Statistical Modeling of Branching Probabilities for Tree-structured Data Objects

Hsin-Wen Chang, Hari Iyer, Elizabeth Bullitt and Haonan Wang

Version of May 24, 2011

Abstract

Study of factors affecting the functioning of the human brain has been of considerable interest for more than a century. In this paper, we focus on the structure of brain artery systems in humans and study how this might be related to factors such as age, gender or handedness (left or right handed). To facilitate this study we first represent brain artery systems using tree-structured objects and construct stochastic systems whose realizations are such tree-structured objects. We show that the parameters of the stochastic system, primarily the branching probabilities, may be effectively studied using a logistic regression framework. This appears to be a fruitful approach for understanding tree-structured data. Applying this novel approach to actual data collected on 98 subjects, we are able to conclude that age and gender do seem to influence brain artery branching patterns. Most brain arteries have decreasing branching probabilities with increasing age, and brain arteries of females are slightly more likely to branch than those of males.

Key words: binary tree, generalized linear model, object oriented data

1 Introduction

Four major brain artery systems — the anterior cerebral artery, the posterior cerebral artery, and the left and right middle cerebral arteries [1] — that supply nutrients and oxygen to our brain, are studied through Magnetic Resonance Angiography (MRA) scans of the brain. A 3-dimensional (3-D) image of the brain artery systems can be constructed, via multiple layers of transverse MRA scans, using methods described in [2]. In this paper, we are interested in how certain factors, such as age, gender or handedness, can influence the branching structure of these brain artery systems. Previous studies have shown that age, gender and handedness have an impact on human brains. [3] and [4] suggested that aging is related to brain atrophy, and that there are gender differences on this atrophy mechanism in different parts of the brain. Also, [5] showed evidence that anatomical asymmetries in the motor cortex are related to handedness and gender. Recently, [6] demonstrated gender differences and aging effect on brain blood vessels. These findings motivate our study on the effect of aging and gender difference on the brain artery systems.

A challenge to the statistical analysis is how to represent a brain artery system as a manipulatable response variable. A convenient approach is to represent each artery system as a tree-structured object especially when our primary interest concentrates on the branching pattern (i.e., topological structure) of these artery systems. Those tree-structured objects live in a non-Euclidean space where linear operations such as addition and scalar multiplication are not well defined. Consequently, traditional Euclidean-based statistical methods can not be directly applicable.

Rigorous analyses of tree-structured data have been recently developed by [7] and [8]. The methodology developed in these papers is called object oriented date analysis [OODA; 7], where the element in the analysis is a *data object*. In [7], the authors proposed empirical measures of centrality and variability for a sample of tree-structured objects. Moreover, [7] generalized the notion of the principal component analysis to the space of tree-structured objects. An efficient algorithm as well as the set of complete solutions can be found in [8]. In addition, [8] have conducted linear regression analysis to study the relationship between the principal component scores and ages. Here, one of our major contributions is to model directly, using the toolkit of generalized linear regression, the branching probabilities associated with blood vessel segments at various locations of each brain artery system. More specifically, our goal is to build an appropriate logistic regression model for these branching probabilities using age, gender or handedness as covariates.

As a first step in our effort to study the branching pattern of brain artery systems, we adopt the framework developed by [7]; that is, each artery system is represented as a binary *tree* with both topological properties such as branching patterns, and geometric properties such as location and thickness of each artery segment. Each vessel segment corresponds to a node in a tree-structured object. Moreover, each vessel segment either branches into two segments or stops branching. Thus, we can use 1 to denote "having two branches" and 0 to denote "having no branch". In other words, each node of the tree has an attribute (response) taking the value 1 if there are two branches emanating from that node and 0 otherwise. This representation of the branching information using 1's and 0's suggests that, it may be possible to adapt a logistic regression framework to study the branching patterns in brain artery networks.

To facilitate the application of a logistic regression model in the context of binary trees we view the binary trees in our sample as realizations of a stochastic Bernoulli process. In particular, each potential node of a tree has an associated probability of having two branches (as opposed to no branches) and, given that the earlier levels of the tree have a branching pattern that provides a path from the root node to the node in question, the branching at the current node is modeled as a Bernoulli random variable with its own success probability (probability of branching). Using such ideas we eventually develop a logistic regression model with mixed effects for binary trees and apply the model to our brain artery data set. A model selection procedure using AIC (Akaike's information criterion) is implemented, and several interesting results are discovered.

The paper is organized as follows. Details regarding our brain artery dataset are provided in Section 2. This section also describes some tools for analyzing tree-structured data. A logistic regression model for studying branching patterns of binary trees is developed in Section 3. The actual data analysis and results are presented in Section 4 along with interpretations of the results. We also provide concluding remarks and scientific discussions in Section 5.

2 Binary Tree Representation and Data Description

Each brain artery system can be represented by a tree-structured object that has both topological and geometric properties associated with it. In this section, we provide a brief introduction to mathematical trees and their branching patterns, often referred to as tree topology. We also describe a method of numerical representation of trees. For more details, see [9] and [10].

2.1 Graphs and Trees

A tree is a collection of nodes and edges where there is a unique path (a sequence of distinct edges) between each pair of nodes. In a *rooted* tree, one node is designated as the root, and the other nodes are the descendants of the root. Each edge connects two distinct nodes, and the node closer to the root is called the *parent* and the other is the *child.* The node with no child is called a *terminal* (*leaf*) node. Here, we will focus on rooted *binary* trees in which each node has either no children or two children, the left child and the right child. In the rest of this paper, the term tree is used to refer to a rooted binary tree for the sake of simplicity of terminology.

For mathematical convenience, each node ω is uniquely labeled by an integer called its *level-order index* [7], and is denoted by $\mathrm{ind}(\omega)$. It is defined in a recursive way as follows:

(i) The level-order index of the root node is defined to be 1;

(ii) Any node ω other than the root has a parent. Suppose that the parent's index is m. Then the index of the node in question is $2m$ if it is the left child and $2m+1$ if it is the right child.

The topological structure of a tree t can be characterized by its *level-order index set*, consisting of the level order indices of all the nodes. The level order index set of a tree t is denoted by $\mathcal{I} \equiv IND(t)$. In Figure 2, a tree topology has been depicted for the anterior cerebral artery system from one of the 98 subjects. Its level-order index set is

$$\{1,2,3,4,5,6,7,8,9,12,13,16,17,18,19,24,25,32,33,64,65,66,67\}.$$

Another important notion is the *level* of a node. This is defined as the total number of edges of the path to the root. Straightforward calculation shows that the level of a node is given by $\lfloor \log_2 (\mathrm{ind}\,(\omega)) \rfloor$, where $\lfloor x \rfloor$ is the greatest integer not exceeding x. For instance, the level of the root node is 0.

The binary tree can also be coded by many other ways. For instance, each binary tree can be represented as a sequence $\{s_n\}$ of zeros and ones such that $s_k = 1$ if a node exists whose level order index equals k, and $s_k = 0$ otherwise.

In the next subsection, we describe the brain artery dataset that we consider in this paper.

2.2 The Brain Artery Dataset

The dataset used in this paper consists of information extracted from MRA brain images of 98 healthy volunteers, collected by the *CASILab* at the University of North Carolina; see [11] and [12] for more details. The subjects range from 19 to 79 years of age, and in each decade group (19-29, 30-39, 40-49, 50-59, and 60 above) there are about twenty subjects, with roughly equal numbers of males and females. For each subject, multiple layers of transverse MRA scans were produced, and these were used as building blocks for the reconstruction of a 3-D image of the artery system [2]. One such MRA image can be download from an internet site [12], where white regions indicate strong blood flows; an example of a 3-D image of the artery system is shown in Figure 1. Four major artery systems are displayed, each with a different color — the anterior cerebral artery in `red`, the posterior cerebral artery in `gold`, the left middle cerebral artery in `cyan`, and the right middle cerebral artery in `blue`. Moreover, each artery system roots from the bottom and then gradually grows upward towards the upper part of the head.

Each of the four brain artery systems is discretized by sequences of voxels (volume elements) and digitized by 3-D coordinate information of these voxels. The information extracted for each artery consists of topological properties such as branching patterns as well as geometric properties such as location and thickness of each of the artery segments. In addition, other information, including age, gender (`male/female`), handedness (`left/right/ambidextrous`) and ethnicity (`white/Asian/African American`) of each of the subjects is also recorded.

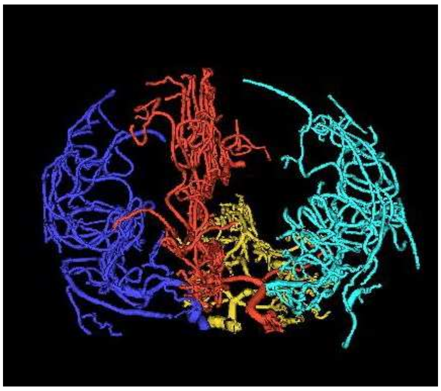

Figure 1: The 3-D image of the artery systems for subject 14 constructed from multiple transverse MRA slices, and four major artery systems are displayed with different colors — the anterior cerebral (`red`), the posterior cerebral (`gold`), the left middle cerebral (`cyan`), and the right middle cerebral (`blue`) brain artery systems.

Analysis of such *data objects* poses a serious statistical challenge since each element of data, i.e., data from a single individual, possesses both topological and geometric properties. In [7], a *tree* (see Figure 2) structure is used to represent the topology of each artery system and attributes are associated with each node of the tree to represent the geometric properties.

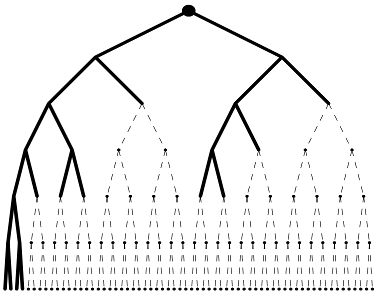

Figure 2: A graphical illustration of the tree topology for the anterior cerebral tree in Figure 1. It is truncated to the top 7 levels in this display.

Trees provide a natural way to represent brain artery systems. Such systems begin as an initial artery segment that branches into two segments. Each of these branches may further branch into two additional segments, and so on. Every *segment* in the artery system is mapped to a *node* in the tree representation. The initial segment is represented by the root node. Each of the branches of the initial segment is represented as a *child* node. These child nodes are connected to the root node by edges thus recording the branching hierarchy information. These second level branches may themselves branch further, giving rise to additional nodes in the tree representation, and so on. Notice that in representing each brain artery system by a tree, each node represents an artery segment, while each edge indicates the connectivity property of two arteries (i.e., nodes). For example, the root node, the top node of the tree in Figure 2, represents the starting thick red artery segment shown in Figure 1.

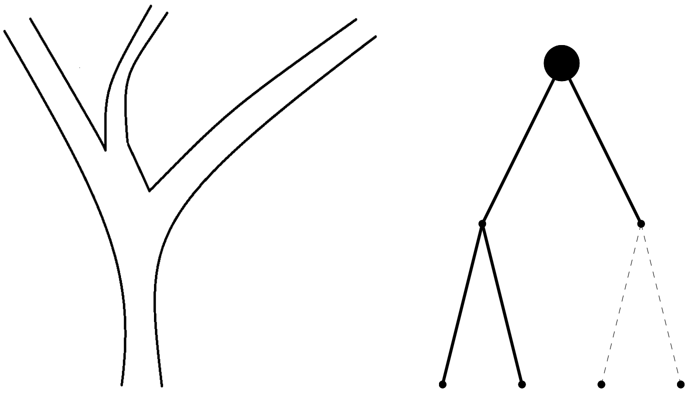

Figure 3: A graphical illustration of the tree representation (right panel) of blood vessels (left panel).

Figure 3 shows, more clearly, the correspondence between an artery system and the associated tree representation. It is apparent that a given segment either subdivides to produce two branches or does not branch at all, in which case it is a terminal segment.

Consequently, the corresponding tree representation has the property that each node has two child nodes or it is a terminal node. Such trees belong to the class of binary trees.

In our construction of the tree representation of an artery system, ambiguity arises when deciding which branch will be depicted as the left node and which as the right node. This is rather difficult to tell from a 3-D brain artery image. We remove this ambiguity by using the *thickness correspondence* convention, whereby one uses the left child node to represent the thicker of the two branches (based on average thickness of the segment), and the right child node to represent the thinner of the two branches. Based on such a tree representation, the branching structure of each artery system can be conveniently recorded by a set of integer numbers, the level order index set; see Section 2.1 for details. As for the geometrical properties, only arterial thickness information is used in further data analysis, and other information (e.g. location of voxels) is ignored in this paper. The average thickness of each segment of the brain artery system is associated with the corresponding node in the tree representation and forms one of the *nodal attributes* of the node in question. Each node may have several attributes associated with it. For instance, the level of a node is a nodal attribute.

2.3 Branching Probabilities of a Brain Artery System

Our main interest here centers on the branching property of the artery system. For mathematical convenience, we will refer to the corresponding binary tree t and study the branching structure of this tree. Let $\mathcal{I}$ denote the level-order index set of this tree. Let m denote the cardinality of $\mathcal{I}$ and write $\mathcal{I} = \{I_1, \ldots, I_m\}$. With each binary tree t having level-order index set $\mathcal{I}$, we associate a corresponding vector $\boldsymbol{y} = (y_1, \ldots, y_m)$, a vector of 0's and 1's of length

m, as follows: for $k = 1, \dots, m$,

$$y_k = \begin{cases} 1, & \text{if the node } I_k \text{ has two children} \\ 0, & \text{otherwise.} \end{cases}$$

Note that, for $k > 1$, $y_k = 1$ if and only if $\{2I_k, 2I_k+1\} \subseteq \mathcal{I}$; it is zero otherwise. In addition, it can be seen that y_k is zero if and only if the node I_k is a terminal node. For an illustration, consider the tree t in Figure 2. The vector $\boldsymbol{y}$ corresponding to this tree is given by

$$\boldsymbol{y} = (1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0).$$

Since we are interested in modeling the tree topology as a function of the thickness of the arterial segments as well as the attributes of the subject (covariates such as age, gender, ethnicity, and handedness), the vector $\boldsymbol{y}$ will serve as the response vector to be modeled.

3 Mixed Effects Logistic Regression Model on Trees

We begin by considering a population $\mathcal{P}$ of binary trees and introduce some simple stochastic models for describing the topological properties of the trees in $\mathcal{P}$. Later we introduce a class of mixed effects logistic regression models for modeling the brain artery system dataset which is the primary focus of this paper.

3.1 Simple Generative Stochastic Models for a Population of Binary Trees

Any tree t in the population of trees $\mathcal{P}$ may be viewed as a realization of a simple stochastic process described below. In this description, a tree t will be identified with its level-order index set $\mathcal{I}$. Let $\boldsymbol{x}_k$ denote the vector of nodal attributes associated with node k. We write

$\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_m^T)^T$ to denote the vector of concatenated nodal attributes for all the nodes of tree t.

A Stochastic Generative Model for Binary Trees.

A tree t is generated inductively as follows.

(a) The element 1 (root node) belongs to $\mathcal{I}$ with probability 1.

(b) Suppose a node with the label k belongs to $\mathcal{I}$. Let $\ell(k)$ denote the level of the node k. Generate a nodal attribute vector $\boldsymbol{x}_k$ from a distribution with cumulative distribution function (cdf) F_k where F_k depends only on the level $\ell(k)$ of k. We assume that $\boldsymbol{x}_j, j \geq 1$ are jointly independent random vectors.

(c) Given a node k which is already part of the tree, and its nodal attributes $\boldsymbol{x}_k$, conduct an independent bernoulli experiment with probability of success equal to $\mu_k = \mu(\boldsymbol{x}_k)$. Let y_k denote the outcome of this experiment (1 for success and 0 for failure). It is assumed that $\{y_k, k \geq 1\}$ is a sequence of independent bernoulli random variables.

- If $y_k = 1$, then append two nodes, with labels $2k$ and $2k + 1$, to the level-order index set $\mathcal{I}$.
- If $y_k = 0$, then node k becomes a terminal node and nodes $2k$ and $2k + 1$ are nonexistent in the tree being generated.

Observe that the probabilities of a node branching or not branching depend only on the attributes of the node in question. If two nodes have identical attributes, then the branching probabilities for these two nodes are identical. The stochastic model is fully described by $\{(F_k, \mu_k) | k \geq 1\}$.

The process of generating a tree according to this model may also be viewed in the following way. First generate the nodal attributes sequence $\boldsymbol{x}_k, k \geq 1$. Given the nodal attributes, we then generate the tree topology by generating the sequence $y_k, k \geq 1$. The topology of the realized tree is completely determined by the y_k sequence. In particular, for a given $\boldsymbol{x}_k$ sequence, different y_k sequences may be realized, giving rise to different trees.

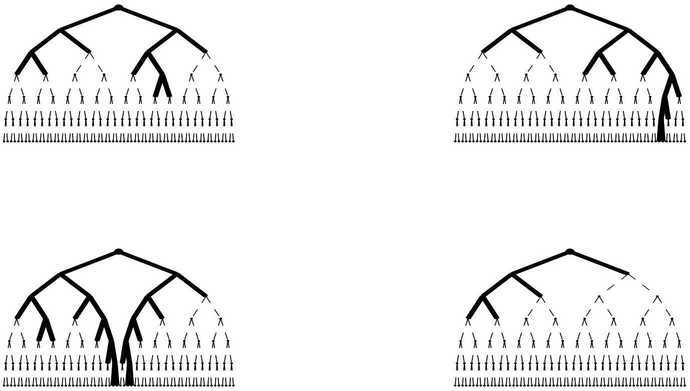

Figure 4: A sample of trees generated according to the proposed stochastic model.

The following illustrative example sheds some light on this generative model.

Example. Suppose each node k of a tree has associated with it a single attribute x_k representing the average thickness of the corresponding arterial segment. For $r \geq 1$, let the distribution of average thickness for nodes at level r be the uniform distribution on the interval $\left[\frac{16}{2^{r+2}}, \frac{16}{2^{r+1}}\right]$. The average thickness for the root segment is set to be 10. Furthermore, let $\mu(x) = \frac{e^{x-1}}{1+e^{x-1}}$ denote the branching probability when the nodal attribute (thickness) has the value x. Figure 4 shows four different realizations of trees generated according to the above model. The corresponding level-order index sets and the nodal attributes vector (thickness vector) are given below.

Tree 1: $\mathcal{I} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 26, 27\}$ and $\boldsymbol{X} = ($10.000, 2.816, 3.237, 1.429, 1.114, 1.937, 1.578, 0.736, 0.936, 0.628, 0.981, 0.334, 0.400)^T.

Tree 2: $\mathcal{I} = \{1, 2, 3, 4, 5, 6, 7, 12, 13, 14, 15, 30, 31, 60, 61, 120, 121\}$ and $\boldsymbol{X}$ = (10.000, 2.515, 2.547, 1.637, 1.962, 1.381, 1.877, 0.957, 0.752, 0.644, 0.957, 0.440, 0.500, 0.148, 0.149, 0.102, 0.087)^T.

Tree 3: $\mathcal{I} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 18, 19, 22, 23, 24, 25, 46, 47, 48, 49, 94, 95, 98, 99\}$ and $\boldsymbol{X}$ = (10.000, 2.497, 2.713, 1.939, 1.566, 1.243, 1.604, 0.791, 0.674, 0.815, 0.814, 0.751, 0.844, 0.494, 0.436, 0.382, 0.458, 0.312, 0.340, 0.218, 0.224, 0.227, 0.153, 0.095, 0.107, 0.122, 0.111)^T.

Tree 4: $\mathcal{I} = \{1, 2, 3, 4, 5, 8, 9\}$ and $\boldsymbol{X}$ = (10.000, 2.648, 3.861, 1.513, 1.100, 0.502, 0.604)^T.

We use this stochastic generative model as the basis for defining a class of generalized linear logistic random coefficient regression models that may be suitable for modeling the topological structures of populations of tree-structured objects.

3.2 Modeling Tree-Structured Data Using Generalized Linear Logistic Random Coefficient Regression Models

Consider a population of subjects with a tree-structured response associated with each subject. For a given subject i, let the associated tree-structured object be denoted by t_i. Let $\boldsymbol{x}_{i,k}$ and $y_{i,k}$, $k \geq 1$, denote the vector of nodal attributes and binary response variable associated with each *potential node* k respectively. Note that node k may or may not be actually present. Given that node q is present, the probability that node q has two branches (it is not a terminal node) is denoted by $\mu_{i,q}$. This probability is modeled as a function of $\boldsymbol{x}_{i,q}$

alone, and is given by

$$g(\mu_{i,q}(\boldsymbol{x}_{i,q})) = \boldsymbol{x}_{i,q}^T \boldsymbol{b}_i$$

where $g(\cdot)$ is a *link function.* For convenience, we let $y_{i,q}$ be a random variable which takes the value 1 if node q has two branches and 0 otherwise. Then $\mu_{i,q} = \mathbb{E}(y_{i,q}|\boldsymbol{x}_{i,q})$.

In this paper we consider the logistic link function given by

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

so that we have

$$\mu_{i,q} = \frac{\exp\left(\boldsymbol{x}_{i,q}^T \boldsymbol{b}_i\right)}{1 + \exp\left(\boldsymbol{x}_{i,q}^T \boldsymbol{b}_i\right)}. \tag{1}$$

Here, $\boldsymbol{b}_i$ is thought of as a vector of regression coefficients, specific to subject i. If $\boldsymbol{b}_i = \boldsymbol{b}$, regardless of the specific subject under consideration, then we have a fixed effects generalized logistic regression model. If $\boldsymbol{b}_i$ is in fact subject-dependent, then we have a mixed-effects generalized logistic regression model. We shall consider the latter, more general case. We model the collection $\{\boldsymbol{b}_i\}$ as iid random vectors with a common multivariate normal distribution $N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. Letting $\boldsymbol{v}_i = \boldsymbol{b}_i - \boldsymbol{\beta}$, we can write the probabilities $\mu_{i,q}$ as

$$g(\mu_{i,q}) = \boldsymbol{x}_{i,q}^T \boldsymbol{\beta} + \boldsymbol{x}_{i,q}^T \boldsymbol{v}_i.$$

One or more of the diagonal elements of the matrix $\boldsymbol{\Sigma}$ is allowed to be zero. In particular, if the (l,l) diagonal element of $\boldsymbol{\Sigma}$ is zero, this will imply that the l^{th} element in the regression coefficient $\boldsymbol{b}_i$ is a constant and is equal to β_l, the l^{th} element of $\boldsymbol{\beta}$. Hence the l^{th} element of $\boldsymbol{v}_i$ will be zero with probability one. Thus, some of the regression coefficients may be considered as fixed for all subjects, whereas the remaining regression coefficients may be considered as subject-specific (random). These are standard considerations in the area of mixed models.

Let $\boldsymbol{u}_i$ be the subvector of $\boldsymbol{v}_i$ corresponding to only the random effects (the entries corresponding to fixed effects will be zeros). The corresponding subvector of $\boldsymbol{x}_{i,q}$ will be denoted by $\boldsymbol{z}_{i,q}$. The revised model, allowing for some of the regression coefficients to be fixed and others to be random, may be written as

$$g(\mu_{i,q}) = \boldsymbol{x}_{i,q}^T\boldsymbol{\beta} + \boldsymbol{z}_{i,q}^T\boldsymbol{u}_i.$$

If none of the regression coefficients are fixed, then we will have $\boldsymbol{u}_i = \boldsymbol{v}_i$ and $\boldsymbol{x}_{i,q} = \boldsymbol{z}_{i,q}$.

Suppose a random sample of (binary) tree-structured data objects, $\mathcal{S} = \{t_1, t_2, \ldots, t_n\}$ is obtained from a population. For $i = 1, \ldots, n$, denote the level-order index set of tree t_i by $\mathcal{I}_i = IND(t_i)$. Let the branching pattern of tree t_i be represented by the binary response vector $\boldsymbol{y}_i$. If tree t_i has m_i nodes, we write

$$\mathcal{I}_i = \{I_{i,1}, \ldots, I_{i,m_i}\} \qquad \text{and} \quad \boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,m_i})\,.$$

In our example of brain artery system, the vector $\boldsymbol{x}_{i,k}$ of covariates, for node k of the tree for subject i, consists of the nodal attributes such as `age`, `gender`, `handedness`, and `ethnicity`, for each subject, and average `thickness` for each artery segment (each node in our tree representation). Observe that some of the nodal attributes (age, gender, handedness, ethnicity) are common to all nodes but thickness will generally vary from one node to the next. The covariate vector may also include interaction terms involving these attributes.

Under the model specified above, the joint likelihood function for the model parameters, based on sample data, is equal to

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}_{i,k}, \boldsymbol{x}_{i,k}, i = 1, \ldots, n; k = 1, \ldots, m_i) =$$
$$\int\!\!\int_{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n} \!\!\cdots \int \left(\prod_{i=1}^{n}\prod_{k=1}^{m_i} \mu_{i,k}^{y_{i,k}}(1-\mu_{i,k})^{1-y_{i,k}}\right) \left(\prod_{i=1}^{n} f_i(\boldsymbol{u}_i)\right) \; d\boldsymbol{u}_1\, d\boldsymbol{u}_2\, \ldots, d\boldsymbol{u}_n$$

where $f_i(\cdot)$ is the pdf of $\boldsymbol{u}_i$, and $\mu_{i,k}$ is defined implicitly in Equation (1), or equivalently,

$$\mu_{i,k} = \frac{\exp(\boldsymbol{x}_{i,k}^T \boldsymbol{\beta} + \boldsymbol{z}_{i,k}^T \boldsymbol{u}_i)}{1 + \exp(\boldsymbol{x}_{i,k}^T \boldsymbol{\beta} + \boldsymbol{z}_{i,k}^T \boldsymbol{u}_i)}.$$

Here, the optimization of the likelihood function is carried out using the `lme4` package in `R`, where the penalized iteratively reweighted least squares algorithm (PIRLS) is used to determine the conditional modes, and then the Laplace approximation to the likelihood is optimized. See the `lme4` package manual [13] for more details.

3.3 Model Selection

Variable selection, for both fixed effects and random effects, is implemented to help choose a parsimonious model. Various selection methods have been proposed and widely used in the literature. Different (empirical) discrepancies have been considered, such as AIC (Akaike's information criterion), BIC (Bayesian information criterion), Mallow's C_p, etc. Moreover, based on the choice of discrepancy, automatic selection procedures, e.g. the best subsets or the stepwise procedures, can be used for variable selection. Ideally, one can consider models with all possible combinations of main fixed effects, main random effects, and interactions. It is not feasible to compute the empirical discrepancies for all such models. In this paper, we will only consider main effects and two-way interaction terms. Let the set consisting of the intercept term, all main effects and all two-way interaction terms, be denoted by $\mathcal{A}$. We propose a two-step model selection procedure based on AIC.

1. *Selection of random effects.* Consider the family of models in which each model consists of predictors in set $\mathcal{A}$ as fixed effects, and subsets of all predictor variables in set $\mathcal{A}$ as random effects. The random and fixed intercept terms will always be included.

Compute the AIC value for each candidate model, and choose the model with the smallest AIC value. The random effects included in the resulting model are chosen to be the random effects to use in the next step.

2. *Selection of fixed effects.* Here, consider the family of models in which each model consists of the random effects selected from Step 1, and subsets of all predictor variables in set $\mathcal{A}$ as fixed effects. The random and fixed intercept terms will always be included. Compute the AIC value for each candidate model, and choose the model with the smallest AIC value.

The proposed two-step procedure greatly reduces the computational load, increases the computing efficiency and avoids the algorithmic convergence problems.

4 Data Analysis of Brain Artery Systems — a Case Study

The mixed effects logistic regression model on tree-structured outcomes is implemented in the statistical analysis of the brain artery dataset. We analyze all four brain artery systems — the left middle, the right middle, the anterior, and the posterior cerebral artery systems separately, and use the proposed AIC-based model selection procedure to choose four different regression models for the arterial branching probability of each system.

4.1 Numerical Results

The numerical results are presented in Table 1. An estimated regression coefficient (denoted as coef.) is provided if the predictor is in the model, or otherwise the cell is left blank. The p-value for the test of whether each estimated regression coefficient is significantly different from 0 is also provided. For simplicity, `left`, `right`, `front`, and `back` are used to represent the left middle, right middle, anterior and posterior cerebral artery systems.

Table 1 near here

From Table 1, it can be seen that the model selection procedure described in Section 3.3 provides similar random-effect results but different fixed-effect results for different artery systems. In particular, the random effect of thickness is present for every cerebral artery system. This means that the effect of arterial thickness on arterial branching probability varies from subject to subject. In contrast, different sets of fixed-effect predictor variables are selected for different artery systems. For the left middle, right middle, and posterior cerebral artery systems, the set of predictors consists of an intercept, the continuous variable `age`, the continuous variable `thickness`, the interaction `age` × `thickness`, and the indicator variable `male` which distinguishes arteries of male subjects from those of females. For the anterior cerebral arteries, the set of predictors consists of an intercept, the continuous variable `thickness`, and the indicator `male` for arteries of male subjects.

In short, despite the natural fact that arterial thickness affects arterial branching in every cerebral artery system, the effect varies from subject to subject. Also, age has an impact on the left middle, right middle and posterior cerebral arteries, and gender plays an important role in every artery system. More detailed interpretations will be given in Section 4.2.

4.2 Graphical Results and Interpretations

In this section we present the graphical results and their corresponding interpretations for each cerebral artery system.

The Left Middle, Right Middle, and Posterior Cerebral Artery Systems

The fitted surfaces of the branching probability are displayed in Figure 5 for brain arteries of male subjects (left column) and female subjects (right column), and for the left middle (top row), right middle (middle row) and posterior (bottom row) artery systems. To facilitate our further interpretations, the projected fit on each continuous variable is displayed in Figure 6.

From the subplots of the left column in Figure 6, it can be observed that regardless of gender, as age increases, the branching probability of thin arteries increases and that of thick arteries decreases. For any given age, thicker arteries are more likely to branch, and female arteries are slightly more likely to branch than male ones.

From the right panel of Figure 6, it can be seen that as thickness increases, the branching probability of arteries increases regardless of age. Also, thin arteries are more likely to branch in the brains of elderly people, while thick arteries are more likely to branch in the brains of young people. Again, female arteries are slightly more likely to branch than male ones.

To sum up, for the left middle, right middle, and posterior cerebral artery systems, higher age is associated with higher branching probability for thin arteries and with lower branching probability for thick arteries. Notice that, for any given age, thicker arteries are more likely to branch, and so even though thin arteries become more likely to branch as age increases, the branching probability is still smaller than that of thick arteries. Also, for any given age, female arteries are slightly more likely to branch than male ones.

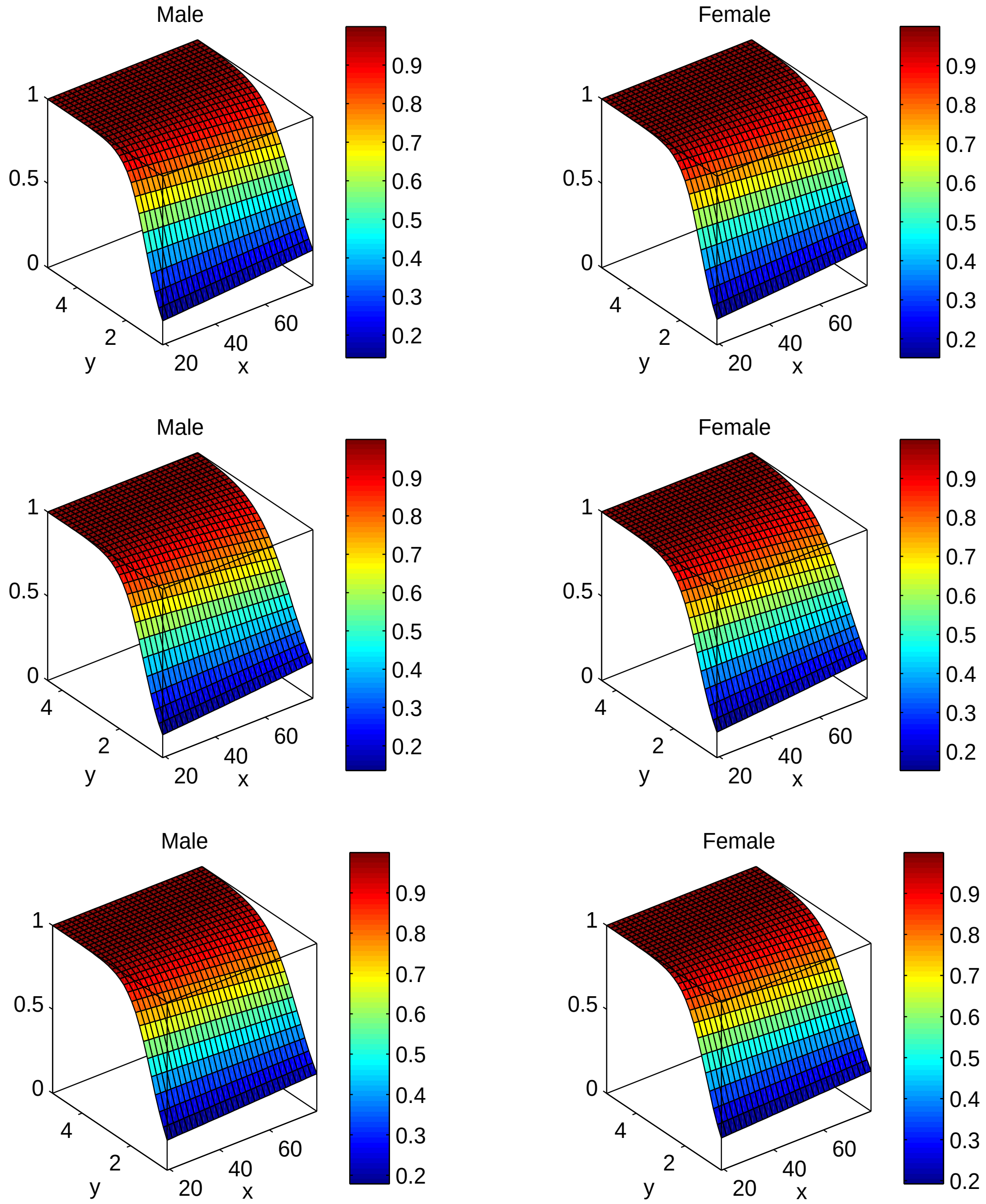

Figure 5: The fitted probability surfaces for the left middle (top row), right middle (central row) and posterior (bottom row) artery systems. In each row, the fitted probability surfaces for the vessel segments from male subjects (left) and female subjects (right) are depicted respectively. In each subplot, the x-axis is the age, the y-axis is the arterial thickness, and the vertical axis is the fitted branching probability. From these plots we can see that arteries become more likely to branch as they become thicker.

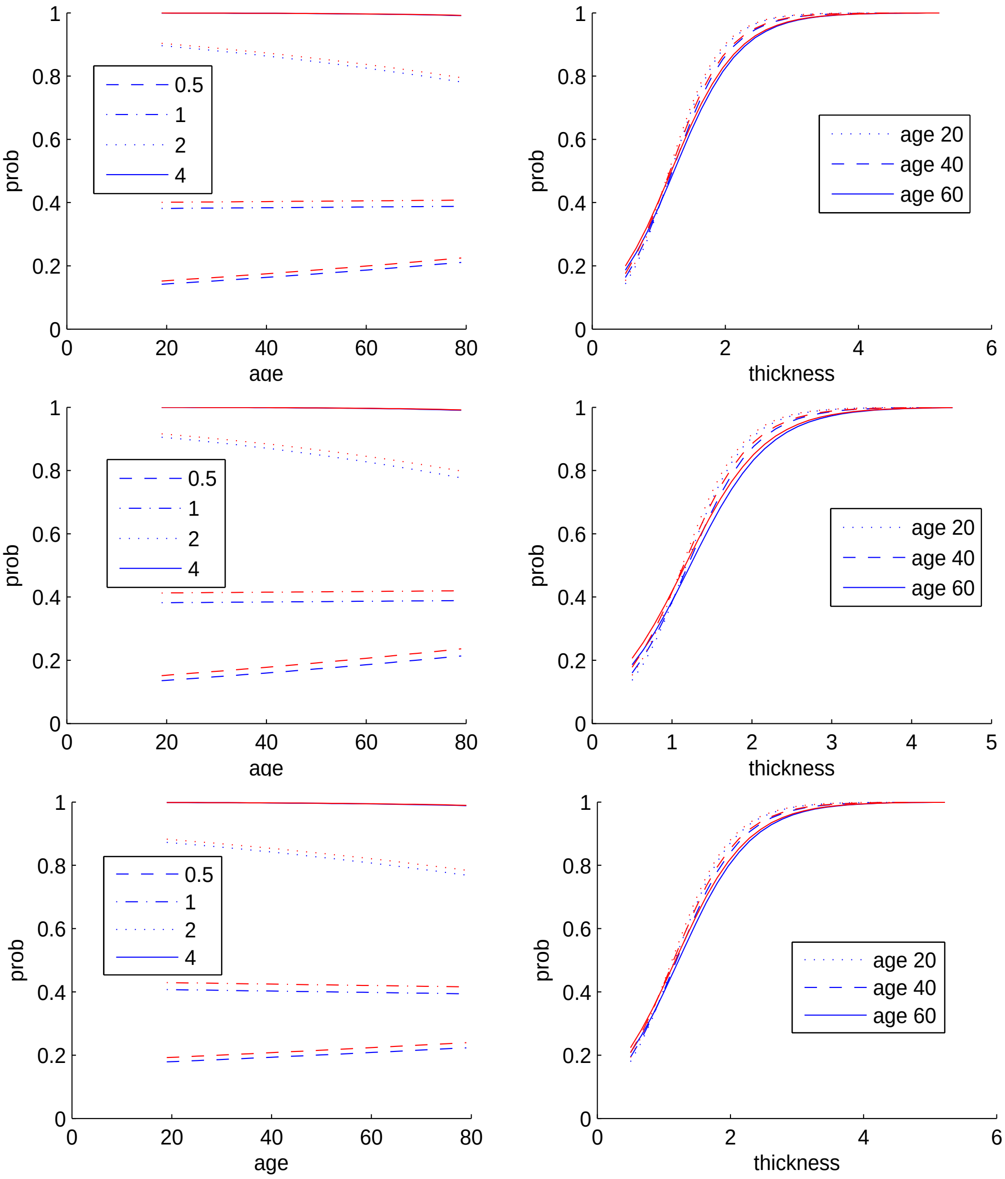

Figure 6: The fitted branching probability on age (the left column) and thickness (the right column) for the left middle (top row), right middle (central row) and posterior (bottom row) artery systems. In each subplot, red lines are for females and blue lines are for males. In the left panel of each row, the relationship between the branching probability (y-axis) and age (x-axis) is depicted for various choices of arterial thickness. We can see that as age increases, thicker arteries become less likely to branch while thinner vessel segments become more likely to branch, regardless of gender. In the right panel of each row, the relationship between the branching probability (y-axis) and thickness (x-axis) is depicted for various choices of age. It can be seen that as the arteries become thicker, they are more likely to branch. From both panels we can see that female subjects have slightly higher artery branching probability than that of males.

The Anterior Cerebral Artery System

Figure 7 shows that, as arterial thickness increases, the branching probability of arteries increases regardless of gender. In addition, given any level of thickness, brain arteries of female subjects are more likely to branch than those of male ones. Moreover, age doesn't have a significant impact on arterial branching in the anterior cerebral artery system in our dataset.

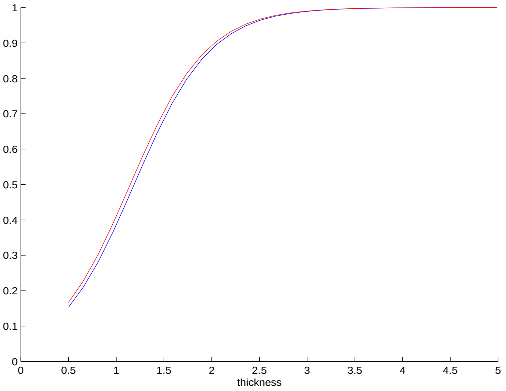

Figure 7: A graphical display of the relationship between the fitted branching probability (y-axis) and arterial thickness (x-axis) for the anterior cerebral artery system. The red line is for females and the blue line is for males. It can be seen that, regardless of gender, arteries become more likely to branch as they become thicker. For any given thickness level, arteries of female subjects are more likely to branch than those of males.

5 Conclusions and Discussion

The main contribution of this paper is to provide a rigorous framework for the analysis of the branching structure when the responses are tree-structured with binary branching outcomes. Together with some previous work [e.g. 7, 8], theories of statistical analysis on tree-structured objects can be made more complete. Also, while the model considered in this paper is for trees with binary branching outcomes, it can be extended to trees with more than two branching outcomes using polytomous logistic regression.

The logistic regression model framework for analyzing branching patterns in binary trees is an application of generalized linear mixed models. Thus the analysis can be carried out using existing software packages directly; see the R package `lme4` by [13] for more details.

This paper also provides a useful example of how information contained in medical images can be subjected to statistical analyses. In fact, [2] developed a fast and accurate method for constructing the 3-D images of vessels, but the analysis of these 3-D images has become a great challenge. This paper not only provides a useful approach for studying the topological structure behind these 3-D brain artery images, but also provides insights into how gender and age might affect our brain artery systems. These results might help in advancing research in theories of human brains, and further efforts in integrating medical image developments with appropriate statistical approaches might expedite advancements of medical theories as well.

For the particular dataset analyzed in this paper, we conclude that, except for the anterior cerebral arteries, the other cerebral artery systems will be affected by age. In the left middle, right middle, and posterior cerebral artery systems, thick arteries become less likely to branch as age increases. This is consistent with the results found in [6], which showed healthy aging is associated with brain vessel losses. However, our statistical analysis also suggests that as age increases, thin arteries become more likely to branch, but the branching probability is still lower than that of thick arteries. One explanation to this might be that, for older people, vessels can get thinner due to blockage or other causes, so it could be that these splits were in vessels that were previously thicker, and the thickness diminished over time. Another possible explanation is that as one gets older, the brain becomes more “specialized”, and so more branches from thinner arteries are needed to take care of one's strong brain areas, while artery segments in one's weak areas die out. Also, interestingly, in all of the four major cerebral artery systems, males have smaller probability of arterial branching than females, though the difference appears to be small. While there are many possible explanations to this phenomenon, it suggests that gender differences exist even in brain artery topology.

The brain artery dataset also contains the handedness and ethnicity information of each subject. However, the variable selection procedure trims out these variables suggesting that right-handed, left-handed and ambidextrous people do not differ much with respect to brain arterial branching, as are people from different ethnic groups. However, this could be just due to the fact that sample sizes for left-handed, ambidextrous and non-white subjects are relatively small.

Acknowledgments

The authors thank Dr. Ladha and Dr. Aydin for their help with postprocessing the brain artery dataset. The research of Hari Iyer was partially supported by the National Science Foundation, DMS-0707037 and DMS-1007542, and Caterpillar, Inc. The research of Elizabeth Bullitt was partially supported by NIH grants R01EB000219-NIH-NIBIB and R01 CA124608- NIH-NCI. The research of Haonan Wang was partially supported by the National Science Foundation, DMS- 0706761 and DMS-0854903. The MR brain images from healthy volunteers used in this paper were collected and made available by the CASILab at The University of North Carolina at Chapel Hill and were distributed by the MIDAS Data Server at Kitware, Inc.

References

[1] A.G. Osborn and K.A. Tong. *Handbook of Neuroradiology: Brain and Skull*. Mosby-Year Book, Inc., second edition, 1996.

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

[3] J. Xu, S. Kobayashi, S. Yamaguchi, K. Iijima, K. Okada, and K. Yamashita. Gender effects on age-related changes in brain structure. *American Journal of Neuroradiology*, 21(1):112–118, 2000.

[4] C.D. Good, I.S. Johnsrude, J. Ashburner, R.N.A. Henson, K.J. Friston, and R.S.J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*, 14(1):21–36, 2001. Part 1.

[5] K. Amunts, L. Jancke, H. Mohlberg, H. Steinmetz, and K. Zilles. Interhemispheric asymmetry of the human motor cortex related to handedness and gender. *Neuropsychologia*, 38(3):304–312, 2000.

[6] E. Bullitt, D. Zeng, B. Mortamet, A. Ghosh, S.R. Aylward, Lin W., Marks B.L., and Smith K. The effects of healthy aging on intracerebral blood vessels visualized by magnetic resonance angiography. *Neurobiology of Aging*, 31(2):290–300, 2010.

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

[8] B. Aydin, G. Pataki, H. Wang, E. Bullitt, and J.S. Marron. A principal component analysis for trees. *The Annals of Applied Statistics*, 3:1597–1615, Oct 2009.

[9] R. Diestel. *Graph Theory.* Springer, third edition, 2005.

[10] D.E. Knuth. *The Art of Computer Programming Volume 1. Fundamemtal Algorithms.* Addison-Wesley Longman, Inc., third edition, 1997.

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

[12] E. Bullitt, J.K. Smith, and W. Lin. Internet site: http://hdl.handle.net/1926/594, 2008.

[13] D. Bates and M. Maechler. Internet site: http://cran.r-project.org/web/packages/lme4/index.html; see the reference manual and the vignettes, 2009.

Dataset	Left		Right		Back		Front	
predictors:								
	coef.	p-value	coef.	p-value	coef.	p-value	coef.	p-value
intercept	0.106	<0.001	0.135	<0.001	0.108	<0.001	0.087	0.006
age (standardized continuous variable)	-0.041	0.045	-0.045	0.025	-0.047	0.011		
thickness (standardized continuous variable)	0.861	<0.001	0.873	<0.001	0.779	<0.001	0.764	<0.001
age×thickness	-0.083	0.068	-0.096	0.031	-0.063	0.123		
male (indicator for male)	-0.082	0.044	-0.129	0.001	-0.091	0.014	-0.102	0.021
$\sigma_{\text{intercept}} =$	0.000		0.000		0.000		0.000	
$\sigma_{\text{thickness}} =$	0.351		0.346		0.322		0.329	

Table 1: The estimated regression coefficients and p-values for the four brain artery systems. From this table, we can see which variables are important in each part of the brain. For continuous variables (notice that they are standardized to have mean 0 and standard deviation of 1), if the coefficient is greater than 0, then as the value of the variable becomes larger, the branching probability becomes larger, and vice versa. For indicator variables, if the coefficient is less than 0, then the branching probability for this category of subjects is smaller than that for subjects not in the category. The presence of the interaction term age×thickness reflects the fact that the effects of age on branching probability differ across arteries of different thicknesses.