Geometry of the Space of Phylogenetic Trees

MATH285K - Spring 2010

Presented by: Aliz Raksi

Motivation

Trying to build phylogenetic trees from different subsets of our data often gives us different trees. How can we deal with this?
1. Combine all our data sets into one (using majority rule, strict consensus, or Bayesian combination) and build a single tree from the combined data.
2. Subdivide our data or do bootstrap resampling, build the different phylogenetic trees, then average them out.
Second method works better in finding original tree. How do we average out trees? Novel way: geometric context.

Also, phylogenetic trees we obtain have uncertainty, so we need to look at statistics. We can do this in a geometrical context. Recall that there are

$$(2n-3)!! = (2n-3)\times(2n-5)\times\ldots\times 3\times 1 = \frac{(2n-2)!}{2^{n-1}(n-1)!}$$

rooted binary semi-labeled trees with n leaves. Computing *best* tree using maximum parsimony and maximum likelihood is NP complete, so we use optimization methods that introduce random moves, and we might not get the best tree. Geometric context allows us to quantitatively compare different solutions. E.g a maximum likelihood tree can represent a point in a space of trees with branch lengths; it should then be possible to define isocontour regions around the estimated tree to build the desired confidence regions.

1 A Preliminary Attempt

Take a convex polytope called the *associahedron.* Identify vertices with the set of planar rooted binary trees with n leaves in a fixed order or, equivalently, with the set of triangulations of an $(n + 1)$-gon. These trees are linked by *rotation*. By “gluing” associahedra together, one can construct a space of planar labeled trees with n leaves, where each associahedron corresponds to a different ordering of the labels. Denote this space $\overline{M}_{0,n+1}$. This works, but we can simplify this model, since we are not interested in the orientation of trees embedded in the planes.

2 Construction of the Space of Trees

Consider a tree T, with interior edges $e_1,\ldots, e_r$ of lengths $l_1,\ldots, l_r$ respectively. If T is binary, then $r = n - 2$; otherwise $r < n - 2$. The vector $(l_1,\ldots, l_r)$ specifies a point in the positive open orthant $(0, \infty)^r$. To each other point in this orthant, we associate the unique metric n-tree which is combinatorially the same as T but has different edge lengths, specified by the coordinates of that point. Points on the boundary of the orthant, i.e., length vectors with at least one coordinate equal to zero, correspond to metric *n*-trees which are obtained from *T* by shrinking some interior edges of *T* to 0; thus each point in the orthant $[0, \infty)^r$ corresponds to a unique metric *n*-tree.

n=3 $(2 \cdot 3 - 3)!! = 3$ binary trees
Therefore, we have three 1-dimensional "orthants"

n=4 $(2 \cdot 4 - 3)!! = 15$ binary trees
Therefore, we have 15 2-dimensional quadrants

All quadrants share the same origin. Take line segment $x + y = 1$ in each quadrant to obtain graph called *link of the origin*.

For n=4, the entire space $\mathcal{T}_4$ is an infinite cone, and the link of the origin is the *Peterson graph*.

3 Combinatorics of the Space of Trees

LEM 3.1 (Relation to the Associahedron and Moduli Spaces) *Link of edges in* $\mathcal{T}_\mathrm{n}$ *correspond to the dual of the associahedron on* n *letters.*

Proof: The associahedron parameterizes the set of planar rooted trees with *n* leaves in a fixed order. ■

3.2 Combinatorics of the Link of the Origin

The link of the origin L_n has the homotopy type of a wedge of $(n - 1)!$ spheres of dimension $(n{-}3)$ Each of these spheres corresponds to the boundary of an associahedron embedded in $\mathcal{T}_\mathrm{n}$.

3.3 Tree Rotations

The distance between binary trees is measured by counting the number of rotations needed to change one tree to another. Here a *rotation* is a move which collapses an interior edge to zero and then expands the resulting degree 4 vertex into an edge and two degree 3 vertices in a new way. This move is also called a nearest neighbor interchange (NNI).

The maximal rotation distance between two trees on *n* leaves is $O(n \log n)$, while the maximal rotation distance between two trees contained in the same associahedron is exactly $2n - 6$.

4 Geometry of the Space of Trees

4.1 Non-positive curvature

A metric space *X* is said to have *non-positive curvature* if triangles in *X* are “at least as thin” as Euclidean triangles. More precisely, *X* is said to be CAT(0) if the following is true: given any three points *a*, *b*, and *c* in *X*, with distances $d_1 = d(b, c)$, $d_2 = d(a, c)$, $d_3 =$

$d(a, b)$, form a “comparison triangle” in the Euclidean plane with vertices *a*, *b*, and *c* with side lengths $d_1 = d(b', c')$, $d_2 = d(a', c')$, $d_3 = d(a', b')$. If *x* is a point on the geodesic from *a* to *b*, at distance *d* from *a*, find the corresponding point *x* on the straight line from *a'* to *b'* at distance *d* from *a'*. Then $d(x, c) \leq d(x', c')$.

LEM 4.1 T_n is a CAT(0) space.

Proof: We first subdivide each orthant into the unit cubes having integral vertices. The space $\mathscr{T}_n$ is then a cubical complex. A theorem of Gromov states that a cubical complex is CAT(0) if and only if the link of every vertex is a *flag* complex, i.e., a simplicial complex in which a simplex belongs to the complex if and only if its entire 1-skeleton does.

Let *v* be an arbitrary vertex of the cube complex, which lies in the interior of a (unique) orthant of dimension *k*. This orthant corresponds to a tree with *k* interior edges, and thus to a set *S* of *k* pairwise compatible partitions of $\{0,..., n\}$. If *k* is maximal, i.e., $k = n - 2$, the link of *v* is a triangulated sphere, which we think of as the *k*-fold suspension of the empty set. In general, the link of *v* is the *k*-fold suspension of the subcomplex of L_n spanned by all partitions compatible with *S*. Since this itself is a flag complex, and since the suspension of a flag complex is again flag, this completes the proof. ■

4.2 Geodesics

When measuring the distance between two points in Tn, the distance between any two points in the same orthant is simply the usual Euclidean distance. If two points are in different orthants, we can join them by a sequence of straight segments, with each segment lying in a single orthant, and add up the lengths of the segments. A segmented path giving the smallest distance between two points is called a *geodesic*.

The path between two trees T and T' in Tn that we obtain by connecting T to the origin by a straight line segment, then connecting the origin to T' by another straight line segment, is called the *cone path* from T to T'. The cone path may or may not be a geodesic.

4.3 and 5

The rest of the paper covers lemmas about geodesics, and also describes the geometric properties of centroids in the tree space. Finally, it presents a real data example using the phylogenetic tree of primates. However, due the length of the paper, I will not go into these. Please refer to the original paper for further details.

References

Billera, L., S. Holmes, and K. Vogtman. 2001. Geometry of the space of phylogenetic trees. Adv. Appl. Math. 27:733–767.