all 11 trees on the attribute treeline passing through the median-mean tree (shown in Figure 17). 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 middle two frames in Figure 15, with a very short root, as can be seen in the raw data in Figure 11.

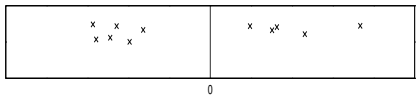

Fig. 17. Projection coefficients of 11 trees on the attribute treeline passing through the median-mean tree

Figure 16 shows the attribute treeline passing through the full support tree. Similar to Figure 15, the six frames show that the main root also has a tendency of flipping over and the length of main root becomes shorter (three subplots on the top row) then becomes longer (three subplots on the bottom row). A projection plot, similar to Figure 17, shows the same structure.

In this example, we saw that the tree version PCA found a surprising characteristic of the population: there are two different orientations about the blood flow in the data set. This dominates the total variation, perhaps obscuring population features of more biological interest. Future work with Dr. Bullitt will investigate these.

References

1. Yushkevich, Paul, et al (2001), "Intuitive, Localized Analysis of Shape Variablity".
2. Pizer, S.M., A Thall, Chen, D.. (1999). "M-Reps: A New Object Representation for Graphics".
3. Margush, T. (1982), "Distances Between Trees", Discrete Applied Mathematics 4:281-290.
4. Banks, D. and Constantine, G. M. (1998), "Metric Models for Random Graphs", Journal of Classification 15:199-223.
5. Shannon, William and Banks, David (1999), "Combining Classification Trees Using MLE", Statistics in Medicine 18:727-740.
6. Bullitt E, Aylward S. (2003) "Volume rendering of segmented image objects". In press IEEE-TMI.
7. Wang, Haonan, (2003), "Functional Data Analysis of Populations of Tree-structured Objects".Dissertation in progress.