Tree-space statistics and approximations for large-scale analysis of anatomical trees

Aasa Feragen[1,2], Megan Owen[3], Jens Petersen[1], Mathilde M. W. Wille[4], Laura H. Thomsen[4], Asger Dirksen[4], and Marleen de Bruijne[1,5]

Department of Computer Science, University of Copenhagen, Denmark[1], Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, Tübingen, Germany[2], David R. Cheriton School of Computer Science, University of Waterloo, Canada[3], Lungemedicinsk Afdeling, Gentofte Hospital, Denmark[4], Erasmus MC - University Medical Center Rotterdam, The Netherlands[5]
{aasa,phup,marleen}@diku.dk,
WWW home page: http://www.image.diku.dk/aasa

Abstract. Statistical analysis of anatomical trees is hard to perform due to differences in the topological structure of the trees. In this paper we define statistical properties of leaf-labeled anatomical trees with geometric edge attributes by considering the anatomical trees as points in the geometric space of leaf-labeled trees. This tree-space is a geodesic metric space where any two trees are connected by a unique shortest path, which corresponds to a tree deformation. However, tree-space is not a manifold, and the usual strategy of performing statistical analysis in a tangent space and projecting onto tree-space is not available. Using tree-space and its shortest paths, a variety of statistical properties, such as mean, principal component, hypothesis testing and linear discriminant analysis can be defined. For some of these properties it is still an open problem how to compute them; others (like the mean) can be computed, but efficient alternatives are helpful in speeding up algorithms that use means iteratively, like hypothesis testing. In this paper, we take advantage of a very large dataset ($N = 8016$) to obtain computable approximations, under the assumption that the data trees parametrize the relevant parts of tree-space well. Using the developed approximate statistics, we illustrate how the structure and geometry of airway trees vary across a population and show that airway trees with Chronic Obstructive Pulmonary Disease come from a different distribution in tree-space than healthy ones. Software is available from http://image.diku.dk/aasa/software.php.

1 Introduction

Anatomical trees, such as vessels, airways or dendrites, are transportation networks that play an important role in the development of diseases. In order to better understand disease and its interaction with the anatomical tree geometry and structure, one needs to be able to perform statistical analysis of sets of anatomical trees, including both the topological structure of the trees and the shape of the branches. In particular, detection of disease or disease phenotype

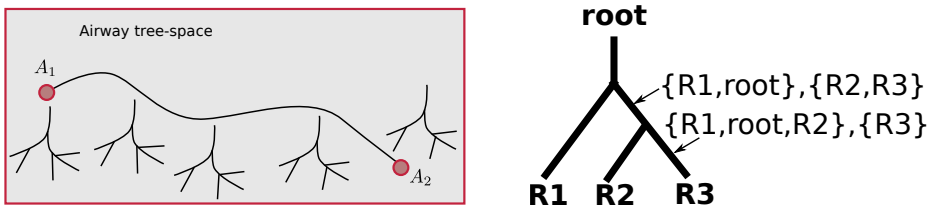

Fig. 1. Left: Tree-space is path connected space,i.e., any two trees can be joined by a path in tree-space. Moving along such a path corresponds to a deformation of trees. **Right:** Tree edges are defined by partitions of the leaf label set.

based on anatomical tree structure and geometry could improve computer-aided tools for diagnosis and prognosis of disease. However, since anatomical trees have different topological structures, it is not obvious how they should be compared. In particular, two trees can have different topologies but be geometrically and functionally similar, like the trees along the tree-space path in Fig. 1.

Background and related work. There are several tree-space constructions that treat trees as a continuous family of objects where tree topology changes as part of a continuous deformation of the tree [4, 9]. However, the geometry of these tree-spaces does not allow for an easy transfer of standard statistical properties. For instance, tree-spaces have corners and branching points, and so manifold statistics do not apply. For this reason, most statistical measurements are still not available for tree-structured data, although the statistical analysis of tree- and graph-structured data is increasingly studied in mathematical and applied statistics [10,13,23]. Recent results include the existence, uniqueness and computation of Fréchet means [2, 7, 16, 21] and first results on principal components analysis [17]; however, how principal components should be defined and computed remains an open problem. Most available tools have so far only been used to analyze small datasets [7]. *Tree kernels* [8] form an alternative method for analyzing tree-structured data, which gives access to machine learning algorithms for e.g., classification or regression. However, tree kernels do not operate in a true space of trees, and cannot produce tree-valued solutions, such as an average tree, or the variation along a principal component or between classes.

Our contributions. We choose a set of statistics that in different ways are useful in understanding dataset and class variation, although at present we do not know how to compute them all exactly: *principal components (PCA), two-sample hypothesis testing, linear discriminant analysis*. Hypothesis tests can already be computed to a high precision using available algorithms for means [16], whereas the others cannot. Using automatic airway branch labeling [10] on a database of airway trees, we obtain a large set of leaf-labeled airway trees. Treating the large dataset as a discretization of the relevant parts of tree-space, we define approximations of the listed statistics, which can be computed for large datasets in very limited time. Using the developed methods, we perform a large-scale study on a real dataset consisting of 8016 airway trees from 1692 individuals, of which 842 are diagnosed with Chronic Obstructive Pulmonary Disease (COPD). With

our newly developed tools, we can quantify and visualize statistical properties of the airway population such as means and variance, show that the airway tree-shape differs significantly between COPD patients and healthy individuals, and visualize this difference.

Organization. The paper is organized as follows: In Sec. 2 we briefly introduce the tree-space used in this paper. In Sec. 3 we review known methods for computing mean trees, and then define and compute or approximate geodesic PCA, hypothesis tests for means and variances, and geodesic linear discriminant analysis. A detailed description of the airway dataset is given in Sec. 4, and the presented methods and results are discussed in detail in Sec. 5.

2 Tree-space

The tree-space $\mathcal{T}$ used here is a straight-forward generalization of the space of phylogenetic trees originally defined in [4], where scalar edge length attributes have been generalized to multi-dimensional edge shape vectors in $(\mathbb{R}^3)^d$, consisting of d equidistantly sampled points along the edge (in this paper $d = 5$, giving an edge shape space $\mathbb{R}^{15}$). The space $\mathcal{T}$ is *path connected*, i.e., any two trees can be joined by a path in tree-space, corresponding to a tree deformation, see Fig. 1. Any two trees in tree-space have a unique *shortest* path joining them [4], whose length defines a distance between the two trees, giving a metric d on $\mathcal{T}$.

Each point in $\mathcal{T}$ is a leaf-labeled tree with root r and 20 leaves labeled by the 20 airway segmental branch labels $\mathcal{L} = \{L1, ..., L10, R1, ..., R10\}$. Each edge in the tree is combinatorially represented as a partition of $\mathcal{L} \cup \{r\}$ into the leaves descending from the edge, and the remaining leaves (including r), see Fig. 1. If S is the set of possible partitions of $\mathcal{L} \cup \{r\}$, then each tree uniquely corresponds to a vector in $(\mathbb{R}^{15})^S$, where each consecutive set of 15 coordinates corresponding to a possible edge (identified with a partition of $\mathcal{L} \cup \{r\}$). If the edge associated with that partition appears in the tree, then those 15 coordinates will be its branch vector; otherwise they are all 0. Certain edges can never appear in a tree together (e.g., an edge that splits $\{R1, R2\}$ off from the rest of the tree and an edge that splits $\{R1, R3\}$ off), so not all vectors are possible trees. *Tree-space* $\mathcal{T}$ is precisely those vectors in $(\mathbb{R}^{15})^S$ that correspond to trees; thus, $\mathcal{T}$ is a proper subset of Euclidean space. The shortest-path distance between two trees is the length of the shortest path between them which stays within $\mathcal{T}$, measured in the ambient Euclidean space. There is no analytic formula for this distance, but it can be computed recursively in polynomial time [18].

Figure 3 shows portions of the spaces of leaf-labeled trees with 3 and 4 leaves and edge length attributes. Tree-space is not a manifold, because it has a branching structure and corners: it can be decomposed into *orthants* where tree structure is constant, and where the orthants are glued together along subspaces containing contracted versions of the orthant trees, see Fig. 2. This geometric structure complicates statistical analysis. First, while the concept of "direction" makes sense locally within an orthant, where the space locally looks Euclidean, it does not make sense on a global level. In linear spaces, directions are defined

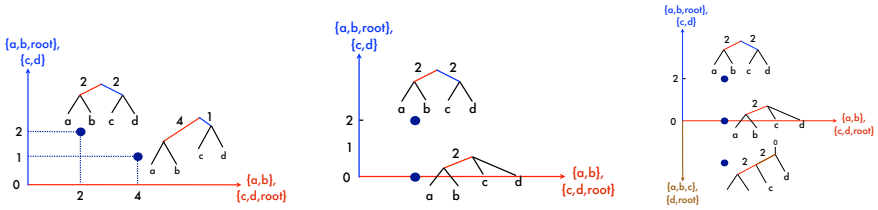

Fig. 2. Tree-space is a union of *orthants*, where each orthant is the non-negative part of a Euclidean space, corresponding to a specific leaf-labeled tree topology. **Left:** Within orthants the leaf-labeled tree topology is constant. **Middle:** In trees at the boundary of an orthant, at least one edge is contracted and described by a zero vector. **Right:** Orthant boundaries correspond to intermediate tree topologies. For simplicity, tree-space is illustrated using trees with edge length attributes rather than $3D$ shape. The same behavior carries over to edges with shape-vector attributes.

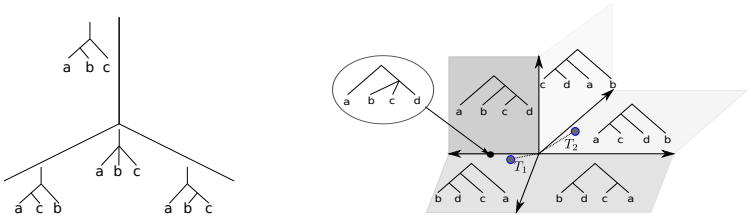

Fig. 3. Portions of tree-space $\mathcal{T}$ for trees with 3 and 4 leaves. At orthant boundaries, a large number of orthants may meet, creating *self-intersections* or *corners*.

by lines, and on manifolds, lines and directions are defined by geodesics and their tangent vectors. In tree-space, whenever a geodesic curve hits a corner or a branching point, there will be multiple geodesic extensions of the curve beyond the corner or branching point. This makes the extension of notions like PCA difficult, as we will return to below. A second difficulty occurs with optimization problems – iterative procedures like gradient descent become computationally intractable when they step close to the orthant boundaries, which can have many neighboring orthants in which the gradient has to be evaluated. In the worst case, if an iterative algorithm goes near the zero tree, the number of gradients that need to be evaluated is exponential in the number of leaves.

While the understanding of tree-space geometry is important, it is not a contribution of this paper, and we refer to [4, 16, 18] for a detailed description.

3 Tree-space statistics: theory and results

We will formulate statistical properties as solutions to optimization problems in the tree-space $\mathcal{T}$ [4]. Whereas optimization in tree-space is a very hard problem, we make the assumption that our finite dataset $\mathcal{X} \subset \mathcal{T}$ is large enough to give

a good approximate parametrization of the regions of tree-space in which our solutions will be found. More precisely, given the optimization problem

$$S = \text{argmin}_{x \in \mathcal{T}} f(x),$$

where $f \colon \mathcal{T} \to \mathbb{R}$ is an objective function whose minimizer is a statistic that we would like to compute, we shall instead solve the optimization problem

$$S = \text{argmin}_{x \in \mathcal{X}} f(x),$$

where we only consider solutions that are themselves part of the dataset $\mathcal{X}$. The resulting statistics are called *set statistics*.

We define a set of tree-space statistics covering the most basic types of data analysis, along with computable approximations. We demonstrate the usefulness of our algorithms by analyzing a set of 8016 airway trees extracted from repeated CT scans of 1692 subjects. For a detailed description of the dataset, see Sec. 4.

3.1 The mean airway tree

Given a subset $X = \{x_i\}_{i=1}^N \subset \mathcal{X}$, an iterative optimization scheme, *Sturm's algorithm* [2, 16, 21], already exists for computing the *Frèchet mean* of X in tree-space, defined as the minimizer

$$\mu = \text{argmin}_{x \in \mathcal{T}} \sum_{i=1}^{N} d^2(x, x_i).$$

Sometimes a faster approximation may be useful, and we define the set mean:

$$\mu = \text{argmin}_{x \in \mathcal{X}} \sum_{i=1}^{N} d^2(x, x_i)$$

Experiments. We compute set and Sturm means for our entire dataset $X = \mathcal{X}$ with $\sharp\mathcal{X} = 8016$. Plotting the Sturm mean and set mean together (Fig. 4) supports our basic hypothesis (*set statistics are good approximations*) since the set mean is visually a close approximation to the Sturm mean.

3.2 Analysis of variance: Tree-space PCA

Principal component analysis (PCA) is a basic tool for dimensionality reduction and analysis of variance in Euclidean spaces. In Euclidean and manifold PCA, the first principal component is often defined as the line, or more generally the geodesic curve γ, that minimizes the squared projection error [11, 12]:

$$PC1 = \text{argmin}_{\gamma} \sum_{i=1}^{N} d^2(x_i, \text{pr}_{\gamma}(x_i)), \tag{1}$$

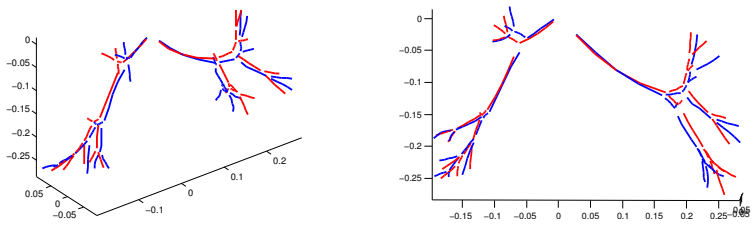

Fig. 4. Left and middle: The similarity of Sturm's Fréchet mean [16] (blue) to the approximate set mean (red) for 8016 airway trees supports the hypothesis that the dataset approximates the relevant parts of tree-space well near the population center.

where projection[1] is defined as $\mathrm{pr}_\gamma(x_i) = \mathrm{argmin}_{x\in\gamma} d(x, x_i)$, which is computed in tree-space using a golden ratio search along the geodesic, as done by Nye [17].

On manifolds the best-fit geodesic line is found by optimizing over all geodesic lines passing through the Fréchet mean, parameterized by their tangent directions. However, since the concept of direction is not well defined in tree-space, this approach does not carry over. Nye [17] computes a version of the first principal component by requiring it to pass through the *majority consensus tree* (a summary tree used in phylogenetics [3]), and considering only a subset of geodesics that have unique extensions beyond corners. Even this is computationally infeasible, and MCMC simulation methods are used to find the geodesic.

We choose to find the first principal component for $X = \{x_i\}_{i=1}^N \subset \mathcal{X}$ by optimizing over geodesic segments connecting pairs of trees, parametrized by the endpoint trees:

$$PC1 = \mathrm{argmin}_{x,x'\in\mathcal{T}} \sum_{i=1}^{N} d^2(x_i, \mathrm{pr}_{\gamma_{x,x'}}(x_i)), \tag{2}$$

where $\gamma_{x,x'}$ is the (unique) geodesic joining x to x'. However, since tree-space optimization is hard, finding the optimal line segment is also an open problem.

Proposition 3 *PC1 exists, but is not unique.*

Proof. Let $\bar{B}(\bar{0}, r) = \{x \in \mathcal{T} | d(x, \bar{0}) \leq r\}$ be the smallest closed ball about the tree-space origin which contains all data points. It is enough to optimize over geodesics whose endpoints lie in the compact ball $\bar{B}(\bar{0}, r)$, and a minimizer of (2) exists by compactness. The minimizing solution can be extended to a longer geodesic segment which is also a PC1; thus, the PC1 is not unique. □

Even though PC1 is not unique, it will usually contain a unique minimal segment containing all projected data points. In order to achieve a computable solution, we randomly sample endpoint pairs from our dataset and select the optimal geodesic segment, see Algorithm 1.

[1] The projection of a point onto a geodesic is unique in any $CAT(0)$ space ($\mathcal{T}$ is $CAT(0)$ [4]), as follows directly from the $CAT(0)$ property. This is also noted in [17].

Algorithm 1 Computing set PCA in tree-space

Input: Dataset $\mathcal{X} \subset \mathcal{T}$, subset $X = \{x_i\}_{i=1}^N \subset \mathcal{X}$ for which PC1 is computed; number M of random endpoint samples.
$m = 1$
while $m \leq M$ **do**
  Select random $x, x' \in \mathcal{X}$; endpoints$(m) = (x, x')$;
  Compute geodesic $\gamma_{x,x'}$.
  Compute projected dataset $\mathrm{pr}_{\gamma_{x,x'}}(X)$.
  score$(m) = \sum_{i=1}^N d^2(x_i, \mathrm{pr}_{\gamma_{x,x}}(x_i))$.
  $m = m + 1$
end while
$(x_0, x_0') =$ endpoints$(\mathrm{argmin}_m\{\mathrm{score}(m)\})$; $PC1 = \gamma_{x_0,x_0'}$.

Note that in $\mathbb{R}^d$, PCA serves many purposes: visualization of variance along PCs, dataset visualization (like multidimensional scaling), and dimensionality reduction. Our version of tree-space PC1 primarily serves the first purpose.

Experiments. Fig. 5 show $PC1$ computed from $X = \mathcal{X}$ consisting of 8016 airway trees belonging to 1692 subjects ($M = 18286$). Note that along PC1 the shape changes both in terms of vertical scale and angle of the lungs, which is consistent with deformation due to differences in inspiration level. In addition, there are topological changes arising from topological variance in the data.

3.3 Hypothesis testing in tree-space

We define one hypothesis test for the sample means and two for sample variance.

Hypothesis test for the mean. Let $A = \{a_i\}_{i=1}^{N_1}$ and $B = \{b_j\}_{j=1}^{N_2}$ be two samples from tree-space. To test for difference in means we use the univariate approach of Terriberry et al [22], with test statistic $T(A, B) = d(\hat{\mu}_A, \hat{\mu}_B)$, where $\hat{\mu}_A, \hat{\mu}_B$ are the Sturm or set means from Sec. 3.1. Under the null hypothesis the samples A and B are drawn from the same distribution on $\mathcal{T}$, and randomly permuting the elements of A and B should not affect the value of T.

Form the two-class data set $X = A \cup B \subset \mathcal{X}$ and consider partitions of X into datasets of size N_1 and N_2. Due to the size of X we cannot check all possible permutations, but compute the test statistics $T_m = d(\hat{\mu}_{A_m}, \hat{\mu}_{B_m})$, $m = 1, \ldots, M$, for M random partitions $X = A_m \cup B_m$, with $|A_m| = N_1$ and $|B_m| = N_2$. Comparing the T_m to the original statistic value $T_0 = d(\hat{\mu}_A, \hat{\mu}_B)$ we obtain a p-value approximating the probability of observing T_0 under the null hypothesis:

$$p = \frac{1 + \sum_{T_m \geq T_0, m \in \{1,\ldots,M\}} 1}{M + 1},$$

where the additional 1 is added to avoid $p = 0$, which is impossible in the limit where all permutations are tested [14].

Hypothesis test for the variance. Again, let A and B be two tree-space samples. Testing the equality of the variances σ_A and σ_B is formulated as a permutation test based on tree-space distances and means, using the test statistics

$$S_1(A,B) = \| \tfrac{1}{N_1} \textstyle\sum_{i=1}^{N_1} d(x_i, \hat{\mu}_A)^2 - \tfrac{1}{N_2} \sum_{j=1}^{N_2} d(y_i, \hat{\mu}_B)^2 \|,$$
$$S_2(A,B) = \| \tfrac{1}{N_1^2} \textstyle\sum_{i=1}^{N_1} \sum_{j=1}^{N_1} d(x_i, x_j)^2 - \tfrac{1}{N_2^2} \sum_{i=1}^{N_2} \sum_{j=1}^{N_2} d(y_i, y_j)^2 \|,$$

where S_1 tests variance about the mean, and S_2 tests the dataset spread.

Experiments. The defined test statistics were applied to samples from COPD patients and healthy subjects ($\sharp Z = 1692$, 842 with COPD and 850 healthy), one scan from each subject. There were 732 women and 960 men. The pairs of classes females/males and COPD/healthy were shown to come from tree-space distributions with different means and variances (Table 1). The tests used three different mean computations: The set mean optimized over the set $X = A \cup B$ with one airway tree per subject ($\sharp X = 1692$); the set mean optimized over the set $\mathcal{X}$ of all available airway trees ($\sharp \mathcal{X} = 8016$) and the Sturm mean. Only the latter two detected significant differences for means (5% significance level). Note that the set $\mathcal{X}$ of 8016 airway trees was only used as a tree-space discretization, while the samples $X = A \cup B$ were the same in all tests.

	Gender class separation			COPD class separation		
Test statistic	Set mean ($N = 1692$) $m = 10\,000$	Set mean ($N = 8016$) $m = 10\,000$	Sturm mean $m = 1\,000$	Set mean ($N = 1692$) $m = 10\,000$	Set mean ($N = 8016$) $m = 10\,000$	Sturm mean $m = 1\,000$
T	0.12	**0.0011**	**0.00099**	0.49	1.0	**0.00099**
S_1	**0.0034**	**0.045**	**0.00099**	**0.00099**	**0.000099**	**0.00099**
S_2	**0.0084**			**0.000099**		

Table 1. Computed p-values for class separation for tests on mean (T) and variance (S_1 and S_2). Recall that S_2 does not use the mean value. The $p = 1$ appears because the two set means coincide in this case.

Fig. 5. Left: Trees sampled along PC1. Note the changing tree topology. **Right:** The LDA geodesic segment γ_{LDA} shows trees with long branches in the COPD cluster, most likely due to missed branches in the segmentation.

3.4 Classification: Linear Discriminant Analysis in tree-space

A basic classification method in Euclidean space $\mathbb{R}^d$ is *Fisher's linear discriminant analysis (LDA)* [1,5], which searches for a linear codimension 1 classification boundary. The classification boundary is determined by any line which is orthogonal to it, up to a translation along the line, and by visualizing the data objects found along such a perpendicular line one can visualize how between-class variation affects the data objects. Whereas the concept of a linear classification boundary is not well-defined in tree-space, we can use our well-defined geodesic lines and projections onto them to formulate a version of tree-space LDA. One of the advantages of working in tree-space as opposed to analyzing trees using tree kernels is that every point in tree-space corresponds to a tree. Thus, with a version of Fisher's LDA in tree-space, the change in the geometric trees between the extremes of the classes can be visualized.

In Euclidean space $\mathbb{R}^d$, optimizing over LDA classification boundaries is equivalent to optimizing over lines $L \in \mathscr{L}$ orthogonal to the classification boundary, where $\mathscr{L}$ is the set of all lines in $\mathbb{R}^d$. The translation of the classification boundary is equivalent to a classification threshold for projected datapoints onto the line. When $N > d$ for dataset size N, given a training set $X = A \cup B$ consisting of two classes A and B, the optimal LDA line is defined as maximizing the distance between projected class means, normalized by projected class scatter $\hat{s}^2(\mathrm{pr}_l(A)) = \sum_{x \in A}(\mathrm{pr}_l(x) - \mu_{\mathrm{pr}_l(A)})^2$, etc. [5]:

$$L = \mathrm{argmax}_{l \in \mathscr{L}} \frac{d^2\left(\hat{\mu}(\mathrm{pr}_l(A)), \hat{\mu}(\mathrm{pr}_l(B))\right)}{\hat{s}^2(\mathrm{pr}_l(A)) + \hat{s}^2(\mathrm{pr}_l(B))}. \tag{4}$$

We define a version of tree-space LDA analogous to (4):

$$\mathrm{argmax}_{x,x' \in \mathcal{T}} \frac{d^2(\hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(A), \hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(B))}{\hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(A)) + \hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(B))}, \tag{5}$$

which we approximate, as with PCA, by only considering geodesic segments that pass between elements of a dataset $\mathcal{X}$ (often bigger than the training set X):

$$\gamma_{LDA} = \mathrm{argmax}_{x,x' \in \mathcal{X}} \frac{d^2(\hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(A), \hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(B))}{\hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(A)) + \hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(B))}. \tag{6}$$

Projected means and scatters on the geodesic segment can be reformulated as computing means and scatters on the real line, which is fast. Searching over all possible point pairs is not feasible, so as with PCA, we only use a set of randomly selected pairs of geodesic endpoints from the dataset. Given the LDA geodesic segment γ_{LDA}, we classify new points x_0 by assigning the class whose projected class mean is closer to the projection $\mathrm{pr}_{\gamma_{LDA}}(x_0)$. Note, however, that more refined methods for classification on the 1-dimensional geodesic line segment can easily be incorporated.

Note: In the Euclidean setting, when $d > N$, there will exist a line in $\mathbb{R}^d$, called *the maximal data piling direction* (MDP) [1], such that the classes A and

Algorithm 2 Computing set LDA in tree-space

Input: Classes A, B; number M of geodesic endpoint permutations.
$m = 1$
while $m \leq M$ **do**
  Select random $x, x' \in \mathcal{X}$, endpoints$(m) = (x, x')$
  Compute geodesic $\gamma_{x,x'}$
  Compute projected samples $\mathrm{pr}_{\gamma_{x,x'}}(A)$, $\mathrm{pr}_{\gamma_{x,x'}}(B)$
  Compute projected sample means $\hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(A))$, $\hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(B))$ and scatters $\hat{s}(\mathrm{pr}_{\gamma_{x,x'}}(A))$ and $\hat{s}(\mathrm{pr}_{\gamma_{x,x'}}(B))$
  $\mathrm{score}(m) = \frac{d^2(\hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(A), \hat{\mu}(\mathrm{pr}_{\gamma_{x,x'}}(B))}{\hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(A)) + \hat{s}^2(\mathrm{pr}_{\gamma_{x,x'}}(B))}$
  $m = m + 1$
end while
$(x_0, x_0') = \mathrm{endpoints}(\mathrm{argmin}_m\{\mathrm{score}(m)\})$; $\gamma_{LDA} = \gamma_{x_0,x_0'}$

B project onto two separate, single points. The MDP direction will coincide with the line L defined by (4) (as opposed to Fisher's LDA [1]), since this line gives zero denominators in (4). For $N \geq d$, LDA = MDP [1]. Thus, our version of LDA is really a generalization of MDP, and a generalization of LDA when $N > d$.

Experiments. We apply LDA by dividing the set of 1692 airway trees (one from each of 1692 subjects, 842 with COPD) into a training set (846, 421 with COPD) and a test set (846 individuals, 421 with COPD). Note that the dimension of tree-space is $d = 585$, so we have $d < N$. The LDA algorithm was ran with $M = 11829$, and a classification accuracy of 55.3% was obtained. Since LDA does not just give a binary classification, but in fact a whole geodesic segment, it is also useful for visualizing the differences between classes. The LDA geodesic segment γ_{LDA} is shown in Fig. 5.

4 Data

The developed techniques were applied to the analysis of the shape of airway trees classified by gender and COPD diagnosis. The airway trees were segmented [20] from 8016 CT-scans of 1692 subjects from a national lung cancer screening trial. Centerlines were extracted from the interior surface using the front propagation method of [15]. As the resulting centerlines are disconnected at bifurcation points, the end points were connected using a shortest path search within an inverted distance map of the interior surface. All images of the same subject were registered and the resulting deformation fields were used to remove centerline errors, such as spurious branches, following the approach of [19]. Based on the centerline shape, 20 segmental leaf labels ($R1 - R10, L1 - L10$) were automatically assigned to each airway tree using a geodesic labeling scheme [10], and branches below the segment level were discarded, producing a set of leaf-labeled airway trees with a fixed leaf label set. Airways for which not all leaf labels could be assigned were left out. For classification and hypothesis testing

experiments, a set of 1692 trees were selected, one from each subject, of which 842 were diagnosed with COPD and 732 were women.

5 Discussion and conclusion

We have defined a series of new tree-space statistics along with computable approximations of them. The main advantage of our approach is the ability to perform large-scale statistical analysis of real-world datasets and learn about connections between illness and anatomical tree geometry and structure in new ways. We show that the distribution of airway tree-shape is different in patients with COPD compared to healthy individuals. We provide a visual demonstration of mean tree-shape and tree-shape variability along the principal component. We perform LDA both for classification and visualization of the classification mechanism, and visualize the variation along the LDA component, which is consistent with increased difficulty of segmentation in COPD patients. There are no exact algorithms for computing some of our defined statistics, such as PCA or LDA, but the approximations give a rough estimate of the expected behavior of the optimal solutions. This may help in further developing how tree-space statistics should ideally be defined.

There are several disadvantages to our approach. Set statistics require a large dataset, which is not always available, and even when such a dataset is available it does not always give a good discretization of tree-space. Furthermore, due to the computational complexity of computing distances between large, unordered trees [6], we work with airway trees which have been labeled and cut off at the segmental level, reducing trees with $100 - 500$ branches to trees with ~ 40 branches, most likely discarding relevant information.

These disadvantages have potential solutions, which will be topics of future work. Small sample sizes can be helped using sampling methods in tree-space, e.g. similar to those used by Nye [17], ideally initialized by the original data set. Methods should be developed to take advantage of larger parts of the tree, either by developing heuristics for computing distances between unordered trees, or by using alternative distance measures.

In summary: We propose a set of tree statistics along with computable approximations and show that they work on a real medical dataset. Software will be made available online upon publication of the paper.

Acknowledgements

This research was supported by the Danish Council for Independent Research | Technology and Production Sciences; the Lundbeck Foundation; AstraZeneca; The Danish Council for Strategic Research; Netherlands Organisation for Scientific Research. M.O. was supported by a Fields-Ontario Postdoctoral Fellowship. The authors wish to thank Steve Marron and Tom Nye for insightful discussions.

References

1. J. Ahn and J.S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010.
2. M. Bacak. A novel algorithm for computing the Fréchet mean in Hadamard spaces. *Preprint, http://arxiv.org/abs/1210.2145*, 2012.
3. J.P. Barthélémy. The median procedure for n-trees. *J. Class.*, 3:329–334, 1986.
4. L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001.
5. R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification, 2. ed.* Wiley. 2001.
6. A. Feragen. Complexity of computing distances between geometric trees. In *SSPR/SPR*, pages 89–97, 2012.
7. A. Feragen, S. Hauberg, M. Nielsen, and F. Lauze. Means in spaces of tree-like shapes. In *ICCV*, 2011.
8. A. Feragen, Petersen J., D. Grimm, A. Dirksen, J.H. Pedersen, K. Borgwardt, and M. de Bruijne. Geometric tree kernels: Classification of copd from airway tree geometry. In *IPMI*, 2013.
9. A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, and F. Lauze. Towards a theory of statistical tree-shape analysis. *IEEE TPAMI, in press*, 2013.
10. A. Feragen, J. Petersen, M. Owen, P. Lo, L.H. Thomsen, M.M.W. Wille, A. Dirksen, and M. de Bruijne. A hierarchical scheme for geodesic anatomical labeling of airway trees. In *MICCAI (3)*, pages 147–155, 2012.
11. P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *TMI*, 23:995–1005, 2004.
12. S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58, 2010.
13. B. J. Jain and K. Obermayer. Structure spaces. *JMLR*, 10:2667–2714, 2009.
14. T. A. Knijnenburg, L. F. A. Wessels, M. J. T. Reinders, and I. Shmulevich. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.
15. P. Lo, B. van Ginneken, J.M. Reinhardt, and M. de Bruijne. Extraction of Airways from CT (EXACT09). In *2. Int. WS. Pulm. Im. Anal.*, pages 175–189, 2009.
16. E. Miller, M. Owen, and J. S. Provan. Averaging metric phylogenetic trees. *Preprint, http://arxiv.org/abs/1211.7046*, 2012.
17. T. M. W. Nye. Principal components analysis in the space of phylogenetic trees. *Ann. Statist.*, 39(5):2716–2739, 2011.
18. M. Owen and J.S. Provan. A fast algorithm for computing geodesic distances in tree space. *ACM/IEEE Trans. Comp. Biol. Bioinf.*, 8:2–13, 2011.
19. J. Petersen, V. Gorbunova, M. Nielsen, A. Dirksen, P. Lo, and M. de Bruijne. Longitudinal analysis of airways using registration. In *4. Int. WS. Pulm. Im. Anal.*, 2011.
20. J. Petersen, M. Nielsen, P. Lo, Z. Saghir, A. Dirksen, and M. de Bruijne. Optimal graph based segmentation using flow lines with application to airway wall segmentation. In *IPMI*, LNCS, pages 49–60, 2011.
21. K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. volume 338 of *Contemp. Math.*, pages 357–390. 2003.
22. T.B. Terriberry, S.C. Joshi, and G. Gerig. Hypothesis testing with nonlinear shape models. In *IPMI*, pages 15–26, 2005.
23. H. Wang and J. S. Marron. Object oriented data analysis: sets of trees. *Ann. Statist.*, 35(5):1849–1873, 2007.