

A PRINCIPAL COMPONENT ANALYSIS FOR TREES

BY BURCU AYDIN^{*}, GÁBOR PATAKI, HAONAN WANG[†], ELIZABETH
BULLITT[‡] AND J.S. MARRON[§]

University of North Carolina and Colorado State University

The active field of Functional Data Analysis (about understanding the variation in a set of curves) has been recently extended to Object Oriented Data Analysis, which considers populations of more general objects. A particularly challenging extension of this set of ideas is to populations of tree-structured objects. We develop an analog of Principal Component Analysis for trees, based on the notion of tree-lines, and propose numerically fast (linear time) algorithms to solve the resulting problems to proven optimality. The solutions we obtain are used in the analysis of a data set of 73 individuals, where each data object is a tree of blood vessels in one person's brain. Our analysis revealed a significant relation between the age of the individuals and their brain vessel structure.

1. Introduction. Functional data analysis has been a recent active research area: we refer the reader to Ramsay and Silverman (2002, 2005) for a good introduction and overview, and Ferraty and Vieu (2006) for a more recent viewpoint. A major difference between this approach, and more classical statistical methods is that curves are viewed as the *atoms* of the analysis, i.e. the goal is the statistical analysis of a *population of curves*.

Wang and Marron (2007) recently extended functional data analysis to Object Oriented Data Analysis (OODA), where the atoms of the analysis are allowed to be more general data objects. Examples studied there include images, shapes and tree structures as the atoms, i.e. the basic data elements of the population of interest. Other recent examples are populations of movies, as in functional magnetic resonance imaging. A major contribution of Wang and Marron (2007) was the development of a set of tree-population analogs of standard functional data analysis techniques, such as Principal Component Analysis (PCA). The foundations were laid via the formulation of particular

^{*}Partially supported by NSF grant DMS-0606577, and NIH Grant RFA-ES-04-008.

[†]Partially supported by NSF grant DMS-0706761.

[‡]Partially supported by NIH grants R01EB000219-NIH-NIBIB and R01 CA124608-NIH-NCI.

[§]Partially supported by NSF grant DMS-0606577, and NIH Grant RFA-ES-04-008.

AMS 2000 subject classifications: Primary 62H35, 62H35; secondary 90C99

Keywords and phrases: Object Oriented Data Analysis, Population Structure, Principal Component Analysis, Tree-Lines, Tree Structured Objects

optimization problems, whose solution resulted in that analysis method (in the same spirit in which ordinary PCA can be formulated in terms of an optimization problem).

Here the focus is on the challenging OODA case of tree structured data objects. A limitation of the work of Wang and Marron (2007) was that no general solutions appeared to be available for the optimization problems that were developed. Hence, only limited toy examples (three and four node trees, which thus allowed manual solutions) were used to illustrate the main ideas (although one interesting real data lesson was discovered even with that strong limitation on tree size).

One of our main contributions is that, through a detailed analysis of the underlying optimization problem, and a complete solution of it, a linear time computational method is now available. This allows the first actual OODA of a production scale data set of a population of tree structured objects. Clinical findings resulting from this OODA include significant correlation of age and structure in left sub-population. Comparison across sub-populations was consistent with the expected symmetry.

Our ideas are illustrated in Section 2 using a set of blood vessel trees in the human brain, collected as described in Aylward and Bullitt (2002). In the present paper, we choose to consider only variation in the *topology* of the trees, i.e. we consider only the branching structure and ignore other aspects of the data, such as location, thickness and curvature of each branch.

Even with this topology only restriction, there is still an important *correspondence* decision that needs to be made: which branch should be put on the left, and which one on the right, see Section 2.1. Later analysis will also include location, orientation and thickness information, by adding attributes to the tree nodes being studied. A useful set of ideas for pursuing that type of analysis was developed by Wang and Marron (2007).

In Subsection 2.2 we define our main data analytic concept, the *tree-line*, and the notion of principal components based on tree-lines. Here we also state, and illustrate our main result, Theorem 2.1, which will allow us to quickly compute the principal components. Subsection 2.3 is devoted to our data analysis using the blood vessel data: we carefully compare the correspondence approaches, and present our findings based on the computed principal components. In Section 3 we prove Theorem 2.1 along with a host of necessary claims.

We finish the introduction by listing some relevant references on the use of trees in statistics, and the statistical analysis of tree populations. Both are relatively new and attractive areas. A likelihood approach to the analysis of tree populations is developed by Banks and Constantine (1998). Breiman

et. al. (1984) worked on classification and regression tree analysis. Breiman (1996) and Everitt et. al. (2001) studied the use of trees in cluster analysis. Some examples of statistical analysis of phylogenetic trees are in Holmes (1999) and Li et. al. (2000). We also refer to Pachter and Sturmfels (2005) for a comprehensive account of various uses of trees in biological statistics.

Another widely investigated approach to PCA of structured data is the family of *kernel methods*. These map each data point in a Non-Euclidean space to a vector space, where linear PCA methods can be applied. The details of these methods, together with some commonly used kernel functions for tree space can be found in Shawe-Taylor and Cristianini(2004).

The use of kernel methods for tree structured data commonly appears in the context of text categorization, where the parsing of sentences can be modeled via trees. Collins and Duffy (2002) develop some useful kernels for this purpose, and Eom et. al. (2006) use tree kernels to mine the biomedical literature for protein interactions. Another field where tree kernel ideas are applied is bioinformatics. See Yamanishi et. al. (2007) for an example of the use of the tree kernels approach for classifying carbohydrate sugar chains modeled as trees, and Vert (2002), where an application of tree-kernel PCA is used to measure similarities between the phylogenetic profiles of proteins.

2. Data and Analysis. The data analyzed here are from a study of Magnetic Resonance Angiography brain images of a set of 73 human subjects of both sexes, ranging in age from 18 to 72, which can be found at Handle (2008). One slice of one such image is shown in Figure 1. This mode of imaging indicates strong blood flow as white. These white regions are tracked in 3 dimensions, then combined, to give trees of brain arteries.

The set of trees developed from the image of which Figure 1 is one slice is shown in Figure 2. Trees are colored according to region of the brain. Each region is studied separately, where each tree is one data point in the data set of its region. The goal of the present OODA is to understand the population structure of 73 subjects through 3 data sets extracted from them: Back data set (gold trees), left data set (cyan) and right data set (blue). One point to note is that the front trees (red) are not studied here. This is because the source of flow for the front trees is variable, therefore this subpopulation has less biological meaning. For simplicity we chose to omit this sub-population.

The stored information for each of these trees is quite rich (enabling the detailed view shown in Figure 2). Each colored tree consists of a set of branch segments. Each branch segment consists of a sequence of spheres fit to the white regions in the MRA image (of which Figure 1 was one slice), as described in Aylward and Bullitt (2002). Each sphere has a center (with

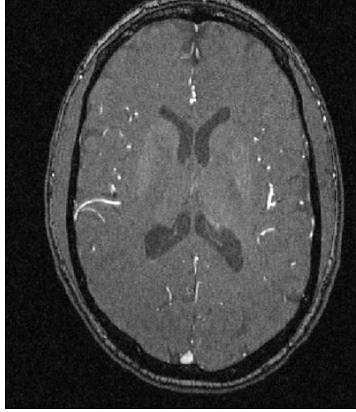


FIG 1. *Single Slice from a Magnetic Resonance Angiography image for one patient. Bright regions indicate blood flow. These regions in many MRA slices from each patient are tracked by a computer software to construct Figure 2.*

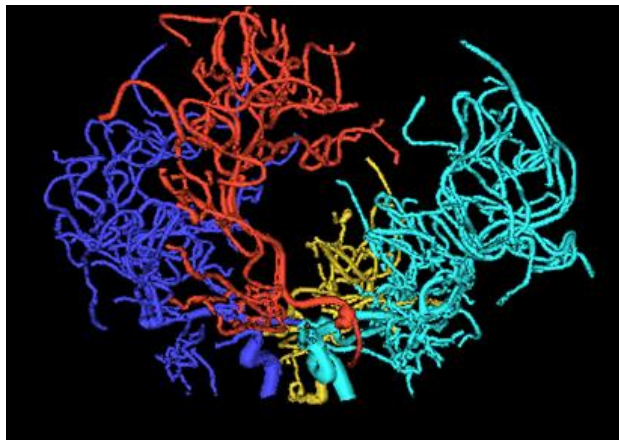


FIG 2. *Reconstructed set of trees of brain arteries for the same patient as shown in Figure 1. The colors indicate regions of the brain: Gold (back), Right (blue), Front (red), Left (cyan).*

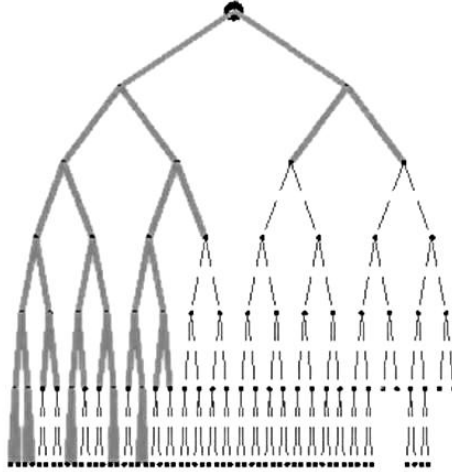


FIG 3. *Thick line segments show the topology only representation of the gold (back tree) from Figure 2. Only branching information is retained for the OODA. Branch location and thickness information is deliberately ignored. Thin dashed lines show the union over all trees in the sample.*

x , y , z coordinates, indicating location of a point on the center line of the artery), and a radius (indicating arterial thickness).

2.1. *Tree Correspondence.* Given a single tree, for example the gold colored (back) tree in Figure 2, we reduce it to only its topological (connectivity) aspects by representing it as a simple binary tree. Figure 3 is an example of such a representation. Each node in Figure 3 is best thought of as a branch of the tree, and the thick line segments simply show which child branch connects to which parent. The root node at the top represents the initial fat gold tree trunk shown near the bottom of Figure 2. The thin dashed lines show the support tree, which is just the union of all of the back trees, over the whole data set of 73 patients.

There is one set of ambiguities in the construction of the binary tree shown in Figure 3. That is the choice, made for each adult branch, of which child branch is put on the left, and which is put on the right. The following two ways of resolving this ambiguity are considered here. Using standard terminology from image analysis, we use the word *correspondence* to refer to this choice.

- **Thickness Correspondence:** Put the node that corresponds to the

child with larger median radius (of the sequence of spheres fit to the MRA image) on the left. Since it is expected that the fatter child vessel will transport the most blood, this should be a reasonable notion of *dominant branch*.

- **Descendant Correspondence:** Put the node that corresponds to the child with the most descendants on the left.

These correspondences are compared in Subsection 2.3.

Other types of correspondence, that have not yet been studied, are also possible. An attractive possibility, suggested in personal discussion by Marc Niethammer, is to use location information of the children in this choice. E.g. in the back tree, one could choose the child which is physically more on the left side (or perhaps the child whose descendants are more on average to the left) as the left node in this representation. This would give a representation that is physically closer to the actual data, which may be more natural for addressing certain types of anatomical issues.

2.2. *Tree-Lines.* In this section we develop the tools of our main analysis, based on the notion of *tree-lines*. We follow the ideas of Wang and Marron (2007), who laid the foundations for this type of analysis, with a set of ideas for extending the Euclidean workhorse method of PCA to data sets of tree structured objects. The key idea (originally suggested in personal conversation by J. O. Ramsay) was to define an appropriate *one dimensional representation*, and then find the one that best fits the data. The tree-line is a first simple approach to this problem.

First we define a binary tree:

DEFINITION 2.1. A **binary tree** is a set of nodes that are connected by edges in a directed fashion, which starts with one node designated as **root**, where each node has at most two children.

Using the notation t_i for a single tree, we let

$$(2.1) \quad T = \{t_1, \dots, t_n\}$$

denote a data set of n such trees. A toy example of a set of 3 trees is given in Figure 4.

To identify the nodes within each tree more easily, we use the level-order indexing method from Wang and Marron (2007). The root node has index 1. For the remaining nodes, if a node has index ω , then the index of its left child is 2ω and of its right child is $2\omega + 1$. These indices enable us to identify a binary tree by only listing the indices of its nodes.



FIG 4. Toy example of a data set of trees, T , with three data points ($n = 3$). This will be used to illustrate several issues below.

The basis of our analysis is an appropriate metric, i.e. distance, on tree space. We use the common notion of Hamming distance for this purpose:

DEFINITION 2.2. Given two trees t_1 and t_2 , their **distance** is

$$d(t_1, t_2) = |t_1 \setminus t_2| + |t_2 \setminus t_1|,$$

where \setminus denotes set difference.

Two more basic concepts are defined below; the notion of support tree has already been shown in Figure 3 (as the thin dashed lines).

DEFINITION 2.3. For a data set T , given as in (2.1), the support tree, and the intersection tree are defined as

$$\begin{aligned} \text{Supp}(T) &= \cup_{i=1}^n t_i \\ \text{Int}(T) &= \cap_{i=1}^n t_i. \end{aligned}$$

Figure 7 shows the support trees of the data sets used in this study. Figure 8 includes the corresponding intersection trees.

The main idea of a tree-line (our notion of one dimensional representation) is that it is constructed by adding a sequence of single nodes, where each new node is a child of the most recent child:

DEFINITION 2.4. A **tree-line**, $L = \{\ell_0, \dots, \ell_m\}$, is a sequence of trees where ℓ_0 is called the starting tree, and ℓ_i comes from ℓ_{i-1} by the addition of a single node, labeled v_i . In addition each v_{i+1} is a child of v_i .

An example of a tree-line is given in Figure 5. Insight as to how well a given tree-line fits a data set is based upon the concept of projection:

DEFINITION 2.5. Given a data tree t , its **projection** onto the tree-line L is

$$P_L(t) = \arg \min_{\ell \in L} \{d(t, \ell)\}.$$

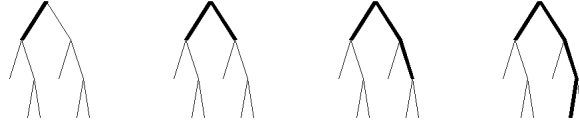


FIG 5. Toy example of a tree-line. Each member come from adding a node to the previous. Each new node is a child of the previously added node. Starting point (ℓ_0), the first tree in the example, is the intersection tree of the toy data set of Figure 4.

Wang and Marron (2007) show that this projection is always unique. This will also follow from Claim 3.1 in Section 3, whose characterization of the projection will be the key in computing the principal component tree-lines, defined shortly.

The above toy examples provide an illustration. Let t_2 be the second tree shown in Figure 4. Name the trees in the tree-line, L , shown in Figure 5, as $\ell_0, \ell_1, \ell_2, \ell_3$. The set of distances from t_2 to each tree in L is tabulated as

j	0	1	2	3
$d(t_2, \ell_j)$	5	4	3	2

The minimum distance is 2, achieved at $j = 3$, so the projection of t_2 onto the tree-line L is ℓ_3 .

Next we develop an analog of the first principal component (PC1), by finding the tree-line that best fits the data.

DEFINITION 2.6. For a data set T , the **first principal component tree-line**, i.e. PC1, is

$$L_1^* = \arg \min_L \sum_{t_i \in T} d(t_i, P_L(t_i))$$

In conventional Euclidean PCA, additional components are restricted to lie in the subspace orthogonal to existing components, and subject to that restriction, to fit the data as well as possible. For an analogous notion in tree space, we first need to define the concept of the union of tree-lines, and of a projection onto it.

DEFINITION 2.7. Given tree-lines $L_1 = \{\ell_{1,0}, \ell_{1,1}, \dots, \ell_{1,p_1}\}, \dots, L_q = \{\ell_{q,0}, \ell_{q,1}, \dots, \ell_{q,p_q}\}$, their **union** is the set of all possible unions of members of L_1 through L_q :

$$L_1 \cup \dots \cup L_q = \{\ell_{1,i_1} \cup \dots \cup \ell_{q,i_q} \mid i_1 \in \{0, \dots, p_1\}, \dots, i_q \in \{0, \dots, p_q\}\}$$

Given a data tree t , the projection of t onto $L_1 \cup \dots \cup L_q$ is

$$(2.2) \quad P_{L_1 \cup \dots \cup L_q}(t) = \arg \min_{\ell \in L_1 \cup \dots \cup L_q} \{d(t, \ell)\}.$$

In our non-Euclidean tree space, there is no notion of orthogonality available, so we instead just ask that the 2nd tree-line fit as much of data as possible, when used in combination with the first, and so on.

DEFINITION 2.8. For $k \geq 1$ the k th principal component tree-line is defined recursively as

$$(2.3) \quad L_k^* = \arg \min_{\ell \in L} \sum_{t_i \in T} d(t_i, P_{L_1^* \cup \dots \cup L_{k-1}^* \cup L}(t_i)),$$

and it is abbreviated as *PCk*.

For the concept of PC tree-lines to be useful, it is of crucial importance to be able to compute them efficiently. We need two more notions:

DEFINITION 2.9. Given a tree-line

$$L = \{\ell_0, \ell_1, \dots, \ell_m\}$$

we define the path of L as

$$V_L = \ell_m \setminus \ell_0.$$

Intuitively, a tree-line that well fits the data “should grow in the direction that captures the most information”. Furthermore, the k th PC tree-line should only aim to capture information that has not been explained by the first $k - 1$ PC tree-lines. This intuition is made precise in the following theorem, which is the main theoretical result of the paper:

THEOREM 2.1. Let ℓ_0 be a given starting point, $k \geq 1$, and L_1^*, \dots, L_{k-1}^* be the first $k - 1$ PC tree-lines. For $v \in \text{Supp}(T)$ define

$$(2.4) \quad w_k(v) = \begin{cases} 0, & \text{if } v \in V_{L_1^*} \cup \dots \cup V_{L_{k-1}^*}, \\ \sum_i \delta(v, t_i), & \text{otherwise} \end{cases}$$

Then the k th PC tree-line L_k^* is the tree-line whose path maximizes the sum of w_k weights in the support tree, i.e. $\sum_{v \in V_{L_k^*}} w_k(v)$.

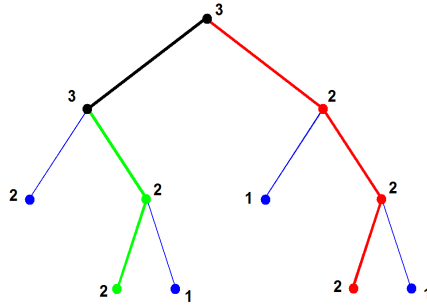


FIG 6. *Weighted support tree illustrating Theorem 2.1. Intersection tree is shown in black. Nodes that are added to construct PC1 are red. Nodes that make up PC2 are shown in green. The rest of the nodes in the support tree are blue.*

Here, the *delta* function $\delta(v, t_i)$ is equal to 1 if v is a node that exists in tree t_i , and 0 otherwise. The proof of Theorem 2.1 is given in Section 3. Figure 6 is an illustration: the weight of a node is the number of times the node appears in the trees of Figure 4. The black edge is the intersection tree of the same data set. The maximum weight path attached to $\text{Int}(T)$ is the red path, which gives rise to the tree-line of Figure 5, which is thus the first principal component of the data set of Figure 4.

After setting the weights of the nodes on the red path to zero, the maximum weight path attached to $\text{Int}(T)$ becomes the green path, which by Theorem 2.1 gives rise to *PC2*. The usefulness of these tools is demonstrated with actual data analysis of the full tree data set.

2.3. Real Data Results. This section describes an exploratory data analysis of the set of $n = 73$ brain trees discussed above using these tree-line ideas. The principal component tree-lines are computed as defined in Theorem 2.1. Both correspondence types, defined in Section 2.1 are considered and compared.

The different brain location types (shown as different colors in Figure 2) are analyzed as separate populations (i.e. the $n = 73$ blue trees are first considered to be the population, then the $n = 73$ gold trees, etc.), called *brain location sub-populations*. This reveals some interesting contrasts between the brain location types in terms of symmetry.

We first compare the two types of correspondence defined in Section 2.1 using the concept of the support tree. This is done by displaying the support trees each type of correspondence, and for each of the three tree location types (shown with different colors in Figure 2), in Figure 7. Note that all

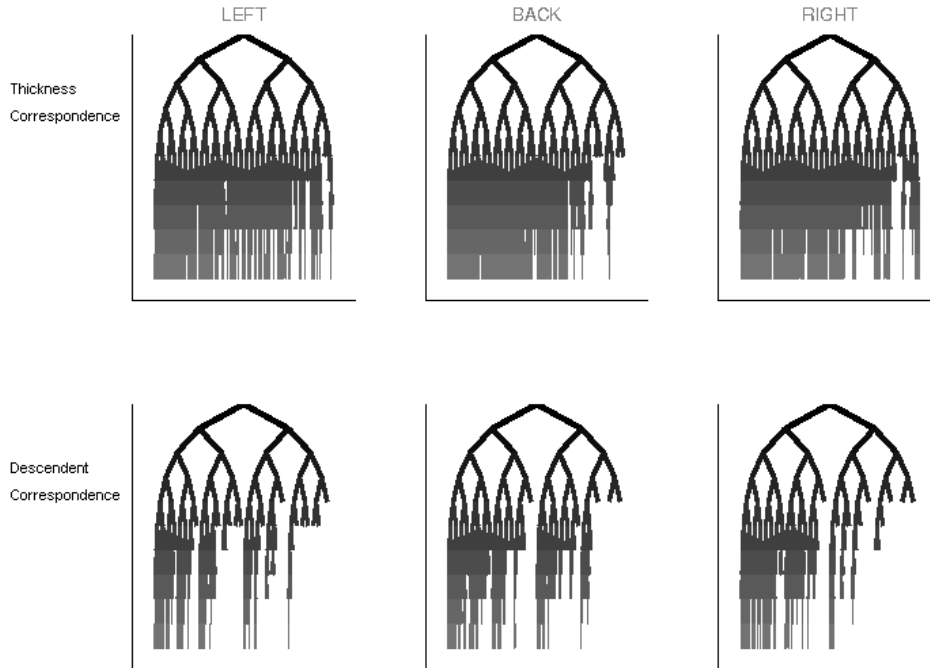


FIG 7. Support trees, for both types of correspondence (shown in the rows), and for three brain location tree types (shown in columns, corresponding to the colors in Figure 2). Shows that the descendant correspondence gives a population with more compact variation than the thickness correspondence.

of the support trees for the descendant correspondence (bottom) are much smaller than for the thickness correspondence (top), indicating that the descendant correspondence results in a much more compact population. This seems likely to make it easier for our PCA method to find an effective representation of the descendant based population.

Figure 7 already reveals an aspect of the population that was previously unknown: there is not a very strong correlation between median tree thickness of a branch, and the number of children.

Figure 8 shows the first 3 PC tree-lines, for the three sub-populations (shown as rows), with the intersection tree as the starting tree, for the descendant correspondence.

In the human brain, the back circulation (gold) arises from a single vessel

(the basilar artery) and immediately splits into two main trunks, supplying the back sides of the left and right hemispheres. These two parts of the back circulation are expected to be approximately mirror-image symmetrical with both sides containing one main vessel and other branches stemming from that. Consequently, for each tree on the back data set if we imagine a vertical axis that goes through the root node, we expect the subtrees on both sides of the axis to be symmetrical with each other.

The results of our model for the back subpopulation are consistent with this expectation. The main vessel of one of the hemispheres can be seen in the starting point (intersection tree) as the leftmost set of nodes, while the other main vessel becomes the first principal component.

As for the left and right circulations (cyan and blue trees) of the brain, they are expected to be close to mirror images of each other. Unlike the case of the back subpopulation, in each of these circulations there is a single trunk from which smaller branches stem. For this reason the bilateral symmetry observed within the back trees is not expected to be found here.

The fact that $PC1$'s for left and right subpopulations are at later splits suggest that the earlier splits tend to have relatively few descendants. The remaining $PC2$ and $PC3$ tree-lines do not contain much additional information by themselves. However, when we consider PC 's 1,2 and 3 together and compare left and right subpopulations, i.e. compare the second and third rows of Figure 8, the structural likeliness is quite visible. It should also be noted that for both of the subpopulations all PC 's are on the left side of the root-axis, indicating a strong bilateral asymmetry, as expected.

The tree-lines, and insights obtained from them, were essentially similar for the thickness correspondence, so those graphics are not shown here.

Next we study the tree-line analog of the familiar *scores plot* from conventional PCA (a commonly used high dimensional visualization device, sometimes called a *draftsman's plot*. In that case, the scores are the projection coefficients, which indicate the size of the component of each data point in the given eigen-direction. Pairwise scatterplots of these often give a set of useful two dimensional views of the data. In the present case, given a data point and a tree-line, the corresponding *score* is just the length (i.e. the number of nodes) of the projection. Unlike conventional PC scores, these are all integer valued.

Figure 9 shows the scores scatterplots for the set of left trees, based on the descendant correspondence. The data points have been colored in Figure 9, to indicate age, which is an important covariate, as discussed in Bullitt et al (2008). The color scheme starts with purple for the youngest person (age 20) and extends through a rainbow type spectrum (blue-cyan-green-yellow-

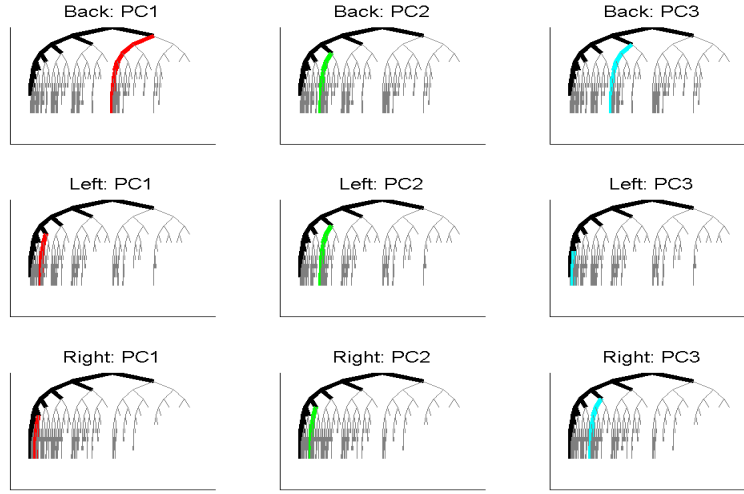


FIG 8. Best fitting tree-lines, for different sub-populations (rows), and PC number (columns). Intersection trees are shown in black. Support trees are shown in gray.

orange) to red for the oldest (age 72). An additional covariate, of possible interest, is sex, with females shown as circles, males as plus signs, and two transgender cases indicated using asterisks.

It was hoped that this visualization would reveal some interesting structure with respect to age (color), but it is not easy to see any such connection in Figure 9. One reason for this is that the tree-lines only allow the very limited range of scores. A simple way to generate a wider range of scores is to project not just onto simple tree-lines, but instead onto their union, as defined in (2.2). Figure 10 shows the scatterplots of several union PC scores, in particular $PC1$ vs. $PC1 \cup 2$ (shorthand for $PC1 \cup PC2$) vs. $PC1 \cup 2 \cup 3$. This combined plot, called the *cumulative scores scatterplot*, shows a better separation of the data than is available in Figure 9. The PC unions show a banded structure, which again is an artifact that follows from each PC score individually having a very limited range of possible values. This seems to be a serious limitation of the tree-line approach to analyzing population structure.

As with Figure 9, there is unfortunately no readily apparent visual connection between age and the visible population structure. However, visual impression of this type can be tricky, and in particular it can be hard to see

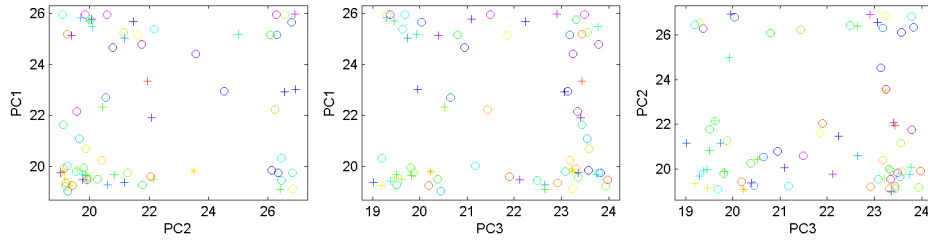


FIG 9. *Scores Scatterplot for the Descendant Correspondence, Left Side sub-population. The axes are PC scores for each data point. Colors show age, with cold colors corresponding to young subjects whereas warm colors are older subjects. No clear visual patterns are apparent with respect to age. Symbols indicate gender: circles are females, plus signs are males and asterisks are transgenders.*

some subtle effects.

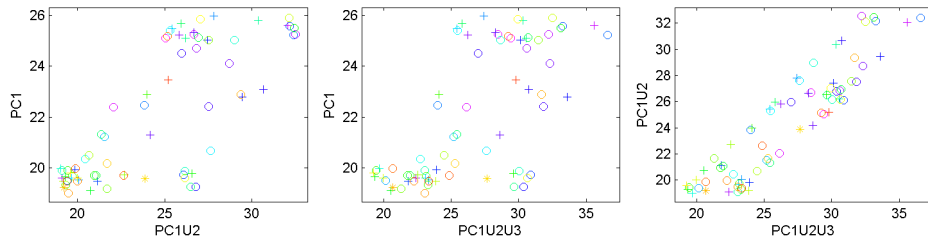


FIG 10. *Cumulative Scores Scatterplot for the Descendant Correspondence, Left Side sub-population. The axes are cumulative PC scores for each data point. Colors show age, with cold colors corresponding to young subjects whereas warm colors are older subjects. No clear visual patterns are apparent with respect to age. Symbols indicate gender.*

Figure 11 shows a view that more deeply scrutinizes the dependence of the $PC1$ score on age, using a scatterplot, overlaid with the least squares regression fit line. Note that most of the lines slope downwards, suggesting that older people tend to have a smaller $PC1$ projection than younger people. Statistical significance of this downward slope is tested by calculating the standard linear regression p -value for the null hypothesis of 0 slope. For the left tree, using the descendant correspondence, the p -value is 0.0025. This result is strongly significant, indicating that this component is connected with age. This is consistent with the results of Bullitt et al (2008), who noted a decreasing trend with age in the total number of nodes. Our result is the first location specific version of this.

Similar score versus age plots have been made, and hypothesis tests have

been run, for other PC components, and the resulting p -values, for the left tree using the descendent correspondence are summarized in this table:

$PC1$	$PC2$	$PC3$	$PC4$	$PC1 \cup 2$	$PC1 \cup 2 \cup 3$	$PC1 \cup \dots \cup 4$
0.003	0.169	0.980	0.2984	0.003	0.004	0.007

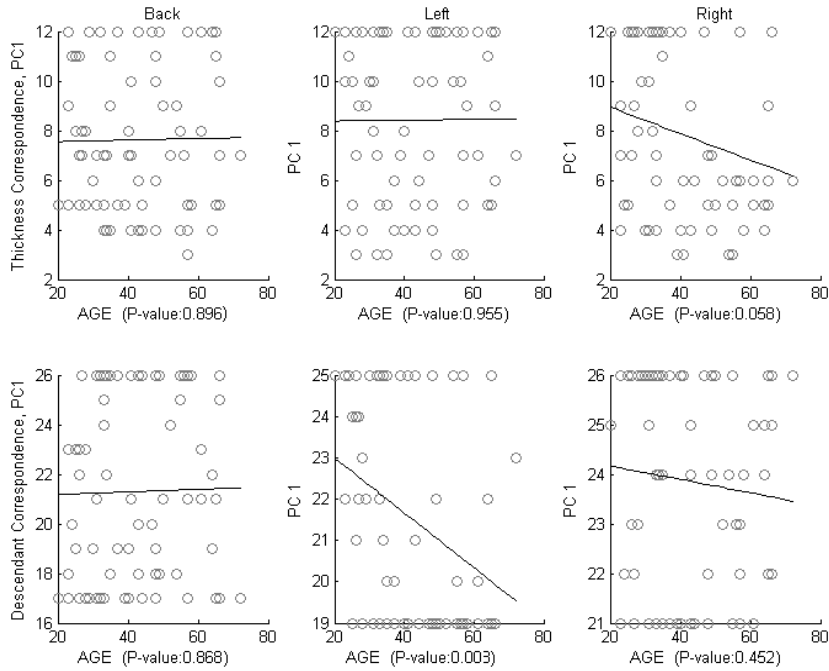


FIG 11. Scatterplot of $PC1$ score versus age. Least squares fit regression line suggests a downward trend in age. Trend is confirmed by the p -value of 0.003 (for significance of slope of the line).

Note that for the individual PCs, only $PC1$ gives a statistically significant result. For the cumulative PCs, all are significant, but the significance diminishes as more components are added. This suggests that it is really $PC1$ which is the driver of all of these results.

To interpret these results, recall from Figure 8, that for the left trees, $PC1$ chooses the left child for the first 3 splits, and the right child at the 4th split. This suggests that there is not a significant difference between the ages in the tree levels closer to the root, however, the difference does show up when

one looks at the deeper tree structure, in particular after the 4th split. This is consistent with the above remark, that for the left brain sub-population, the first few splits did not seem to contain relevant population information. Instead the effects of age only appear on splits after level 4.

We did a similar analysis of the back and right brain location sub-populations, but none of these found significant results, so they are not shown here. However, these can be found at the web site (19).

We also considered parallel results for the thickness correspondence, which again did not yield significant results (but these are on the web site (19)). The fact that descendant correspondence gave some significant results, while thickness never did, is one more indication that descendant correspondence is preferred.

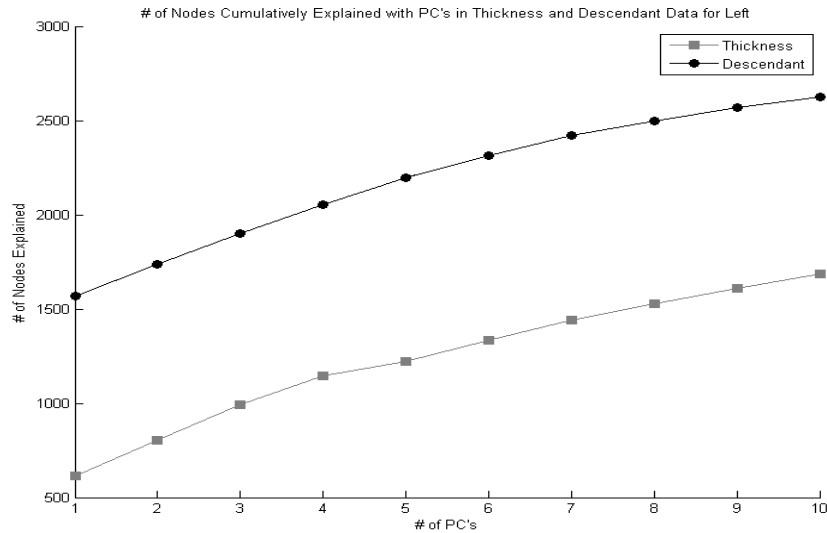


FIG 12. *Total number of nodes explained, as a function of Cumulative PC Number. Shows that the descendant correspondence allows PCA to explain a much higher proportion of the variation in the population than the thickness correspondence.*

One more approach to the issue of correspondence choice is shown in Figure 12. This shows the amount of variation explained, as a function of the order of the Cumulative Union PC, for both the thickness and the descendant correspondences, for the left brain location sub-population. The *amount of variation explained* is defined to be the sum, over all trees in the sub-population of the lengths of the projections. There are 5023 nodes in total for both correspondences. (The correspondence difference affects the locations

of nodes, total count remains the same.)

It is not surprising that these curves are concave, since the first PC is designed to explain the most variation, which each succeeding component explaining a little bit less. But the important lesson from Figure 12 is that the descendant correspondence allows PCA to explain much more population structure, at each step, than the thickness correspondence.

In summary, there are several important consequences of this work:

- In real data sets with branching structure, tree PCA can reveal interesting insights, such as symmetry.
- The descendant correspondence is clearly superior to the thickness correspondence, and is recommended as the default choice in future studies.
- As expected, the back sub-population is seen to have a more symmetric structure.
- For the left sub-population there is a statistically significant structural age effect.
- There seems to be room for improvement of the tree-line idea for doing PCA on populations of trees. A possible improvement is to allow a richer branching structure, such as adding the next node as a child of one of the last 2 or 3 nodes. We are exploring this methodology in our current research.

The data set used in this study have been expanded and improved during the course of the study. Preliminary analysis of the new data set shows that the age effect seen in left sub-population has become visible in all sub-populations. This issue will be handled in detail in future work.

3. Optimization proofs. This section is devoted to the proof of Theorem 2.1 with some accompanying claims.

CLAIM 3.1. *Let $L = \{\ell_0, \dots, \ell_m\}$ be a tree-line, and t a data tree. Then*

$$(3.1) \quad P_L(t) = \ell_0 \cup (t \cap V_L).$$

Proof: Since $\ell_i = \ell_{i-1} \cup v_i$, we have

$$(3.2) \quad d(t, \ell_i) = \begin{cases} d(t, \ell_{i-1}) - 1 & \text{if } v_i \in t; \\ d(t, \ell_{i-1}) + 1 & \text{otherwise.} \end{cases}$$

In other words, the distance of the tree to the line decreases as we keep adding nodes of V_L that are in t , and when we step out of t , the distance begins to increase, so Claim (3.1) follows. □

CLAIM 3.2. *Let L_1, \dots, L_q be tree-lines with a common starting point, and t a data tree. Then*

$$P_{L_1 \cup \dots \cup L_q}(t) = P_{L_1}(t) \cup \dots \cup P_{L_q}(t).$$

Proof: For simplicity, we only prove the statement for $q = 2$. Assume that

$$\begin{aligned} L_1 &= \{\ell_{1,0}, \ell_{1,1}, \dots, \ell_{1,p_1}\} \\ L_2 &= \{\ell_{2,0}, \ell_{2,1}, \dots, \ell_{2,p_2}\} \end{aligned}$$

with $\ell_0 = \ell_{1,0} = \ell_{2,0}$, and

$$(3.3) \quad V_{L_1} = \{v_{1,1}, \dots, v_{1,p_1}\}, V_{L_2} = \{v_{2,1}, \dots, v_{2,p_2}\}.$$

Also assume

$$(3.4) \quad P_{L_1}(t) = \ell_{1,r_1},$$

$$(3.5) \quad P_{L_2}(t) = \ell_{2,r_2}.$$

For brevity, let us define

$$(3.6) \quad f(i, j) = d(t, \ell_{1,i} \cup \ell_{2,j}) \text{ for } 1 \leq i \leq p_1, 1 \leq j \leq p_2.$$

Using Claim 3.1, (3.4) means

$$(3.7) \quad v_{1,i} \in t, \text{ if } i \leq r_1, \text{ and } v_{1,i} \notin t, \text{ if } i > r_1,$$

hence

$$(3.8) \quad \begin{aligned} f(i, j) &\leq f(i-1, j) \text{ if } i \leq r_1; \\ f(i, j) &\geq f(i-1, j) \text{ if } i > r_1. \end{aligned}$$

By symmetry, we have

$$(3.9) \quad \begin{aligned} f(i, j) &\leq f(i, j-1) \text{ if } j \leq r_2; \\ f(i, j) &\geq f(i, j-1) \text{ if } j > r_2. \end{aligned}$$

Overall, (3.8) and (3.9) imply that the function f attains its minimum at $i = r_1, j = r_2$, which is what we had to prove. \square

CLAIM 3.3. *Let S be a subset of $\text{Supp}(T)$ which contains ℓ_0 . For $v \in \text{Supp}(T)$ define*

$$(3.10) \quad w_S(v) = \begin{cases} 0, & \text{if } v \in S, \\ \sum_i \delta(v, t_i), & \text{otherwise} \end{cases}$$

Then among the tree-lines with starting tree ℓ_0 the one which maximizes

$$\sum_{t_i \in T} |(V_L \cup S) \cap t_i|$$

is the one whose path V_L maximizes the sum of the w_S weights: $\sum_{v \in V_L} w_S(v)$.

Proof: For $v \in \text{Supp}(T)$, and a subtree t of $\text{Supp}(T)$, we have:

$$\begin{aligned} \arg \max_{\ell \in L} \sum_{t_i \in T} |(V_L \cup S) \cap t_i| &= \arg \max_{\ell \in L} \sum_{t_i \in T} \sum_{v \in V_L \cup S} \delta(v, t_i) \\ &= \arg \max_{\ell \in L} \sum_{v \in V_L \cup S} \sum_{t_i \in T} \delta(v, t_i) \\ &= \arg \max_{\ell \in L} \sum_{v \in V_L \cup S} w_\emptyset(v) \\ &= \arg \max_{\ell \in L} \sum_{v \in V_L} w_S(v). \end{aligned}$$

□

Finally, we prove our main result:

Proof of Theorem 2.1: For better intuition, we first give a proof when $k = 1$. Using Claim 3.1 in Definition 2.6, we get

$$L_1^* = \arg \min_L \sum_{t_i \in T} d(t_i, \ell_0 \cup (t_i \cap V_L)).$$

Since V_L is disjoint from ℓ_0 ,

$$L_1^* = \arg \max_L \sum_{t_i \in T} |V_L \cap t_i|,$$

the statement follows from Claim 3.3 with $S = \emptyset$.

We now prove the statement for general k . For an arbitrary data tree t , and tree-line L , we have

$$\begin{aligned} (3.11) \quad P_{L_1^* \cup \dots \cup L_{k-1}^* \cup L}(t) &= P_{L_1^*}(t) \cup \dots \cup P_{L_{k-1}^*}(t) \cup P_L(t) \\ &= \ell_0 \cup (V_{L_1^*} \cap t) \cup \dots \cup (V_{L_{k-1}^*} \cap t) \cup (V_L \cap t) \\ &= \ell_0 \cup [(V_{L_1^*} \cup \dots \cup V_{L_{k-1}^*} \cup V_L) \cap t], \end{aligned}$$

with the first equation from Claim 3.2, the second from Claim 3.1, and the third straightforward.

Combining (3.11) with (2.3) we get

$$(3.12) \quad L_k^* = \arg \min_L \sum_{t_i \in T} d(t_i, \ell_0 \cup [(V_{L_1^*} \cup \dots \cup V_{L_{k-1}^*} \cup V_L) \cap t_i]).$$

Again, the paths of L_1^*, \dots, L_{k-1}^* and L are disjoint from ℓ_0 , so (3.12) becomes

$$(3.13) \quad L_k^* = \arg \max_L \sum_{t_i \in T} |(V_{L_1^*} \cup \dots \cup V_{L_{k-1}^*} \cup V_L) \cap t_i|,$$

so the statement follows from Claim 3.3 with $S = V_{L_1^*} \cup \dots \cup V_{L_{k-1}^*}$. \square

References.

- [1] Banks, D. and Constantine, G. M. (1998). Metric Models for Random Graphs. *J. Classification* 15 199-223.
- [2] Bullitt, E. and Bullitt, E. (2002) Initialization, noise, singularities and scale in height ridge traversal for tubular object centerline extraction, *IEEE Transactions on Medical Imaging*, 21, pp. 61-75
- [3] Bullitt, E., Zeng, D., Ghosh, A., Aylward, S. R., Lin, W., Marks, B. L., Smith, K. (2008) The effects of healthy aging on intracerebral blood vessels visualized by magnetic resonance angiography, submitted to *Neurobiology of Aging*.
- [4] Breiman, L., Friedman, J. H., Olshen, J. A., Stone, C. J. (1984), *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- [5] Breiman, L., (1996), Bagging Predictors, *Machine Learning*, vol 24, Number 2, 123-140.
- [6] Collins, M., and Duffy, N. (2002) Convolution Kernels for Natural Language, *Advances in Neural Information Processing Systems 14*, MIT Press, pp. 625-632.
- [7] Eom, J.-H., Kim, S., Kim, S.-H., and Zhang, B.-T. (2006). A tree kernel-based method for protein-protein interaction mining from biomedical literature. *Knowledge Discovery in Life Science Literature*, PAKDD 2006 International Workshop, Proceedings, volume 3886 of Lecture Notes in Computer Science, Singapore. Springer.
- [8] Everitt, B. S., Landau, S., Leese, M. (2001), *Cluster Analysis (4th edition)*, Oxford University Press, New York.
- [9] Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*, Berlin, Springer.
- [10] Handle (2008) <http://hdl.handle.net/1926/594>
- [11] Holmes, S. (1999). Phylogenies: An Overview, IMA series, vol 112, on Statistics and Genetics, (ed. Halloran and Geisser), 81-119 Springer Verlag, New York.
- [12] Li, S., Pearl, D. K., Doss, H. (2000) Phylogenetic Tree Constructure Using Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 95:493- 508.
- [13] Pachter, L., and Sturmfels, B. (2005) *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, United Kingdom.
- [14] Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis*, New York: Springer-Verlag.
- [15] Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*, New York: Springer-Verlag (2nd edition).
- [16] Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*, New York: Cambridge University Press
- [17] Vert, J.P. (2002) A Tree Kernel to Analyse Phylogenetic Profiles, *Bioinformatics*, Vol.18 Suppl.1 2002 pp. 276-284.
- [18] Wang, H. and Marron, J. S. (2007) Object oriented data analysis: Sets of trees, *The Annals of Statistics*, 35, pp. 1849-1873.
- [19] Wang, H. (2008) <http://www.stat.colostate.edu/~wanghn/tree.htm>

- [20] Yamanishi, Y., Bach, F., and Vert, J.P. (2007) Glycan Classification With Tree Kernels, *Bioinformatics*, Vol.23 no.10 2007 pp. 1211-1216.

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-3260
E-MAIL: aydin@email.unc.edu
pataki@email.unc.edu
marron@email.unc.edu

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523-1877
E-MAIL: wanghn@stat.colostate.edu

DEPARTMENT OF SURGERY
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-3260
E-MAIL: bullitt@med.unc.edu