The Annals of Statistics
2018, Vol. 46, No. 4, 1513–1540
https://doi.org/10.1214/17-AOS1593

PARETO QUANTILES OF UNLABELED TREE OBJECTS

BY ELA SIENKIEWICZ AND HAONAN WANG[1]

Colorado State University

In this paper, we consider a set of unlabeled tree objects with topological and geometric properties. For each data object, two curve representations are developed to characterize its topological and geometric aspects. We further define the notions of topological and geometric medians as well as quantiles based on both representations. In addition, we take a novel approach to define the Pareto medians and quantiles through a multi-objective optimization problem. In particular, we study two different objective functions which measure the topological variation and geometric variation, respectively. Analytical solutions are provided for topological and geometric medians and quantiles, and in general, for Pareto medians and quantiles, the genetic algorithm is implemented. The proposed methods are applied to analyze a data set of pyramidal neurons.

1. Introduction. When studying functional aspects of the brain, the hippocampus region is of particular interest. It is associated with long term memory and learning, and it is highly sensitive to pathological changes (e.g., disease, brain injuries). There is an ongoing effort to understand the dynamic behavior of hippocampal neuron cells, specifically their connectivity and firing activity (also known as spike trains). The information transmission between two regions of the hippocampus, CA3 (input) and CA1 (output), has been extensively modeled in an effort to develop, among others, a neural prosthesis [36]. Morphological aspects of neurons in these two regions have also been studied; see Johnston and Wu [18], Vida [37], Migliore and Shepherd [23] among others. Neural morphology is an important determinant of neural functions. Different types of neurons, or even the same type of neurons from different brain regions, show distinct forms of morphologies.

Pyramidal neurons from the hippocampus typically consist of a soma, an axon and two types of dendrites (see Figure 1). The tree-like dendritic structures, also referred to as arborizations, are commonly associated with the functional complexity of the brain. The current "synaptotropic hypothesis," as stated in Cline and Haas [10], describes the growth of dendritic branches as "dynamic and exploratory." The branches can live for as short as 10 minutes, as they "sample the environment to detect the appropriate cells" [10]. This dynamic process cannot be directly

Received November 2016; revised March 2017.
[1]Supported in part by NSF Grants DMS-1106975 and DMS-1521746.
MSC2010 subject classifications. Primary 62G99; secondary 62P10.
Key words and phrases. Data object, genetic algorithm, multi-objective optimization, object oriented data, tree-structured data.

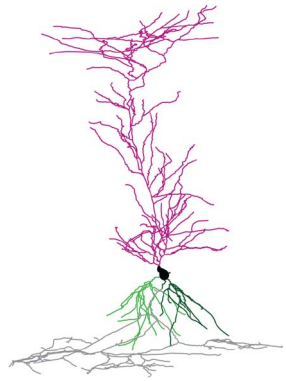

FIG. 1. *Graphical display of a pyramidal neuron cell, named after its pyramid-like shape. All arborizations grow out of the* soma, *which is depicted in black. Other components include the* axon *shown in grey*, apical dendrites *shown in magenta and* basal dendrites *shown in green* (*two shades of green are used to depict two disjoint arborizations*). *The basal dendrites often form a forest of several disjoint binary trees. The axon is ignored in our analysis.*

observed, and the data available for analysis usually provide one snapshot in the lifetime of a neuron. However, mathematical models capable of quantifying and generating neural morphologies are needed; see Ascoli et al. [4], Hendrickson et al. [17]. Given a set of static reconstructions of neural cells with different topological and geometric properties, our goal is to characterize the distribution of these properties in the population of neurons, and thus provide a quantitative description of neural morphology. This could potentially lead to the establishment of a novel bio-marker for diagnosis of various neural diseases and damages. It can also serve as a tool that identifies subtle differences between morphologies under normal and abnormal conditions, which may potentially enable the disease diagnosis at early phases. For instance, changes in neural structure were observed in degenerative brain disorders, such as Alzheimer's and Parkinson's diseases; see [15, 24, 34] for details.

In statistical modeling, each neuron can be regarded as a *data object*, a complex entity that is generally outside the scope of classical statistics. The class of data objects can include images, trees, graphs and often curves; see Marron and Alonso [22] for a recent review of objects and related statistical methods. The term *Object Oriented Data Analysis*, a class of tools for the analysis of complex data objects, was introduced to statistics by Wang and Marron [39]. Since then, there has been a great deal of research to extend traditional statistical methods, for example, regression and principal component analysis, to the space of complex data objects [9, 31, 40].

In classical statistics, descriptive measures, such as mean and deviation from the mean, have been widely used to describe and summarize information from data. But those statistics may not be sufficient to highlight the characteristics of complex

data objects. For instance, neurons from different brain regions exhibit topological "heterogeneity." The multitude of shapes, sizes and branching patterns observed in neural cells calls for a more comprehensive depiction of the population distribution. For a univariate random variable, a comprehensive characterization can be established through a quantile function, which provides an intuitive, probabilistic way to measure centrality, dispersion, skewness and the tail behavior of the distribution. In particular, the quantile function is defined through the cumulative distribution function. However, the definition of quantiles becomes nontrivial for a multivariate random vector due to the lack of a natural order in high-dimensional space. Liu [20] introduced the notion of *simplicial depth* and showed that it can be used as an analog of multivariate order statistics. Serfling [30] provided a survey of different approaches to multivariate quantile definitions and useful criteria for their evaluation. The most notable methods are based on depth function and norm minimization. Functional data provide even more challenges, because standard approaches for a finite dimension do not translate well to functional space. Walter [38] offered a thorough study of the properties of functional quantiles and their empirical analogs backed up by a case study of financial data. In that study, the author employed pointwise quantiles, which are biased estimators of population quantiles, but they are consistent under some weak conditions.

The challenges increase even more for complex data objects, such as tree data, which can be characterized as extremely non-Euclidean; see Wang and Marron [39]. There have been previous attempts to define a median of a population of such objects. Some examples come from the work on classification trees [6, 25]. The tree-structured data objects discussed in these papers are of a binary form, and their nodes can be uniquely labeled for correspondence between trees. The median tree is thus defined as a majority tree, that is, a tree consisting of nodes found in the majority of trees in the set. Node labels are important and natural for classification trees or phylogenetic trees [7], but for some tree-structured objects, for example, brain arteries or neural dendrites, there is no established labeling scheme. The labeling of nodes can be crucial in answering many important research questions, but different labeling choices could lead to different results; see Aydin et al. [5] for a discussion on thickness correspondence and descendant correspondence between brain artery systems.

In this paper, we propose a novel approach to evaluate quantiles of tree objects that does not rely on labeling of nodes or edges. We base our approach on a stochastic process view of a tree, which can be interpreted as a birth and death process. The connection between a tree and a stochastic process has been examined before. For instance, Harris [16] studied curves generated by the depth-first traversal of trees. Such curves were instrumental in producing asymptotic results, for example, related to a convergence of a stochastic process, but are not very well suited for comparing trees [31].

As noted by Wang and Marron [39], a tree-structured data object can have both topological and geometric attributes. Topological attributes can be described generally as branching patterns, for example, the number of nodes at any specific level.

Geometric attributes could include distances between nodes, radiuses of edges or angles between edges. In this paper, we focus our attention on the length of edges. Here, we propose two functional representations of each tree-structured object encompassing its topological and geometric properties, respectively. We define a quantile of tree objects by taking both properties into account; in particular, the quantile can be formulated as a solution of a multi-objective optimization problem. We also find empirical quantiles of tree distributions using a genetic algorithm.

This paper is organized as follows. In Section 2, we introduce two new functional representations for each unlabeled tree-structured object, which summarize the topological and geometric properties. In Section 3.1, we define topological and geometric median trees through optimization problem. In Section 3.2, we introduce a novel notion of a Pareto median tree object as solutions to the multi-objective optimization problem. Next, in Section 3.3, we extend this idea to define Pareto quantiles of tree objects. In Section 3.4, we discuss the genetic algorithm used to find the solutions of the optimization problems. In Section 4, we examine our proposed method through a simulation study. Finally, Section 5 provides a case study of a set of neurons using our proposed methods. The proofs and additional data analysis results, as well as details about simulation strategies, are included in the Supplementary Material [33].

2. Data object and its curve representation.

2.1. *Data*. In this paper, our motivating example is a set of neuron cells from the brains of rodents. The original dataset consists of digital reconstructions of neurons obtained from the online inventory site neuromorpho.org [2], which includes more than 8000 neurons from various brain regions. For details on the data and data collection process, see Pyapali et al. [27], Pyapali and Turner [28, 29]. Our primary interest centers on pyramidal neurons from two areas of the hippocampus, regions CA1 and CA3. Here, a set of $n = 187$ pyramidal neurons, including 119 and 68 from CA1 and CA3 regions, respectively, is used. It is known that neurons from the CA3 region receive input signals from other cells in the brain, while neurons from the CA1 region form the output from the hippocampus.

A sample of neurons from CA1 and CA3 regions is shown in Figure 12. Each subplot depicts a pyramidal neuron which has three major components, apical dendrites (colored in magenta), basal dendrites (colored in green) and a soma (colored in black) in between. The soma can be a single point or a line, and it is very small compared with apical and basal dendrites. In addition, the top row of Figure 12 shows three neurons from CA1, and the bottom row shows three neurons from CA3. From all six subplots, the basal dendrites seem to be shorter than the apical dendrites; whereas the difference in branching of both groups of neurons is apparent. In particular, the initial segments of apical and basal dendrites are shorter for CA1 neurons than those of CA3 neurons, and basal structures of CA3 neurons are larger than those in CA1 neurons.

In Section 2.2, we will discuss a tree representation of dendritic structures and further propose two new curve representations in Section 2.3.

2.2. *Graph as a data object*. In mathematical graph theory, a *tree* is a simple graph with a set of nodes and edges, and there is a unique sequence of edges between any two nodes. A *forest* is a collection of trees. For any *rooted tree*, the *root* is a specific node which can be designated based on the application. The level of a node is the number of edges of the path to the root node. For any two adjacent nodes connected by an edge, the node that is closer to the root node is called a parent node, and the other node is a child node. A node with no children is called a leaf node or a terminal node. For a tree object, if each node has at most two children, namely a left child and a right child, it is called a *binary tree*, and, if it has exactly two children, it is called a *full binary tree*.

In many scientific applications, binary trees have been used to model tree-structured objects. For instance, Wang and Marron [39] proposed to use binary trees to represent human brain blood vessel systems. In our study, as can be seen in Figure 1, the apical dendrites emerge from the apex of a soma, and branch like a single tree. Basal dendrites are somewhat different; in general, several dendritic trees grow out of the base of a soma and form the basal dendrites. Here, we model the apical dendrites as a binary tree and the basal dendrites as a forest of binary trees. The term *forest* is referring to a disjoint union of binary trees. The construction of tree objects from the data is discussed in Section A of the Supplementary Material [33]. The procedure is straightforward, but ambiguity may arise when identifying the left and right child nodes. Most recent work on tree-structured objects has focused on sets of labeled trees. The term *labeled tree* refers to a tree in which each node has a well-specified label. In practice, as suggested by Aydin et al. [5] and followed by Wang et al. [40], two approaches can be considered to establish a labeling system, namely, *thickness correspondence* and *descendant correspondence*. In the first approach, at each split point the thicker dendritic segment is denoted as the left child node of its parent node. Alternatively, in the second approach the dendritic segment with more subsequent segments is denoted as the left child node. These two approaches may lead to two distinct curves; see an example in Section B of the Supplementary Material [33]. The distinct resulting data objects may potentially lead to different scientific conclusions.

In this paper, our main focus is a set of rooted unlabeled trees or forests which has not been addressed before.

2.3. *Curve representations for unlabeled trees and forests*. For labeled binary trees, dyadic tree representation provides an intuitive way to visualize the topological property. Such representation may not be suitable to depict a sample of tree-structured objects due to space limitation. In probability literature, tree-structured objects are usually modeled as branching processes. Harris [16] established a correspondence, called *Harris correspondence*, between trees and random

walks; see Section B of the Supplementary Material [33] for examples. The Harris path provides insightful information regarding the topological property of a single tree-structured object. An alignment issue arises when comparing Harris paths obtained from a set of trees. Shen et al. [31] proposed the modified Harris path and the *branch length representation* (BLR) to overcome this problem. Those authors conducted principal component analysis on the set of Dyck paths and the set of branch length curves, and certain important scientific findings have been reported. The success of their approach relies on the descendant correspondence and the corresponding labeling system of binary trees. However, we might reach different conclusions using the same data and different types of correspondence. This issue becomes even more serious when the data objects are forests. When comparing two forests with different numbers of tree components, a well-defined order is usually not available. To circumvent this problem, we propose two new tree/forest representations, which are independent of the choice of correspondence and the labeling system. Moreover, certain nodal attributes, for example, the length of segments, can also be incorporated.

For a rooted tree, we introduce a function $g(x)$, $x \in [0, \infty)$ defined as the number of distinct segments at distance x from the root. An illustrative example is given in panel (C) of Figure 2. Such function $g(x)$ provides a *geometric curve representation* of a tree-structured object. Note that $g(x)$ is a piecewise constant function with $g(0) = 0$ and $g(\infty) = 0$. In particular, $g(x)$ is left continuous on $(0, \infty)$. If we further assume that *no two segments start or end at the same distance from the root*, the size of a jump is either 1 or -1. One can also notice, in panel (C), that the number of jumps in the range $(0, \infty)$ represents the number of nodes in the tree, a positive jump corresponds to an internal node, a negative jump corresponds to a leaf.

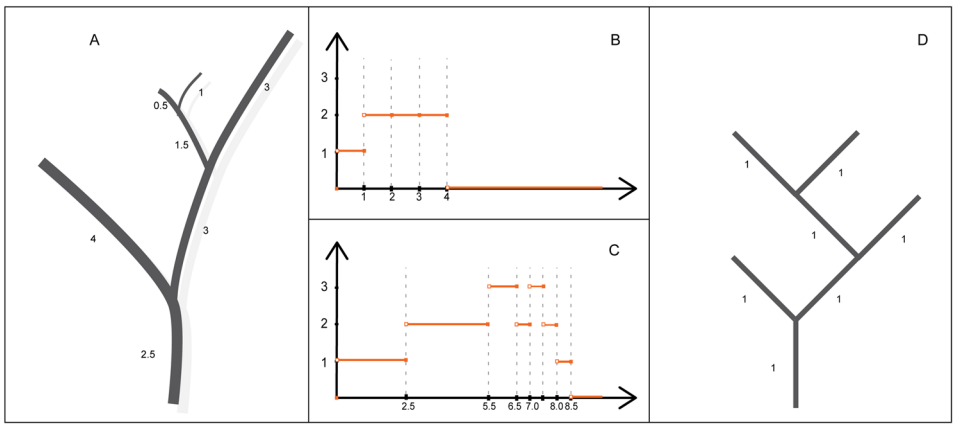

FIG. 2. *An example of a tree object* (A) *and its corresponding length-scaled tree* (D). *There are two curve representations of the tree in* (A): *the topological curve* (B) *and the geometric curve* (C).

Next, to describe the topological property, we introduce the *length-scaled* tree for tree-structured data object. That is, we assume all segments have length equal to 1. An illustrative example is given in panel (D) of Figure 2, which is the length-scaled tree for the tree object depicted in panel (A). The topological tree representation, denoted by $\ell(x)$, $x \in [0, \infty)$ is defined as the number of distinct segments at distance x from the root in the corresponding length-scaled tree; see panel (B) of Figure 2. The number of nodes in the tree can be retrieved as $\sum_{i=1}^{\infty} \ell(i)$, which is always an odd number. In contrast to the Harris path, the tree curve mimics the *breadth-first search* algorithm in graph theory in the sense that we would like to count the number of branches at any given radius x. For a forest with k distinct binary trees, the curve representation is defined as the sum of $g^{(1)}(x), \ldots, g^{(k)}(x)$, where $g^{(i)}(x)$ is the curve representation associated with the ith tree. Note that such representations depend only on the counts of branches at a distance from the root, so they can be extended to more general rooted trees or forests.

In our study, we will represent both apical and basal dendrites using tree curves. For our convenience, we will display a *joint tree curve* for both apical and basal dendrites. Specifically, we show a tree curve for the apical dendrites and the mirror-view of a tree curve for the basal dendrites in one plot. An example of joint tree curves is given in Figure 3. Here, the raw data is depicted in the upper panel, the

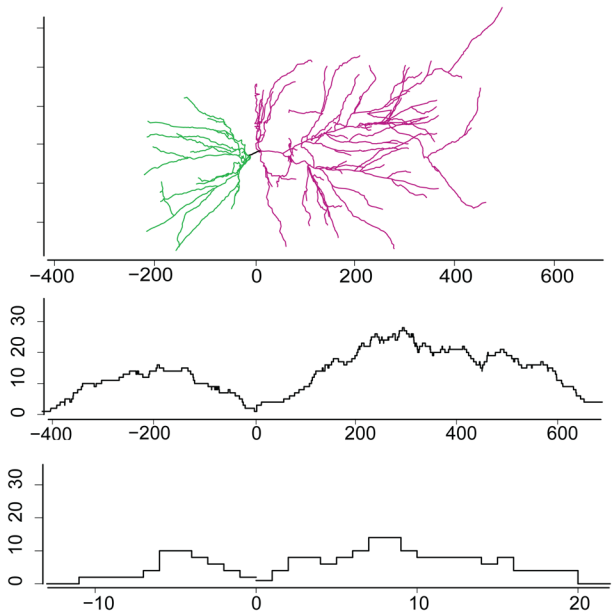

FIG. 3. *A graphical display of a joint geometric tree curve* (*middle*) *and a joint topological* (*bottom*) *for the corresponding neuron object* (*top*) *with apical dendrites* (*colored in magenta*) *and basal dendrites* (*colored in green*).

corresponding (joint) geometric and topological tree curves are depicted in the middle and the bottom panels.

2.4. *Equivalence classes of topology and geometry.* For each tree or forest, the geometric tree curve provides a functional representation which, in fact, is *not* a one-to-one mapping from the space of binary trees or forests, labeled or unlabeled, to the space of (piecewise constant) functions. In Figure 2, given a tree curve $g(x)$ in panel (C), we can reconstruct a tree; however, such reconstruction is generally not unique. Two trees, say t_1 and t_2, are geometrically *equivalent* if they have the same geometric tree curve, and hence can be written as $t_1 \overset{\mathrm{G}}{\sim} t_2$. The geometric equivalence class of tree t is the set of trees that are equivalent to t and is denoted by $[t]_{\mathrm{G}}$. Analogously, we define the topological equivalence class of a tree t, and denote it by $[t]_{\mathrm{T}}$. All trees in $[t]_{\mathrm{G}}$ and $[t]_{\mathrm{T}}$ have the same number of nodes, which is equal to $2m_t + 1$, including m_t internal nodes and $m_t + 1$ leaves.

Next, we will define an operation, called *implant*, for trees and forests. In particular, for any tree (or forest) t, an *implant* of t is defined by swapping any two subtrees at the same distance from the root. Note that, for topological equivalence, the level plays a role of a distance. It can be seen that two trees (forests) are equivalent if and only if one tree (forest) can be obtained by a sequence of implant operations from the other. Thus, there may not be a unique tree reconstruction from a tree curve, or even from geometric and topological curves combined. In this paper, we often reconstruct a tree with the procedure as described in Section E.1 of the Supplementary Material [33].

3. Methodology.

3.1. *Median trees and L_1 distance.* The notion of *median* tree has been previously studied by Phillips and Warnow [25], Banks and Constantine [6] and Wang and Marron [39]. Median trees have been developed for classification trees [6] and phylogenetic trees [25]. Wang and Marron [39] took the first step to consider a set of tree-structured objects motivated by medical imaging analysis. In particular, for a sample of labeled binary trees, $t_1, \ldots, t_n$, the authors proposed a (topological) median as the minimizer tree of

$$\min_t \sum_{i=1}^{n} d_I(t, t_i), \tag{3.1}$$

where d_I is the *integer tree metric* defined in (3.1) of Wang and Marron [39] for labeled binary trees. This metric is, essentially, the cardinality of a symmetric difference of two sets of node indices. This notion of center point in tree space can be viewed as a special case of the Fréchet median [40]. For general metric space, Fréchet [13] proposed to define the center point, or Fréchet median, as the minimizer of (3.1) for any given metric.

In Section 2.3, two curve representations, topological and geometric tree curves, have been proposed for unlabeled trees. Consequently, an intuitive idea to measure the distance between two unlabeled trees is to use the L_1 metric between the corresponding curves. Note that each equivalence class has a unique curve representation. Thus the L_1 metric between tree curves in fact provides a distance between equivalence classes of trees.

Let $\mathcal{T}$ be the collection of all full labeled binary trees, as considered in Wang and Marron [39]. We further assume that no two segments start or end at the same distance from the root.

First, we will consider topological tree curves. Note that the equivalence classes form a partition of $\mathcal{T}$. For any two trees $s, t \in \mathcal{T}$ with topological tree curves $\ell_s(x)$ and $\ell_t(x)$, the distance between the equivalence classes $[s]_{\mathrm{T}}$ and $[t]_{\mathrm{T}}$ is defined as

$$d\big([s]_{\mathrm{T}}, [t]_{\mathrm{T}}\big) = \|\ell_s - \ell_t\|_1 \equiv \int_0^\infty \big|\ell_s(x) - \ell_t(x)\big|\,dx. \tag{3.2}$$

Such distance metric is independent of any labeling. Theorem 3.1 establishes the connection between the L_1 distance in (3.2) and the integer tree metric of [39].

THEOREM 3.1. *For any two trees $s, t \in \mathcal{T}$, we have*

$$d\big([s]_{\mathrm{T}}, [t]_{\mathrm{T}}\big) = \min_{s' \in [s]_{\mathrm{T}}, t' \in [t]_{\mathrm{T}}} d_I\big(s', t'\big). \tag{3.3}$$

From now on, let $\{t_1, \ldots, t_n\}$ be a random sample of trees, and let $\ell_i(x)$ be the topological curve representation of t_i. Similar to (3.1), we can formulate the median tree through an optimization problem described as

$$\operatorname*{argmin}_{\ell} \sum_{i=1}^{n} \|\ell - \ell_i\|_1, \tag{3.4}$$

where $\ell(\cdot)$ runs over the collection of topological tree curves.

If we relax the constraint in (3.4) and consider all possible functions $\ell(\cdot)$, the solution is the pointwise median function, that is, $m_0(x) = \text{median}\{\ell_1(x), \ldots, \ell_n(x)\}$. When n is odd, such pointwise median function is always unique. When n is even, the pointwise median function may not be unique for some x, and $m_0(x)$ takes the smallest value to break the tie. In Theorem 3.2, we will prove that such pointwise median function $m_0(x)$ corresponds to an equivalence class in which all elements are called topological median trees.

THEOREM 3.2. *Assume that $\{t_1, \ldots, t_n\}$ is a sample of trees with finite levels, that is, the number of edges to the root node. Let $\ell_i(x)$ be the topological curve representation of t_i. The pointwise median $m_0(x)$ corresponds to an equivalence class of trees, and hence is the minimizer of* (3.4).

Our primary interest is a sample of trees with nodal attributes, for example, the lengths of dendritic segments. In the literature, for trees with nodal attributes, Wang and Marron [39] proposed a median-mean tree whose topology is determined by the topological median and whose nodal attributes can be obtained by averaging corresponding nodal attributes. For a set of unlabeled trees, their notion of "median-mean" cannot be generalized. In this paper, enlightened by (3.4), for a sample of trees $t_1, \ldots, t_n$ with geometric tree curves $g_1(x), \ldots, g_n(x)$, respectively, the geometric median tree can be defined through

$$\operatorname*{argmin}_{g} \sum_{i=1}^{n} \|g_i - g\|_1, \tag{3.5}$$

where $g(\cdot)$ runs over all possible geometric tree curves. Similar to Theorem 3.2, we will show that the pointwise median, denoted as $m_1(x)$, is a geometric tree curve.

THEOREM 3.3. *A pointwise median of a finite sample of geometric tree curves $g_i(x)$ represents a valid tree class.*

To better illustrate the topological and geometric median trees, we will consider two examples, as shown in Figures 4 and 5. In each figure, a sample of three tree-structured objects are depicted in the top row (panels A–C). In Figure 4, the three

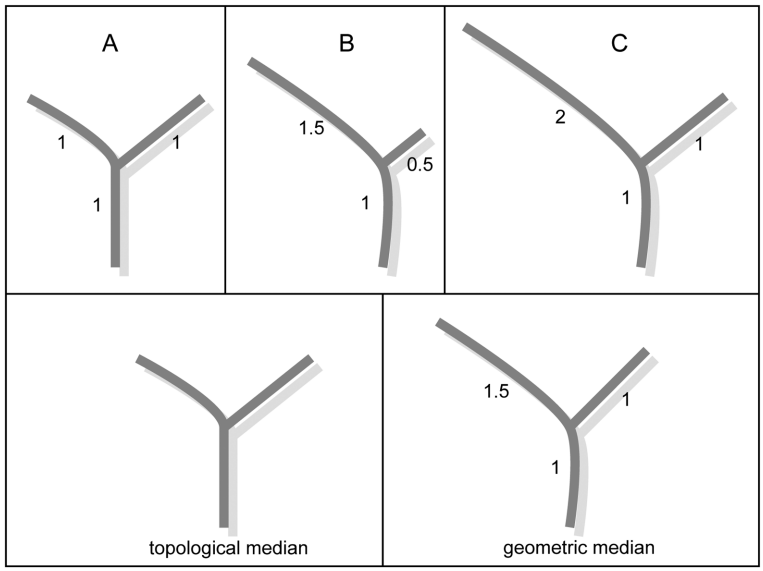

FIG. 4. *A graphical display of the topological median (lower-left panel) and the geometric median (lower-right panel) of a sample of three tree-structured objects (top row). The number associated with each branch segment is the segment length, and is referred to as a geometric attribute. Here, both median trees have the same topological structure with three branch segments.*

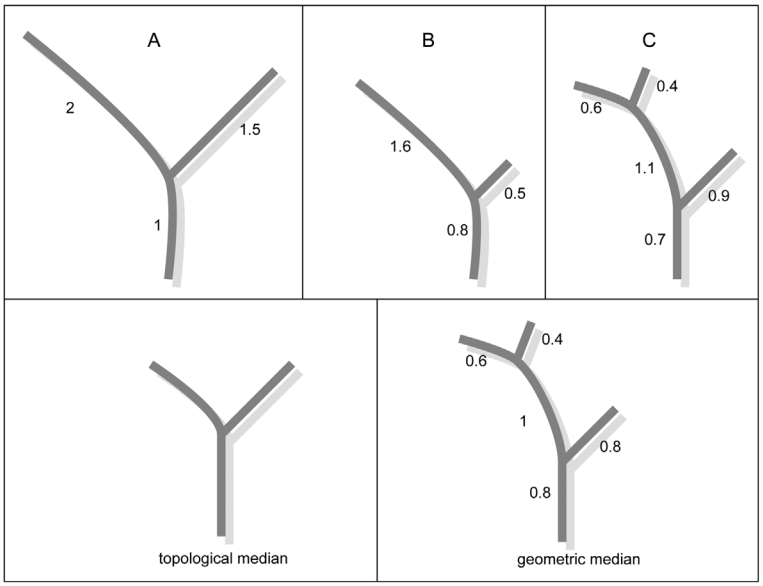

FIG. 5. *A graphical display of the topological median and the geometric median of a sample of three tree-structured objects. Here, the topological median and the geometric median have different topologies.*

trees have the same topological structure, including one root segment and two offspring segments among which one is relatively longer than the other one. The topological and geometric median trees are displayed in the lower-left and lower-right panels. It can be seen that the topological median also has the same topology as all three trees. From the geometric median tree, it can be seen that two offspring segments have unequal length. In Figure 5, trees A and B have the same topology, and tree C has more segments than the other two trees. Surprisingly, the topological median and geometric median have different tree structures. The reason is that the topological median only characterizes the centrality of topological properties, while the geometric median tree is influenced by the length of segments.

In the next section, we will introduce a new notion of median, called Pareto median, which will take both topological and geometric information into consideration.

3.2. *Pareto median trees—a multi-objective approach.* We continue to let $\{t_1, \ldots, t_n\}$ be a random sample of trees. Let $\ell_i(x)$ and $g_i(x)$ be the topological and geometric curve representations of a tree t_i, respectively.

In panel (A) of Figure 6, a sample of 21 trees is depicted. All trees have the same simple topology, a trunk and two branches, and randomly generated geometric attributes. The topological and geometric median trees are shown in panels (B) and (C), respectively. It can be seen that the geometric median has a more complex

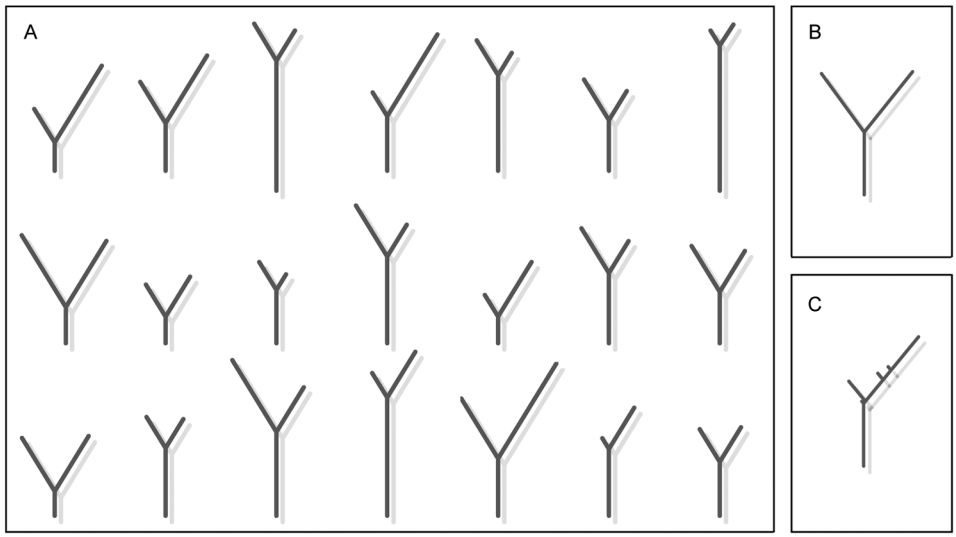

FIG. 6. (A) *Graphical display of a sample of* 21 *simulated trees with the same topology and different geometric attributes*; (B) *topological median tree*; (C) *geometric median tree*.

topological structure than the topological median. The complexity of the geometric median reflects the diversity of geometric attributes of the tree set. By contrast, the topological median is the manifestation of the topological homogeneity of the data. Preferably, we would like to find a median tree that takes into account both topological and geometric attributes together. In other words, we would like to find a tree to minimize both (3.4) and (3.5) simultaneously, which, in fact, is a multi-objective optimization problem. Mathematically, it can be formulated as

$$\operatorname*{argmin}_{t} \big(T_n(t), G_n(t)\big), \tag{3.6}$$

where t runs over the space of binary trees,

$$T_n(t) = \sum_{i=1}^{n} \|\ell_t - \ell_i\|_1 \quad \text{and} \quad G_n(t) = \sum_{i=1}^{n} \|g_t - g_i\|_1.$$

Here, $\ell_t(\cdot)$ and $g_t(\cdot)$ are the topological and geometric tree curves of t, respectively.

In multi-objective optimization, there is no guarantee of the existence of a solution which minimizes both $T_n(t)$ and $G_n(t)$. An alternative is the Pareto optimum; see Coello Coello, Lamont and Van Veldhuizen [11] for a formal definition. A Pareto set contains all feasible solutions such that there is no other solution that improves one of the criteria without worsening another. In other words, a Pareto optimal set is a set of feasible solutions which are not dominated by any other solution.

In our problem considered here, for any two trees s and s', s' is *dominated* by s if $T_n(s) \leq T_n(s')$ and $G_n(s) \leq G_n(s')$, and at least one inequality is strict. In addition, a tree is *Pareto optimal* or a *Pareto median tree* if it is not dominated by any other trees. Let $\mathcal{P}$ be the collection of all Pareto median trees. There are two trivial Pareto median trees, namely, the topological Pareto median and the geometric Pareto median trees. For the topological Pareto median, we minimize $G_n(t)$ among the subclass of trees whose topological tree curves minimize $T_n(t)$. On the other hand, for the geometric Pareto median, we minimize $T_n(t)$ among the subclass of minimizer trees of $G_n(t)$. If a common solution exists, it is called an *ideal* tree.

Recall that, in Section 3.1, the topological median trees and the geometric median trees are defined by minimizing $T_n(t)$ and $G_n(t)$, respectively. The geometric medians and the geometric Pareto medians have the same geometric curve; however, in terms of topology, the former is less restrictive than the latter. This is due to the fact that the geometric Pareto median minimizes T_n within a subclass of trees possessing the same geometric tree curve. Similarly, the topological medians and the topological Pareto medians have the same topology, but the former can have any geometric properties or attributes, and the latter has the geometry that minimizes G_n in a subclass of trees possessing the same topological curve.

For the example in Figure 6, there are four elements in the Pareto optimal set. All four Pareto median trees are shown in Figure 7. In particular, the first tree (panel A) is the geometric Pareto median, and the last tree (panel D) is the topological Pareto median. In addition, for each Pareto median tree, the corresponding values of T_n and G_n are depicted as a point in Figure 7. In this example, there is a unique tree corresponding to each pair of T_n and G_n, but it is not the case in general.

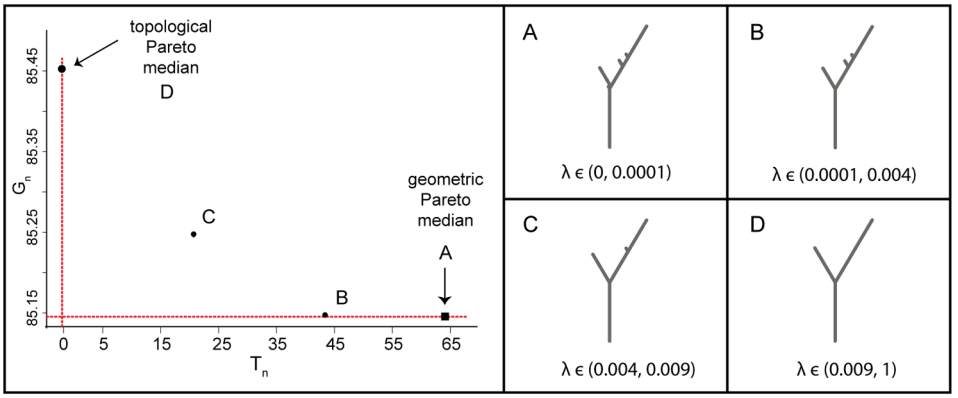

FIG. 7. *A graphical display of the T_n-values (horizontal axis) and G_n-values (vertical axis) for all Pareto median trees. The solutions form a* Pareto front. *In particular, the tree objects corresponding to those values are shown in panels* (A)–(D). *Each Pareto median is also the solution of* (3.7) *with corresponding values of* λ.

The geometric Pareto medians can be found efficiently using convex optimization techniques; see Antoniou and Lu [1]. However, the number of topologies grows with the number of nodes in the tree. Finding an optimal topology can be viewed as a combinatorial problem of high dimension, which, for small and medium size trees, can be efficiently solved using a genetic algorithm. In general, computation for a multi-objective optimization problem (3.6) can be very complicated. However, as will be seen in Section 3.4, this problem can also be solved with a genetic algorithm.

In the literature, a widely used approach to multi-objective optimization is the *weighted-sum* method [11]. Specifically, consider the following optimization problem, for $0 < \lambda < 1$,

$$\min_t \lambda T_n(t) + (1-\lambda) G_n(t). \tag{3.7}$$

This criterion is a linear combination of T_n and G_n. The solution takes both topological and geometric information into consideration. Moreover, note that the solutions of (3.7) are Pareto optimal of (3.6), hence are Pareto median trees. For a preselected λ, (3.7) is a single-objective optimization problem, which can be efficiently solved using a standard genetic algorithm. In addition, a range of λ may correspond to a single Pareto optimal solution. This can be observed in Figure 7, where for each of four Pareto median trees, the corresponding range of λ is specified. In particular, the topological Pareto median corresponds to the largest values of λ, and the geometric Pareto median corresponds to the smallest values of λ. However, the multi-objective optimization in (3.6) may not be equivalent to the single-objective optimization in (3.7); that is, some Pareto optimal trees may not be solutions of (3.7) regardless of the choice of λ. In the literature, given a preselected λ, the weighted sum method is often referred to as an *a priori* method, in contrast to the *a posteriori* method that finds many Pareto solutions and selects the best solution after the search is completed [11]. In this paper, given enough computing time, our approach will yield all or most of the Pareto solutions. More details are available in Section 3.4.

3.3. *Pareto quantiles of unlabeled trees.* In this section, we will extend the notion of Pareto median trees to *Pareto quantile trees*. To motivate our discussion, we first consider a random variable X. Finding the sample quantile of X based on a random sample $\{X_1, \ldots, X_n\}$ can be formulated as the optimization problem

$$\operatorname*{argmin}_x \sum_{i=1}^{n} \rho_\tau(X_i - x),$$

where $\rho_\tau(z) = z(\tau - I(z < 0))$ and $\tau \in (0, 1)$. See Koenker and Hallock [19] for more details.

Here, we consider a set of random tree objects rather than random variables. Enlightened by the problem above, we can generalize the formulation in (3.6) and define the Pareto quantiles through a multi-objective optimization problem; that is,

$$\operatorname*{argmin}_{t} \big(T_n^\tau(t), G_n^\tau(t)\big), \tag{3.8}$$

where

$$T_n^\tau(t) = \sum_{i=1}^{n} \int_0^\infty \rho_\tau\big(\ell_i(x) - \ell_t(x)\big)\, dx \quad \text{and}$$

$$G_n^\tau(t) = \sum_{i=1}^{n} \int_0^\infty \rho_\tau\big(g_i(x) - g_t(x)\big)\, dx.$$

In the special case of $\tau = 0.5$, this problem is equivalent to (3.6), and yields the Pareto median trees. Here as well, we take the smallest value to break the tie, thus ensuring the uniqueness of the quantiles. The formulation of the sample Pareto quantiles in (3.8) can be generalized to the population by replacing the finite summation with the expectation in $T_n^\tau(t)$ and $G_n^\tau(t)$.

Similar to the topological and geometric medians, we first minimize $T_n^\tau(t)$ and $G_n^\tau(t)$ individually to obtain topological and geometric quantiles. The analogs of Theorems 3.2 and 3.3 also hold, and are stated as follows. Both theorems play essential roles in the identification of Pareto quantiles.

THEOREM 3.4. *Assume that $\{t_1, \ldots, t_n\}$ is a sample of trees with finite levels, that is, the number of edges to the root node. Let $\ell_i(x)$ be the topological curve representation of t_i. The pointwise topological quantile of $\ell_i(x)$ represents a valid tree class.*

THEOREM 3.5. *A pointwise quantile of a finite sample of geometric tree curves $g_i(x)$ represents a valid tree class.*

In general, for a sample of forests, it can be shown that a pointwise quantile represents a valid forest. Next, as in (3.6), the existence of the minimizer of (3.8) is not guaranteed. Here, we intend to find the Pareto optimal set for (3.8). Each element in this Pareto optimal set is called a 100τth Pareto quantile. As with the topological and geometric Pareto medians, there are two trivial elements in the Pareto optimal set of (3.8), namely, the topological Pareto quantile and the geometric Pareto quantile. For illustration, the Pareto optimal sets for the 25th and 75th quantiles for the example from Figure 6 are depicted in Figure 8. Both sets consist of just two solutions, the geometric Pareto quantile and the topological Pareto quantile. Note that for this toy example, all solutions can be obtained using the weighted sum method in a similar fashion as defined in (3.7).

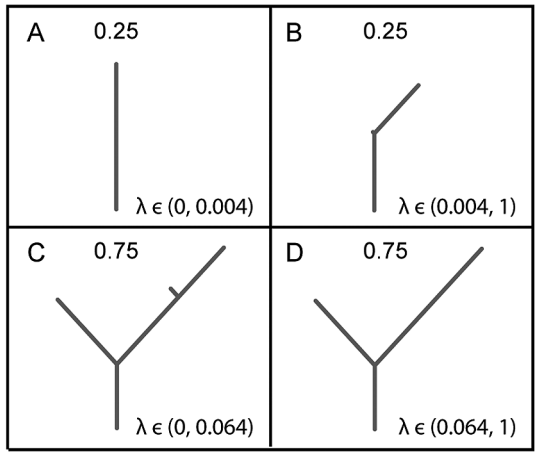

FIG. 8. *Top row*: *Pareto set for the* 25*th quantile*; *bottom row*: *Pareto set for the* 75*th quantile*. *Trees* A *and* C *are the geometric Pareto quantiles*, *and trees* B *and* D *are the topological Pareto quantiles*.

3.4. *Genetic algorithm.* A genetic algorithm provides a useful tool to solve combinatorial problems that do not have an analytic solution. By imitating the mechanism of genetic selection acting on chromosomes and genes, the algorithm finds the *fittest* elements in the population. The interpretation of fitness depends on the optimization goal. In single-objective optimization, there is generally one best element. In multi-objective optimization, there is a set of best elements defined through nondominance, as discussed in Section 3.2. The algorithm starts with an initial population of objects, designed to provide sufficient *genetic diversity* for the natural selection to work, and creates new individuals stochastically via random *crossovers* and *mutations* applied to the fittest (and occasionally less fit) elements of the previous generation. Theoretical results regarding the convergence of the genetic algorithm are based on the schema theory [14]. Practical advice on algorithm design is available in Sivanandam and Deepa [35]. In general, a genetic algorithm performs better than a random search, and it does so by exploiting accumulated information about the features that improve the overall capabilities of the chromosomes.

The implementation of a genetic algorithm is generally nontrivial, but one can usually find a suitable, configurable framework in the programming language of choice. Some important elements of the design are required, including the encoding of data as *chromosomes*, and the definitions of *crossover* and *mutation*. Here, we employ a genetic algorithm to solve the optimization problem (3.6). Recall that the geometric curves are piecewise constant functions on a finite set of jump points. Therefore, a solution can be found among piecewise constant functions on the finite set of jump points formed by the union of all jump points of tree curves in the sample. We encode the population of geometric tree curves as chromosome-like strings, and we define crossover and mutations over those tree curves. Note

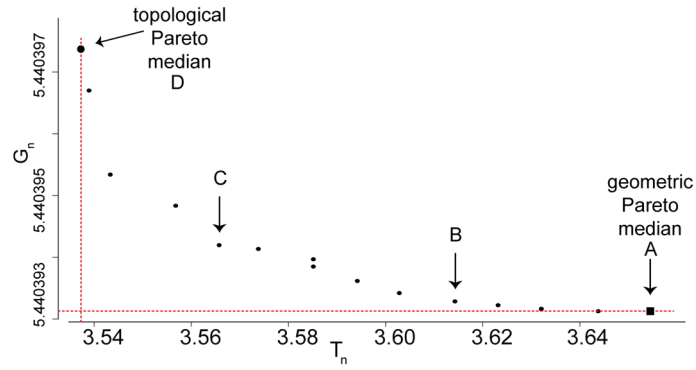

FIG. 9. *A graphical display of the subset of a Pareto set for the median of apical dendritic trees from CA*1. *The solutions form a Pareto front. The geometric Pareto median* (A) *and topological Pareto median* (D) *as well as two other Pareto medians*, (B) *and* (C), *are highlighted. All four are depicted in Figure* 15. *Here*, *both axes are shown in log-scale*.

that a mutation can create, delete or move a single branch, and a crossover can swap two subtrees from two parents. For details regarding these operations on tree objects, see Sienkiewicz [32].

For the toy example in Figure 6, the genetic algorithm finds the entire Pareto set of median trees (Figure 7) as well as the 25th and the 75th quantile trees (Figure 8). For the real data, the problem is much more complex. Figure 9 shows a collection of solutions found in a single run of the algorithm.

We outline the algorithm as follows:

1. Generate the initial population, including $100p$th geometric quantiles of the real data, for p in the neighborhood of τ.
2. Encode each individual (a geometric curve) in the population as a "chromosome," for example, using a sequence of counts from a geometric curve.
3. For each generation $i = 1, 2, \ldots, K$
   (1) For each individual t, calculate two scores: $G_n^\tau(t)$ and $T_n^\tau(t)$.
   (2) Calculate a rank of each individual, for example, the number of individuals dominating it (in the Pareto sense) from the same generation.
   (3) Repeat until the generation $i + 1$ is created:

      (a) Draw two members from the population i with probability inversely proportional to the ranks.

      (b) Perform a crossover of these individuals at a random location creating two new individuals.

      (c) For each of the two new individuals, perform a mutation with a probability p, which increases or decreases a count by one.

      (d) Add these new individuals to the generation $i + 1$.

   (4) Copy all individuals with rank 0 from generation i to $i + 1$.

The number of iterations K used in step 3 is user-specific and can be tuned for particular trees. The minimum values of T_n^τ and G_n^τ correspond to the topological and geometric quantiles, respectively, and can be computed using Theorems 3.4 and 3.5. The value of K can be increased if the algorithm does not produce trees with sufficiently small $T_n^\tau(t)$. Note that the above algorithm can also be used to find solutions for the optimization problem (3.7). In particular, we can calculate the weighted score in Step (3–1).

4. Simulation study.

4.1. *Simulation methods*. In this section, we conduct a simulation study to demonstrate the performance of our proposed method. The first step is to simulate a population of trees with topological and geometric characteristics corresponding to those in the real data. Specifically, we focus on neurons from CA1 region. In Section D of the Supplementary Material [33], we describe two simulation methods in detail, Topology-Geometry Strategy (*topo-geo*) and Geometry-Topology Strategy (*geo-topo*), named after the order in which properties are generated. To summarize, in the *topo-geo* strategy, we first randomly generate a tree topology, and then, for each tree branch, we assign a length (weight) generated from an appropriate distribution. Here, we use a Gamma distribution, with the parameters estimated from the data for each tree level se parately. In the *geo-topo* strategy, we consider a doubly stochastic Poisson process (also known as a Cox process), which governs the splits and terminations of branches as a function of the distance from the root of the tree. Then, for the established tree geometry, we randomly select a compatible tree topology.

For tree topology, we consider conditioned binary Galton–Watson trees, which are trees with fixed number of nodes; see [26] for the introduction. Here, the size of a tree is characterized by the number of internal nodes m, and it can be determined by data-driven methods. In the *geo-topo* method, each realization of the randomly stopped Cox process determines the size of the resulting tree. In the *topo-geo* method, we estimate the tree size using the probability mass function estimation proposed by Canale and Dunson [8]. Without considering topological equivalence, there are $\mathbb{C}_m$ full binary trees with m internal nodes, where $\mathbb{C}_m = \binom{2m}{m}/(m+1)$ is the Catalan number. The number of different equivalence classes, denoted by $\mathbb{T}_m$, is much smaller. For comparison, Table 1 contains a list of the first 10 Catalan numbers $\mathbb{C}_m$ and corresponding $\mathbb{T}_m$.

A tree topology can be randomly selected from one of the $\mathbb{C}_m$ topologies according to an underlying tree distribution. Two distributions are commonly discussed in the literature, a uniform tree distribution (each tree is equally likely), and a distribution of binary-search trees; see Mäkinen [21], Flajolet et al. [12] among others. We also adapted the distribution of binary search trees to the collection of full binary trees. We refer to this new distribution as *leaf-uniform* distribution.

TABLE 1
Comparison of $\mathbb{C}_m$ and $\mathbb{T}_m$ for $m = 1, \ldots, 10$. Here, $\mathbb{C}_m$ represents the Catalan number, and $\mathbb{T}_m$ represents the number of topological equivalence classes

m	1	2	3	4	5	6	7	8	9	10
$\mathbb{C}_m$	1	2	5	14	42	132	429	1430	4862	16,796
$\mathbb{T}_m$	1	1	2	3	5	9	16	28	50	89

More discussions, including the differences between these distributions, the simulation methods and the results of model fits, can be found in Section C of the Supplementary Material [33].

In this section, we will present the results from the *topo-geo* method. The simulation algorithm for the *topo-geo* method is outlined as follows:

1. Randomly generate the tree size according to the probability mass function estimated from the data.
2. For a given tree size, generate tree topology using the uniform distribution for apical trees and leaf-uniform tree distribution for basal trees.
3. Assign weights (lengths) to tree branches by drawing from a Gamma distribution, with parameters estimated from the data for each tree level.

The results from the *geo-topo* method are provided in the Supplementary Material [33].

4.2. *Simulation results*. Here, our goal is to examine sample-to-sample variation in Pareto quantiles of a finite set of trees. First, we generate a population of $N = 1000$ trees $\{t_1, t_2, \ldots, t_N\}$ using methods described in Section 4.1 and in the Supplementary Material [33]. Consider a random element following a uniform distribution on this set of trees. Similar to (3.8), the population Pareto quantiles can be defined as

$$\operatorname*{argmin}_t \big(T_N^\tau(t), G_N^\tau(t)\big),$$

which can be obtained using the algorithms discussed in Section 3.4 with $K = 200$. In our simulation study, the population Pareto medians serve as the target "parameters" and are shown in black dashed linetype in Figure 10. Additional Pareto quantiles are depicted in Figures S.10 and S.11 of the Supplementary Material [33].

A study with 100 repetitions is conducted. For each repetition, a sample of $n = 200$ tree objects is drawn from the population. The sample topological and geometric Pareto quantiles are calculated based on our proposed methods. Figure 10 shows sample-to-sample variation in calculated medians. Panels (A) and (B) of Figure 10 depict two tree curves corresponding to the geometric Pareto

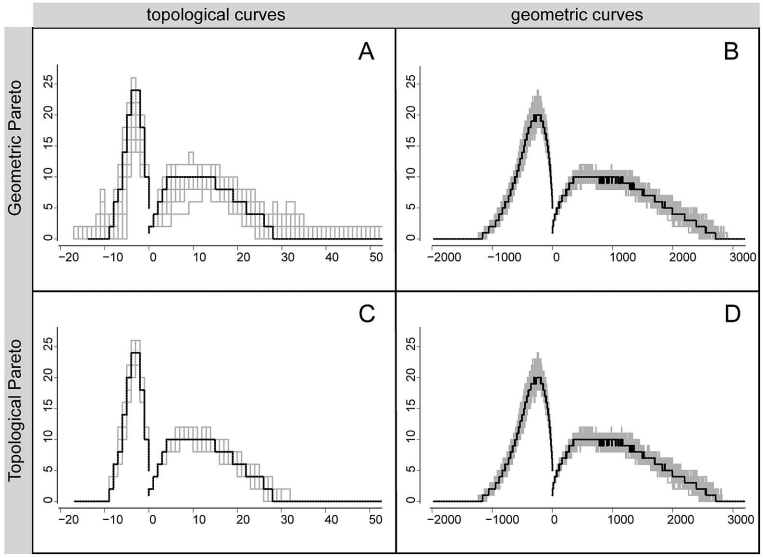

FIG. 10. *Top row*: *topological* (A) *and geometric* (B) *tree curves corresponding to the geometric Pareto median for the population* (*black*) *and each sample* (*grey*). *Bottom row*: *topological* (C) *and geometric* (D) *tree curves corresponding to the topological Pareto median for the population* (*black*) *and each sample* (*grey*).

median tree. Panels (C) and (D) depict two tree curves corresponding to the topological Pareto median tree. Note that the topological tree curve of the geometric Pareto quantile (panel A) shows far more sample-to-sample variability than the topological tree curve in panel (C). There is no such difference in variability for the geometric tree curves in panels (B) and (D). Overall, the topological Pareto median shows less sample-to-sample variation than the geometric Pareto median.

4.3. *Discussion*. As a comparison, we consider an alternative method for obtaining the median tree based on node labeling using descendant correspondence [31]. In Figure 11, three different topological medians for each sample from the simulated tree population are depicted, including BLR (panel A), the topological Pareto median (panel B) and the geometric Pareto median (panel C) using our proposed method. The modified Harris path method was excluded due to its infeasibility for such large data. The plot shows the number of dendritic branches (y-axis) at each level (x-axis) in each estimated median in each sample, and the number of branches in the population. It can be seen that BLR for labeled trees yield trees with more dendritic segments in lower levels and fewer dendritic segments in higher levels, which may potentially underestimate the topological complexity.

From the computational perspective, the BLR median is evaluated per each node label. As a consequence, it is fast to compute; however, the results in Figure 11

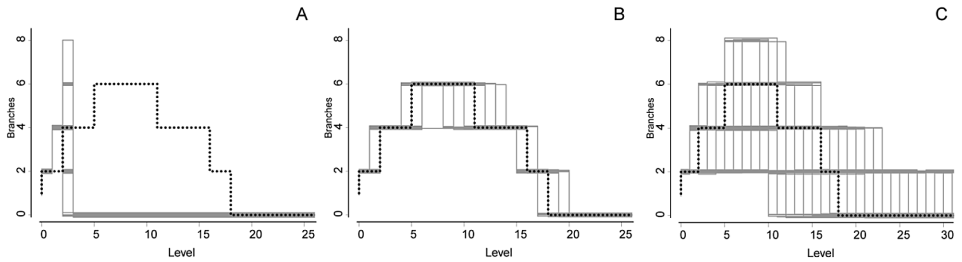

FIG. 11. *Topology for the simulated trees*: *BLR labeled method* (A), *topological Pareto quantile calculated with our proposed method* (B), *geometric Pareto quantile calculated with our proposed method* (C). *The number of branches in the population is depicted with a black dotted line in each image. Here, the x-axis represents tree levels and the y-axis represents the number of branches at any given level x.*

show a rather large discrepancy with the population median. The computation of Pareto medians is carried out using stochastic optimization, which may require more time depending on the size of the sample and the length of the curves. In our case, the computation of each quantile for a single set of neurons varies between 5–30 minutes.

5. Real data analysis. In this section, our proposed methods are applied to a set of pyramidal neurons as described in Section 2.1. This data set consists of 119 digital reconstructions of neurons from the CA1 region (three of these are depicted in the top row of Figure 12) and 68 from the CA3 region of the hippocampus (three of these are depicted in the bottom row of Figure 12). In general, each neuron consists of approximately 3000 interconnected voxels, and each voxel is associated with a type (soma, axon, basal dendrite, apical dendrite) and a radius (not used in this study). The first step of the analysis involves the extraction of the tree object for both dendritic structures of every neuron. The geometric properties of each branch can be extracted in multiple ways. Ascoli and Krichmar [3] provided a comprehensive survey of studies of the relationship between branch length, branch radius at bifurcation points and branching angles. The authors pointed out that approximating a branch length by a straight line between the bifurcation points leads to a much smaller tree. An alternative approach is to approximate the branch length by summing the distances between voxels in each branch, which could potentially lead to a larger tree. Here, we take the latter approach. The topological and geometric curve representations can be constructed based on the digitally reconstructed neurons.

Figure 13 shows both geometric (left column) and topological (right column) curve representations for neurons in CA1 (top row) and CA3 (bottom row). In each panel, joint tree curves, as defined in Section 2.3, are depicted. It can be observed that neurons from the CA3 region have a much more developed basal section (left portions of the tree curves) than neurons from CA1. The apical sections of neurons

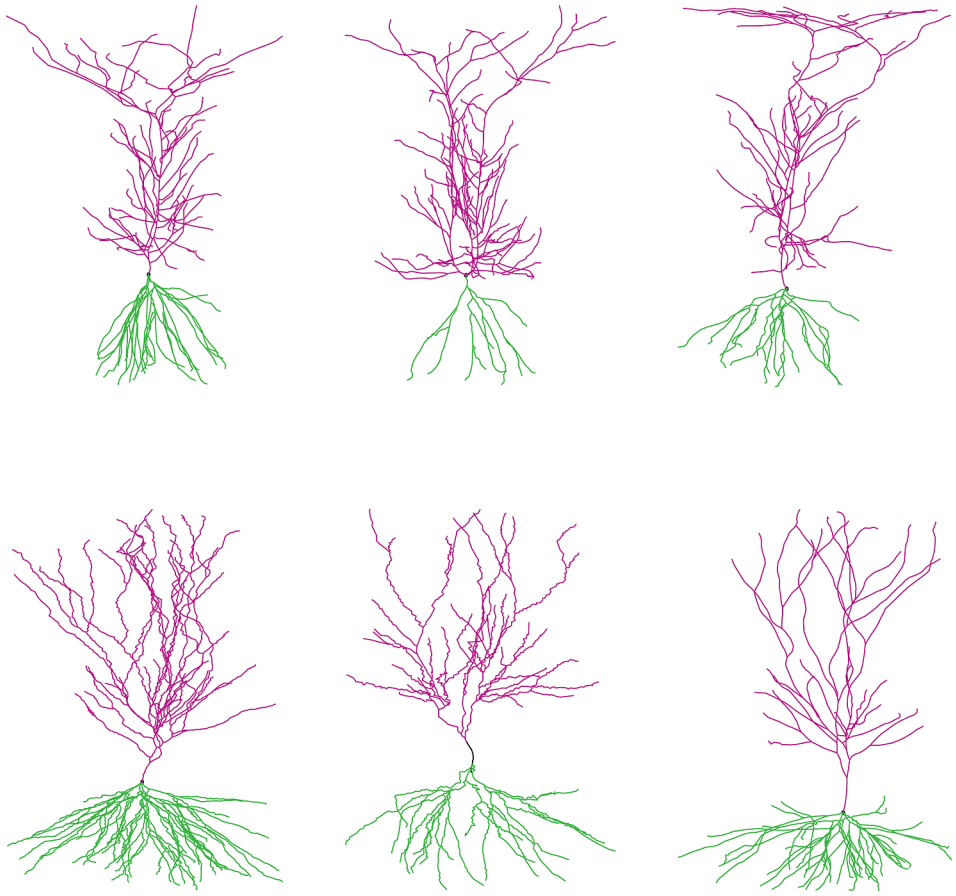

FIG. 12. *Graphical display of six neurons. Top row: three neurons from the CA1 region of the hippocampus; Bottom row: three neurons from the CA3 region of the hippocampus. In each subplot, apical dendrites are shown in magenta, and basal dendrites are shown in green.*

from both regions differ substantially. For instance, in panels (A) and (C), the geometric apical tree curves from CA1 (the right portions of the curves) seem to be longer than the ones form CA3. In fact, many CA1 tree curves are longer than 1000 (in units of micrometers), whereas most CA3 tree curves are less than that. In addition, the largest branch counts for tree curves from CA1, on the y-axis, are bigger than the branch counts for tree curves from CA3. The topological curves of the apical trees, in panels (B) and (D), indicate that apical trees from CA1 are taller than those from CA3. Specifically, many CA1 curves reach levels 30 or higher, while all apical trees from CA3 end before level 20.

For each choice of τ, we implement the genetic algorithm as discussed in Section 3.4 to obtain quantiles of apical and basal dendritic trees, with the number of iterations $K = 200$. In Figure 9, the Pareto set for the median apical dendritic

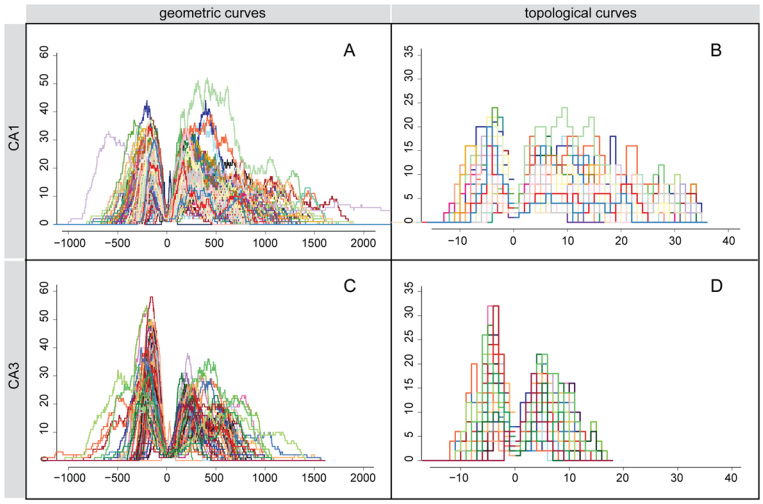

FIG. 13. (A) *Joint geometric curves for all neurons from the CA*1 *region*; (B) *joint topological curves for all neurons from the CA*1 *region*; (C) *joint geometric curves for all neurons from the CA*3 *region*; (D) *joint topological curves for all neurons from the CA*3 *region*.

trees is depicted. Each element in the Pareto set will correspond to a Pareto median tree. In particular, the topological Pareto median and geometric Pareto median are highlighted in Figure 9, and their corresponding tree representations are included in Figure 14. The branching angles of reconstructed tree objects, depicted in Figure 14 and subsequent figures, are selected for better visualization of a large number of branches, given limited space.

Recall that each Pareto solution consists of two curves, a topological curve and a geometric curve, and for each pair a tree can be reconstructed following the procedure outlined in Section E.1 of the Supplementary Material [33]. The geometric curves corresponding to both Pareto medians, in panels (B) and (F) of Figure 14, are very similar. The topological curves, in panels (A) and (E), reveal some topological differences between Pareto medians; in particular, around level 10, the geometric Pareto median tree tends to have more branches than the topological Pareto median.

In panel (B) of Figure 14, the joint curves for the geometric Pareto median trees from CA1 (solid line) and CA3 (dashed line) are compared. The median basal dendrites from both regions are of the same overall length, but basal dendrites from CA3 have substantially more branches in the middle section. In fact, the maximal number of branches for both basal dendrites occurs roughly at the same distance from the root. For the apical dendrites, the median geometric curves (the positive portions in panel B) are closer in maximum branch count (y-axis) as well as tree height (x-axis). The CA1 median tree is only slightly longer, and the maximal numbers of branches for both apical dendrites are aligned at the same distance. This observation is quite interesting compared with the depiction of raw data in

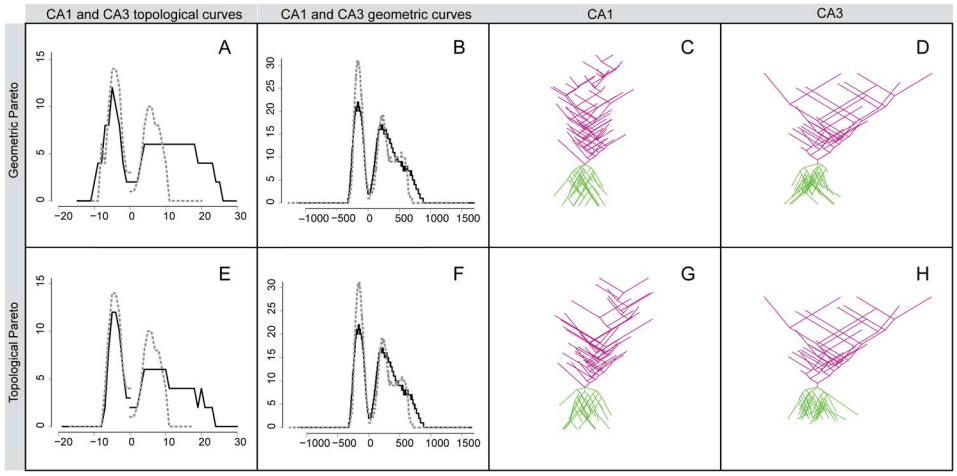

FIG. 14. *Graphical display of topological and geometric tree curves, as well as corresponding tree objects, for the geometric Pareto median (top row) and the topological Pareto median (bottom row). (A) topological tree curves for the geometric Pareto median from CA1 (solid) and CA3 (dashed); (B) geometric curves for the geometric Pareto median from CA1 (solid) and CA3 (dashed); (C) tree object for the CA1 geometric Pareto median; (D) tree object for the CA3 geometric Pareto median; (E) topological tree curves for the topological Pareto median from CA1 (solid) and CA3 (dashed); (F) geometric curves for the topological Pareto median from CA1 (solid) and CA3 (dashed); (G) tree object for the CA1 topological Pareto median; (H) tree object for the CA3 topological Pareto median.*

Figure 13, which implied that apical dendrites from CA1 seem to be longer. Panels (C) and (D) contain simplified depictions of the geometric Pareto median trees from CA1 and CA3.

In panel (E), the topological Pareto median curves from CA1 (solid line) and CA3 (dashed line) are compared. The basal topological median trees have the same height (x-axis), but CA3 maxima are larger than CA1 maxima. In fact, the median basal dendrites for both CA1 and CA3 are forests with three and four trees, respectively. In contrast, the apical dendrites differ considerably in topology. The CA3 apical median tree is much shorter, but fuller, and it reaches the maximum number of branches at about level 5, and from there, the number drops steeply. The CA1 median tree exhibits a different growth pattern. The branch maximum is lower than that of CA3, but the number diminishes slowly, which results in a much taller tree. Panels (G) and (H) contain simplified depictions of the topological Pareto median trees from CA1 and CA3. Our observations regarding differences in apical topologies and similarities in apical geometries match the conclusions reached by Vida [37]. Figure 14 shows two members of the Pareto optimal set, the geometric Pareto median and the topological Pareto median. Figure 15 shows two additional reconstructed Pareto median trees from region CA1. All trees are indeed similar.

Next, we implement our proposed method to compute both topological and geometric Pareto quantiles. Figures S.13 and S.14 in the Supplementary Material

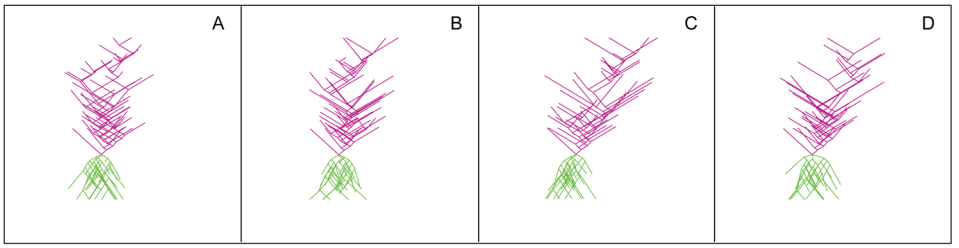

FIG. 15. *Pareto median trees from CA*1, *including the geometric Pareto median* (*panel* A), *the topological Pareto median* (*panel* D), *and two other Pareto median trees* (*panels* B *and* C). *The* T_n *and* G_n *values for these four median trees are highlighted in Figure* 9.

[33] show the 10th and 90th Pareto quantiles of neurons from the CA1 and CA3 regions.

To summarize, our proposed quantiles highlight the difference in distributions of populations of neurons from CA1 and CA3. In fact, the geometric and topological differences between dendritic trees from CA1 and CA3 can be observed by analyzing the 10th, 50th and 90th Pareto quantiles of both regions. For instance, the basal dendrites from both regions are very similar geometrically and topologically, from very small and simple trees to larger and more complex trees. The basal trees from CA3 appear to be forests with more component trees than basal forests from CA1. The apical parts reveal bigger differences between the two regions. Apical trees from CA1 are slightly longer, and they exhibit a different branching pattern compared to the apical trees from CA3. The apical trees from CA3 have more branches at lower levels. The apical trees from CA1 have fewer branches at lower levels, and more branches at higher levels, and topologically, they form much taller trees.

6. Potential application and discussion. In this paper, we developed a notion of Pareto quantiles of sets of unlabeled tree-structured objects. Such development was based on two functional representations for trees, preserving their topological and geometric properties. The driving example was a set of pyramidal neurons from the hippocampus, particularly regions CA1 and CA3. Our proposed methodology enables automatic, computer-aided classification, which has not been studied previously.

Another potential application is related to neuromorphological disorders or diseases. There is an ongoing research to compare regions of the brain and individual neurons in groups of human subjects suffering from degenerative brain diseases, for example, Parkinson's, Creutzfeldt-Jakob's or Alzheimer's; see [15, 24, 34, 41] for details. A major challenge in those studies was to distinguish changes, including morphological changes, related to normal aging from changes related to diseases, which often, as for instance Alzheimer's disease, predominantly affect older individuals. The reduction of neuronal density in regions CA1 and CA3 of

the hippocampus was observed. In addition, Padurariu et al. [24] reported "reduction in dendritic branching," Šimić et al. [34] noticed "structural degeneration of neurons," and Grutzendler et al. [15] observed "abrupt branching endings," "breakage of nearby dendrites" and sprouting "unusually long, thin and not resembling dendritic sprouting." These type of changes in neural structure could be observed early using our methodology, before the disease impairs cognitive functions of affected individuals. The empirical evidence can be strengthened by adopting more rigorous statistics for complex tree data. Our proposed tree quantiles provide an essential toolkit to meet the demand for the diagnosis of degenerative brain diseases.

Acknowledgments. We are grateful to the Editor, Associate Editor and the anonymous referees for their helpful and constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Pareto quantiles of unlabeled tree objects" (DOI: 10.1214/17-AOS1593SUPP; .pdf). This document includes the description of the data object construction, proofs, and additional details regarding simulation and data analysis.

REFERENCES

[1] ANTONIOU, A. and LU, W.-S. (2007). *Practical Optimization*: *Algorithms and Engineering Applications*. Springer, New York. MR2364120

[2] ASCOLI, G. A., DONOHUE, D. E. and HALAVI, M. (2007). NeuroMorpho.Org: A central resource for neuronal morphologies. *J. Neurosci*. **27** 9247–9251.

[3] ASCOLI, G. A. and KRICHMAR, J. L. (2000). L-Neuron: A modeling tool for the efficient generation and parsimonious description of dendritic morphology. *Neurocomputing* **32** 1003–1011.

[4] ASCOLI, G. A., KRICHMAR, J. L., SCORCIONI, R., NASUTO, S. J., SENFT, S. L. and KRICHMAR, G. L. (2001). Computer generation and quantitative morphometric analysis of virtual neurons. *Anat. Embryol*. **204** 283–301.

[5] AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. S. (2009). A principal component analysis for trees. *Ann. Appl. Stat*. **3** 1597–1615. MR2752149

[6] BANKS, D. and CONSTANTINE, G. M. (1998). Metric models for random graphs. *J. Classification* **15** 199–223. MR1665974

[7] BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math*. **27** 733–767. MR1867931

[8] CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc*. **106** 1528–1539. MR2896854

[9] CHANG, H.-W., IYER, H., BULLITT, E. and WANG, H. (2013). Generalized linear mixed models for branching probabilities of brain artery systems. *Model Assist. Stat. Appl*. **8** 121–133.

[10] CLINE, H. and HAAS, K. (2008). The regulation of dendritic arbor development and plasticity by glutamatergic synaptic input: A review of the synaptotrophic hypothesis. *J. Physiol*. **586** 1509–1517.

[11] COELLO COELLO, C. A., LAMONT, G. B. and VAN VELDHUIZEN, D. A. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed. Springer, New York. With a foreword by David E. Goldberg. MR2350880

[12] FLAJOLET, P., GAO, Z., ODLYZKO, A. and RICHMOND, B. (1993). The distribution of heights of binary trees and other simple trees. *Combin. Probab. Comput.* **2** 145–156. MR1249127

[13] FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10** 215–310. MR0027464

[14] GOLDBERG, D. (1989). *Genetic Algorithms in Optimization, Search and Machine Learning*. Addison Wesley Publishing Company, New York.

[15] GRUTZENDLER, J., HELMIN, K., TSAI, J. and GAN, W.-B. (2007). Various dendritic abnormalities are associated with fibrillar amyloid deposits in Alzheimer's disease. *Ann. N.Y. Acad. Sci.* **1097** 30–39.

[16] HARRIS, T. E. (1952). First passage and recurrence distributions. *Trans. Amer. Math. Soc.* **73** 471–486. MR0052057

[17] HENDRICKSON, P. J., GENE, J. Y., SONG, D. and BERGER, T. W. (2016). A million-plus neuron model of the hippocampal dentate gyrus: Critical role for topography in determining spatiotemporal network dynamics. *IEEE Trans. Biomed. Eng.* **63** 199–209.

[18] JOHNSTON, D. and WU, S. M.-S. (1994). *Foundations of Cellular Neurophysiology*. MIT Press, Cambridge.

[19] KOENKER, R. and HALLOCK, K. (2001). Quantile regression. *J. Econ. Perspect.* **15** 143–156.

[20] LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400

[21] MÄKINEN, E. (1999). Generating random binary trees—a survey. *Inform. Sci.* **115** 123–136. MR1671899

[22] MARRON, J. S. and ALONSO, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56** 732–753. MR3258083

[23] MIGLIORE, M. and SHEPHERD, G. M. (2005). An integrated approach to classifying neuronal phenotypes. *Nat. Rev., Neurosci.* **6** 810–818.

[24] PADURARIU, M., CIOBICA, A., MAVROUDIS, I., FOTIOU, D. and BALOYANNIS, S. (2012). Hippocampal neuronal loss in the CA1 and CA3 areas of Alzheimer's disease patients. *Psychiatr. Danub.* **24** 152–158.

[25] PHILLIPS, C. and WARNOW, T. J. (1996). The asymmetric median tree—a new model for building consensus trees. *Discrete Appl. Math.* **71** 311–335. MR1420306

[26] PITMAN, J. (2006). *Combinatorial Stochastic Processes*: *Lectures from the* 32*nd Summer School on Probability Theory Held in Saint-Flour, July* 7–24, 2002. *Lecture Notes in Math.* **1875**. Springer, Berlin. With a foreword by Jean Picard. MR2245368

[27] PYAPALI, G. K., SIK, A., PENTTONEN, M., BUZSAKI, G. and TURNER, D. A. (1998). Dendritic properties of hippocampal CA1 pyramidal neurons in the rat: Intracellular staining in vivo and in vitro. *J. Comp. Neurol.* **391** 335–352.

[28] PYAPALI, G. K. and TURNER, D. A. (1994). Denervation-induced dendritic alterations in CA1 pyramidal cells following kainic acid hippocampal lesions in rats. *Brain Res.* **652** 279–290.

[29] PYAPALI, G. K. and TURNER, D. A. (1996). Increased dendritic extent in hippocampal CA1 neurons from aged F344 rats. *Neurobiol. Aging* **17** 601–611.

[30] SERFLING, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Stat. Neerl.* **56** 214–232. Special issue: Frontier Research in Theoretical Statistics, 2000 (Eindhoven). MR1916321

[31] SHEN, D., SHEN, H., BHAMIDI, S., MUÑOZ MALDONADO, Y., KIM, Y. and MARRON, J. S. (2014). Functional data analysis of tree data objects. *J. Comput. Graph. Statist.* **23** 418–438. MR3215818

[32] SIENKIEWICZ, E. (2015). Analysis of big data and structured data with application in neuroscience. Ph.D. thesis, Colorado State Univ.

[33] SIENKIEWICZ, E. and WANG, H. (2017). Supplement to "Pareto quantiles of unlabeled tree objects." DOI:10.1214/17-AOS1593SUPP.

[34] ŠIMIĆ, G., KOSTOVIĆ, I., WINBLAD, B. and BOGDANOVIĆ, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *J. Comp. Neurol.* **379** 482–494.

[35] SIVANANDAM, S. N. and DEEPA, S. N. (2008). *Introduction to Genetic Algorithms*. Springer, Berlin. MR2441025

[36] SONG, D., CHAN, R. H. M., MARMARELIS, V. Z., HAMPSON, R. E., DEADWYLER, S. A. and BERGER, T. W. (2007). Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses. *IEEE Trans. Biomed. Eng.* **54** 1053–1066.

[37] VIDA, I. (2010). Morphology of hippocampal neurons. In *Hippocampal Microcircuits* 27–67. Springer, Berlin.

[38] WALTER, S. (2011). Defining quantiles for functional data. Ph.D. thesis, The Univ. Melbourne.

[39] WANG, H. and MARRON, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35** 1849–1873. MR2363955

[40] WANG, Y., MARRON, J. S., AYDIN, B., LADHA, A., BULLITT, E. and WANG, H. (2012). A nonparametric regression model with tree-structured response. *J. Amer. Statist. Assoc.* **107** 1272–1285. MR3036394

[41] WEST, M. J., COLEMAN, P. D., FLOOD, D. G. and TRONCOSO, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *Lancet* **344** 769–772.

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