Nonparametric Regression Model with Tree-structured Response

Yuan Wang[1], J.S. Marron[2], Burcu Aydın[3],

Alim Ladha[2], Elizabeth Bullitt[2] and Haonan Wang[1]

[1]Colorado State University, USA,

[2]University of North Carolina, Chapel Hill, USA,

[3]HP Labs, USA

Abstract

Highly developed science and technology from the last two decades motivated the study of complex data objects. In this paper, we consider the topological properties of a population of tree-structured objects. Our interest centers on modeling the relationship between a tree-structured response and other covariates. For tree objects, this poses serious challenges since most regression methods rely on linear operations in Euclidean space. We generalize the notion of nonparametric regression to the case of a tree-structured response variable. In addition, a fast algorithm with theoretical justification is developed. We implement the proposed method to analyze a data set of human brain artery trees. An important lesson is that smoothing in the full tree space can reveal much deeper scientific insights than the simple smoothing of summary statistics.

Key words: binary tree, data object, object oriented data, smoothing

1 Introduction

Complex data objects, including tree-structured data, manifold data and curve data, are frequently encountered in many modern statistical applications. Using terminology introduced by Wang and Marron (2007), we call such data types *Object Oriented Data*. Often, object oriented data live in non-Euclidean spaces, in which addition and scalar multiplication are typically ill-defined. Thus, traditional statistical methods, most of which are based upon Euclidean analysis, can not be directly implemented.

In medical image analysis, tree-structured objects are found to be an efficient data representation when the focus of the medical study involves variation in branching structures. Object oriented data analysis (OODA) on tree-structured data objects has been studied in Wang and Marron (2007) and Aydın et al. (2009). The first paper proposed measures of centrality and variability for populations of tree-structured objects. In addition, an analog of principal component analysis has been developed for tree space starting with the formulation of an appropriate optimization problem. A detailed study, including a fast and complete solution, of this optimization can be found in the second paper.

Aydın et al. (2009) took OODA on trees further by studying the dependence of the principal component scores on age through a simple linear regression analysis. Here, we take a more direct approach to modeling the relationship between tree-structured objects and age. In the case of a scalar response, various linear and nonlinear regression models have been widely studied; see Davison (2003) for a recent overview. The main contribution of the present paper is the first generalization of the notion of locally weighted smoother to tree space. Our new approach is formulated as a particular optimization problem. Another contribution is an efficient algorithm with which a complete solution can be obtained in linear time.

Our motivating example is a set of human brain artery trees; see Aylward and Bullitt (2002) for a detailed description of the data collection. The study of brain artery trees has many target applications, including study of potential stroke victims, as well as screening for loci of pathologies such as brain tumors, see Aydın et al. (2011). In the present data set, only normal brains (determined by pre-screening) are considered, and the main goal is to understand general tendencies of change in the brains of adults, over the approximate age range 20 to 70.

The tree data are very rich, with many types of information, including connectivity, location, shape and thicknesses of the branches, all of which deserve further study. In this early analysis, we deliberately choose to directly target just one important aspect: connectivity. To avoid confounding this with other aspects, we reduce the data to purely topological structures, and analyze only those. Study of the many other interesting aspects will be an important goal for future work.

Even when studying only topological structure of the trees, there are still major mathematical challenges. Not only are these data objects non-Euclidean, the space they reside in is even less Euclidean than the space of manifold data. In particular, manifolds admit approximating tangent planes, that form the basis of many suggested statistical analyses (Bhattacharya and Patrangenaru, 2003; Fletcher et al., 2004; Bhattacharya and Patrangenaru, 2005). No such approximation is available in topological tree space, so we term this "strongly non-Euclidean". A reviewer made the interesting related comment that nonlinear smoothing in this space appears to be more straightforward to implement than simple linear regression.

The rest of this paper is organized as follows. Section 2 mainly describes the brain artery data and the tree representation of it, in which we focus on the connectivity. The nonparametric tree smoothing methodology is stated in Section 3, which yields an easily computed algorithm. In Section 4, a case study involving brain artery data is discussed. Our smoothing method reveals much deeper scientific insights than is available from nonparametric regression on the number of nodes. A simulation study demonstrating the effectiveness of our tree smoothing method is

given in Section 5. Section 6 describes some future work and finally Section 7 contains proofs of the theorems.

2 Tree Representation and Brain Artery Data

2.1 Data

In this paper, we study human brain artery systems. As noted above, there are several potential applications. In the present study, we have only healthy adults, so we focus on changes in arterial structure as a function of age.

This data set was collected by the *CASILab* at The University of North Carolina at Chapel Hill. Detailed description can be found in Aylward and Bullitt (2002) and Bullitt et al. (2005). Magnetic resonance angiography (MRA) scans have been collected for each participant, and the resulting images constitute a 3-D image of the brain artery system. One slice of such a 3-D image is shown in the left panel of Figure 1. MRA is good for finding arteries because motion (e.g., blood flow through arteries) shows up as white. The white regions in the left panel of Figure 1 are thus slices of arteries. Aylward and Bullitt (2002) used a tube tracking algorithm to find artery pieces, represented as a sequence of spheres. These pieces were then manually combined into trees. An example is shown in the right panel of Figure 1. In order to maintain locality of blood systems, the arteries have been separated into four sub-systems that feed different brain regions. These are colored in the right panel of Figure 1 as anterior (`red`), posterior (`gold`), left middle (`cyan`) and right middle (`blue`) cerebral artery systems.

In this study, there are 98 healthy human subjects involved. Other covariates are also available, such as gender, handedness, and ethnicity for each subject.

The tree data are very rich, involving many types of information, such as location, branching structure, and thickness of branches. While all of these will ultimately be of interest, in this

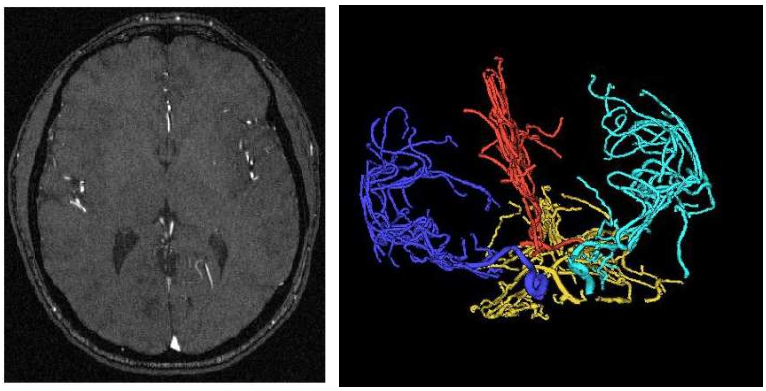

Figure 1: Left: One slice of an MRA scan for one subject; Right: a 3-D graphical illustration of the corresponding brain artery system. There are four major components: anterior (`red`), posterior (`gold`), left (`cyan`) and right (`blue`). The original MRA images are publicly available at http://hdl.handle.net/1926/594.

early study we choose to focus solely on the population variation of the branching structure topology. For this reason, we reduce each tree to only its branching structure, as in Wang and Marron (2007) and Aydın et al. (2009).

In the next subsection, we will briefly introduce some needed basic concepts of graph theory. We will continue the discussion of the tree representation of brain artery systems in Section 2.3.

2.2 Binary Tree

A tree is a collection of nodes (or vertices) and edges, where there exists exactly one simple path (a sequence of edges) between every pair of nodes. In a rooted tree, one node is designated as the *root*, and the *level* of a node is the total number of edges along the path to the root. For instance, the level of the root node is zero. Between each pair of nodes connected by an edge, the one with higher level is the *child*, and the other one is the *parent* of the child node. A node with no children is called a *leaf* node.

A *binary tree* is a special type of tree, and has been widely used in many scientific fields.

Every node of a binary tree has at most two children, a left child and a right child. In this paper, only rooted binary trees are considered. For our convenience, the set of all possible binary trees is referred to as *binary tree space*, and is denoted by $\mathcal{T}$.

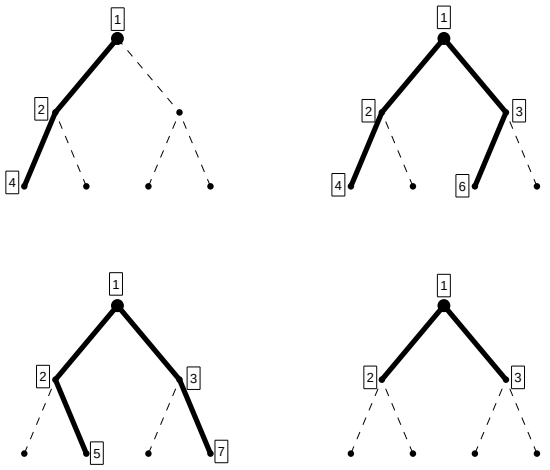

Figure 2: A graphical example of four binary trees t_1 (upper left), t_2 (upper right), t_3 (lower left), t_4 (lower right). The numbers in boxes are the level-order indices and the solid line segments indicate the edges in each tree.

To uniquely identify each node of a binary tree, we use the labeling system, from Wang and Marron (2007), called *level-order index*. Simple understanding of this system comes from Figure 2, where four example trees are depicted with different level-order index sets:

$$\mathrm{IND}(t_1) = \{1, 2, 4\}, \quad \mathrm{IND}(t_2) = \{1, 2, 3, 4, 6\}, \quad \mathrm{IND}(t_3) = \{1, 2, 3, 5, 7\}, \quad \mathrm{IND}(t_4) = \{1, 2, 3\}.$$

For two binary trees t_1 and t_2, the Hamming metric based on their level-order index sets is defined as

$$d_I(t_1, t_2) = \sum_{k=1}^{\infty} I\{k \in \mathrm{IND}(t_1) \triangle \mathrm{IND}(t_2)\} \tag{1}$$

where $I\{\cdot\}$ is the indicator and $\triangle$ is the symmetric difference between two sets. This metric is called the *integer tree metric* in Wang and Marron (2007), and it counts the total number of

noncommon nodes between two trees. As simple examples, the trees t_1 and t_2 in Figure 2 have distance 2, and t_2 and t_3 have distance 4. A reviewer has pointed out that, in some situations, it may not be reasonable to assume that every node can be treated equally. It is straightforward to extend this integer tree metric by adding a weight to each node. In particular, our results could be generalized using a weighted tree metric defined as

$$d_I^w(t_1, t_2) = \sum_{k=1}^{\infty} \alpha_k I\{k \in \mathrm{IND}(t_1) \triangle \mathrm{IND}(t_2)\} \tag{2}$$

where α_k is a positive weight of node k. d_I is the special case where $\alpha_k = 1$ for all k. It is easy to check that the weighted metric is also a metric on binary tree space.

2.3 Tree-structured Objects

In this subsection, we describe how the connectivity structure is extracted to form each purely topological *data object*, as in Wang and Marron (2007) and Aydın et al. (2009). An illustrative toy example is shown in Figure 3. The parts of arteries between splits are called artery segments, which are labeled with corresponding circled letters in each panel. Note that the artery system in the left panel initially starts from a thick artery segment at the bottom. It grows upward, and then branches into two segments. Each of these artery segments may continue branching into additional artery segments. In our tree representation in the right panel, each artery segment is denoted by a node, labeled with the same circled letters. The initial segment Ⓐ is designated as the root node. The two artery segments Ⓑ and Ⓒ, connected to the root node, are its children. For definiteness, the node with more descendants is put on the left. Thus, a tree representation for each artery system is extracted, in which each artery segment is a node and the edges indicate connectivity of artery segments.

Two visualizations of a real data topological tree (displayed using the cyan color in the right panel of Figure 1) are plotted in Figure 4. The left panel shows a dyadic view whose format is the same as Figure 2. While this view is very intuitive, the dyadic nature of the display limits

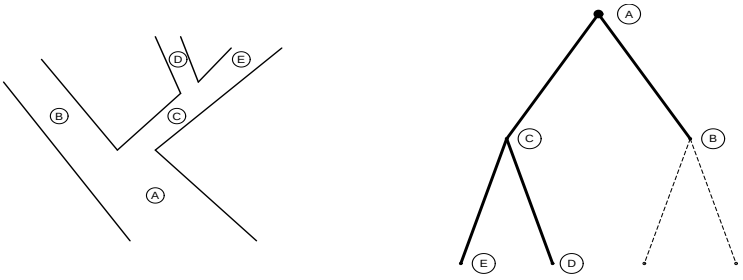

Figure 3: An example of brain artery binary tree construction. Left: An artery system with 5 segments; Right: Corresponding tree representation. The circled letters represent artery segments, and the level order index set is $\{1, 2, 3, 4, 5\}$.

the number of levels to about 8. This is a serious limitation for our data which frequently has 20 to 30 levels. The right panel shows the more sophisticated *D-L view* (Descendant-Level view) proposed by Aydın et al. (2011). The D-L view shows the same nodes and the same connected line segments, where the nodes are positioned in $\mathbb{R}^2$ according to quantities of interest. The vertical coordinate of the node positions is the logarithm (base 2) of the number of descendants, and the horizontal coordinate is the level of the node. This allows easy viewing of the full tree, and focusses on these important aspects, as illustrated in Section 4. While these coordinates are very informative, they result in substantial over-plotting because several nodes can have the same number of nodes on the same level. For enhanced visualization, jittering of a normal perturbation with mean 0 and standard deviation 0.05 is added to the y coordinate of each node.

3 Methodology

Regression analysis is one of the most commonly used techniques in statistics for modeling the relationship between a dependent variable and independent variables.

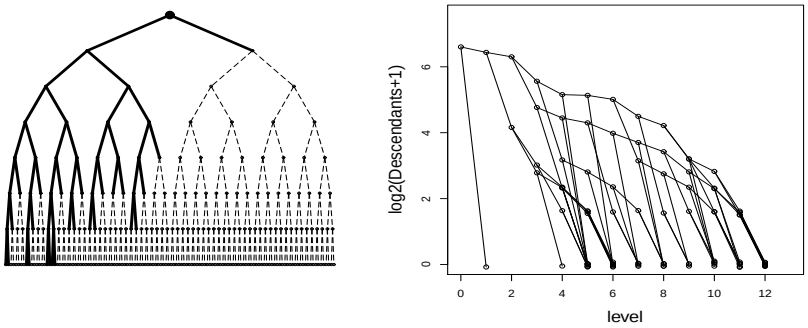

Figure 4: Tree visualizations for the left middle cerebral artery system depicted using the cyan color in Figure 1. Left: Dyadic tree representation, truncated to the first 8 levels for illustration; Right: D-L view showing the full structure of the same tree.

In general, a regression can be characterized as a mapping

$$f : X \mapsto E[Y|X]$$

where X is the predictor variable and Y is the response variable whose distribution depends on X. In this paper, the regression problem between a tree-structured response, $Y \in \mathcal{T}$, and the covariate age, $X \in \mathbb{R}^1$, is considered. In view of the strongly non-Euclidean nature of the tree space, it is not straightforward to develop simple linear regression. As pointed out by a reviewer, it is perhaps surprising that nonlinear regression by smoothing seems to be more feasible.

3.1 Nadaraya-Watson Estimate

Let $\{(X_i, Y_i), i = 1, \ldots, n\}$ be a random sample from a population $(\mathcal{X}, \mathcal{Y}) \subset \mathbb{R}^2$. The Nadaraya-Watson estimate, see Härdle (1990) for a good introduction, of the function f is a moving local average of the form

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)},$$

where h is called a bandwidth, $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function which can be any probability density function symmetric with respect to zero. We will also use the fact that $\widehat{f}(x)$ is a

minimizer, with respect to β, of the locally weighted least squares $\sum_{i=1}^{n}(Y_i-\beta)^2K_h(X_i-x)$ for any given x.

Several important issues need to be addressed here including the choices of kernel function and bandwidth. Here, we use the Gaussian kernel in our computations throughout this paper for the reasons given in Chaudhuri and Marron (2000). The choice of bandwidth for our tree smoother is rather challenging since most existing techniques can not be easily adapted to tree space. In the theory of classic nonparametric regression, various choices of bandwidth are suggested to optimize some expected discrepancy or its empirical/asymptotic approximation. While the bandwidth selection in tree space is still under development, we recommend the scale space viewpoint on bandwidth selection, as suggested by Chaudhuri and Marron (1999, 2000). We also implemented a cross-validation approach, see the supplemental material Section A. As expected from classical nonparametric regression (Härdle et al., 1988), this bandwidth choice is unreliable.

3.2 Measure of Centrality in Tree Space

A fundamental concept to all types of classical Euclidean regression analysis is conditional expectation. Here, we develop an appropriate analog in tree space. Let T be a tree-structured random element of $\mathcal{T}$. Assume that T has a probability distribution $P(\boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta}$. For instance, Banks and Constantine (1998) considered a family of probability measures on a finite set of graphs (including trees). An interesting question is, what is the *central tree* of $\mathcal{T}$ under the probability distribution $P(\boldsymbol{\theta})$? Fréchet (1948) proposed the Fréchet median which is the minimizer of $\mathbb{E}d(X,m)$. In the case of Euclidean distance on $\mathbb{R}^1$, this is the conventional median. More generally, it gives a useful notion of median in other Euclidean spaces.

Here we take our centerpoint to be the Fréchet median tree denoted by μ_F; that is,

$$\mu_F = \arg\min_{m\in\mathcal{T}} \mathbb{E}d_I(T,m) \tag{3}$$

where T is a random tree drawn from the distribution $P(\boldsymbol{\theta})$ on $\mathcal{T}$, and where d_I is the integer tree metric as defined in (1). The *Fréchet Variation* about the center is quantified by

$$V_F = \mathbb{E} d_I(T, \mu_F) \tag{4}$$

In Euclidean space, V_F is the usual mean absolute deviation from the median, a common robust dispersion measure.

When working with data, the empirical version of (3) and (4) are more convenient. In particular, define the sample Fréchet median and variation as

$$\widehat{\mu}_F = \arg\min_{m \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} d_I(t_i, m) \tag{5}$$

$$\widehat{V}_F = \frac{1}{n} \sum_{i=1}^{n} d_I(t_i, \widehat{\mu}_F) \tag{6}$$

where $\{t_1, \ldots, t_n\}$ is a random sample from $P(\boldsymbol{\theta})$. This problem has been considered by Wang and Marron (2007), who proposed an algorithm, *the majority rule*: a median tree contains all the nodes that appear more than $n/2$ times in the tree sample, and some or all nodes that appear exactly $n/2$ times. As a simple example, in Figure 2, tree t_4 is a median tree of $\{t_1, t_2, t_3, t_4\}$.

3.3 Tree Smoother

Continue to let T be a tree-structured random element. Our main focus with the blood artery data is the dependence of the topological structure of T on the covariate age. This is a regression problem in which the response is a tree-structured random element. As noted above, classical linear regression techniques are hard to implement directly due to the non-Euclidean nature of tree space. This is because linear operations such as addition and scalar multiplication are not well defined. We approach this problem using nonparametric smoothing. In particular, we express the regression problem as a general locally weighted optimization problem. The regression relationship between the tree-structured object (T) and the covariate of age ($x \in \mathbb{R}^1$) is formulated as a mapping from $\mathbb{R}^1$ to $\mathcal{T}$. This mapping is a *functional tree-structured object*

$T(\cdot)$, indexed by the covariate x. Moreover, we assume a heuristic notion of *smoothness* for this map in the sense that, for any pair of close enough x_1 and x_2, the corresponding mapped tree objects are also close with respect to the tree metric. Thus, for each x, the "conditional expected" tree $T(x)$ given x can be estimated by the solution to the following optimization problem: for any x, minimize over m

$$\sum_{i=1}^{n} d_I(t_i, m) K_h(x - x_i) \tag{7}$$

based on the sample data $(x_1, t_1), \ldots, (x_n, t_n)$. In other words, for each x, we obtain the locally weighted sample Fréchet center, and use it to estimate the corresponding conditional central tree. If the random variables $d_I(t_i, m), i = 1, \ldots, n$, are i.i.d., this weighted sum is an estimator of $\mathbb{E}d_I(T, m(x))$ up to a scale factor. In the much different domain of diffeomorphisms as data objects, Davis (2008) developed a related smoothing approach, called *manifold kernel regression*. They worked on a manifold, which allows local Euclidean approximation, and hence is easier to deal with than the extremely non-Euclidean tree space treated here.

Fast calculation of the minimizing tree is enabled by characterizations, based on locally weighted average nodal occurrences, of the solutions of (7) developed in Theorems 1 and 2. In particular, an easy-to-compute threshold value ensures a linear-time algorithm for fast computation of all solutions of this optimization problem. See Section B of the supplemental material for a detailed algorithm to solve the minimization problem (7).

For convenience of discussion, we first introduce a data dependent score function; that is, for a fixed node k, let

$$D_k(x) = \sum_{i=1}^{n} K_h(x - x_i) I\{k \in \mathrm{IND}(t_i)\} - \sum_{i=1}^{n} K_h(x - x_i)/2. \tag{8}$$

Note that, for any x, this score function measures the difference between the weighted average number of occurrences (the first quantity on the right hand side of (8)) and a normalization term depending on x (the second quantity on the right hand side of (8)). Theorem 1 provides a necessary condition for a minimizer tree of (7). In particular, if a node is included in the

minimizer tree, then the weighted average number of occurrences is above the normalization term, i.e., $D_k(x) \geq 0$.

Theorem 1. *For any $x \in \mathbb{R}^1$, let $t(x)$ be a minimizer tree of (7). All the nodes of the tree $t(x)$ must satisfy the following inequality: for any node $k \in \mathrm{IND}(t(x))$, $D_k(x) \geq 0$.*

For any x, let $S(x)$ be the collection of all nodes with nonnegative scores; that is,

$$S(x) = \{k : D_k(x) \geq 0\}.$$

Using an argument similar to the proof of Theorem 1, we can demonstrate that $S(x)$ is indeed a topological tree by proving the following statement: for any x, if $D_k(x) \geq 0$ then $D_{k'}(x) \geq 0$ for any ancestor node k' of k. As a direct consequence of Theorem 1, a minimizing tree $t(x)$ is a topological subtree of tree $S(x)$.

Moreover, we can further decompose $S(x)$ as $S(x) = S_1(x) \cup S_2(x)$, where $S_1(x)$ contains all nodes k with positive scores, and $S_2(x) = S(x) - S_1(x)$ is the set of nodes with zero scores. Next, a sufficient condition for a minimizing tree, based on such a partition, is provided in Theorem 2.

Theorem 2. *For any $x \in \mathbb{R}^1$, any minimizer tree $t(x)$ can be expressed as*

$$\mathrm{IND}(t(x)) = S_1(x) \cup S_2^*(x)$$

where $S_2^*(x)$ is a subset of $S_2(x)$.

Proofs of both theorems are included in Section 7.

It can be seen that every $S_2^*(x)$, that results in a tree, corresponds to a different minimizer tree. Thus, the minimizing tree is unique if and only if $S_2(x)$ is empty. When $S_2^*(x)$ is empty, we call $t(x)$, with the level-order index set $S(x)$, the unique minimal solution. When $S_2^*(x) = S_2(x)$, the tree $t(x)$, with the level-order index set $S_1(x) \cup S_2(x)$, gives the unique maximal solution. The minimal tree, which has the fewest nodes among all minimizing trees, is recommended as a device for breaking any ties. It is straightforward to show that Theorems 1 and 2 continue to

hold for a weighted analog of (7), based on a weighted distance d_I^w as defined at (2). Hence, the minimizer tree does not depend on the choice of weight sequence.

The majority rule in Wang and Marron (2007) is a special case of Theorems 1 and 2, in which we set $K_h(x - x_i) \equiv 1$, i.e., assigning equal weight to each tree in the sample.

3.4 Alternative Smoothing Representation

Additional insight to our tree smoother comes from representing it as a conventional smoother of indicator functions. Suppose that, for each node k, we want to model the relationship between the occurrence of the node, $Y = I\{k \in \text{IND}(T)\}$ (T is a random tree-structured object), and some covariate X. A simple regression can be characterized as a mapping

$$m_k : x \mapsto E[Y|X = x] = P(k \in \text{IND}(T)|X = x).$$

In particular, we have a random sample, $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i is the ith observation of the covariate X and $y_i = I\{k \in \text{IND}(t_i)\}$ is the corresponding response. Thus, the Nadaraya-Watson estimate for the mapping m_k can be written as

$$\widehat{m}_k(x) \equiv \frac{\sum_{i=1}^n K_h(x - x_i) I\{k \in t_i\}}{\sum_{i=1}^n K_h(x - x_i)}.$$

Note that, compared with the score function $D_k(x)$ in (8),

$$D_k(x) = 0 \Leftrightarrow \widehat{m}_k(x) = \frac{1}{2} \quad \text{and} \quad D_k(x) > 0 \Leftrightarrow \widehat{m}_k(x) > \frac{1}{2}.$$

In fact, the function $\widehat{m}_k(x) - 1/2$ can be used as the (rescaled) score function, in the sense that the sign indicates when the node is included. Thus, our proposed tree smoother can be simply viewed as a Nadaraya-Watson estimator node by node. Our tree smoother will keep those nodes whose Nadaraya-Watson estimator is greater than 1/2. Furthermore, it is worth noting that, $\{k : \widehat{m}_k(x) \geq 1/2\}$ and $\{k : \widehat{m}_k(x) > 1/2\}$ correspond to the maximal and minimal solution trees of (7), respectively.

4 A Case Study

In this section, our proposed method is implemented on the brain artery data, as described in Section 2. There are $n = 98$ subjects in the study, and each individual has four artery systems. Here, our interest centers on the relationship between the topological structure of each artery system and age. Aydın et al. (2009) considered a principal component approach in tree space to explore that relationship. They established a significant linear relationship between principal component scores and age. Our tree smoother makes three important contributions. First, this analysis directly targets the dependence on age instead of indirectly through principal components. Second, we allow more flexible nonlinear dependence on time. Third, we regress directly on tree objects instead of on numerical summaries, e.g., number of nodes.

To demonstrate the value of regressing the full tree objects on age, we first show a simple linear regression of the number of nodes of each tree on age in the first panel of Figure 5. Results from the left middle cerebral artery system are shown here. The other artery systems give similar results as seen in Section C of the supplemental material. The other panels show the Nadaraya-Watson estimators for different bandwidths. The slope of the linear regression line is significantly negative, and the Nadaraya-Watson estimators all suggest an overall decreasing pattern.

Now, we consider the deeper tree regression problem of modeling the relationship between full tree topology and age. We focus on the left middle cerebral artery system to save space. The results for the other brain artery systems are shown in Section D of the supplemental material. We implemented our smoothing method for the set of bandwidths, $\mathcal{C} = \{2, 3, 4, 5, 6, 8, 12, 18\}$. These are approximately equally spaced on a log scale, with 2 (18) clearly undersmoothed (oversmoothed, respectively). For each bandwidth h in the candidate set $\mathcal{C}$, the predicted tree $\widehat{t}_{h,i}$ is obtained for each observation (x_i, t_i). For our discussion, let $\widehat{t}_h(x)$ denote the fitted tree object at age x with bandwidth h. The change in characteristics of tree structure, such as branching pattern, as a function of age, is shown using a sequence of smoothed tree objects.

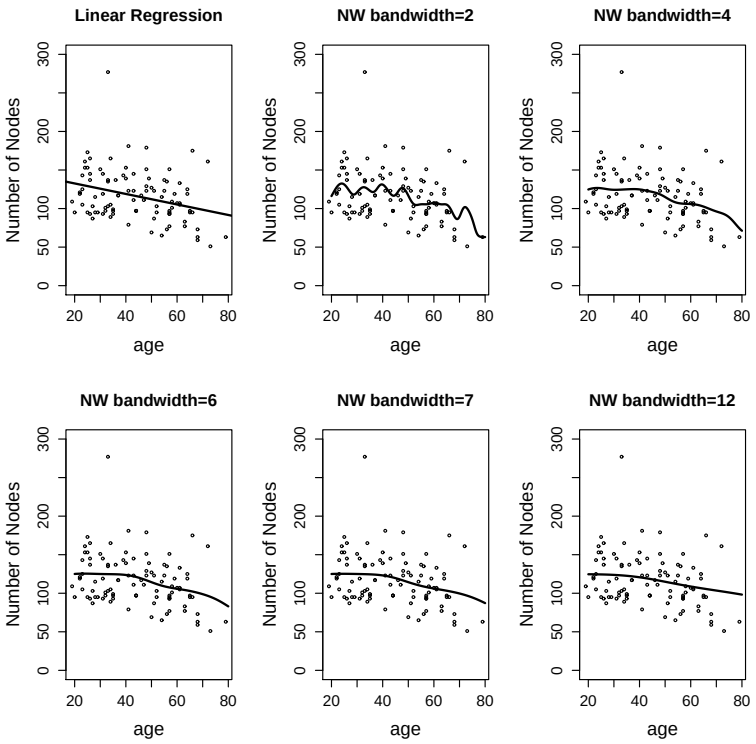

Figure 5: Scatterplots of the number of nodes of the left middle cerebral artery trees. First panel: Simple linear regression with a significantly negative slope. Panels 2 to 6: Nadaraya-Watson estimators for the bandwidths $2, 4, 6, 7, 12$, respectively. Note an overall decreasing trend in the number of nodes as age increases.

This is analogous to the conventional view of the smoothed curve in nonparametric regression analysis. The result is shown in Figure 6, which contains six subplots. In each subplot, a D-L view of the topological tree structure of a fitted tree at a given age, obtained by our smoothing method with bandwidth $h = 6$, is depicted. The reason for choosing $h = 6$ will be stated later. These subplots show the tendency of change in the topological structure of artery trees. Figure 6 shows that the topological structures corresponding to the trees in the upper row (younger ages) are more complicated than those of the trees in the lower row (older ages), which is also suggested in Figure 8. Moreover, compared with $\widehat{t}_6(20)$, $\widehat{t}_6(30)$ has more branching structure at higher levels. As age increases, $\widehat{t}_6(40)$ has less descendants than $\widehat{t}_6(30)$ at levels 1, 5 and 8. This indicates a tendency for artery segments to diminish or shrink over time. Moreover, the maximum level of the three estimated central trees, $\widehat{t}_6(50)$, $\widehat{t}_6(60)$ and $\widehat{t}_6(70)$, continues to decrease. Note also that, a tendency towards a larger number of descendants as age increases from $x = 50$ to $x = 70$. This suggests that some of the artery segments have diminished while others have split further. In addition, for older people, the increasing number of artery segments in some subtrees is consistent with the need to fill a greater area in response to artery blockage. A useful diagnostic for regression in Euclidean space is the residual plot. Here we develop an analog, called the *absolute deviation plot*, which is shown in Figure 7 for $h = 6$. Age is on the horizontal axis, and the distance between observations and fitted values is shown on the vertical axis. Note that the absolute deviations are spread between 50 and 150 fairly randomly, which suggests no heteroscedasticity problems for this data set.

Similar graphics, for other bandwidths, are shown in Section E of the supplemental material. A useful summary of the results is shown in Figure 8, where tree smooths are summarized by the number of nodes in each smoothed tree object. For each bandwidth, the corresponding panel shows a scatterplot of the number of nodes of each predicted tree $\widehat{t}_{h,i}$ versus age x_i for all $h \in \mathcal{C}$. When the bandwidth is 2, the fitted tree seems to be "undersmoothed", since the numbers of nodes of the fitted tree fluctuate strongly. As the bandwidth increases (3, 4 or

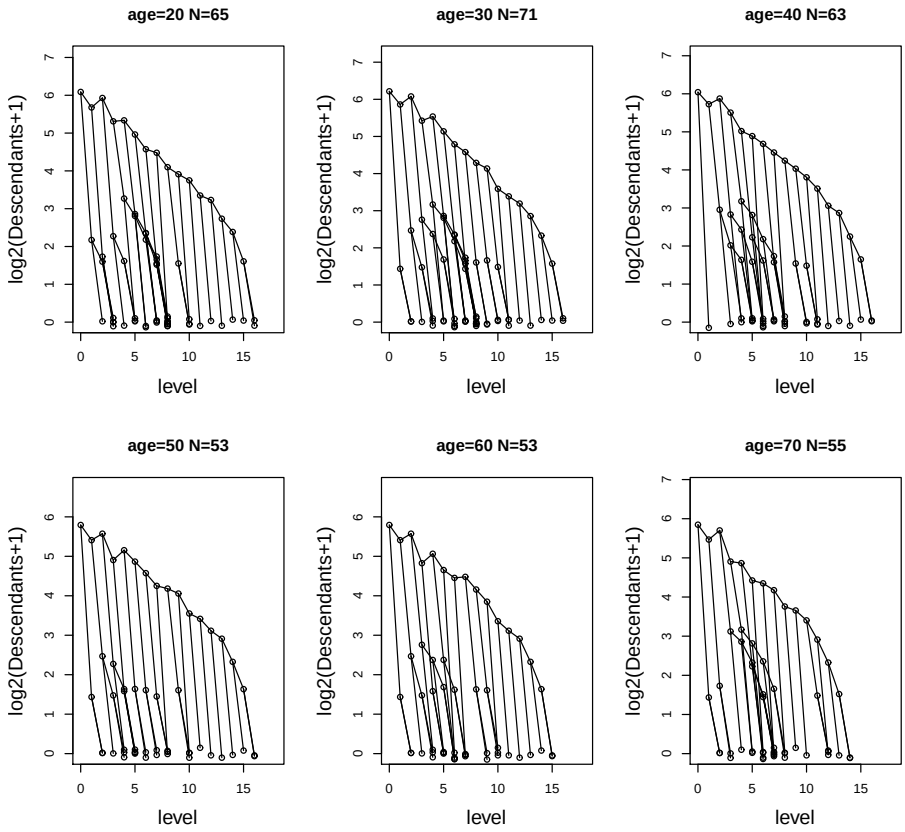

Figure 6: A graphical illustration of the topological structures of the fitted tree-structured objects at age $20, 30, 40, 50, 60, 70$ when the bandwidth is 6. In each subplot, a D-L view of the fitted tree at a given age is depicted. The variable N indicates the number of nodes of each tree. This clearly shows an increase in structure for younger ages, followed by a decrease in structure for older people.

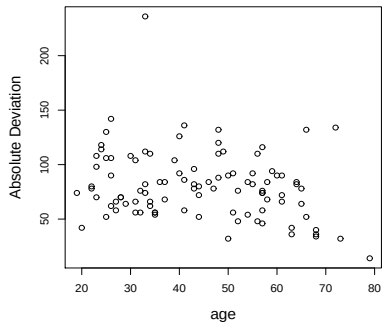

Figure 7: Absolute deviation plot for bandwidth $h = 6$. Vertical axis: the distance between each observation and the corresponding fitted tree; Horizontal axis: age. This suggests no heteroscedasticity and reasonable performance of our tree smoother.

5), there is a change in the decreasing pattern around age 50. In fact, the pattern becomes flat beyond 50. When the bandwidth is about 6 or 8, the overall pattern is very smooth and clear. Surprisingly, there are three distinct trends: one is increase from 20 to about 30, and one is decrease from 30 to 50, and the third is essentially flat after 50. On the other hand, when the bandwidth is 12 or 18, the local smoother tends to oversmooth the tree data. In fact, the number of nodes remains constant for a range of consecutive x values. For these large bandwidths, the dominant pattern is decreasing through the entire domain, which is different from the behavior for smaller bandwidths. Notice that the number of nodes of the smoothed trees are much smaller than the observed data, which suggests the smoothed estimator indeed captures important underlying population structure. Among those bandwidths considered, $h = 6$ seems to represent a reasonable trade-off between variance and bias.

As discussed in the supplemental material, we also considered cross-validation for bandwidth selection, but the result was quite oversmoothed at $h = 18$. This is consistent with the large sample variability for cross-validation in Euclidean smoothing discussed by Härdle et al. (1988). It is also consistent with the results of Chiu and Marron (1990), who showed how autocorrelation among data observations may cause such phenomena. In our tree regression problem, the driver

of the cross-validation bandwidth remains unclear.

The new anatomical phenomena discovered with bandwidth $h = 6$ demonstrates the value of smoothing in the full topological tree space as opposed to simply smoothing summary statistics in Euclidean space. Figure 8 reveals deep non-monotone behavior in age. In particular the increasing number of nodes from age 20 to age 30, seen in bandwidth $h = 6$, is not visible in simple smoothing of the summary statistics.

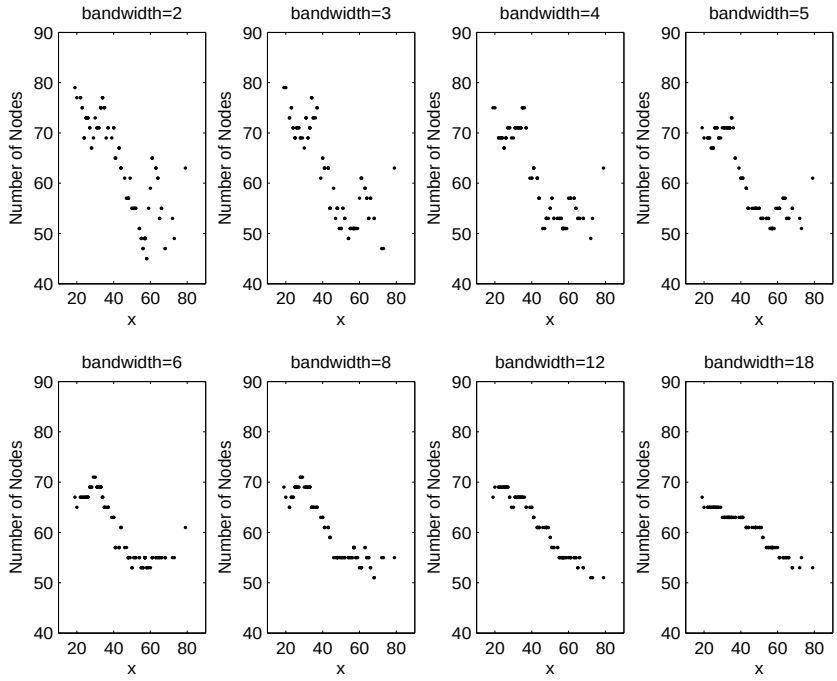

Figure 8: Scatterplots of the number of nodes of the smoothed (left middle cerebral artery) tree versus age. Note mostly decreasing trends, except increasing for younger people at intermediate bandwidths.

5 A Simulation Study

In Section 4, our proposed method has revealed three potential trends in the number of nodes of predicted tree objects using our proposed tree smoothing method. To investigate whether

these are spurious artifacts of the sample noise or important underlying structure, we conduct a simulation study. We will first introduce a probability distribution in tree space.

Following the development of probability measures on a finite set of graphs by Banks and Constantine (1998), we consider the following probabilistic framework on binary tree space $\mathcal{T}$. Define a notion of *circle* of radius k around the center μ in the tree space

$$S_k(\mu) = \{t \in \mathcal{T} : d_I(t, \mu) = k\}.$$

In our probabilistic framework, we make the following assumptions on a family of probability measures, denoted by $\{P_{\mu,\lambda}\}$:

1. The probability measure $P_{\mu,\lambda}$ is supported on $\mathcal{T}(\mu)$, where $\mathcal{T}(\mu) \subset \mathcal{T}$ is the collection of binary trees that contain the tree μ as a subtree; that is, in the notation of Wang and Marron (2007), $\mathcal{T}(\mu) = \{t \in \mathcal{T} : \text{IND}(\mu) \subset \text{IND}(t)\}$.

2. For any $t \in \mathcal{T}(\mu)$, the probability that it belongs to the circle of radius k is taken to be geometric(q), i.e. $P_{\mu,\lambda}(t \in S_k(\mu)) = pq^k, k = 0, 1, \ldots,$where $q = \exp\{-\lambda\}$ and $p = 1 - q$.

3. Within each circle, trees are chosen uniformly.

Combining all three assumptions, it can be seen that this probability measure assigns, for each tree $t \in \mathcal{T}(\mu)$,

$$P_{\mu,\lambda}(t) = \frac{1 - \exp\{-\lambda\}}{N(d_I(t, \mu))} \exp\{-\lambda d_I(t, \mu)\} \tag{9}$$

where $N(d_I(t, \mu))$ is the total number of trees in the circle of radius $d_I(t, \mu)$. Note that λ plays the role of a precision parameter, and the tree μ is the unique modal tree.

In general, exact computation of the theoretical Fréchet median tree, μ_F, is rather challenging. In this simulation study, we use a numerical approximation of μ_F. We did this by drawing a sample of size 701 from $P_{\mu,\lambda}$ and taking the sample Fréchet median. This was rapidly computed using the majority rule from Wang and Marron (2007). The size 701 gave the same answer in each of twenty replications.

To demonstrate the performance of our smoothing method, we simulated random trees from the distribution (9) in four different scenarios. In each smoothing scenario, the parameters μ and λ could be functions of some covariate, for instance, age. Here, we only consider constant λ, thus only μ changes over time, as illustrated in Figure 9. In all four scenarios, the precision parameter λ is 0.04 for all ages, chosen to make the variation of the simulated trees visually similar to the real data.

Scenario 1: The tree function $\mu(x)$ first grows as age increases until age $x = 55$, and then shrinks.

Scenario 2: The parameter $\mu(x)$ first shrinks as age increases until age $x = 55$, and then grows. This V-shape mimics the pattern we observed in the data analysis of the anterior cerebral artery system, as shown in Section D.3 of the supplemental material.

Scenario 3: The parameter $\mu(x)$ is a constant for young ages until age $x = 45$, and then grows. The flat pattern over ages 20 to 40 is motivated by what we observed in the right cerebral artery system, as shown in Section D.1 of the supplemental material.

Scenario 4: The parameter $\mu(x)$ shrinks until age $x = 35$, and grows until age $x = 55$, then shrinks.

For each of the four scenarios, we conduct a Monte Carlo experiment of $R = 100$ replicates. In each iteration, $n_x = 4$ trees are generated for each age x. We first generate geometric random variables $k_1, k_2, \ldots, k_{n_x}$ with $p = 1 - e^{-\lambda}$. Then, for $i = 1, \ldots, n_x$, a tree is randomly selected from the circle S_{k_i}. Figure 9 depicts the number of nodes in the tree $\mu_F(x)$ versus age x, obtained by numerical approximation, with the number of nodes from one realization overlaid.

When choosing the bandwidth, we implement a scale-space approach by looking at several bandwidths and then choosing the one which gives a reasonable compromise between variance and smoothness. These manually selected bandwidths for scenarios 1 to 4 are $4, 4, 5$ and 3, respectively. First, we will show the simulation results for one realization in Scenario 1. The top row of Figure 10 shows the D-L view of all four realizations (long-dashed) at age $x = 36$. Each

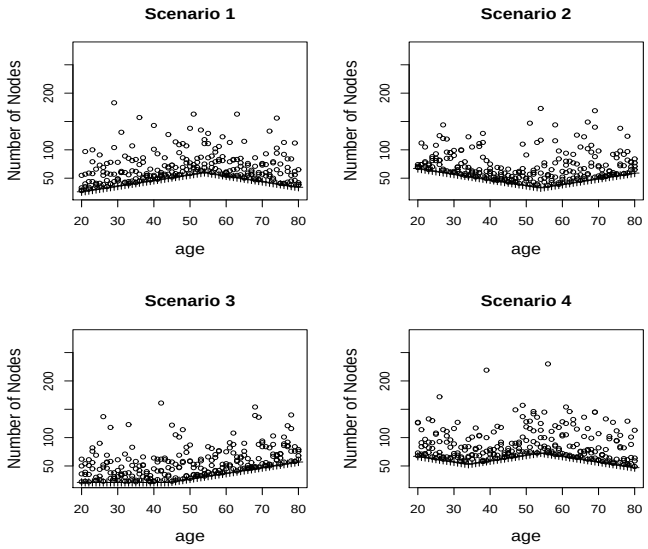

Figure 9: A graphical illustration of parameters $\mu(x)$ for each age x (shown as "+" symbols) in the four scenarios, together with one realization of the data (shown as "$\circ$" symbols). The horizontal axis is the covariate age and the vertical axis is the number of nodes at age x. In all four scenarios, at every age x, the numerically approximated Fréchet median tree $\mu_F(x)$ is identical to the true parameter $\mu(x)$. Hence "+" symbols also represents the number of nodes of $\mu_F(x)$.

panel also contains the population Fréchet median tree $\mu_F(x)$ (solid) and the resulting tree from our smoothing method (dotted). It can be seen that while the observations tend to be larger than the tree $\mu_F(x)$, the fitted tree captures the pattern of $\mu_F(x)$ quite well. The second row shows the corresponding results for age $x = 66$.

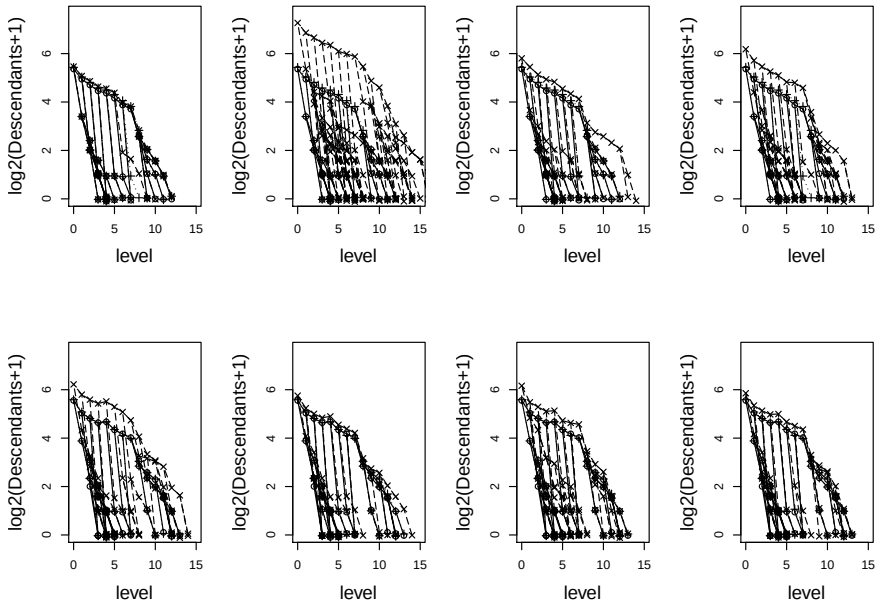

Figure 10: Performance of our smoothing method in the simulation study. In the top row, the D-L view of four realizations (long-dashed) at age $x = 36$ are given in each subplot. The bottom row shows corresponding plots for age $x = 66$. The tree μ_F (solid) and our smooth predicted tree (dotted) at the same age are overlaid. These show good performance of our proposed smoothing method for this relatively high noise level.

To save space, corresponding scatterplots, similar to Figure 10, for the other scenarios are included in Section F of the supplemental material. For each scenario, a movie version of Figure 10 that provides a better visualization is Internet available, and is included with this submission. A diagnostic for heteroscedasticity is provided by the absolute deviation plot in Figure 11. As in Figure 7, the height of the circles shows the distance between each simulated

observation and the corresponding fitted tree at the same age. We also overlaid the distance between the fitted trees and the true Fréchet median tree μ_F as "+" symbols. It can be seen that the fitted trees capture μ_F very well even for this high noise level.

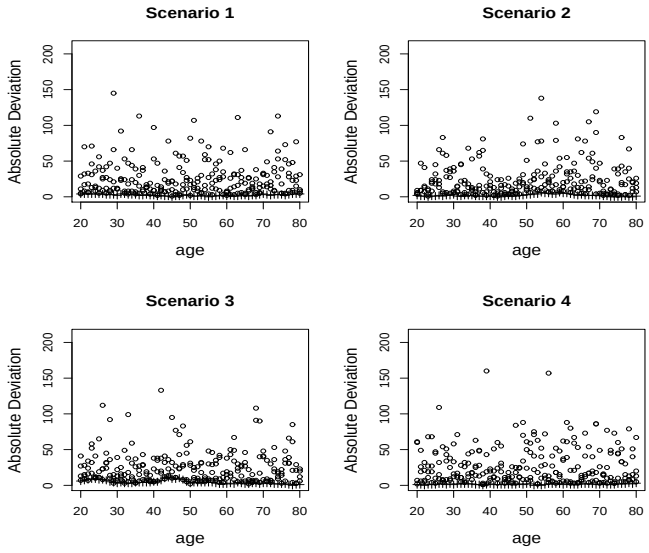

Figure 11: Absolute deviation plots for all four simulation scenarios indicate homoscedasticity in the data, shown by "∘" symbols (deviations for the raw data), and good estimation performance, shown by "+" symbols (deviations between the smooth, $\widehat{\mu}_F(x)$, and the median, $\mu_F(x)$).

Next, we will summarize the results from $R = 100$ replications. To demonstrate the performance of our proposed method, a sensible measure is the expected distance between $\mu_F(x)$ and $\widehat{\mu}_F(x)$ for each x, i.e., $\mathbb{E}d_I(\mu_F(x), \widehat{\mu}_F(x))$, which is called the *absolute estimation error*. Note that, this measure is defined in a similar way as in the classical linear regression analysis. Moreover, the Monte Carlo estimate of the absolute estimation error is

$$\frac{1}{R}\sum_{r=1}^{R} d_I(\mu_F(x), \widehat{\mu}_F^r(x))$$

where $\widehat{\mu}_F^r(x)$ is the estimated Fréchet median tree at age x in the rth iteration.

Another quantity to measure the overall estimation performance, which can be used together with the estimation error, is the Fréchet variation. In our regression setting, for each x, the

Fréchet variation $V_F(x)$ can be similarly defined as in (4) with the expectation taken over the probability distribution $P_{\mu(x),\lambda(x)}$. The empirical estimate of such pointwise Fréchet variation is

$$\widehat{V}_F(x) = \frac{1}{n_x} \sum_{i=1}^{n_x} d_I(t_i(x), \widehat{\mu}_F(x)).$$

In addition, the Monte Carlo estimate of the pointwise Fréchet variation can be calculated as an average of the empirical Fréchet variations, as defined above, for all 100 replicates.

In Figure 12, the Monte Carlo estimates of both estimation error and pointwise Fréchet variation are depicted, shown as "$\circ$" and "$*$" respectively. It can be seen that the estimation errors are quite small in contrast to the Fréchet variation, which suggests that our proposed tree smoother captures the topological structure of the Fréchet median tree. Moreover, errors near the boundaries and the changepoints tend to be higher, as is familiar from classical nonparametric regression. We also constructed boxplots, for each x, of the estimated Fréchet variation based on all replicates, which is included in Section F of the supplemental material. These indicate a constant variation in all four scenarios.

6 Future Work

Statistical analysis on tree space is still in its infancy, and we envision many future improvements over the work in this paper. First, the integer tree metric we used throughout treats all the nodes equally, which will be improved by incorporating suitable dependence on length and thickness of the artery segments. Second, 3-D spatial location and branch curvature will be included. A major challenge will be correspondence between different artery systems. Third, assessment of the statistical significance of the trends, observed in Section 4, of the brain artery system as a function of age is another problem of great interest. Some analogs of confidence bands or a SiZer approach may be useful here. Fourth, in Section 5, we showed empirically that, under the given probability distribution, the parameter tree μ seems to be the Fréchet median tree. Work

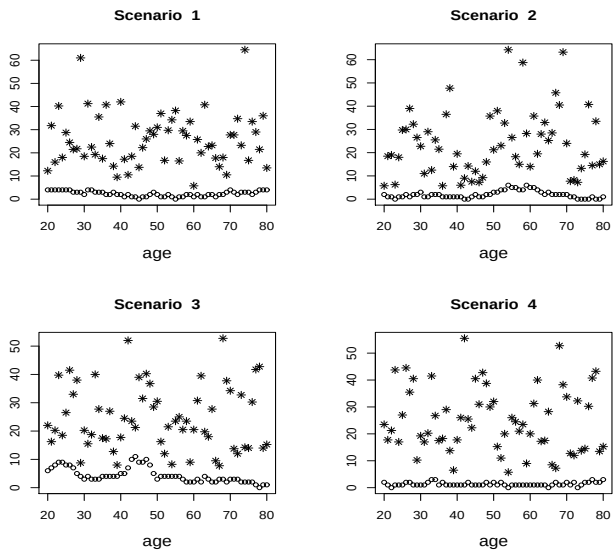

Figure 12: Absolute estimation error plot for all four simulation scenarios, shown by "$\circ$" symbols, and estimated Fréchet variation, shown by "$*$" symbols. Note that the error is small compared with the noise level of the data. As expected, error tends to be higher near the boundaries and the changepoints.

is in progress on a theoretical formulation of this. Finally, it will be of interest to develop some goodness-of-fit tests to explore potential distributional models for tree space.

7 Proofs of Theorems

Write $\omega_i = K_h(x - x_i)$. Recall from (8), the score function, for any node k, can be written as $D_k(x) = \sum_{i=1}^n \omega_i I\{k \in \text{IND}(t_i)\} - \sum_{i=1}^n \omega_i/2$.

Proof of Theorem 1

Suppose that, there exist some nodes with negative score and assume that k_0 is the node with the largest level among these nodes. Note that the node k_0 is a leaf node. Otherwise, considering one child of k_0, say k_1, we have, for any tree t_i, $I\{k_1 \in \text{IND}(t_i)\} \leq I\{k_0 \in \text{IND}(t_i)\}$. Hence,

$$\sum_{i=1}^n \omega_i I\{k_1 \in \text{IND}(t_i)\} \leq \sum_{i=1}^n \omega_i I\{k_0 \in \text{IND}(t_i)\} < \sum_{i=1}^n \omega_i/2.$$

Thus, the score of the node k_1 is negative, which contradicts the maximality of the level of node k_0.

Next, consider the tree $t'(x)$ such that $\text{IND}(t'(x)) = \text{IND}(t(x)) \setminus \{k_0\}$. It can be seen that

$$\begin{aligned}
\sum_{i=1}^n \omega_i d_I(t', t_i) &= \sum_{i=1}^n \omega_i d_I(t, t_i) - \sum_{i=1}^n \omega_i I\{k_0 \notin \text{IND}(t_i)\} + \sum_{i=1}^n \omega_i I\{k_0 \in \text{IND}(t_i)\} \\
&= \sum_{i=1}^n \omega_i d_I(t, t_i) - \sum_{i=1}^n \omega_i + 2\sum_{i=1}^n \omega_i I\{k_0 \in \text{IND}(t_i)\} \\
&< \sum_{i=1}^n \omega_i d_I(t, t_i),
\end{aligned}$$

which is a contradiction with the assumption that $t(x)$ is a minimizing tree.

Proof of Theorem 2

We have proved that any minimizer tree $t(x)$ has to be a subtree of $S(x)$ in Theorem 1. Suppose that there exist some nodes in $S_1(x)$ not contained in $t(x)$. Moreover, assume that j_0 is the node with smallest level among those nodes. Similar to the proof of Theorem 1, we can prove that j_0's parent has to be contained in $t(x)$ and this makes $t''(x)$, with $\mathrm{IND}(t''(x)) = \mathrm{IND}(t(x)) \cup \{j_0\}$, a topological tree. It can be seen that

$$\begin{aligned}
\sum_{i=1}^{n} \omega_i d_I(t'', t_i) &= \sum_{i=1}^{n} \omega_i d_I(t, t_i) + \sum_{i=1}^{n} \omega_i I\{j_0 \notin \mathrm{IND}(t_i)\} - \sum_{i=1}^{n} \omega_i I\{j_0 \in \mathrm{IND}(t_i)\} \\
&= \sum_{i=1}^{n} \omega_i d_I(t, t_i) + \sum_{i=1}^{n} \omega_i - 2\sum_{i=1}^{n} \omega_i I\{j_0 \in \mathrm{IND}(t_i)\} \\
&< \sum_{i=1}^{n} \omega_i d_I(t, t_i),
\end{aligned}$$

which is a contradiction with the assumption that $t(x)$ is a minimizing tree.

Acknowledgement

This paper is the thesis work of Yuan Wang, written under the supervision of Haonan Wang. The research of Elizabeth Bullitt was partially supported by NIH grants R01EB000219-NIH-NIBIB and R01 CA124608- NIH-NCI. The research of J.S. Marron was partially supported by NSF grants DMS-0606577 and DMS-0854908, and NIH Grant RFA-ES-04-008. The research of Haonan Wang was partially supported by NSF grants DMS-0706761, DMS-0854903 and DMS-1106975, and by the Air Force Office of Scientific Research under contract number FA9550-10-1-0241.

References

Aydın, B., Pataki, G., Wang, H., Bullitt, E., and Marron, J. (2009). A principal component analysis for trees. The Annals of Applied Statistics, 3:1597–1615.

Aydın, B., Pataki, G., Wang, H., Ladha, A., Bullitt, E., and Marron, J. (2011). Visualizing the structure of large trees. Electronic Journal of Statistics, 5:405–420.

Aylward, S. and Bullitt, E. (2002). Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction. IEEE Transactions on Medical Imaging, 21(2):61–75.

Banks, D. and Constantine, G. (1998). Metric models for random graphs. Journal of Classification, 15(2):199–223.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. i. Annals of Statistics, 31:1–29.

Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. ii. Annals of Statistics, 33:1225–1259.

Bullitt, E., Zeng, D., Gerig, G., Aylward, S., Joshi, S., Smith, J., Lin, W., and Ewend, M. (2005). Vessel tortuosity and brain tumor malignancy: A blinded study. Academic Radiology, 12:1232–1240.

Chaudhuri, P. and Marron, J. (1999). Sizer for exploration of structures in curves. Journal of the American Statistical Association, 94:807–823.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. The Annals of Statistics, 28:408–428.

Chiu, S.-T. and Marron, J. (1990). The negative correlations between data-determined bandwidths and the optimal bandwidth. Statistics & Probability Letters, 10:173–180.

Davis, B. (2008). Medical image analysis via Fréchet means of diffeomorphisms. Ph. D. Dissertation, the University of North Carolina at Chapel Hill.

Davison, A. (2003). Statistical Models. Cambridge University Press, United Kingdom.

Fletcher, P. T., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Medical Imaging, 23:995–1005.

Fréchet, M. (1948). Les éléments alétories de nature quelconque dans un espace distancié. Ann. Inst. H. Poincare, 10:215–310.

Härdle, W. (1990). Applied nonparametric regression. Cambridge University Press, United Kingdom.

Härdle, W., Hall, P., and Marron, J. (1988). How far are automatically chosen regression smoothing parameters from their optimum? J. Amer. Statist. Assoc., 83:86–95.

Wang, H. and Marron, J. (2007). Object oriented data analysis: Sets of trees. The Annals of Statistics, 35(5):1849–1873.