Dept. of Computer Science, University of Copenhagen

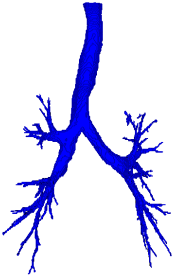
Statistical analysis of geometric trees

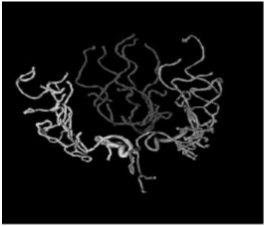
Aasa Feragen
aasa@diku.dk

Summer School on
Graphs in Computer Graphics, Image and Signal Analysis
Rutsker, Bornholm, Denmark, August 15, 2011

Geometric trees?

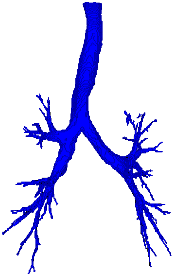
- A tree is a graph with no cycle

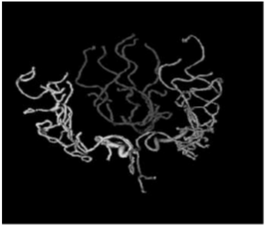




Geometric trees?

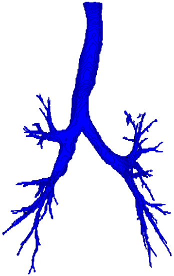
- A tree is a graph with no cycle
- In this talk, all trees have a root

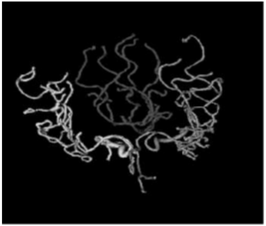




Geometric trees?

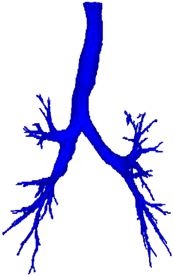
- A tree is a graph with no cycle
- In this talk, all trees have a root
- Algorithmic advantages over graphs

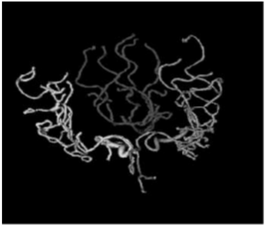




Geometric trees?

- A tree is a graph with no cycle
- In this talk, all trees have a root
- Algorithmic advantages over graphs
- Still difficult enough!





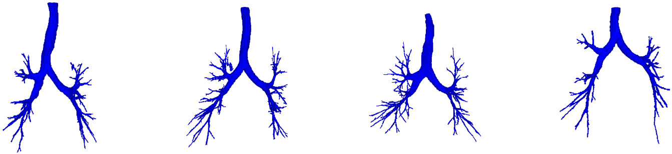
Outline

- Motivation through examples
- Modeling geometric trees
- Classical example: Tree edit distance
- Approach 1: The object-oriented data analysis of Marron et al
- Approach 2: Phylogenetic trees and their like
- Approach 3: Statistical tree-shape analysis
- Conclusions and open problems

Motivation through examples

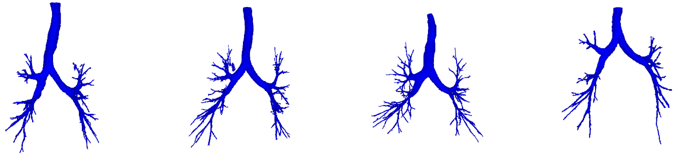
Example 1: Human airway trees

What does the average human airway tree look like? Nobody knows!



Example 1: Human airway trees

What does the average human airway tree look like? Nobody knows!



Properties of airway trees:

- Topology, branch shape, branch radius
- Somewhat variable topology (combinatorics) in *anatomical* tree
- Substantial amount of noise in *segmented* trees (missing or spurious branches), especially in COPD patients, *i.e. inherently incomplete data*

Example 1: Human airway trees

The raw segmented data is a tree embedded in $3D$

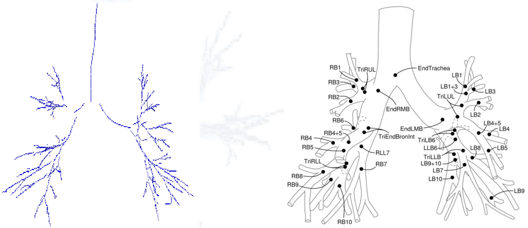


Figure: Right: Shamelessly borrowed from Tschirren, TMI 2005

- Computational problem: comparing unordered branches
- Can we attach anatomical labels to the branches?
- Related question: Can we order the branches?
- If yes, then the tree-structures are far less complex!

Example 1: Human airway trees

With statistical methods for tree-data, we could find out:

- how is the average airway tree, and how do the airway trees vary in different populations?
- are there different types of airway tree geometry, where some are more prone to illness than others?
- does the airway tree geometry change when you get ill?
- how do you distinguish a funny healthy structure from an ill structure? That is, how to analyze variation in tree data?

Example 2: Blood vessels

Figure: Left: Shamelessly borrowed from Wang and Marron, Ann. Statistics, 2007

Properties:

- Different vessel types, very different complexity
- Connectivity, branch length, branch shape
- Easier to segment than airways, hence more precise data.

Example 2: Blood vessels

With tree-statistical methods, we can:

- Find average vessel structure and variation in different populations
- Look for correlation between illness and tree geometry

Difference from airways:

- In general, more variable structure from person to person
- Properties depend highly on vessel type

Example 3: Phylogenetic trees

Properties of phylogenetic trees:

- Combinatorial tree with leaf labels
- branch lengths (describing time before division into species)
- Fixed leaf labels

Example 3: Phylogenetic trees

- Given a set of leafs

  (i.e. { human, gorilla, orangutan, computer scientist }),

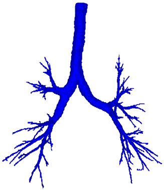
  different methods for establishing their phylogenetic tree will give different result. An average tree would be a bid for "the correct" phylogenetic tree.

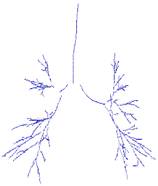
Modeling geometric trees

More general concept: Geometric trees

A geometric tree can be described as a combination of

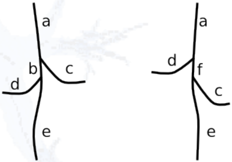
- tree topology (connectivity / combinatorics)
- geometric branch descriptors (branch shape, length, parametrization, weight, other attributes)





More general concept: Geometric trees

So why don't you just collect the edge information in a long vector and compute averages? Consider the *rather similar* trees:



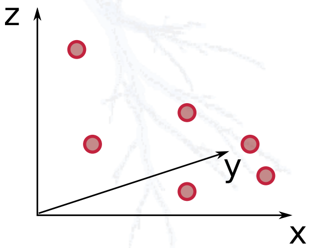
which are represented by the *rather different* vectors

$$(a, b, c, d, e) \text{ and } (a, d, f, e, c).$$

We need methods which can handle topological differences.

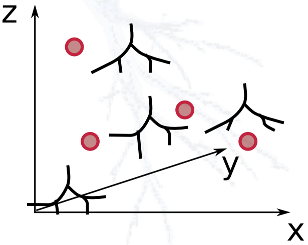
A thought:

- Usually: statistics in Euclidean space of n dimensions $\mathbb{R}^n$



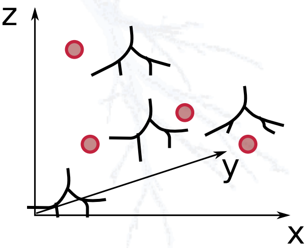
A thought:

- Usually: statistics in Euclidean space of n dimensions $\mathbb{R}^n$
- Imagine a "space of geometric trees"



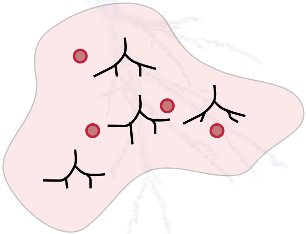
A thought:

- Usually: statistics in Euclidean space of n dimensions $\mathbb{R}^n$
- Imagine a "space of geometric trees"
- Each point represents a tree



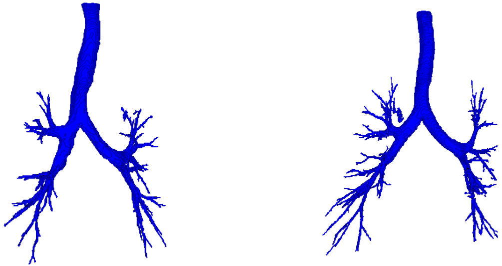
A thought:

- Usually: statistics in Euclidean space of n dimensions $\mathbb{R}^n$
- Imagine a "space of geometric trees"
- Each point represents a tree
- (And it is not really $\mathbb{R}^n$!)



A thought:

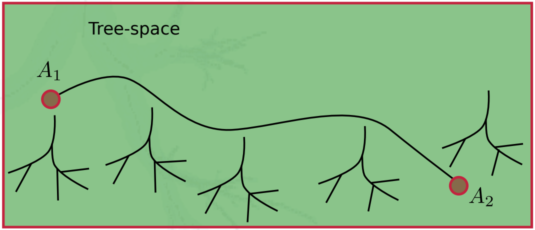
What if we were able to measure a "distance" (a metric) between two trees, which describes how similar (close) or different (far apart) they are?



Such distances would give us geometric tools to study the "space of all trees!"

Hold that thought and bring it further:

- Can we define distances between airway trees that correspond to *traversed distances* in the space of trees?



- We get distance *and* a canonical, shortest deformation (a *geodesic*) from A_1 to A_2.
- Play tree deformation movie

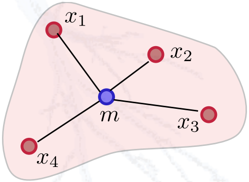
Hold that thought and bring it further:

Redefine statistics geometrically:

Definition

A *mean* of $\{x_1, \ldots, x_n\}$ is the point m which minimizes

$$f(m) = \sum_{i=1}^{n} d(x_i, m)^2.$$

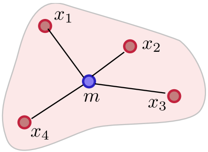


We seek situations where means are unique or locally unique.

What else can we do with a geometric framework?

With (locally) unique geodesic deformations, we can start to define:

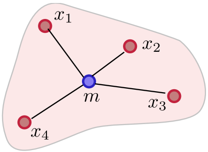
- shape of average tree



What else can we do with a geometric framework?

With (locally) unique geodesic deformations, we can start to define:

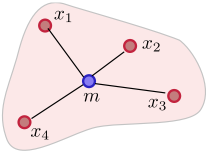
- shape of average tree
- "manifold" learning, dimensionality reduction, analysis of data variance

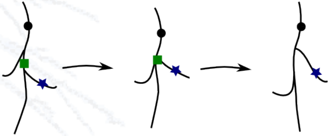


What else can we do with a geometric framework?

With (locally) unique geodesic deformations, we can start to define:

- shape of average tree
- ”manifold” learning, dimensionality reduction, analysis of data variance
- deformation-based registration and labeling





The model we are looking for: qualitative properties

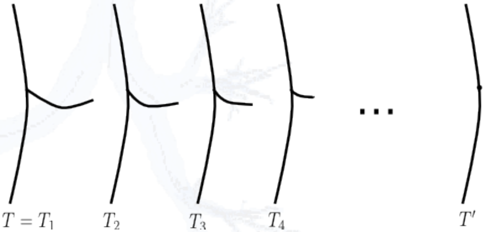


Figure: Tolerance of structural noise.

The model we are looking for: qualitative properties

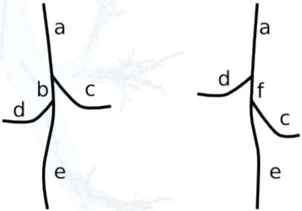


Figure: Tolerance of internal structural differences.

The model we are looking for: qualitative properties

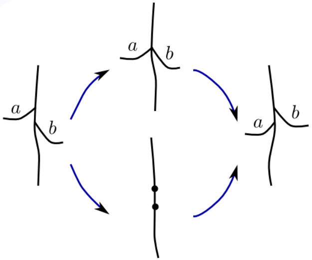


Figure: Top path: the a and b branches correspond to each other. Bottom path: They do not.

The model we are looking for: qualitative properties

Figure: What about these situations?

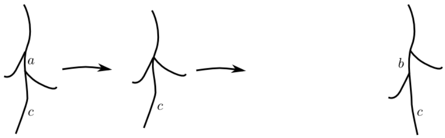
Classical example: Tree edit distance

Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.

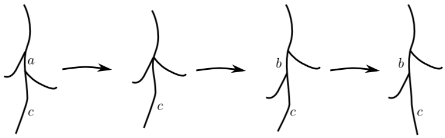
Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



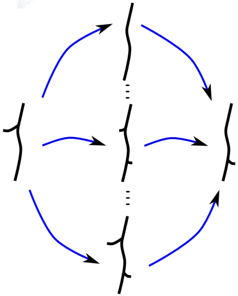
Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



Classical example: Tree edit distance (TED)

- Almost all geodesics between pairs of trees are non-unique (infinitely many).



- Then what is the average of two trees? Many!
- TED is *not* suitable for statistics.

Classical example: Tree edit distance (TED)

Most state-of-the-art approaches to distance measures and statistics on tree- and graph-structured data *are* based on TED!

- Wang and Marron: Object oriented data analysis: sets of trees. Annals of Statistics 35 (5), 2007.
- Ferrer, Valveny, Serratosa, Riesen, Bunke: Generalized median graph computation by means of graph embedding in vector spaces. Pattern Recognition 43 (4), 2010.
- Riesen and Bunke: Approximate Graph Edit Distance by means of Bipartite Graph Matching. Image and Vision Computing 27 (7), 2009.
- Trinh and Kimia, Learning Prototypical Shapes for Object Categories. CVPR workshops 2010.

Classical example: Tree edit distance (TED)

- The problems can be "solved" by choosing specific geodesics.
- Geometric methods can no longer be used for proofs, and one risks choosing problematic paths.[1]

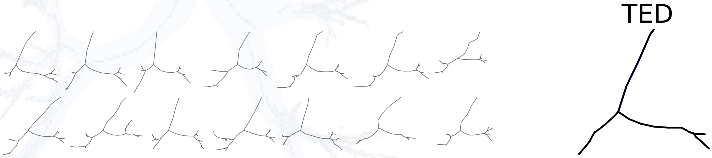


Figure: Right: Average upper airway trees computed using a method by Trinh and Kimia (CVPR workshops 2010) based on TED with the simplest possible choice of geodesics.

[1] Feragen, Lo, de Bruijne, Nielsen, Lauze: Towards a theory of statistical tree-shape analysis, submitted.

Classical example: Tree edit distance (TED)

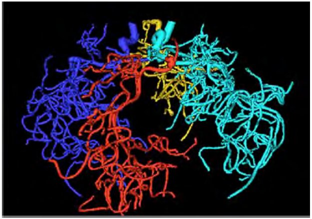
- TED *is* successfully used for other applications, which only require a distance – e.g classification
- TED is computationally demanding (especially between unordered trees, where it is generally NP hard to compute)
- The problem of finding faster algorithms, either heuristic or approximations, is a whole research field in itself.
- For statistics, we need something else – let's get to work!

Approach 1: The object-oriented data analysis of Marron et al [1]

[1] H. Wang and J. S. Marron. Object oriented data analysis: sets of trees. Annals of Statistics, 35(5):1849-1873, 2007.

Tree representation

- Framework built to study brain blood vessels



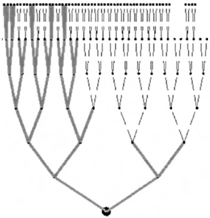
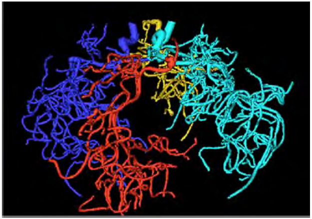


Figure: Figures from Aydin et al, 2009

Tree representation

- Framework built to study brain blood vessels
- "Trees" are rooted, ordered combinatorial trees (vertices connected by branches) with vertex attributes



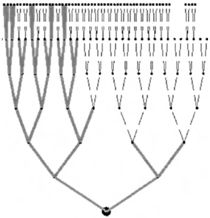
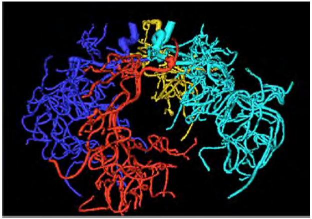


Figure: Figures from Aydin et al, 2009

Tree representation

- Framework built to study brain blood vessels
- "Trees" are rooted, ordered combinatorial trees (vertices connected by branches) with vertex attributes
- Vertices in the representative tree correspond to branches in the vessel tree



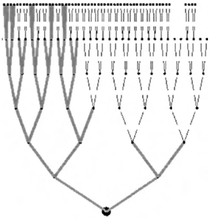
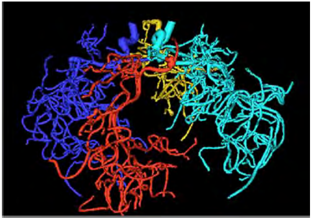


Figure: Figures from Aydin et al, 2009

Tree representation

- Framework built to study brain blood vessels
- "Trees" are rooted, ordered combinatorial trees (vertices connected by branches) with vertex attributes
- Vertices in the representative tree correspond to branches in the vessel tree
- Vertex attributes are geometric branch properties, such as branch start- and endpoint, length, radius etc



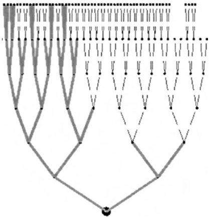
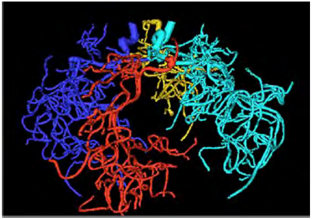


Figure: Figures from Aydin et al, 2009

Tree representation

- Framework built to study brain blood vessels
- "Trees" are rooted, ordered combinatorial trees (vertices connected by branches) with vertex attributes
- Vertices in the representative tree correspond to branches in the vessel tree
- Vertex attributes are geometric branch properties, such as branch start- and endpoint, length, radius etc
- Trees are represented via an ordered, maximal binary tree (a "union" of all the trees in the dataset) T with vertices V



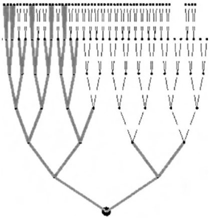
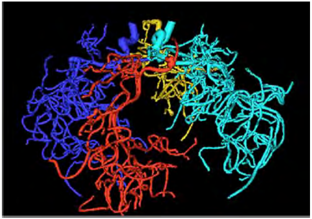


Figure: Figures from Aydin et al, 2009

Tree representation

- Framework built to study brain blood vessels
- "Trees" are rooted, ordered combinatorial trees (vertices connected by branches) with vertex attributes
- Vertices in the representative tree correspond to branches in the vessel tree
- Vertex attributes are geometric branch properties, such as branch start- and endpoint, length, radius etc
- Trees are represented via an ordered, maximal binary tree (a "union" of all the trees in the dataset) T with vertices V
- Vertex attributes form an ordered set of vectors $\{A_v\}_{v \in V}$, one for each vertex.



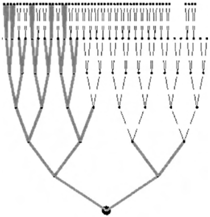
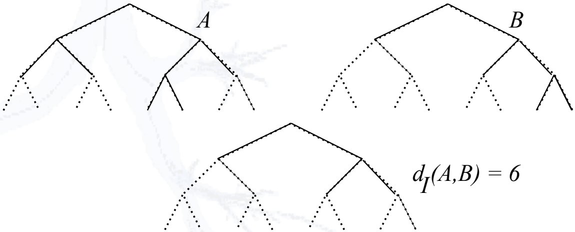


Figure: Figures from Aydin et al, 2009

Tree metric

- Define a metric on the space of trees with vector attributes:

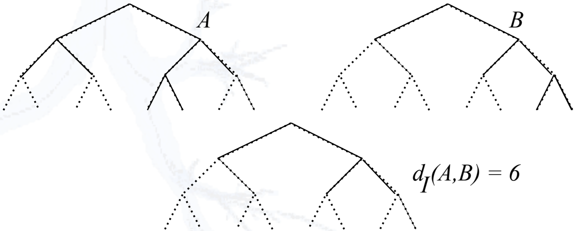
$$d(T_1, T_2) = d_I(T_1, T_2) + d_A(T_1, T_2)$$



Tree metric

- Define a metric on the space of trees with vector attributes:

$$d(T_1, T_2) = d_I(T_1, T_2) + d_A(T_1, T_2)$$

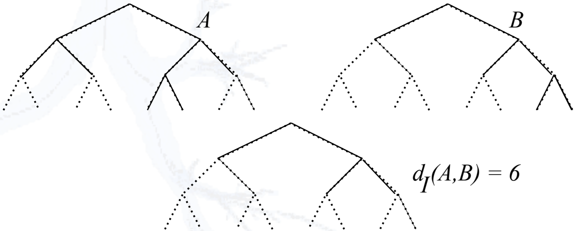


- d_I counts the number of TED leaf deletions/additions needed to turn T_1 into T_2,

Tree metric

- Define a metric on the space of trees with vector attributes:

$$d(T_1, T_2) = d_I(T_1, T_2) + d_A(T_1, T_2)$$

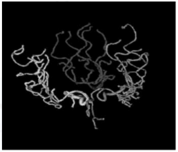


- d_I counts the number of TED leaf deletions/additions needed to turn T_1 into T_2,
- d_A is a weighted Euclidean metric on the attributes:

$$d_A(T_1, T_2) = \sqrt{\sum_{v \in V} c_v \|A_1(v) - A_2(v)\|^2},$$

"Object Oriented Data Analysis"

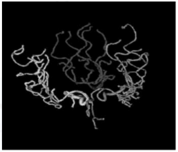
- Metric used for analyzing clinical data (brain blood vessels).



[2] Aydin, Pataki, Wang, Bullitt, Marron: A principal component analysis for trees, 2009

"Object Oriented Data Analysis"

- Metric used for analyzing clinical data (brain blood vessels).

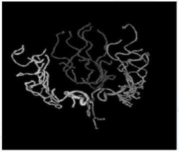


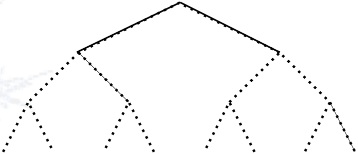
- Primary statistic: median-mean tree (combinatorial median, mean attributes)

[2] Aydin, Pataki, Wang, Bullitt, Marron: A principal component analysis for trees, 2009

"Object Oriented Data Analysis"

- Metric used for analyzing clinical data (brain blood vessels).



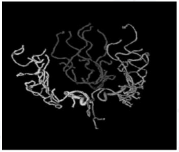


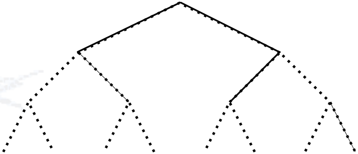
- Primary statistic: median-mean tree (combinatorial median, mean attributes)
- Secondary statistic: form of "PCA" where the principal components are "treelines"; describing directions in the tree where most of the variation is found. [2]

[2] Aydin, Pataki, Wang, Bullitt, Marron: A principal component analysis for trees, 2009

"Object Oriented Data Analysis"

- Metric used for analyzing clinical data (brain blood vessels).



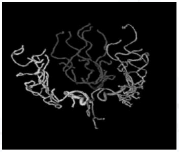


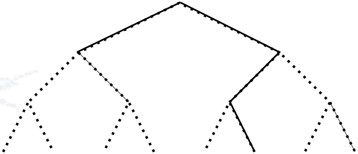
- Primary statistic: median-mean tree (combinatorial median, mean attributes)
- Secondary statistic: form of "PCA" where the principal components are "treelines"; describing directions in the tree where most of the variation is found. [2]

[2] Aydin, Pataki, Wang, Bullitt, Marron: A principal component analysis for trees, 2009

"Object Oriented Data Analysis"

- Metric used for analyzing clinical data (brain blood vessels).





- Primary statistic: median-mean tree (combinatorial median, mean attributes)
- Secondary statistic: form of "PCA" where the principal components are "treelines"; describing directions in the tree where most of the variation is found. [2]

[2] Aydin, Pataki, Wang, Bullitt, Marron: A principal component analysis for trees, 2009

Modeling issues

- The tree representation assumes a common, *ordered* underlying tree-structure

Modeling issues

- The tree representation assumes a common, *ordered* underlying tree-structure
- The metric has discontinuities

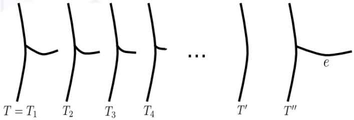


Figure: The sequence T_n with edge length attributes, does not converge. The length of e is 3 and all the c_e are 1/3, $\lim d(T_n, T')$ is the same as $\lim d(T_n, T'') = 1$.

Modeling issues

- The tree representation assumes a common, *ordered* underlying tree-structure
- The metric has discontinuities

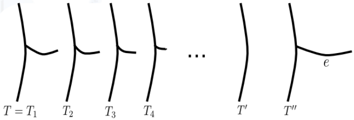


Figure: The sequence T_n with edge length attributes, does not converge. The length of e is 3 and all the c_e are $1/3$, $\lim d(T_n, T')$ is the same as $\lim d(T_n, T'') = 1$.

- The median-means defined are not unique

Modeling issues

- The tree representation assumes a common, *ordered* underlying tree-structure
- The metric has discontinuities

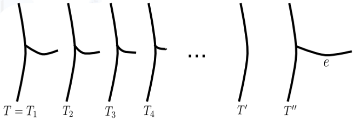


Figure: The sequence T_n with edge length attributes, does not converge. The length of e is 3 and all the c_e are 1/3, $\lim d(T_n, T')$ is the same as $\lim d(T_n, T'') = 1$.

- The median-means defined are not unique
- The treeline PCA is mostly combinatorial

Modeling issues

- The tree representation assumes a common, *ordered* underlying tree-structure
- The metric has discontinuities

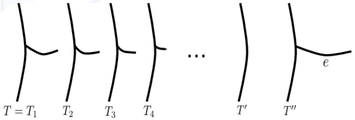


Figure: The sequence T_n with edge length attributes, does not converge. The length of e is 3 and all the c_e are $1/3$, $\lim d(T_n, T')$ is the same as $\lim d(T_n, T'') = 1$.

- The median-means defined are not unique
- The treeline PCA is mostly combinatorial
- Application-specific metric.

Summary

Pros:

- Easy to pass from the data tree to its representation
- Distances and statistical properties are easy and fast to compute
- First formulation of PCA for trees (or graphs?)

Summary

Pros:

- Easy to pass from the data tree to its representation
- Distances and statistical properties are easy and fast to compute
- First formulation of PCA for trees (or graphs?)

Cons:

- Modeling issues: Will not work for continuous, deformable trees, different topological structures
- Noise insensitivity, discontinuities
- No room for topological differences between trees except at leaves
- Statistical properties not well defined – for instance, a given set can have more than one median-mean

Approach 2: Phylogenetic trees and their like

Spaces of phylogenetic trees

- Billera et al. study the metric geometry of spaces of phylogenetic trees[3], which describe genetic development of species.

Figure: Figure borrowed from 3

[3] Billera, Holmes, Vogtmann: *Geometry of the space of Phylogenetic trees*, Adv. in Appl. Math, 2001.

Spaces of phylogenetic trees

- Billera et al. study the metric geometry of spaces of phylogenetic trees[3], which describe genetic development of species.
- Rooted trees with labeled leaves (so ordered trees) and length attributes on all edges.

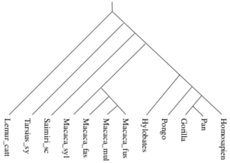


Figure: Figure borrowed from 3

[3] Billera, Holmes, Vogtmann: *Geometry of the space of Phylogenetic trees*, Adv. in Appl. Math, 2001.

Spaces of phylogenetic trees

- Billera et al. study the metric geometry of spaces of phylogenetic trees[3], which describe genetic development of species.
- Rooted trees with labeled leaves (so ordered trees) and length attributes on all edges.
- Metric geometry $\rightsquigarrow$ existence and uniqueness of geodesics and dataset centroids, computation of centroids of a set of phylogenetic trees.

Figure: Figure borrowed from 3

[3] Billera, Holmes, Vogtmann: *Geometry of the space of Phylogenetic trees*, Adv. in Appl. Math, 2001.

Spaces of phylogenetic trees

- Billera et al. study the metric geometry of spaces of phylogenetic trees[3], which describe genetic development of species.
- Rooted trees with labeled leaves (so ordered trees) and length attributes on all edges.
- Metric geometry $\rightsquigarrow$ existence and uniqueness of geodesics and dataset centroids, computation of centroids of a set of phylogenetic trees.
- Centroid computation is NP hard

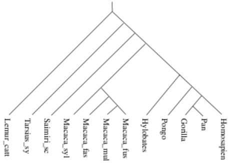


Figure: Figure borrowed from 3

[3] Billera, Holmes, Vogtmann: *Geometry of the space of Phylogenetic trees*, Adv. in Appl. Math, 2001.

Spaces of phylogenetic trees

- Billera et al. study the metric geometry of spaces of phylogenetic trees[3], which describe genetic development of species.
- Rooted trees with labeled leaves (so ordered trees) and length attributes on all edges.
- Metric geometry $\rightsquigarrow$ existence and uniqueness of geodesics and dataset centroids, computation of centroids of a set of phylogenetic trees.
- Centroid computation is NP hard
- Model applies directly to leaf-labeled trees with constant labels sets and edge length attributes

Figure: Figure borrowed from 3

[3] Billera, Holmes, Vogtmann: *Geometry of the space of Phylogenetic trees*, Adv. in Appl. Math, 2001.

Modeling phylogenetic trees

- Fix a set of n leaf labels, e.g. {human, gorilla, orangutan, computer scientist}, or $\{1, 2, 3, 4\}$.

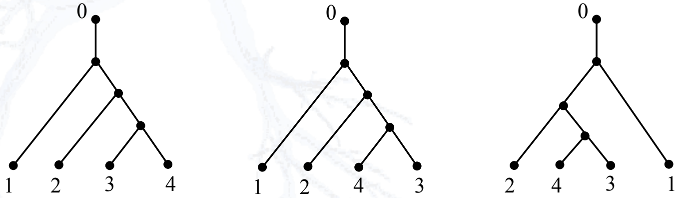


FIG. 6. Three pictures of the same tree.

Figure: Figure borrowed from Billera et al

Modeling phylogenetic trees

- Fix a set of n leaf labels, e.g. {human, gorilla, orangutan, computer scientist}, or $\{1, 2, 3, 4\}$.
- Build the binary tree with the corresponding leaves

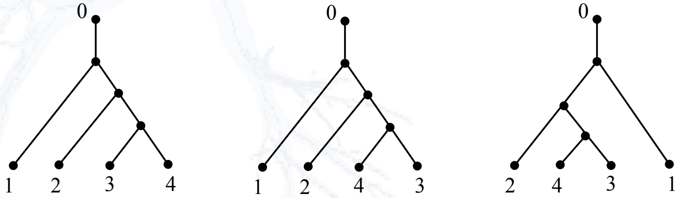


FIG. 6. Three pictures of the same tree.

Figure: Figure borrowed from Billera et al

Modeling phylogenetic trees

- Fix a set of n leaf labels, e.g. {human, gorilla, orangutan, computer scientist}, or $\{1, 2, 3, 4\}$.
- Build the binary tree with the corresponding leaves
- Attach lenghts $\in \mathbb{R}_+ = [0, \infty[$ to all branches, representing evolutionary length

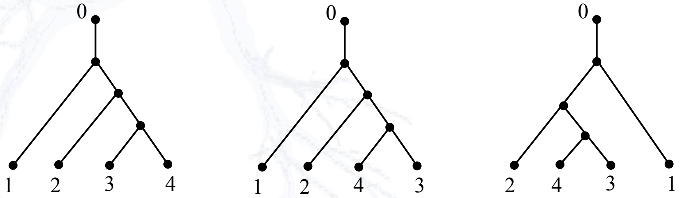


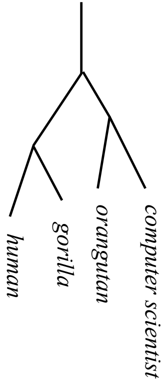
FIG. 6. Three pictures of the same tree.

Figure: Figure borrowed from Billera et al

Modeling phylogenetic trees

This gives a *space of phylogenetic trees*:

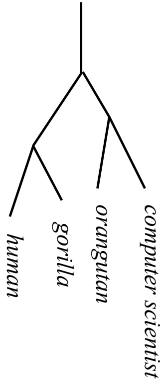
- For each type of binary tree with the given labels, we form a quadrant $\mathbb{R}^N_+ \subset \mathbb{R}^N$ ($N = \sharp$ branches in binary tree with n leaves)



Modeling phylogenetic trees

This gives a *space of phylogenetic trees*:

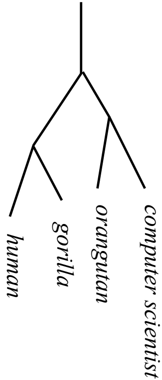
- For each type of binary tree with the given labels, we form a quadrant $\mathbb{R}^N_+ \subset \mathbb{R}^N$ ($N = \sharp$ branches in binary tree with n leaves)
- Now for a point $x \in \mathbb{R}^N_+$ the coordinate $x_i \geq 0$ is the length of the i^{th} branch



Modeling phylogenetic trees

This gives a *space of phylogenetic trees*:

- For each type of binary tree with the given labels, we form a quadrant $\mathbb{R}^N_+ \subset \mathbb{R}^N$ ($N = \sharp$ branches in binary tree with n leaves)
- Now for a point $x \in \mathbb{R}^N_+$ the coordinate $x_i \geq 0$ is the length of the i^{th} branch
- Glue the quadrants together along the natural branch collapses



The space of phylogenetic trees

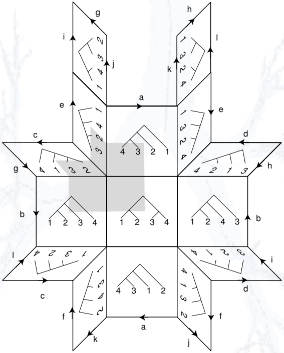


FIG. 4. Cubical tiling of $M_{0,5}$, where the arrows indicate oriented identifications.

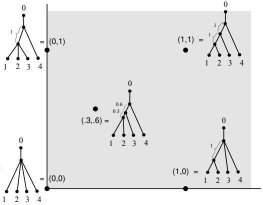


FIG. 8. The 2-dimensional quadrant corresponding to a metric 4-tree.

Figure: Figures shamelessly copied from Billera, Holmes, Vogtmann: Geometry of the space of Phylogenetic Trees

The really cool thing about the space of phylogenetic trees!

Theorem (Billera, Holmes, Vogtmann)

The space of phylogenetic trees is a $CAT(0)$ space

The really cool thing about the space of phylogenetic trees!

Theorem (Billera, Holmes, Vogtmann)

The space of phylogenetic trees is a $CAT(0)$ *space*

What does that mean?

Timeout: $CAT(0)$-spaces, our new favorite statistical playground?

Statistics in metric spaces?

Recall that a metric space is a space X of points with a distance measure d such that

- $d(x, y) = d(y, x)$
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)

Statistics in metric spaces?

Recall that a metric space is a space X of points with a distance measure d such that

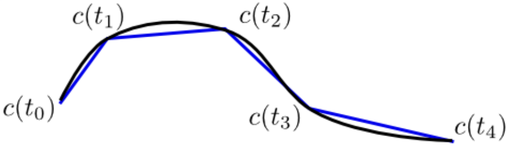
- $d(x, y) = d(y, x)$
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)

- In order to formulate statistics, we want to have geodesics. What does the word "geodesic" even mean in a metric space?

Geodesics in metric spaces

- Let (X, d) be a metric space. The length of a curve $c : [a, b] \to X$ is

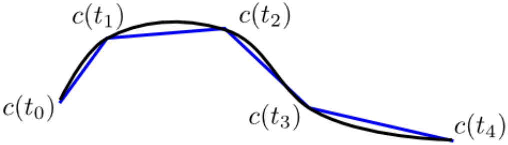
$$l(c) = sup_{a=t_0 \leq t_1 \leq \ldots \leq t_n = b} \sum_{i=0}^{n-1} d(c(t_i, t_{i+1})).$$



Geodesics in metric spaces

- Let (X, d) be a metric space. The length of a curve $c : [a, b] \to X$ is

$$l(c) = sup_{a=t_0 \leq t_1 \leq \ldots \leq t_n = b} \sum_{i=0}^{n-1} d(c(t_i, t_{i+1})).$$

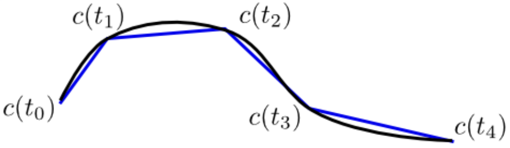


- A *geodesic* from x to y in X is a path $c : [a, b] \to X$ such that $c(a) = x, c(b) = y$ and $l(c) = d(x, y)$.

Geodesics in metric spaces

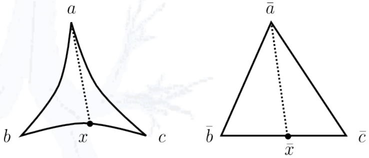
- Let (X, d) be a metric space. The length of a curve $c\colon [a, b] \to X$ is

$$l(c) = sup_{a=t_0 \leq t_1 \leq \ldots \leq t_n = b} \sum_{i=0}^{n-1} d(c(t_i, t_{i+1})).$$



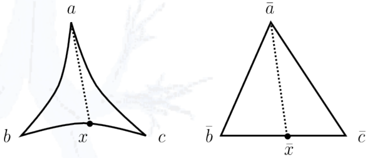
- A *geodesic* from x to y in X is a path $c\colon [a, b] \to X$ such that $c(a) = x, c(b) = y$ and $l(c) = d(x, y)$.
- (X, d) is a *geodesic space* if all pairs x, y can be joined by a geodesic.

Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x,a) \leq d(\bar{x},\bar{a})$.

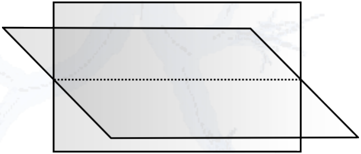
Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.
- A space has non-positive curvature if it is locally $CAT(0)$.

Curvature in metric spaces

Example



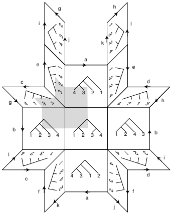


FIG. 4. Cubical tiling of $M_{0,5}$, where the arrows indicate oriented identifications.

Figure: *CAT*(0) spaces.

Curvature in metric spaces

Example

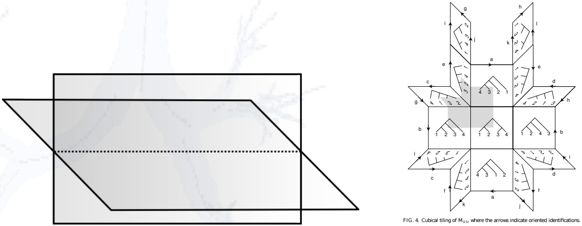


Figure: $CAT(0)$ spaces.

Theorem (see e.g. Bridson-Haefliger)

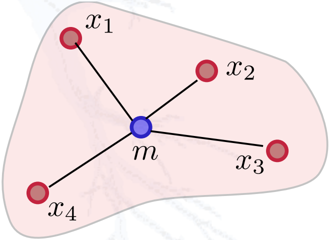
Let (X, d) be a $CAT(0)$ space; then all pairs of points have a unique geodesic joining them. □

Curvature in metric spaces

Subsets $\{x_1, \ldots, x_n\}$ in $CAT(0)$-spaces

Theorem

[4] ...have unique means, defined as $\operatorname{argmin} \sum d(x, x_i)^2$.



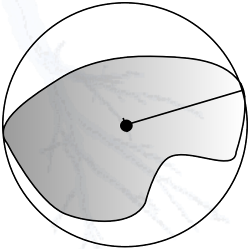
[4] Feragen, Hauberg, Nielsen, Lauze, *Means in spaces of treelike shapes*, ICCV 2011

Curvature in metric spaces

Subsets $\{x_1, \ldots, x_n\}$ in $CAT(0)$-spaces

Theorem (Bridson, Haefliger)

...have unique circumcenters, defined as the center of the smallest sphere containing all the $\{x_i\}_{i=1}^{s}$.



Curvature in metric spaces

Subsets $\{x_1, \ldots, x_n\}$ in $CAT(0)$-spaces

Theorem (Billera, Vogtmann, Holmes)

...have unique centroids, defined by induction on $|S| = n$:

- If $|S| = 2$, then $c(S)$ is the midpoint of the geodesic between the two elements of S.
- If $|S| = n > 2$ and we have defined $c(S')$ for all S' with $|S'| < n$, then denote by $c^1(S)$ the set $\{c(S') | S' \subset S, |S'| = n - 1\}$ and denote by $c^k(S) = c^1(c^{k-1}(S))$ when $k > 1$.
- If $c^k(S) \to p$ for some $p \in \bar{X}$ as $k \to \infty$, then $c(S) = p$ is the centroid of S. □

Timeout over: Back to the phylogenetic trees

What does this mean for the phylogenetic trees?

- We can compute average phylogenetic trees!
- Possible problem: Based on Billera, Holmes, Vogtmann, centroid phylogenetic trees have exponential computation time
- Moreover, geodesics between phylogenetic trees do not have obvious polynomial computation algorithms, either.

Computability?

Using the $CAT(0)$ properties, it is possible to prove:

Theorem

[4] *There is a polynomial time algorithm for computing the geodesic between two phylogenetic trees.*

[4] Owen, Provan: A Fast Algorithm for Computing Geodesic Distances in Tree Space, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011

Summary

Pros:

- A nice mathematical theory
- Computability
- Excellent modeling properties for phylogenetic trees
- $CAT(0)$ property gives potential for more statistical measurements

Summary

Pros:

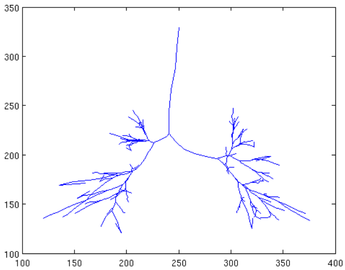
- A nice mathematical theory
- Computability
- Excellent modeling properties for phylogenetic trees
- $CAT(0)$ property gives potential for more statistical measurements

Cons:

- Does not carry directly over to trees with more geometric branch descriptors
- Fixed branch label set
- Ordered trees ($\Leftrightarrow$ leaf labels)
- No noise tolerance

Approach 3: Statistical tree-shape analysis

Motivating application: Airway shape analysis



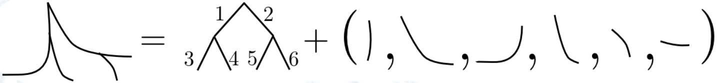
- Unlabeled (unordered) tree in 3D
- Different nr of branches
- Structural noise (missing/extra branches)

Tree representation

How to represent tree-like shapes mathematically?

Tree-like (pre-)shape = pair $(\mathscr{T}, x)$

- $\mathscr{T} = (V, E, r, <)$ rooted, ordered/planar binary tree, describing the tree topology (combinatorics)

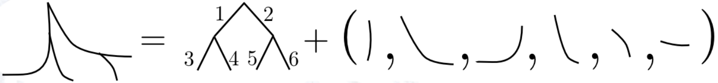


Tree representation

How to represent tree-like shapes mathematically?

Tree-like (pre-)shape = pair $(\mathscr{T}, x)$

- $\mathscr{T} = (V, E, r, <)$ rooted, ordered/planar binary tree, describing the tree topology (combinatorics)
- $x \in \prod_{e \in E} A$ a product of points in attribute space A describing edge shape



Tree representation

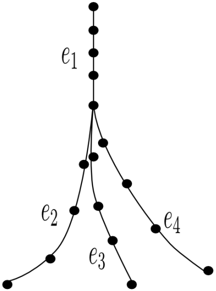
We are allowing collapsed edges, which means that

- we can represent higher order vertices
- we can represent trees of different sizes using the same combinatorial tree $\mathscr{T}$

(dotted line = collapsed edge = zero/constant attribute)

Tree representation

Edge representation through landmark points:
Edge shape space is $(\mathbb{R}^d)^n$, $d = 2, 3$.



The space of tree-like preshapes

Fix a maximal combinatorial $\mathscr{T}$. We use a finite tree; could reformulate for infinite trees.

Definition

Define the space of tree-like *pre*-shapes as the product space

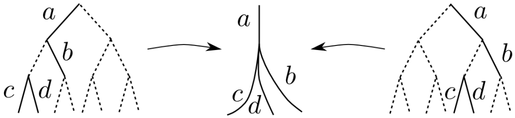
$$X = \prod_{e \in E} (\mathbb{R}^d)^n$$

where $(\mathbb{R}^d)^n$ is the edge shape space.

This is just a space of *pre-shapes*.

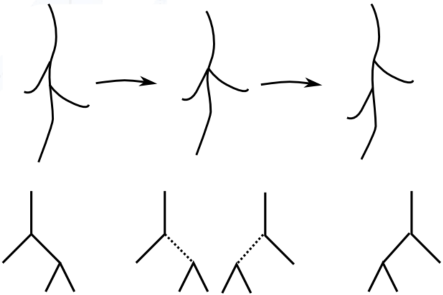
From pre-shapes to shapes

Many shapes have more than one representation



From pre-shapes to shapes

Not all shape deformations can be recovered as natural paths in the pre-shape space:

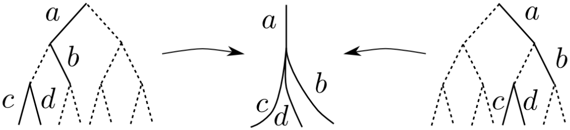


Shape space definition

- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.

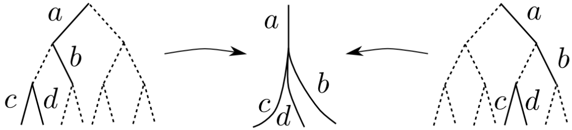
Shape space definition

- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.
- Glue together all points in X that represent the same tree-shape.



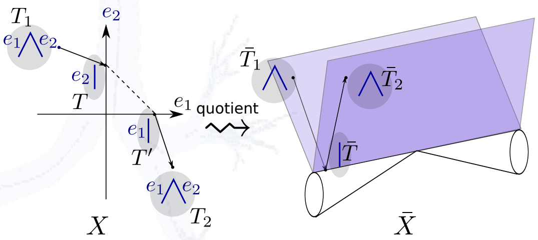
Shape space definition

- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.
- Glue together all points in X that represent the same tree-shape.



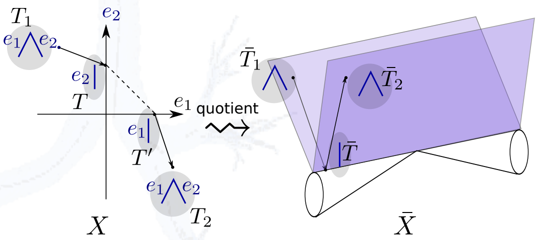
- This corresponds to identifying, or gluing together, subspaces $\{x \in X | x_e = 0 \text{ if } e \notin E_1\}$ and $\{x \in X | x_e = 0 \text{ if } e \notin E_2\}$ in X.

Shape space definition



- For the landmark point shape space this is just a folded Euclidean space; we call it $\bar{X}$.

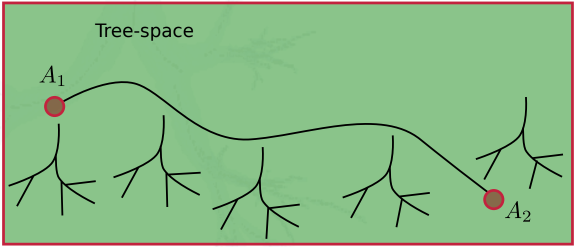
Shape space definition



- For the landmark point shape space this is just a folded Euclidean space; we call it $\bar{X}$.
- The Euclidean norm on X induces a metric on $\bar{X}$, called QED (Quotient Euclidean Distance) metric.

QED properties

It defines a geodesic metric space [5]



[5] Feragen, Lo, de Bruijne, Nielsen, Lauze: Geometries in spaces of treelike shapes, ACCV 2010

QED properties

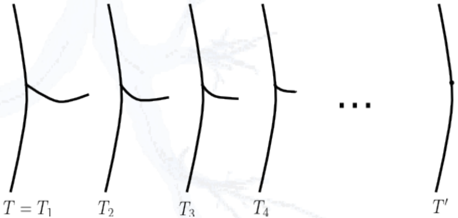
Example of a QED geodesic deformation:

Play movie

Note the tolerance of topological differences and natural deformation.

QED properties

Noise tolerance:



Sequences of trees with disappearing branches will converge towards trees without the same branch.

Curvature of shape space

Theorem

5

- Consider $(\bar{X}, \bar{d}_2)$, shape space with the QED metric.
- At generic points, this space has non-positive curvature, i.e. it is locally $CAT(0)$.
- Its geodesics are locally unique at generic points.
- At non-generic points, the curvature is unbounded.
- Sufficiently clustered datasets in $\bar{X}$ will have unique means, centroids and circumcenters. □

[5] Feragen, Lo, de Bruijne, Nielsen, Lauze: Geometries in spaces of treelike shapes, ACCV 2010

3D trees[6]

So far we talked about ordered tree-like shapes; what about unordered (spatial) tree-like shapes?

[6] Feragen, Lo, de Bruijne, Nielsen, Lauze: Towards a theory of statistical tree-shape analysis, submitted

3D trees[6]

- Unordered trees: Give a random order
- Denote by G the group of reorderings of the edges that do not alter the connectivity of the tree.
- The space of unordered trees is the space $\bar{\bar{X}} = \bar{X}/G$
- There is a (pseudo)metric on $\bar{\bar{X}}$ induced from the Euclidean metric on X.
- $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$ corresponds to considering all possible orders on $\bar{\bar{y}}$ and choosing the order that minimizes $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$.

[6] Feragen, Lo, de Bruijne, Nielsen, Lauze: Towards a theory of statistical tree-shape analysis, submitted

3D trees[6]

Theorem

- For the quotient pseudometric $\bar{\bar{d}}$ induced by either $\bar{d}_1$ or $\bar{d}_2$, the function $\bar{\bar{d}}$ is a metric and $(\bar{\bar{X}}, \bar{\bar{d}})$ is a geodesic space.
- At generic points, $(\bar{\bar{X}}, \bar{\bar{d}}_2)$ has non-positive curvature, i.e. it is locally $CAT(0)$.
- At generic points, geodesics are locally unique-
- At generic points, sufficiently clustered data has unique means, circumcenters, centroids.
- ...so everything we proved for ordered trees, still holds. □

[6] Feragen, Lo, de Bruijne, Nielsen, Lauze: Towards a theory of statistical tree-shape analysis, submitted

Examples [7]

[7] with S. Hauberg, M. Nielsen, F. Lauze, Means in spaces of treelike shapes, ICCV 2011

Averages in the QED metric

Synthetic data:

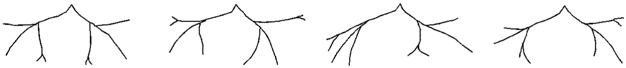


Figure: A small set of synthetic planar tree-shapes.

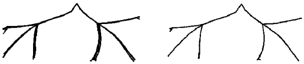


Figure: Left: Mean shape. Right: Centroid shape.

These choices of "average" give rather similar results.

Averages in the QED metric

Leaf vasculature data:

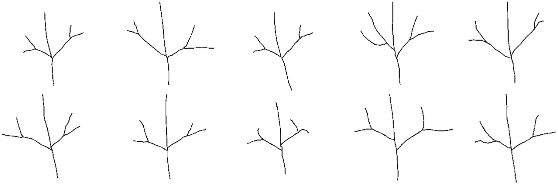


Figure: A set of vascular trees from ivy leaves form a set of planar tree-shapes.

Figure: a): The vascular trees are extracted from photos of ivy leaves. b) The mean vascular tree.

Averages in the QED metric

Airway tree data:

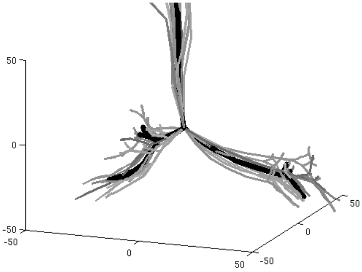


Figure: A set of upper airway tree-shapes along with their mean tree-shape.

Averages in the QED metric

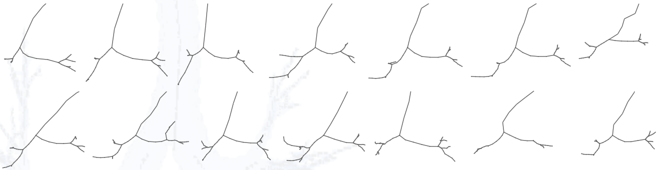


Figure: A set of upper airway tree-shapes (projected).[8]

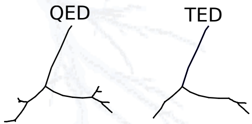


Figure: The QED and TED (algorithm by Trinh and Kimia) means.

[8] with P. Lo, M. de Bruijne, M. Nielsen, F. Lauze, submitted

Summary

Pros:

- Strong modeling properties
- Does not require labels, ordered, or same number of branches
- Continuous topological transitions in geodesics
- Local $CAT(0)$ property $\Rightarrow$ promising for statistical computations
- Good noise-handling properties

Summary

Pros:

- Strong modeling properties
- Does not require labels, ordered, or same number of branches
- Continuous topological transitions in geodesics
- Local $CAT(0)$ property $\Rightarrow$ promising for statistical computations
- Good noise-handling properties

Cons:

- Algorithmic properties
- Computational complexity

Conclusions and open problems

Conclusions

- The interplay between structure/topology/combinatorics and features (geometry) poses a challenging modeling problem
- There is often a tradeoff between modeling properties and computational complexity
- Analysis of tree-structured data can be attacked as a geometric, algorithmic, modeling, statistical, machine learning, ... -problem

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- How does the choice of branch attribute change the tree-space geometry in the different models?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- How does the choice of branch attribute change the tree-space geometry in the different models?
- Can the models be generalized to graphs?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- How does the choice of branch attribute change the tree-space geometry in the different models?
- Can the models be generalized to graphs?
- Can we find efficient algorithms for computing distances and statistical measurements?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- How does the choice of branch attribute change the tree-space geometry in the different models?
- Can the models be generalized to graphs?
- Can we find efficient algorithms for computing distances and statistical measurements?
- Our main goal: Large-scale statistical studies on medical data
  - Geometry-based biomarkers for disease (COPD)?
  - Anatomical modeling?

One more thing!

Means in the Space of Phylogenetic Trees

Talk by Megan Owen
on computational geometry and statistics
for Phylogenetic trees
30. august 2011 kl. 14 - 15 @DIKU