Submitted to Annals of Statistics

OBJECT ORIENTED DATA ANALYSIS: SETS OF TREES[1]

By Haonan Wang and J. S. Marron

Colorado State University and University of North Carolina

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Recent developments in medical image analysis motivate the statistical analysis of populations of more complex data objects which are elements of mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, or of strongly non-Euclidean spaces, such as spaces of tree-structured data objects. These new contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics. This point is illustrated through the careful development of a novel mathematical framework for statistical analysis of populations of tree structured objects.

1. Introduction Object Oriented Data Analysis (OODA) is the statistical analysis of data sets of complex objects. The area is understood through consideration of the *atom of the statistical analysis*. In a first course in statistics, the atoms are numbers. Atoms are vectors in multivariate analysis. An interesting special case of OODA is Functional Data Analysis, where atoms are curves, see Ramsay and Silverman (1997, 2002) for excellent overviews, as well as many interesting analyses, novel methodologies and detailed discussion. More general atoms have also been considered. Locantore, et al (1999) studied the case of images as atoms, and Pizer, et al (1999) and Yushkevich, et al (2001) took the atoms to be shape objects in two and three dimensional space.

An important major goal of OODA is understanding *population structure* of a data set. The usual first step is to find a *centerpoint*, e.g. a mean or median, of the data set. The second step is to analyze the *variation about the center*. Principal Component Analysis (PCA) has been a workhorse method for this, especially when combined with new visualizations as done in Functional Data Analysis. An important reason for this success to date is that the data naturally lie in Euclidean spaces, where standard vector space analyses have proven to be both insightful and effective.

Medical image analysis is motivating some interesting new developments in OODA. These new developments are not in traditional imaging areas, such as the

[1] This work was supported in part by National Science Foundation, Grant DMS-9971649, 1999-2001 and in part by NIH-NCI P01 CA47982.

AMS 2000 subject classifications. Primary 62H99; secondary 62G99.

Key words and phrases. functional data analysis, nonlinear data space, object oriented data analysis, population of tree-structured objects, principal component analysis.

denoising, segmentation and/or registration of a single image, but instead are about the analysis of *populations of images*. Again common goals include finding center-points and variation about the center, but also discrimination, i. e. classification, is important. A serious challenge to this development is that the data often naturally lie in non-Euclidean spaces. A range of such cases has arisen, from mildly non-Euclidean spaces, such as Lie Groups and Symmetric Spaces, to strongly non-Euclidean spaces, such as populations of tree or graph structured data objects. Because such non-Euclidean data spaces are generally unfamiliar to statisticians, there is opportunity for the development of several types of new interfaces between statistics and mathematics. One purpose of this paper is to highlight some of these. The newness of this non-standard mathematical statistics, that is currently under development (and much of which is yet to be developed), is underscored by a particularly deep look at an example of tree structured data objects.

Lie Groups and Symmetric Spaces are the natural domains for the data objects which arise in the medial representation of body parts, as discussed in Section 1.1. Human organs are represented using vectors of parameters, which have both real valued and angular components. Thus each data object is usefully viewed as a point in a Lie Group, or a Symmetric Space, i. e. a curved manifold space. Such representations are often only mildly non-Euclidean, because these curved spaces can frequently be approximated to some degree by tangent spaces, where Euclidean methods of analysis can be used. However the most natural and convincing analysis of the data is done “along the manifold”, as discussed in Section 1.1. Because there already exists a substantial medical imaging literature on this, only an overview is given here.

Data objects which are trees or graphs are seen in Section 1.2 to be important in medical image analysis for several reasons. These data types present an even greater challenge, because the data space is strongly non-Euclidean. Fundamental tools of standard vector space statistical analysis, such as linear subspace, projection, analysis of variance and even linear combination are no longer available. Preliminary ad hoc attempts made by the authors at this type of OODA ended up collapsing in a mass of contradictions, because they were based on trying to apply Euclidean notions in this very non-Euclidean domain. This motivated the development of a really new type of mathematical statistics: a rigorous definition-theorem-proof framework for the analysis of such data, which was the dissertation of Wang (2003). In Section 2 it is seen how these tools provide an analysis of a real data set. Section 3 gives an overview of the mathematical structure that underpins the analysis.

Note that statistics and mathematics (of some non-standard types) meet each other in several ways in OODA. For the Lie Group - symmetric space data, mathematics provides a non-standard framework for conceptualizing the data. For data as trees, an axiomatic system is used as a device to overcome our poor intuition for data analysis in this very non-Euclidean space. Both of these marriages of mathematics and statistics go in quite different directions from that of much of mathematical statistics: the validation and comparison of existing statistical methods through asymptotic analysis as the sample size tends to infinity. Note that this latter type of analysis has so far been completely unexplored for these new types of OODA, and it also should lead to the development of many more interesting connections between mathematics and statistics.

1.1. *OODA on Lie Groups - Symmetric Spaces.* Shape is an interesting and useful characteristic of objects (usually in three dimensions) in medical image analysis. Shape is usually represented as a vector of measurements, so that a data set of shapes can be analyzed as a set of vectors. There are a number of ways to represent shapes of objects. The best known in the statistical literature is landmark based approaches, see Dryden and Mardia (1988) for good overview of this area. While they have been a workhorse for solving a wide variety of practical problems, landmark approaches tend to have limited utility for population studies in medical imaging, because a sufficient number of well defined, replicable landmarks are frequently impossible to define.

Another common approach to shape representation is via various models for the boundary. Popular methods of this type include various types of triangular meshes, the Fourier boundary representations as discussed in Szekely, et al (1996), and the sophisticated active shape / appearance models, see Cootes (2000) and Cootes and Taylor (2001) for good introduction and overview.

A class of convenient and powerful shape representations is *m-reps* (a shortening of “medial representation”), which are based on *medial* ideas, see Pizer, et al (1999) and Yushkevich, et al (2001) for detailed introduction and discussion. The main idea is to find the “central skeletons” of objects, and then to represent the whole object in terms of ”spokes” from the center to the boundary. The central structure and set of spokes to the boundary are discretized and approximated by a finite set of m-reps. The m-rep parameters (location, radius and angles) are the features and are concatenated into a feature vector to provide a numerical summary of the shape. Each data object is thus represented as the direct product (thus a large vector) of these parameters over the collection of m-reps. A major motivation for using m-reps over other types of representation is that they provide a more direct solution to the *correspondence problem*, which is to match parts of one object with corresponding parts of other members of the population.

A simple example of the use of m-reps in OODA is shown in Figure 1, which uses the specific representation of Yushkevich, et al (2001), which studied a set of human corpora callosa, gathered from two dimensional Magnetic Resonance Images. The corpus callosum is the small window between the left and right halves of the brain. The left hand panel of Figure 1 shows a single m-rep decomposition of one corpus callosum. Each large central dot shows the center of an m-rep (5 of which are used to represent this object). The m-reps are a discretization of the medial axis, shown in blue. The boundary of the object is determined by the spokes, which are the shorter green line segments emanating from each m-rep. These spokes are paired, and are determined by their (common) angle from the medial axis, and their length. All of these parameters are summarized into a feature vector which is used to represent each object.

The right hand panel of Figure 1 shows a simple OODA, of a population of 72 corpora callosa. This is done here by simple principal component analysis of the set of feature vectors. The central green shape is the mean of the population. The colored sequence of shapes gives insight into population variation, by showing the first principal component (thus the mode of maximal variation). In particular, each shape shows a location along the eigenvector of the first PC. This shows that the dominant mode of variation in this population is in the direction of more overall bending in one direction, shown by the red curves, versus less overall bending in the opposite direction, shown by the blue curves.

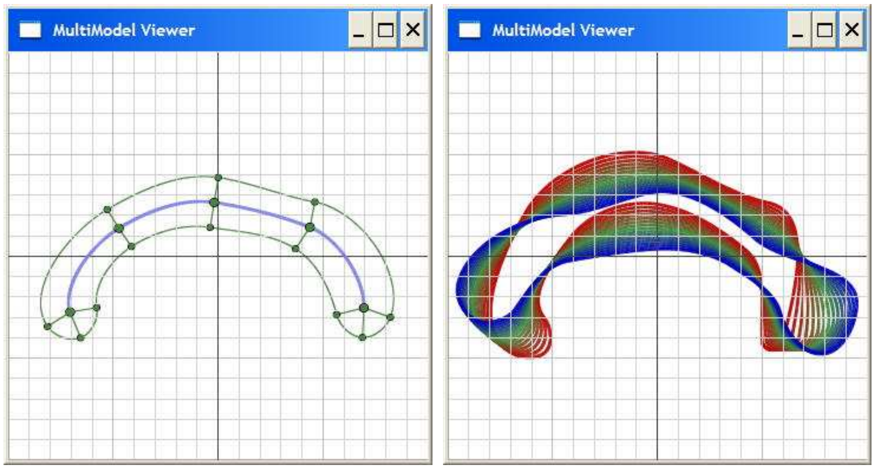

Fig. 1. *Corpus Callosum Data. Left hand panel shows all components of the medial representation of the corpus callosum of one person. Right hand panel shows the boundaries of objects lying along the first PCA eigenvectors, showing the largest component of variation in the population.*

Figure 2 illustrates m-reps for a much more complicated data object, called a *multifigural object*. This time the shapes lie in three dimensions, and each data object consists of the bladder, prostate and rectum (each of which is called a *figure*) of a single patient. The left panel shows the m-rep centers as small yellow spheres. The spokes are shown as colored line segments. A mesh representation of the boundary of each figure is added to the m-reps in the center panel. These boundaries are then rendered as surfaces in the right panel.

A simple approach to OODA for m-rep objects is to simply use Euclidean PCA on the vectors of parameters. However, there is substantial room for improvement, because some parameters are angles, while others are radii (thus positive in sign), and still others are position coordinates. One issue that comes up is that units are not commensurate, so some vector entries could be orders of magnitude different from the others, which will drastically effect PCA. An approach to this problem is to replace the eigen-analysis of the covariance matrix (conventional PCA) with the an eigen-analysis of the correlation matrix (a well known scale free analog of PCA). But this still does not address the central challenge of statistical analysis of angular data. For example, what is the average of a set of angles where some are just above 0°, and the rest are just below 360°? The sensible answer is something very

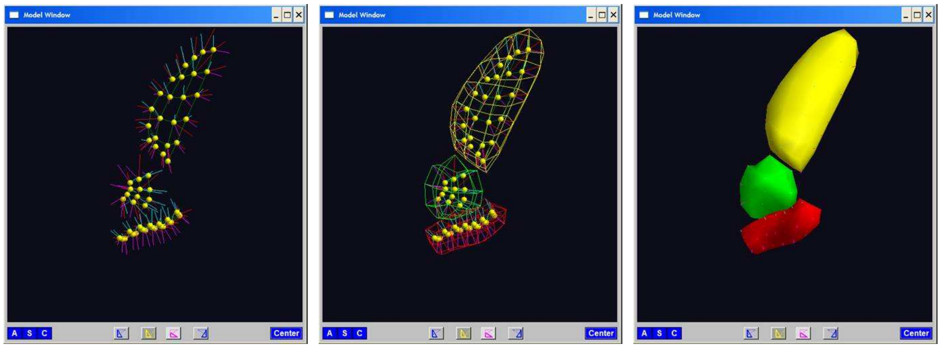

Fig. 2. *Male pelvic data, showing medial representation of the bladder (yellow), prostate (green) and a segment of the rectum (red) for a single person. The left panel shows the medial atoms and spokes. The center panel shows the implied boundary of the three objects represented as a mesh. The right panel shows a surface rendering of the implied boundaries of the three objects.*

close to 0°, but simple averaging of the numbers involved can give a diametrically opposite answer closer to 180°. There is a substantial literature on the statistical analysis of angular data, also called directional data, i. e. data on the circle or sphere. See Fisher (1993), Fisher, Lewis and Emgleton (1987) and Mardia (1972, 2000) for good introduction to this area. A fundamental concept of this area is that the most convenient mathematical representation of angles is as points on the unit circle, and for angles in 3-d, as points on the unit sphere.

For these same reasons it is natural to represent the vectors of m-rep parameters as direct products of points on the circle and/or sphere for the angles, as positive reals for the radii, and as real numbers for the locations. As noted in Fletcher, et al (2005), the natural framework for understanding this type of data object is Lie Groups and/or Symmetric Spaces. Fletcher, et al, go on to develop an approach to OODA for such data. The Frechét approach gives a natural definition of the sample center, and Principal Geodesic Analysis (PGA) quantifies population variation.

The Frechét mean has been a popular concept in robustness, since it provides useful generalizations of the sample mean. It also provides an effective starting point for non-Euclidean OODA. The main idea is that one way to characterize the sample mean is as the minimizer of the sum of squared distances to each data point. Thus the Frechét mean can be defined in quite abstract data spaces, as long as a suitable metric can be found. For Lie Group - symmetric space data, the natural distance is along geodesics, i. e. along the manifold, and this Frechét mean is called the geodesic mean.

Fletcher's Lie Group - symmetric space variation of PCA is PGA. The key to this approach is to characterize PCA as finding lines which maximally approximate the data. On curved manifolds, the analog of lines are geodesics, so PGA searches for geodesics which maximally approximate the data. See Fletcher et al (2005) for detailed discussion and insightful examples.

A different example of OODA on curved manifolds can also be found in Izem, Kingsolver and Marron (2005). That work was motivated by a problem in evolutionary biology, where the data points were curves, but the goal was to quantify nonlinear modes of variation. This was accomplished by the development of an analog of analysis of variance, in the specified directions of evolutionary interest, meaning along a curved manifold surface.

Curved manifolds have also been featured recently in statistics in a completely different context. In the area that is coming to be called *manifold learning*, the main idea is that some high dimensional data sets may lie on low dimensional manifolds. This idea goes back at least to the Principal Curves idea of Hastie and Stuetzle (1989). Some popular current approaches to finding such manifolds are the ISOMap of Tenenbaum, da Silva and Langford (2000), Local Linear Embedding of Saul and Roweis (2004). See also Weinberger and Saul (2004), Donoho and Grimes (2003, 2005). Wang and Marron (2005) addressed the related problem of estimating the dimension of such a low dimensional manifold, using scale space ideas to tolerate a much higher level of noise than most other methods in this area. A fundamental difference between manifold learning, and the above described work is that in the latter, the manifold is fixed and known from the nature of the problem, while in the former, the goal is to find the manifold in the data.

While the above settings, featuring data objects lying in manifolds, present statistical challenges because of the non-Euclidean nature of the data space, there are two senses in which they are relatively mildly non-Euclidean. The first is that when the data are concentrated in a fairly small region, the manifold can be effectively approximated by a Euclidean space called the *tangent bundle*. The second is that Euclidean intuition can still be used via some fairly straightforward generalization, such as replacing lines by geodesics, and Euclidean distance by geodesic distance.

1.2. *OODA on Tree Spaces* A type of data space which is much farther from Euclidean in nature is the set of trees. A simple motivating example of trees as data is the case of multifigural objects, of the type shown in Figure 2. In that example, all three figures are present in every data object. But if some figures are missing, then the usual vector of m-rep parameters has missing values. Thus the natural data structure is trees, with *nodes* representing the figures. For each figure, the corresponding m-rep parameters appear as *attributes* of that node. A more complex and challenging example is the case of blood vessel trees, discussed in Section 2.

In most of the rest of this paper, the focus is on very challenging problem of OODA for data sets of tree-structured objects. Tree-structured data objects are mathematically represented as simple graphs (a collection of *nodes*, and *edges* each of which connects some pair of nodes). Simple graphs have a unique path (a set of edges) between every pair of nodes (vertices). A tree is a simple graph, where one node is designated as the *root node*, and all other nodes are *children* of a *parent* node that is closer to the root, where parents and children are connected by edges. In many applications, a tree-structured representation of each data object is very natural, including medical image analysis, phylogenetic studies, clustering analysis and some forms of classification (i.e. discrimination). Limited discussion, with references of these areas is given in Section 1.2.1. Our driving example, based on a data set of tree-structured blood vessel trees, is discussed in Section 2.

For a data set of tree-structured data objects, it is unclear how to develop notions such as *centerpoint* and *variation about the center*. Our initial ad hoc attempts at this were confounded by the fact that our usual intuitive ideas lead to contradictions. As noted above we believe this is because our intuition is based on Euclidean ideas, such as linear subspaces, projections, etc., while the space of trees is very "non-Euclidean" in nature, in the sense that natural definitions of the fundamental linear operators of addition and scalar multiplication operations do not seem to be available. Some additional mathematical basis for the claim of "non-Euclidean-ness of tree space", in the context of phylogenetic trees, can be found in Billera, Holmes and Vogtmann (2001). This failure of our intuition to give the needed insights, has motivated our development of the careful axiomatic mathematical theory for the statistical analysis of data sets of trees given in Section 3. Our approach essentially circumvents the need to define the linear operations that are the foundations of Euclidean space.

The development essentially starts from a Frechét approach, which is based on a metric. In general, we believe that different data types, such as those listed in Section 1.2.1, will require careful individual choice of a metric. In Sections 3.2 and 3.3, we define a new metric which makes sense for our driving problem of a data set of blood vessel trees.

Once a metric has been chosen, the Frechét mean of a data set is the point which minimizes the sum of the squared distances to the data points. A simple example is the conventional sample mean in Euclidean space (just the mean vector), which is the Frechét mean with respect to Euclidean distance. In Section 3.4, this idea is the starting point of our development of the notion of centerpoint for a sample of trees.

After an appropriate centerpoint is defined, it is of interest to quantify the variability of the sample about this center. Here, an analog of PCA, based on the notion of a *treeline* which plays the role of "one-dimensional subspace", is developed for tree space (see Section 3.5). A key theoretical contribution is a fundamental theory of *variation decomposition in tree space*, a tree version of the Pythagorean Theorem (see Section 3.5), which allows ANOVA style decomposition of sums of squares. In Section 3.6, an example is provided to highlight the difference between the tree version PCA and regular PCA.

The driving problem in this paper is the analysis of a sample of blood vessel trees, in the human brain, see Bullitt and Aylward (2002). We believe that similar methods could be used for related medical imaging problems, such as the study of samples of pulmonary airway systems, as studied in Tschirren, et al (2002). The blood vessel systems considered here are conveniently represented as trees. In our construction of these trees, each node represents a blood vessel, and the edges only illustrate the connectedness property between two blood vessels. For these blood vessel trees, both topological structure (i.e. connectivity properties) and geometric properties, such as the locations and orientations of the blood vessels, are very important. These geometric properties are summarized as the attributes of each node.

Focussing on our driving example of blood vessel trees, and their corresponding attributes, we develop a new metric δ on tree space, see Section 3.3. Margush (1982) gives a deeper discussion of metrics on trees. This metric δ consists of two parts: the integer part d_I, which captures the topological aspects of the tree structure (see Section 3.2 for more detail), and the fractional part f_δ, which captures characteristics of the nodal attributes (see Section 3.3).

The metric δ provides a foundation for defining the notion of centerpoint. A new centerpoint, the median-mean tree is introduced (see Section 3.4). It has properties similar to the median with respect to the integer part metric (see Section 3.2) and similar to the mean with respect to the fractional part metric (see Section 3.3).

In Section 3, methods are developed for the OODA of samples of trees. An interesting question for future research, is how our sample centerpoint and measures of variation about the center correspond to theoretical notions of these quantities, and an underlying probabilistic model for the population. For a promising approach to this problem, see Larget, Simon and Kadane (2002).

1.2.1. *Additional applications of OODA for trees.* Our driving application of OODA for tree structured data objects, to analyze data set of blood vessel trees, is discussed in Section 2. A number of additional important potential applications, which have not been tried yet, are discussed here.

In phylogenetic studies [see, e.g., Holmes (1999) and Li, et al (2000)], biologists build phylogenetic trees to illustrate the evolutionary relations among a group of organisms. Each node represents a taxonomic unit, such as a gene, or such as an individual represented by part of its genome, etc. The branching pattern (topology) represents the relationships between the taxonomic units. The lengths of the branches have meanings, such as the evolutionary time. An interesting metric in this context is the *triples distance*, developed by Critchlow, Li, Nourijelyani and Pearl (2000).

In cluster analysis [see Everitt, et al (2001)], a common practice is to obtain different cluster trees by using different algorithms, or by "bagging" or related methods [see Breiman (1996)], and then seek to do inference on the "central" tree. For cluster trees, the terminal nodes (external nodes, i.e., nodes at the tip of the tree) indicate the objects to be grouped; while the interior nodes indicate deeper level groupings, and the length of the paths indicate how well groups are clustered.

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

2. Tree OODA of a blood vessel data set. In this section, advanced statistical analysis, including centerpoint and variation about the center, of a data set of tree-structured objects, is motivated and demonstrated in the context of human brain blood vessel trees.

An example of arterial brain blood vessels from one person, provided by E. Bullitt, is shown in Figure 3. Because of the branching nature of blood vessel systems, a tree-structured data representation is very natural. See Bullitt and Aylward (2002) for detailed discussion of the data, and the method that was used to extract trees of blood vessel systems from Magnetic Resonance Images. The blood vessel systems considered here have three important components: the left carotid, the right carotid and the vertebrobasilar systems, shown in different colors in Figure 3. Each component consists of one root vessel and many offspring branches (vessels). Each branch is represented as a node in the tree structure. The attributes for each node include both information about that vessel, and also tree connectivity information. The individual information about that branch is coded as a sequence of vessel medial points (essentially a discretization of the medial axis of the blood vessel), where each point has a 3d location and a radius (of the vessel at the point). The connectivity information for each branch (node) records an index of its parent, and also the location of attachment to the parent. All of these attributes are used in the visual rendering shown in Figure 3.

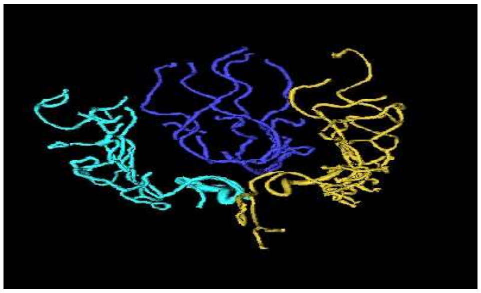

Fig. 3. *The three component blood vessel trees, shown as different colors, from one person. This detailed graphical illustration uses all attributes from the raw data.*

The full data set analyzed here has 11 trees from 3 people. These are the left carotid, right carotid and vertebrobasilar systems from each person, plus two smaller, unattached, components from one of the three people.

For simplicity of analysis, in this paper, we will work with only a much smaller set of attributes, based on a simple linear approximation of each branch. In particular, the attributes of the root node are the 3d locations of the starting and ending medial points. The attributes of the other branches include the index of the parent, together with a connectivity parameter indicating location of the starting point on the linear approximation of the parent, as

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

and the 3d locations of the ending point. An additional simplification is that radial information is ignored.

For computational speed, only a subtree (up to three levels and three nodes) of each element among those 11 trees is considered. There are only two different tree structures in this data set, which are called Type I and Type II, shown in Figure 4. Among these 11 blood vessel trees, seven trees have Type I structure and four trees have Type II structure.

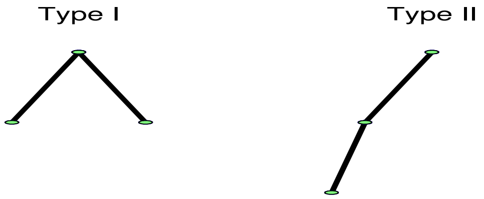

Fig. 4. *Two types of three-node tree structures, that are present in our sample, of simplified blood vessel trees, where 7 are of Type I, and 4 are of Type II.*

Each panel of Figure 5 shows the individual component trees for one person. The three dimensional aspect of these plots is most clearly visible in rotating views, which are internet available from the links "first person", "second person" and "third person" on the web site Wang (2004). These components are shown as thin line trees, which represent each raw data point. Trees are shown using the simplified rendering, based on only the linear approximation attributes, as described above. The root node of each tree is indicated with a solid line type, while the children are dashed. We will first treat each person's component trees as a separate subsample. Each panel of Figure 5 also includes the new notion of *centerpoint* (for that subsample), shown using a thicker line type. This is the *median-mean tree*, as developed in Section 3.4. This tree is central in terms of structure, size, and location, in senses which will be defined there.

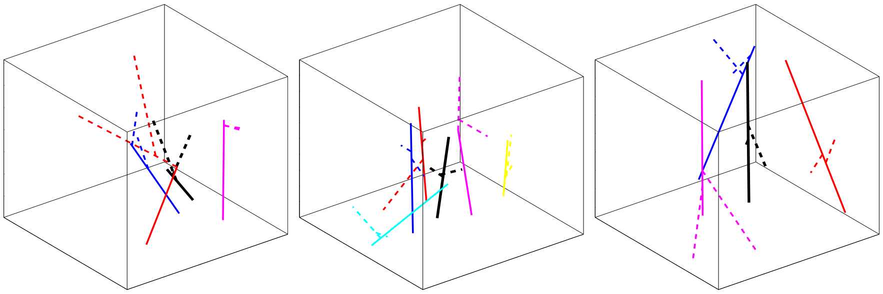

Fig. 5. *Simplified blood vessel trees (thin colored lines), for each person individually, with the individual median-mean trees (thicker black line). Root nodes use solid line types and children are dashed.*

These trees are combined into a single, larger sample in Figure 6. Again a rotating three dimensional view is available at the link "sample" on Wang (2004). This combined sample gives effective illustration of our statistical methodologies. Again, the median-mean tree of the larger sample is shown with a thick black line. This time the median-mean tree is surprisingly small, especially in comparison to the median-mean trees for individual people, shown in Figure 5. This will be explained through a careful analysis of the variation about the median-mean tree.

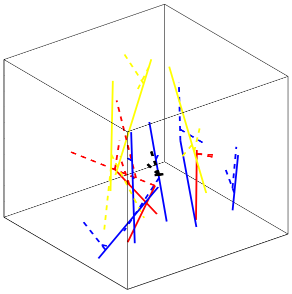

Fig. 6. *Combined sample of simplified blood vessel trees (thin line types) and the median-mean tree (thick line type). The median-mean tree, in the center, has a very short, nearly horizontal root node, and very short branches. The contrast of this, with the individual median-mean trees, shown in the previous figure, will be explained through the analysis of variation.*

Another important contribution of this paper is the development of an approach to analyzing the variation within a sample of trees. In conventional multivariate analysis, a simple first order linear approach to this problem is Principal Component Analysis. We develop an analog, for samples of trees in Section 3.5. Our first approach is illustrated in Figure 7, with an analysis of the dominant mode of tree structure variation, for the full blood vessel tree sample shown in Figure 6.

The generalization of PCA, to samples of tree-structured objects could be approached in many ways, because PCA can be thought of in a number of different ways. After considering many approaches, we found a suggestion by J. O. Ramsay to be the most natural. The fundamental idea is to view PCA as a sequence of one-dimensional representations of the data. Hence, our tree version PCA is based on notions of *one-dimensional representation* of the data set. These notions are carefully developed and precisely defined in Section 3.5. The foundation of this approach is the concept of *treeline*, which plays the role of line (a one-dimensional subspace in Euclidean space) in tree space. Two different types of treelines are developed in Section 3.5. The *structure treeline* which quantifies sample variation in tree structure, is formally defined in Definition 3.1 and is illustrated here in Figures 7 and 8. The *attribute treeline* describes variation within a fixed type of tree structure, is defined in Definition 3.2, and is illustrated here in Figure 9.

The structure treeline which best represents the data set (this will be formally defined in Section 3.5, but for now think in analogy to PCA), is called the *principal structure treeline*. The principal structure treeline for the full simplified blood vessel data is shown in Figure 7 (structure only, without attributes) and Figure 8 (with attributes). In Figure 7, this treeline starts with the tree u_0, which has two nodes. The other trees in this direction are u_1 and u_2, which consecutively add one left child. Generally structure treelines follow the pattern of successively adding single child nodes. This principal structure treeline is chosen, among all treelines that pass through the median-mean tree, to explain as much of the structure in the data as possible (in a sense defined formally in Section 3.5). Hence, this highlights structure variation in this sample, by showing that the dominant component of topological structure variation in the data set is towards branching in the direction of addition of left hand children nodes. Next, we also study how the attributes change as we move along this principal structure treeline, in Figure 8. The three panels show the simplified tree rendering of the trees whose structure is illustrated in Figure 7, with the first treeline member u_0 shown in the left box, the three node tree u_1, which is the median-mean tree, in the center box, and u_2 with four nodes in the right hand box.

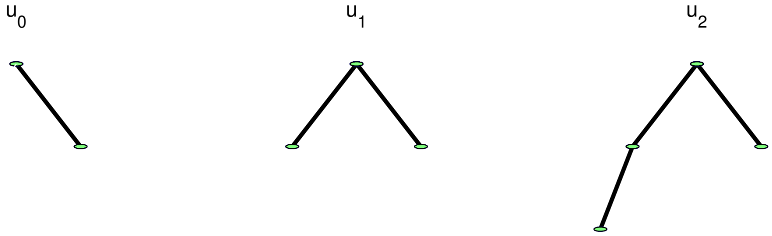

Fig. 7. *The principal structure treeline for the full simplified blood vessel data, without nodal attributes. Shows that the dominant sample variation in structure is towards the addition of left hand children nodes.*

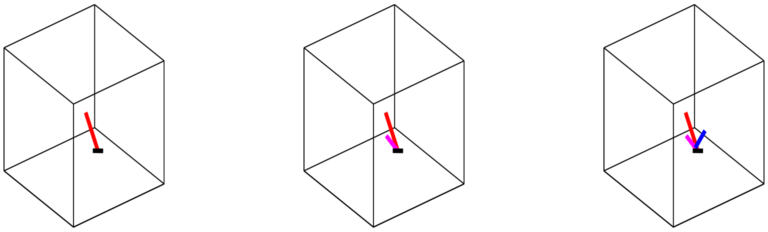

Fig. 8. *The principal structure treeline, with nodal attributes. This shows more about the sample variation, than is available from mere structure information.*

In addition to the principal structure representation, another useful view of the data comes from the principal attribute directions (developed in Section 3.5). *Principal attribute treelines* have a fixed tree structure, and highlight important sample variation within the given tree structure. Since the tree structure is fixed, this treeline is quite similar to the conventional first principal component, within that structure. Here we illustrate this idea, showing the principal attribute treeline which passes through (and thus has the same structure as) the median-mean tree, shown in Figure 9. There are six subplots in this figure. The subplots depict a succession of locations on the attribute treeline, which highlights the sample variation in this treeline direction. These are snapshots which are extracted from a movie version that provides clear visual interpretation of this treeline, and is internet available from the link "median-mean tree" from Wang (2004). A similar movie, showing a different principal attribute treeline can be found at the link "support tree" (this concept is explained in Section 3.1) on Wang (2004).

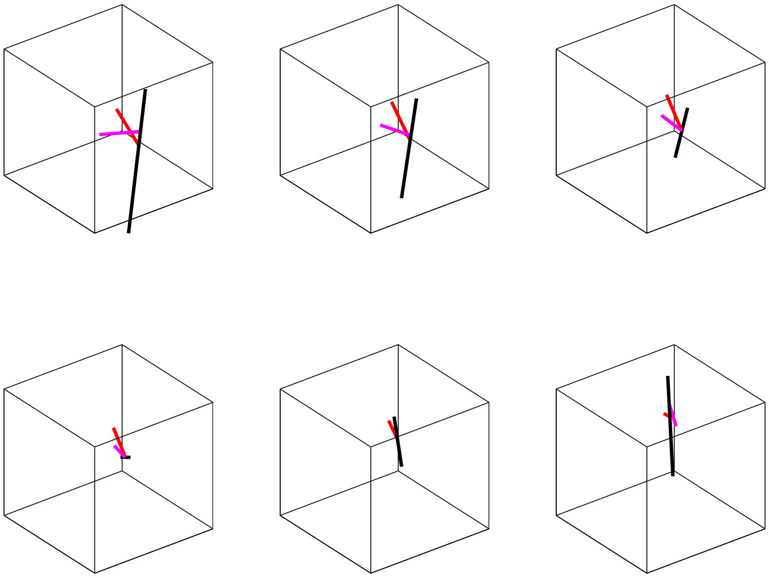

Fig. 9. *The principal attribute treeline passing through the median-mean tree. These are snapshots from a movie which highlights this mode of variation in the sample. The thick black root node flips over.*

In general, Figure 9 shows marked change in the length and orientation of the main root (solid black line). It starts (upper left) as a long nearly vertical segment, which becomes shorter, and moves towards horizontal (upper right). This trend continues in the lower left box, where the root is very short indeed, and is horizontal. In the next plots (lower row) the root begins to grow, this time in the opposite direction. In particular, the root node flips over, with the top and bottom ends trading places. While these trends are visible here, the impression is much clearer in the movie version. The branches also change in a way that shows smaller scale variation in the data. This was a surprising feature of the sample. Careful investigation showed that the given data sets did not all correctly follow the protocol of choosing the coordinate system according to the direction of blood flow. Some of them have the same direction; while, some of them have the inverse direction. A way of highlighting the two different data types is via the *projections* (the direct analogs of the principal component coefficients in PCA) of the 11 trees on this attribute treeline, as shown in Figure 10, and which is formally defined in Section 3.5. Figure

10 is a jitter plot, see Tukey and Tukey (1990), where the projections are shown on the horizontal axis, and a random vertical coordinate is used for visual separation of the points. This shows that there are two distinct groups with a clear gap in the middle, six trees with negative projection coefficients and five with positive ones. This also shows that no trees correspond to the fourth frame in Figure 9, with a very short root, which can also be seen in the raw data in Figure 6. This shows that the surprisingly short root node, for the median-mean tree, resulted from its being central to the sample formed by these two rather different subgroups, that were formed by different orientations of the blood flow in the data set. This dominates the total variation, perhaps obscuring population features of more biological interest.

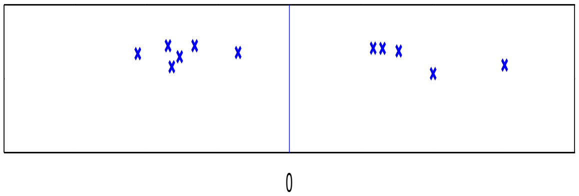

Fig. 10. *Projection coefficients, shown as the horizontal axis, of 11 trees on the principal attribute treeline passing through the median-mean tree. A random height is shown as the vertical axis for better separation of the points. This shows two distinct subgroups, which explains the very short median-mean root node.*

3. Development of the Tree OODA methodology.

In this section, a rigorous mathematical foundation is developed for the OODA of a data set of trees. We will use $\mathcal{S} = \{t_1, t_2, \ldots, t_n\}$ to denote the data set of size n. Careful mathematics are needed because the non-Euclidean nature of tree space means that many classical notions do not carry over as expected. For simplicity, only the case of *binary* trees with finite level, is explicitly studied. A binary tree is a tree such that every node has at most two children (left child and right child). If a node has only one child, it should be designated as one of left and right. In our blood vessel application, we consistently label each single child as left. The set of all binary trees, the *binary tree space*, is denoted by $\mathcal{T}$.

3.1. *Notation and preliminaries.*

This section introduces a labelling system for the nodes of each tree in the sample, i.e. each $t \in \mathcal{S}$. Each tree has a designated node called the *root*. An important indicator of node location in the tree is the *level of the node*, which is the length (number of edges) of the path to the root. In addition, it is convenient to uniquely label each node of a binary tree by a natural number, called the *level-order index*. The level-order index, of the node ω, is denoted by $ind(\omega)$, which is defined recursively as:

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

3. otherwise, if ω is the right child of the node ν, let $ind(\omega) = 2 \times ind(\nu) + 1$.

For a tree t, the set of level-order indices of the nodes is denoted by $IND(t)$. The set $IND(t)$ completely characterizes the topological structure of t, and will be a very useful device for proving theorems about this structure.

An important relationship between trees is the notion of a *subtree*, which is an analog of the idea of subset. A tree s is called a *topological subtree* of a tree t when every node in s is also in t, i.e. $IND(s) \subseteq IND(t)$. Moreover, if for every node $k \in IND(s)$, the two trees also have the same nodal attributes, then s is called an *attribute subtree* of t.

Also useful will be a set operations, such as union and intersection, on the topological binary tree space (i.e. when only structure is considered). For two binary trees t_1 and t_2, the tree t is the *union* (*intersection*) *tree* if $IND(t) = IND(t_1) \cup IND(t_2)$ $(IND(t) = IND(t_1) \cap IND(t_2)$, respectively). A horizon for our statistical analysis is provided by the union of all trees in the sample, which is called the *support tree*. This allows simplification of our analysis, because we only need to consider topological subtrees of the support tree.

The set of all topological subtrees, of a given tree t, is called a *subtree class*, and denoted $\mathcal{T}_t$. The terminology "class" is used because each $\mathcal{T}_t$ is closed under union and intersection.

As noted in Section 1, the first major goal of statistical analysis of samples of tree-structured objects is careful definition of a centerpoint of the data set. For classical multivariate data, there are many notions of centerpoint, and even the simple concept of sample mean can be characterized in many ways. After careful extensive investigation, we have found that approaches related to the *Frechét Mean* seem most natural. This characterizes the centerpoint as the binary tree which is the closest to all other trees, in some sense (sum of squared Euclidean distances gives the sample mean in multivariate analysis). This requires a metric on the space of binary trees. Thus, the second fundamental issue is the definition of a distance between two trees. This will be developed first for the case of topology only, i.e. without nodal attributes, in Section 3.2. In Section 3.3, this metric will be extended to properly incorporate attributes.

3.2. *Metric on the binary tree space without nodal attributes.* Given a tree t, its topological structure is represented by its set of level-order indices $IND(t)$. Two trees have similar (different) topologies, when their level-order index sets are similar (different, respectively). Hence, the non-common level-order indices give an indication of the differences between two trees. Thus, for any two topological binary trees s and t, define the metric

$$d_I(s,t) = \sum_{k=1}^{\infty} 1\{k \in IND(s) \triangle IND(t)\}, \tag{3.1}$$

where $\triangle$ is used to denote the symmetric set difference $(A \triangle B = \big(A \cap \overline{B}\big) \cup \big(\overline{A} \cap B\big)$, where $\overline{A}$ is the complement of A). Note that $d_I(s,t)$ counts the total number of nodes which show up only in either s or t, but not both of them. Another useful view is that this metric is the smallest number of addition and deletion of nodes required to change the tree s into t. Since d_I is always an integer, it is called the *integer tree metric*, hence the subscript of I. This will be extended to trees with attributes, in Section 3.3, by adding a fractional part to this metric.

This metric can also be viewed in another way. Each binary tree can be represented as a binary string using 1 for an existent node and 0 otherwise. Since the metric d_I counts differences between strings of 0s and 1s, it is just the Hamming distance from coding theory.

3.3. *Metric on the binary tree space with nodal attributes.* The integer tree metric d_I captures topological structure of the tree population. In many important cases, including image analysis, the nodes of the trees contain useful attributes (numerical values, see Section 1), which also characterize important features of data objects.

The attributes, contained in the node with level-order index k on the tree t, are denoted by (x_{tk}, y_{tk}), where for simplicity, only the case of two attributes per node is treated explicitly here. For each node, indexed by k, the sample mean attribute vector, $\sum_{t\in\mathcal{S}}(x_{tk}, y_{tk})/\sum_{t\in\mathcal{S}} 1\{k \in IND(t)\}$, can be assumed to be zero in the theoretical development, by subtracting the sample mean from the corresponding attribute vector of every tree which has the node k. Moreover, the upper bound of the absolute values of the attributes, $|x_{tk}|$ and $|y_{tk}|$, can be chosen as $\frac{\sqrt{2}}{4}$. Given any sample $\mathcal{S}$, this assumption can always be satisfied by multiplying each attribute by the scale factors $\frac{\sqrt{2}}{4}\left(\max_{t\in\mathcal{S}} |x_{tk}|\, 1\left\{k \in IND(t)\right\}\right)^{-1}$ and $\frac{\sqrt{2}}{4}\left(\max_{t\in\mathcal{S}} |y_{tk}|\, 1\left\{k \in IND(t)\right\}\right)^{-1}$. This can induce some bias in our statistical analysis, which can be partly controlled through careful choice of weights as discussed below, or by appropriate transformation of the attribute values. But this assumption is important to control the magnitude of the attribute component of the metric, with respect to the topological component. The bound $\frac{\sqrt{2}}{4}$ is used because the Euclidean distance between two-dimensional vectors, whose entries satisfy this bound, is at most 1. For the general nodal attribute vector (e.g., the nodal attribute vectors of the blood vessel trees), a different bound will be chosen to make the attribute difference (between two trees) less than 1.

For any trees s and t with nodal attributes, define the new metric (Theorem 3.1 establishes that this is indeed a metric)

$$\delta(s,t) = d_I(s,t) + f_\delta(s,t), \tag{3.2}$$

where

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

and where $\{\alpha_k\}_{k=1}^{\infty}$ is a non-negative weight series with $\sum_{k=1}^{\infty} \alpha_k = 1$. These weights are included to allow user intervention on the importance of various nodes in the analysis (for example, in some cases, it is desirable for the root node to dominate the analysis, in which case α_1 is taken to be relatively large). When there is no obvious choice of weights, equal weighting, $\alpha_k = \frac{1}{\#(\text{nodes appearing in the sample})}$ for nodes k that appear in the sample, and $\alpha_k = 0$ otherwise, may be appropriate. All the theorems in this paper are developed for general weight sequences. But, in Section 3.6, we consider some toy examples, based on the *exponential weight* sequence, which gives the same weight to nodes within a level, and uses an exponentially decreasing sequence across levels. In particular, the weight

$$\alpha_k = \{2^{-(2i+1)}\}, \text{ where } i = \lfloor \log_2 k \rfloor , \tag{3.4}$$

(where $\lfloor \cdot \rfloor$ denotes the greatest integer function) is used for each node on the i^{th} level, $i = 0, 1, 2, \ldots$. In the analysis of the blood vessel data, different normalization of the attributes is required, because there are as many as six attributes per node. The data analyzed in Section 2, was first recentered to have 0 mean, and rescaled so that the absolute value of the attributes was bounded by $\frac{1}{2\sqrt{7}}$. To more closely correspond to the original data, all of the displays in Section 2 are shown on the original scale.

The last two summations in Equation (3.3) are included to avoid loss of information from those nodal attributes that are in one tree and not the other. This formulation, plus our assumption on the attributes ensures that the second term in Equation (3.2), f_δ (where "f" means fractional part of the metric), is at most 1.

Also, note that f_δ is a square root of a weighted sum of squares. When trees s and t have the same tree structure, $f_\delta(s,t)$ can be viewed as a weighted Euclidean distance. In particular, the nodal attributes of a tree t can be combined into a single long vector called the *attribute vector*, denoted $\overrightarrow{v}$, for conventional statistical analysis. For an attribute subtree of t, the collection of attributes of the nodes of this subtree are a subvector of $\overrightarrow{v}$ which is called the *attribute subvector*.

When trees s and t have different tree structures, it is convenient to replace the non-existent nodal attributes with $(0,0)$. This also allows the nodal attributes to be combined into a single long vector, $\overrightarrow{v}$. Then, $f_\delta(s,t)$ is a weighted Euclidean metric on these vectors.

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

Next, Theorem 3.1 shows that δ is a metric. This requires the following assumption.

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

Theorem 3.1. *Under Assumption 1, δ is a metric on the tree space with nodal attributes.*

A sketch of the proof of Theorem 3.1 is in Section 4. The full proof is in the dissertation Wang (2003), the proof of Theorem 3.1.2 in Section 3.1.

Remark 1. *To understand why the Assumption 1 is critical to Theorem 3.1, consider the integer part d_I. While d_I is a metric on topological tree space, it is not a metric on the binary tree space with nodal attributes. In particular, for any two binary trees s and t with the same topological structure, $d_I(s,t)$ is always equal to zero regardless of their attribute difference. Thus, d_I is only a pseudo-metric on the tree space with nodal attributes. The Assumption 1 ensures that δ is a metric, not just a pseudo-metric.*

3.4. *Central tree.* In the Euclidean space $\mathbb{R}^1$, for a given data set of size n, there are two often-used measurements of the centerpoint, the sample mean and the sample median. Non-uniqueness for the median arises when n is an even number. In this section, the concepts of the sample median and the sample mean will be extended to the binary tree spaces, both with and without nodal attributes.

First, the case with no nodal attributes, i. e. only topological structure, is considered. A sensible notion of centerpoint is the *median tree*, which is defined as the minimizing tree, $\arg\min_t \sum_{i=1}^n d_I(t,t_i)$, taken over all trees t.

This is a modification of the Frechét mean, $\arg\min_t \sum_{i=1}^n d_I(t,t_i)^2$, which is used because it allows straightforward fast computation. This can be done using the characterization of the minimizing tree that is given in Theorem 3.2.

Theorem 3.2. *If a tree s is a minimizing tree according to the metric d_I, then all the nodes of tree s must appear at least $\frac{n}{2}$ times in the binary tree sample $\mathcal{S}$. Moreover, the minimizing 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.*

The proof is given in Section 4.

Non-uniqueness may arise when the sample size is an even number. The *minimal median tree*, which has the fewest nodes among all the median trees, is recommended as a device for breaking any ties.

Banks and Constantine (1998) independently developed essentially the same notion of central tree, and this characterization of the minimizing tree, which is called the *majority rule*. We use this same terminology.

Next the case of nodal attributes is considered. Our proposed notion of centerpoint in this case is called the *median-mean tree*. It has properties similar to the sample median with respect to d_I and similar to the sample mean with respect to f_δ. Its tree structure complies with the majority rule and its nodal attributes can be calculated as the sample mean $\sum_{t\in\mathcal{S}}(x_{tk},y_{tk})/\sum_{t\in\mathcal{S}} 1\{k \in IND(t)\}$. As for the median tree, the median-mean tree may not be unique, and again the minimal median-mean tree (with minimal number of nodes) is suggested for breaking such ties.

The median-mean tree is not always the same as the Frechét mean,

$$\arg\min_t \sum_{s\in\mathcal{S}} \delta\left(t,s\right)^2 .$$

We recommend the median-mean tree because it is much faster to compute. The median-mean tree is also most natural as the centerpoint of the Pythagorean Theorem (i. e. Sums of Squares analysis) developed in Section 3.5.

Another useful concept is the *average support tree*, which consists of all the nodes that appear in the tree sample with nodal attributes calculated as averages, as done in the median-mean tree. Thus the median-mean tree is an attribute subtree of the average support tree.

3.5. *Variation analysis in the binary tree space with nodal attributes.* Now that the central tree has been developed, the next question is how to quantify the variation of the sample about the centerpoint, i.e. about the median-mean tree.

In Euclidean space, the classical analysis of variance approach, based on decomposing sums of squares, provides a particularly appealing approach to quantifying variation. This analysis has an elegant geometric representation via the Pythagorean Theorem.

After a number of trials, we found that the most natural and computable analog of the classical ANOVA decomposition, came from generalizing the usual squared Euclidean norm to the *variation function*:

$$V_\delta(s,t) = d_I(s,t) + f_\delta^2(s,t). \tag{3.5}$$

Note that if every tree has the same structure, then this reduces to classical sums of squares, and the median-mean tree is the Frechét mean, with respect to the variation $V_\delta(s,t) = f_\delta^2(s,t)$, in the sense that it is the minimizing tree, over t, of $\sum_{s\in\mathcal{S}} V_\delta(s,t)$. This is also true in the case of tree samples that are purely topological, i.e. that have no attributes, when the variation becomes $V_\delta(s,t) = d_I(s,t)$. Then d_I is a full metric (not just a pseudo-metric), which can be written as a sum of zeros and ones (see Equation (3.1)). So the metric d_I can be interpreted as a sum of squares, because

$$\sum_{k=1}^{\infty}(1\{k \in IND(s)\triangle IND(t)\})^2 = \sum_{k=1}^{\infty} 1\{k \in IND(s)\triangle IND(t)\} = d_I(s,t). \tag{3.6}$$

In Euclidean space, the total variation of a sample can be measured by the sum of squared distances to its sample mean. For the tree sample $\mathcal{S}$ and the median-mean tree m_δ, the total variation about the median-mean is defined as

$$\sum_{s\in\mathcal{S}} V_\delta(s,m_\delta) = \sum_{s\in\mathcal{S}} d_I(s,m_\delta) + \sum_{s\in\mathcal{S}} f_\delta^2(s,m_\delta).$$

This total variation about the median-mean tree does not depend on how the tie is broken between the median-mean trees (when it is not unique).

In classical statistics, PCA is a useful tool to capture the features of a data set by decomposing the total variation about the centerpoint. In PCA, the first principal component eigenvector indicates the direction in which the data vary the most. Furthermore, other eigenvectors maximize variation in successive orthogonal residual spaces.

In binary tree space, each tree in the sample is considered to be a data point. Unlike Euclidean space, binary tree space is a nonlinear space according to the metric δ defined at (3.2). As noted above, because the space is nonlinear, the generalization of PCA is not straightforward. The foundation of our analog of PCA, is a notion of one-dimensional manifold in binary tree space, which is a set of trees that plays the role of a "line" (a one-dimensional subspace in Euclidean space). There are two important types, defined below, both of which are called *treeline*.

DEFINITION 3.1. *Suppose $l = \{u_0, u_1, u_2, \ldots, u_m\}$ is a set of trees with (or without) nodal attributes in the subtree class $\mathcal{T}_t$, of a given tree t. The set l is called a **structure treeline** (s-treeline) starting from u_0 if for $i = 1, 2, \ldots, m$,*

1. *u_i can be obtained by adding a single node (denoted by ν_i) to the tree u_{i-1} (thus, when attributes exist, they are common through the treeline);*
2. *The next node to be added, ν_{i+1} is the child of ν_i;*
3. *The first tree u_0 is minimal, in the sense that the ancestor node of ν_1 is the root node, or else has another child.*

REMARK 3.1. *Structure treelines are "one-dimensional" in the sense that they follow a single path, determined by u_0, and the sequence of added nodes $\nu_1, \ldots, \nu_m$. In this sense the elements of l are nested. Also when there are attributes, each attribute vector is the corresponding attribute subvector (defined in Section 3.3) of its successor.*

In Definition 3.1, the tree u_{i-1} is a subtree (an attribute subtree, if there are attributes) of the trees u_i, u_{i+1}, etc. Since every element in the s-treeline is a topological subtree of t, the length of the s-treeline cannot exceed the number of levels of the tree t. Illustration of the concept of s-treeline is shown in Figures 11 and 12.

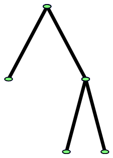

FIG. 11. *Toy example tree t, for illustrating the concept of s-treeline, shown in Figure 12.*

Figure 12 shows the s-treeline in $\mathcal{T}_t$, where t has the tree structure shown in Figure 11. Figure 12 indicates both tree topology, and also attributes. The positive attributes (x, y) are graphically illustrated with a box for each node, where x is shown as the horizontal length and y is the height.

Note that the attributes are common for each member of the treeline. Each succeeding member comes from adding a new node. The starting member, u_0, can not be reduced, because the root node has a child, which does not follow the needed sequence.

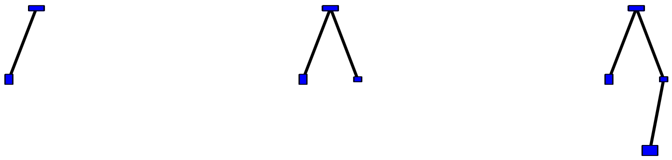

FIG. 12. *An s-treeline in $\mathcal{T}_t$, based on the tree t shown in Figure 11.*

A structure treeline l is said to *pass through* the tree u, when the tree u is an element of the tree set l, i.e., $u \in l$. Recall from Section 2 that, for the blood vessel data, Figure 7 shows the topology of the structure treeline passing through the median-mean tree, and Figure 8 shows the corresponding attributes. The central tree in each figure is the median-mean tree.

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

DEFINITION 3.2. *Suppose $l = \{u_\lambda : \lambda \in \mathbb{R}\}$ is a set of trees with nodal attributes in the subtree class $\mathcal{T}_t$, of a given tree t. The set l is called an **attribute treeline** (a -treeline) **passing through** a tree u^* if*

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

REMARK 3.2. *An a-treeline is determined by the tree u^* and the vector $\vec{v}$. The treeline is "one-dimensional" in this sense, which is essentially the same as a line in Euclidean space.*

Figure 13 shows some members, of an a-treeline from the same subtree class $\mathcal{T}_t$ shown in Figure 11, with $\lambda = 0.5, 1.0, 1.2, 1.5$ and $\vec{v} = [0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.2, 0.2]'$.

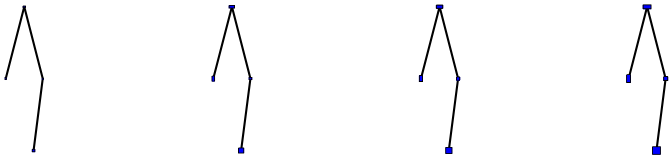

FIG. 13. *Toy example of an a-treeline for the same subtree class $\mathcal{T}_t$ as in Figure 11. Several members of the treeline are shown. The attributes are a linear function of each other.*

The topological structure of all of the trees in Figure 13 are the same. The dimensions of the boxes, illustrating the values of the attributes, change linearly.

In Section 2, Figure 9 illustrated an attribute treeline. That treeline highlighted the strong variation between orientations of the trees in the sample.

From now on, both s-treelines and a-treelines are called treelines. An analogy of the first principal component is the treeline which explains most of the variation in the data. A notion of *projection*, of a tree onto a treeline, needs to be defined, because this provides the basis for decomposition of sums of squares.

For any tree t and treeline l, the projection of t onto l, denoted $P_l(t)$, is the tree which minimizes the distance $\delta(t,\cdot)$ over all trees on the treeline l. The idea of projection is most useful, when it is unique, as shown in the next theorem.

THEOREM 3.3. *Under Assumption 1, the projection of a tree t onto a treeline l is unique.*

The proof is given in Section 4.

The Pythagorean Theorem is critical to the decomposition of the sums of squares in classical analysis of variance (ANOVA). Analogs of this are now developed for tree samples. Theorem 3.4 gives a Pythagorean Theorem for a-treelines and Theorem 3.5 gives a Pythagorean Theorem for s-treelines.

THEOREM 3.4. *(Tree version of the Pythagorean Theorem: Part I) Let l be an a-treeline passing through a tree u in the subtree class $\mathcal{T}_t$. Then, for any $t \in \mathcal{T}_t$,*

$$V_\delta(t,u) = V_\delta(t, P_l(t)) + V_\delta(P_l(t), u). \tag{3.7}$$

REMARK 3.3. *This states that the variation (our analog of squared distance) of a given tree t from a tree u in the treeline l, which is essentially the hypotenuse of our triangle, is the sum of the variation of t from $P_l(t)$, plus the variation of $P_l(t)$ from u, representing the legs of our triangle. This is the key to finding treelines that explain maximal variation in the data, because $V_\delta(t,u)$ is independent of l, so maximizing (over treelines l) a sample sum over $V_\delta(P_l(t),u)$ is equivalent to minimizing the residual sum over $V_\delta(t,P_l(t))$.*

In this paper, only those s-treelines, where every element is an attribute subtree of the average support tree (as defined in Section 3.4), are considered, because this gives a tree version of the Pythagorean Theorem, shown next.

THEOREM 3.5. *(Tree version of the Pythagorean Theorem: Part II) Let $\mathcal{S} = \{t_1, t_2, \ldots, t_n\}$ be a sample of trees. Let l be an s-treeline where every element is an attribute subtree of the average support tree of $\mathcal{S}$. Then, for any $u \in l$,*

$$\sum_{t\in\mathcal{S}} V_\delta(t,u) = \sum_{t\in\mathcal{S}} V_\delta(t, P_l(t)) + \sum_{t\in\mathcal{S}} V_\delta(P_l(t), u). \tag{3.8}$$

REMARK 3.4. *This theorem complements Theorem 3.4, because it now gives a structure treeline version of the Pythagorean Theorem, which simplifies analysis of variance, because minimizing the residual sum $\sum_{t\in\mathcal{S}} V_\delta(P_l(t),t)$ is equivalent to maximizing the sum $\sum_{t\in\mathcal{S}} V_\delta(\mu_\delta, P_l(t))$ over all treelines passing through the minimal* *median-mean tree, μ_δ. In some sense, this theorem is not so strong as Theorem 3.4, because the sample summation is needed, while the Pythagorean Theorem 3.4 is true even term by term.*

Sketches of the proofs of Theorems 3.4 and 3.5 are given in Section 4. Further details are in the proofs of Theorems 3.5.3 and 3.5.4, in Section 3.5 of Wang (2003).

The foundations are now in place to develop variation analysis in binary tree space. There are two main steps to the *PCA on trees* variation analysis.

First, find an s-treeline l_{PS} such that minimizes the sum $\sum_{t\in\mathcal{S}} V_\delta(t, P_l(t))$ over l passing through the minimal median-mean tree μ_δ of the sample $\mathcal{S}$, i.e.,

$$l_{PS} = \operatorname*{arg\,min}_{l:\mu_\delta\in l} \sum_{t\in\mathcal{S}} V_\delta(t, P_l(t)). \tag{3.9}$$

This structure treeline is called a *one-dimensional principal structure representation (treeline)* of the sample $\mathcal{S}$. Because of the Pythagorean Theorem 3.5, the one-dimensional structure treeline l_{PS}, explains a maximal amount of the variation in the data, as is done by the first principal component in Euclidean space. This is illustrated in the context of the blood vessel data in Section 2. Figure 8 shows the principal structure treeline $l_{PS} = \{u_0, u_1, u_2\}$ with nodal attributes, where u_1 is the unique median-mean tree (also the minimal median-mean tree) of the sample. Figure 7 shows the topological tree structures of the principal structure treeline in Figure 8.

Second, a notion of principal attribute treeline direction will be developed. This will complement the principal structure treeline, in the sense that together they determine an analog of a two dimensional subspace of binary tree space. Recall from Definition 3.2, that an attribute treeline, is indexed by a starting tree u^*, with attribute vector $\overrightarrow{v}^*$, and by a direction vector $\overrightarrow{v}$, and has general attribute vector $\overrightarrow{v}^* + \lambda\overrightarrow{v}$, for $\lambda \in \mathbb{R}$. To create the desired two dimensional structure, we consider a family of attribute treelines, indexed by the nested members $\{u_0, u_1, \dots u_m\}$ of the principal structure treeline and their corresponding nested (in the sense of attribute subvectors, as defined in Section 3.3) attribute vectors $\{\overrightarrow{v}^*_0, \overrightarrow{v}^*_1, \dots, \overrightarrow{v}^*_m\}$, and indexed by a set of nested direction vectors $\{\overrightarrow{v}_0, \overrightarrow{v}_1, \dots, \overrightarrow{v}_m\}$.

The union of treelines that are nested in this way is called a *family of attribute treelines*. This concept is developed in general in the following definition.

Definition 3.3. *Let $l = \{u_0, u_1, \dots u_m\}$ be a structure treeline, and let $\vec{c}$ be a vector of attributes, corresponding to the nodes of u_m. The* ***$l, \vec{c}$-induced family of attribute treelines****, $\mathcal{E}_{l,\vec{c}} = \{e_0, e_1 \dots, e_m\}$, is defined, for $k = 0, 1, \dots, m$, as*

$$e_k = \{t_\lambda : t_\lambda \text{ has attribute vector } \overrightarrow{v}_k + \lambda\overrightarrow{c}_k,\ \lambda \in \mathbb{R}\},$$

where $\overrightarrow{v}_k$ is the attribute vector of u_k, and where $\overrightarrow{c}_k$ is the corresponding attribute subvector of $\overrightarrow{c}$.

Next an appropriate family of attribute treelines is chosen to provide maximal approximation of the data (as is done by the first two principal components in Euclidean space). Following conventional PCA, we start with the principal structure treeline l_{PS}, (which we will denote in this paragraph as l simply to save a level of subscripting) and choose the direction vector $\vec{c}$ so that the $l, \vec{c}$-induced family of attribute treelines explains as much of the data as possible. In particular, we define the principal attribute direction vector, $\vec{c}_{PA}$ as

$$\vec{c}_{PA} = \arg\min_{\vec{c}} \sum_{t \in \mathcal{S}} V_\delta \left(t, P_{\mathcal{E}_{l,\vec{c}}}(t)\right),$$

where $P_{\mathcal{E}_{l,\vec{c}}}(t)$ is the projection of the tree t onto the $l, \vec{c}$-induced family of attribute treelines. This is an analog of a two dimensional projection, defined as

$$P_{\mathcal{E}_{l,\vec{c}}}(t) = \arg\min_{s : s \in e(t)} \delta(t, s),$$

where $e(t)$ is the attribute treeline determined by the tree $P_l(t)$ (the projection of t onto the principal structure treeline), and by the direction vector which is the corresponding attribute subvector of $\vec{c}$.

The elements of the $l_{PS}, \vec{c}_{PA}$-induced family of attribute treelines are all called *principal attribute treelines.* As the first two principal components can illuminate important aspects of the variation in data sets of curves, as demonstrated by e. g. Ramsay and Silverman (1997, 2002), the principal structure and attribute treelines can find important structure in a data sets of trees. An example of this, is shown in Figures 9 and 10, where the principal attribute treeline through the median-mean tree revealed the change in orientation among in the data.

For completely different extensions of PCA in nonlinear ways, see the principal curve idea of Hastie and Stuetzle (1989) and the principal geodesic analysis of Fletcher, et al (2003). Principal curves provide an interesting nonlinear decomposition of Euclidean space. Principal geodesics provide a decomposition of data in nonlinear manifolds, including Lie groups and symmetric spaces.

3.6. *Comparison of tree version PCA and regular PCA.* The tree version PCA is a generalization of the regular PCA to the case of trees as data points. 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 find a more appropriate family of attribute treelines. This idea will be illustrated using the following toy example. Here the metric δ is assumed to use exponential weights, as defined at (3.4).

EXAMPLE 3.1. *Let $\mathcal{S} = \{t_1, t_2, \ldots, t_{13}\}$ be a sample of trees with size $n = 13$. Let each member of $\mathcal{S}$ have one of the two structures shown in Figure 14. Let the attributes have the form shown in Table 1, where the x and y values are given in Table 2. In Table 2, the trees without node 3, have a $\star$ shown in the y entry. Thus, trees $t_1, t_2, \ldots, t_7$ have three nodes, while the others have two nodes.*

Conventional PCA, can be applied to this data, if they can be represented by vectors of equal length. A natural approach is to substitute the non-existent nodal attributes $\star$ by the sample average of the corresponding attributes.

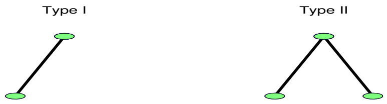

FIG. 14. *The two types of tree structures in the toy data set $\mathcal{S}$.*

TABLE 1
Form of the attributes of the trees in the toy data set $\mathcal{S}$.

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

TABLE 2
Specific values of x and y for each tree in $\mathcal{S}$.

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

The corresponding exponentially weighted attributes are

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

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

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

Thus, for analyzing variation in the sample, $x/\sqrt{8}$ and $y/\sqrt{8}$ are the two important components of the attribute vector. For simple visualization in the following, the principal components will be represented in two-dimensional space of x and y, instead of the full six-dimensional space.

The scatter plot of the attributes, x and y, is shown in Figure 15. It shows that, the attributes of the Type II trees (there are seven, shown with a "+") form a pattern from lower left to upper right. The attributes of the Type I trees (there are six, shown with an "×") have been divided into two groups with a gap in the middle.

Applying the regular PCA to the weighted attribute vectors, gives the first principal direction (first eigenvector), shown as the solid line in Figure 15. This shows that the trees with the Type I structure have a strong effect on the conventional PCA attribute direction, pulling it towards a horizontal line. This clearly is not an effective one dimensional representation of the data.

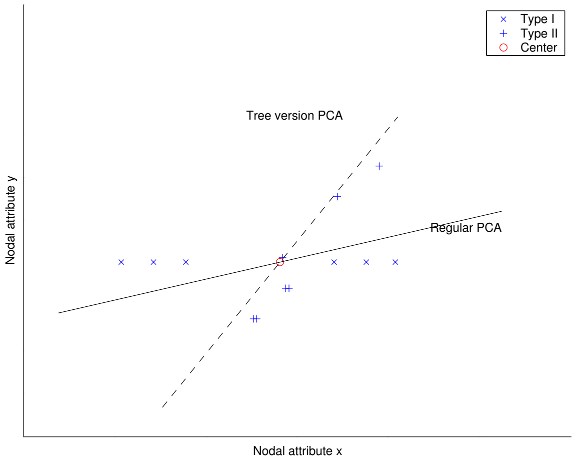

Fig. 15. *Scatter plot of the nodal attributes and principal attribute directions given by Regular PCA and Tree version PCA.*

Next, the tree version PCA is applied to the same toy tree sample $\mathcal{S}$. The tree version PCA has two steps, finding the principal structure treeline and finding the family of principal attribute treelines.

The first two elements (denoted as u_0 and u_1) on the principal structure treeline l_{PS} is shown in Figure 16. Note that u_1 is the median-mean tree of the sample $\mathcal{S}$. Moreover, the elements in $\mathcal{S}$ can be categorized by projections on this treeline. The trees with Type I structure have projection u_0 on the treeline l_{PS}; while, the trees with Type II structure have projection u_1 instead.

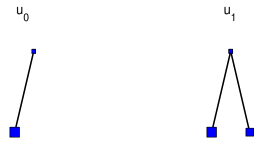

FIG. 16. *Principal structure treeline $l_{PS} = \{u_0, u_1\}$.*

Based on the principal structure treeline, the principal attribute direction is calculated and shown as the dashed line in Figure 15. Comparing with the direction given by regular PCA, it is more appropriate for the reason that it represents the relation of the (weighted) attributes. The Type I elements should not influence the direction because they contain no information about the relationship between the attributes x and y.

Next, the attributes of the six trees with Type I structure will be studied. All these six trees have a common projection on the principal structure treeline, u_0. The projection coefficients of these trees on the attributes treeline passing through u_0 are shown in Figure 17. Note that there is a big jump from the negative coefficients to the positive ones.

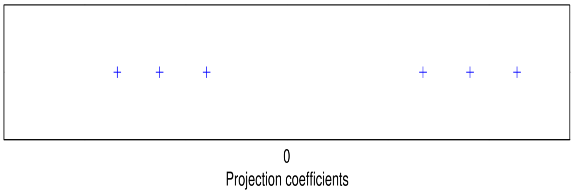

FIG. 17. *Projection coefficients of the trees with Type I structure on the principal attribute direction.*

Note that our tree version PCA only gives an analog of at most the first two principal components. An interesting problem for future research is to find tree analogs of additional components.

4. Derivations of theorems.

A SKETCH OF THE PROOF OF THEOREM 3.1. It follows from the fact that, d_I is a metric on the binary tree space without nodal attributes and f_δ is a weighted Euclidean distance between two attribute vectors (the nodal attributes for non-existent nodes are treated as zeros).

PROOF OF THEOREM 3.2. Let s be a minimizing tree according to the integer tree metric d_I. Suppose some of the nodes in s appear less than $\frac{n}{2}$ times and ν is the node with the largest level among all of those nodes. If a node appears less than $\frac{n}{2}$ times, so do its children. Thus, ν must be a terminal node of s.

For the binary tree s', with $IND(s') = IND(s) \backslash \{ind(\nu)\}$, 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 $\mathcal{S}$}. Since $n_\nu < \frac{n}{2}$,

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

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

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

Finally, a proof is given of the fact that the minimizing 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 $\mathcal{S}$ and $ind(\omega) \notin IND(s)$. Without loss of generality, suppose that ω is a child of some node in the binary tree s. Otherwise, choose one of its ancestor nodes.

For the binary tree s'', with $IND(s'') = IND(s) \cup \{ind(\omega)\}$, 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 $\mathcal{S}$}. Since $n_\omega > \frac{n}{2}$,

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

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

PROOF OF THEOREM 3.3. The proof will be provided for s-treelines and a-treelines separately.

Case 1: l is an s-treeline.

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

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

It will be shown that, for $i \neq p$, $d_I(u_i, t) > d_I(u_p, t)$. The proof will be provided for $i > p$ and $i < p$, respectively.

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

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

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

$$d_I(u_{p+1}, t) = d_I(u_p, t) + 1. \tag{4.3}$$

Iteratively, for $i = 0, \ldots, m - p - 1$, the tree u_{p+i} can be obtained by deleting a node ν_{p+i+1} from the tree u_{p+i+1}. The node ν_{p+i+1} is an offspring node of the node ν_{p+1}. Since $\nu_{p+1} \notin IND(t)$, for $i = 0, \ldots, m - p - 1$, $\nu_{p+i+1} \notin IND(t)$. Hence,

$$d_I(u_{p+i+1}, t) = d_I(u_{p+i}, t) + 1. \tag{4.4}$$

Next, for $p > 0$, consider the two trees u_{p-1} and u_p on the treeline l. The tree u_{p-1} can be obtained by deleting a node ν_p from the tree u_p. It will now be shown that, $\nu_p \in IND(t)$. Otherwise,

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

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

$$d_I(u_{p-1}, t) = d_I(u_p, t) + 1. \tag{4.5}$$

Iteratively, for $i = 0, 1, \ldots, p - 1$, the tree u_{p-i-1} can be obtained by deleting a node ν_{p-i} from the tree u_{p-i}. The node ν_{p-i} is an ancestor node of the node ν_p. Since $\nu_p \in IND(t)$, for $i = 0, 1, \ldots, p - 1$, $\nu_{p-i} \in IND(t)$. Thus,

$$d_I(u_{p-i-1}, t) = d_I(u_{p-i}, t) + 1. \tag{4.6}$$

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

$$d_I(u_i, t) > d_I(u_p, t). \tag{4.7}$$

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

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

Also, from Equation (4.7),

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

The proof will be finished by showing the following inequality

$$|f_\delta(u_i, t) - f_\delta(u_p, t)| < 1. \tag{4.8}$$

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

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

So, if equality holds, then

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

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

Since the fractional part metric is at most 1,

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

In fact, for any two trees on the s-treeline, one of the two trees is an attribute subtree of the other one. Without loss of generality, assume that the tree u_i is an attribute subtree of the tree u_p, and $IND(u_p)\backslash IND(u_i) = K$, where the set K is some proper subset of the positive integers.

Furthermore,

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

which is a contradiction.

Hence, the inequality (4.8) is satisfied. Thus, $\delta(u_i, t) - \delta(u_p, t) > 0$, i.e., u_p is the unique projection.

Case 2: l is an a-treeline.

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

Proof of Theorem 3.4. The projection tree $P_l(t)$ has the same tree structure as the tree u. Therefore, $d_I(P_l(t), u) = 0$ and $d_I(t, P_l(t)) = d_I(t, u)$.

Next, it needs to be shown that

$$f_\delta^2(t, u) = f_\delta^2(t, P_l(t)) + f_\delta^2(P_l(t), u) \tag{4.9}$$

for the a-treeline l.

Note that, for the nodes with level-order index $k \in IND(t)\backslash IND(u)$, the contribution of its nodal attributes to both sides of Equation (4.9) is the same. Thus, without loss of generality, assume that $IND(t) \subseteq IND(u)$. Its attribute vector has the same length as that of the tree u by adding zeroes on $IND(u)\backslash IND(t)$.

The metric δ is the same as the Euclidean distance of two weighted vectors. Thus, it is straightforward that Equation (4.9) follows from the ordinary Pythagorean Theorem.

In the following proof, the relationship *attribute subtree*, where the tree s is an attribute subtree of tree t, is denoted by $s \overset{A}{\subseteq} t$ or $t \overset{A}{\supseteq} s$.

A sketch of the proof of Theorem 3.5. By the definition of d_I and the fact that $P_l(t_i)$ is the member of the treeline l, which is δ-closest, and also d_I-closest, to the tree t_i, for any i,

$$d_I(t_i, u) = d_I(t_i, P_l(t_i)) + d_I(P_l(t_i), u). \tag{4.10}$$

It is necessary to prove that

$$\sum_{i=1}^{n} f_\delta^2(t_i, u) = \sum_{i=1}^{n} f_\delta^2(t_i, P_l(t_i)) + \sum_{i=1}^{n} f_\delta^2(P_l(t_i), u). \tag{4.11}$$

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

$$P_l(t_1) \overset{A}{\subseteq} u, \ldots, P_l(t_K) \overset{A}{\subseteq} u, P_l(t_{K+1}) \overset{A}{\supseteq} u, \ldots, P_l(t_n) \overset{A}{\supseteq} u \tag{4.12}$$

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

First, for $i = 1, 2, \ldots, K$, $P_l(t_i)$ is an attribute subtree of u. Suppose that t is a tree in the set $\{t_1, \ldots, t_K\}$, then $P_l(t) \overset{A}{\subseteq} u$. By the fact that the tree $P_l(t)$ is the projection of the tree t onto the treeline l, the following equality holds

$$IND(t) \cap IND(u) = IND(t) \cap IND(P_l(t)). \tag{4.13}$$

Recalling the fact that $P_l(t) \overset{A}{\subseteq} u$ and applying Equation (4.13), a straightforward calculation [details are given in Section 3.5 of Wang (2003)] shows

$$f_\delta^2(t, u) = f_\delta^2(t, P_l(t)) + f_\delta^2(P_l(t), u). \tag{4.14}$$

For $i > K$, $P_l(t_i) \overset{A}{\supseteq} u$. Again, by the definition of projection and a set theoretical calculation [details are given in Section 3.5 of Wang (2003)], the following equation holds

$$IND(P_l(t_i)) \cap \overline{IND(u)} \cap \overline{IND(t_i)} = \varnothing. \tag{4.15}$$

Thus, using the set relationship of the trees u, $P_l(t_i)$ and t_i and Equation (4.15), the following equations are established,

$$(IND(t_i) \backslash IND(P_l(t_i))) \cup (IND(P_l(t_i)) \backslash IND(u)) = IND(t_i) \backslash IND(u), \tag{4.16}$$

$$(IND(t_i) \backslash IND(P_l(t_i))) \cap (IND(P_l(t_i)) \backslash IND(u)) = \varnothing, \tag{4.17}$$

$$(IND(P_l(t_i)) \backslash IND(u)) \cup (IND(t_i) \cap IND(u)) = IND(t_i) \cap IND(P_l(t_i)), \tag{4.18}$$

and

$$IND(P_l(t_i)) \backslash IND(t) = IND(u) \backslash IND(t_i). \tag{4.19}$$

Calculations show that,

$$\sum_{i=K+1}^{n} f_\delta^2(t_i, u) = \sum_{i=K+1}^{n} f_\rho^2(t_i, P_l(t_i)) + \sum_{i=K+1}^{n} f_\delta^2(P_l(t_i), u). \tag{4.20}$$

Combining Equation (4.14), Equation (4.11) is established, which completes the proof of Theorem 3.5.

Acknowledgments. Much of this paper is the dissertation work of the first author, written under the supervision of the second. S. M. Pizer and the UNC MIDAG project provided financial support for the both authors, and many helpful conversations throughout this research. The term "Object Oriented Data Analysis" came from a conversation with J. O. Ramsay. E. Bullitt provided the blood vessel tree data (gathered with support from the grant R01 EB000219 NIH-NIBIB), and the viewer STree that was used to construct Figure 3. Paul Yushkevich provided the program mmview, that made Figure 1. Ja-Yeon Jeong gave us the program pablo_view_Residue, which was used to create Figure 2. The second author was also supported by NSF Grant DMS-0308331.

REFERENCES

Banks, D. and Constantine, G. M. (1998). Metric Models for Random Graphs. *J. Classification* **15** 199-223.

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

Breiman, L., Friedman, J. H., Olshen, J. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Breiman, L. (1996). Bagging Predictors. *Machine Learning* **24** 123-140.

Bullitt, E. and Aylward, S. (2002). Volume rendering of segmented image objects. *IEEE, Trans. Med. Imag.* **21** 998-1002.

Cootes, T. (2000). An introduction to active shape models. *Model Based Methods in Analysis of Biomedical Images*, (eds. R. Baldock and J. Graham), Oxford University Press, 223-248.

Cootes, T. F. and Taylor, C. (2001). Statistical models of appearance for medical image analysis and computer vision. *Proceedings of SPIE Medical Imaging.*

Critchlow, D. E., Li, S., Nourijelyani, K. and Pearl, D. K. (2000). Some Statistical Methods for Phyolgenetic Trees with Application to HIV Disease. *Mathematical and Computer Modelling* **32** 69-81.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of the Sciences USA*, 100(10), 5591–5596.

Donoho, D. L. and Grimes, C. (2005). Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision*, to appear.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis.* John Wiley and Sons.

Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis.* Oxford Univ. Press, New York.

Fisher, N. I. (1993). *Statistical analysis of circular data.* Cambridge University Press, New York.

Fisher, N. I., Lewis, T. and Emgleton, B. J. J. (1987). *Statistical analysis of spherical data*, Cambridge University Press, New York.

Fletcher, P. T., Lu, C. and Joshi, S. (2003). Statistics of Shape via Principal Geodesic Analysis on Lie Groups. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 95-101.

Fletcher, P.T., Joshi, S., Lu, C., Pizer, S.M. (2005). Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape. To appear *IEEE Transactions on Medical Imaging.*

Hastie, T. and Stuetzle, W. (1989). Principal Curves. *J. Amer. Statist. Assoc.* **84**, 502-516.

Holmes, S. (1999). Phylogenies: An Overview. *IMA series*, vol 112, on Statistics and Genetics, (ed. Halloran and Geisser), 81-119. Springer Verlag, New York.

Izem, R., Kingsolver, J. G. and Marron, J. S. (2005). Analysis of Nonlinear Variation in Functional Data, unpublished manuscript.

LARGET, B., SIMON, D. L. and KADANE, J. B. (2002). Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society, Series B*, 64, 681-693.

LI, S., PEARL, D. K. and DOSS, H. (2000). Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *J. Amer. Statist. Assoc.* **95** 493-508.

MARDIA, K. V. (1972). *Statistics of directional data.* Academic Press, New York.

MARDIA, K. V. (2000). *Directional statistics.* Wiley, New York.

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G. , TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust Principal Component Analysis for Functional Data. *Test*, **8** 1-73.

MARGUSH, T. (1982). Distances Between Trees. *Discrete Appl. Math.* **4** 281-290.

PIZER, S. M., THALL, A. and CHEN, D. (1999). M-Reps: A New Object Representation for Graphics. Submitted to ACM TOG. (See http://midag.cs.unc.edu/pubs/papers/mreps-2000/mrep-pizer.PDF.)

RAMSAY, J. O. and SILVERMAN, B. W. (1997). *Functional Data Analysis.* Springer Verlag, New York.

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis.* Springer Verlag, New York.

SAUL, L. K. and ROWEIS, S. T. (2004). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4, 119–155.

SZEKELY, G., KELEMEN, A., Brechbuhler, C. and Gerig, G. (1996). Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models. *Medical Image Analysis*, 1, 19-34.

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2319-2322.

TSCHIRREN, J., PALÁGYI, K., REINHARDT, J. M., HOFFMAN, E. A. and SONKA, M. (2002). Segmentation, Skeletonization, and Branchpoint Matching — A Fully Automated Quantitative Evaluation of Human Intrathoracic Airway Trees. *Proc. 5th Int. Conf. Medical Image Computing and Computer-Assisted Intervention, MICCAI*, Part II 12-19, Tokyo, Japan.

TUKEY, J., and TUKEY, P. (1990). Strips Displaying Empirical Distributions: Textured Dot Strips. Bellcore Technical Memorandum.

WANG, H. (2003). Functional Data Analysis of Populations of Tree-structured Objects. Ph. D. Dissertation, Dept. Statistics, Univ. of North Carolina at Chapel Hill.

WANG, H. (2004). Internet Site: http://www.stat.colostate.edu/˜wanghn/tree.htm.

WANG, X. A. and MARRON, J. S. (2005). A Scale-Based Approach to Finding Effective Dimensionality, unpublished manuscript.

WEINBERGER, K. Q. and SAUL, L. K. (2004). Unsupervised learning of image manifolds by semidefinite programming. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-04)*, vol. 2, pp. 988–995.

YUSHKEVICH, P., PIZER, S. M., JOSHI, S. and MARRON, J. S. (2001). Intuitive, Localized Analysis of Shape Variablity. *Information Processing in Medical Imaging (IPMI)* 402-408.

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523
E-MAIL: wanghn@stat.colostate.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599
E-MAIL: marron@email.unc.edu