Confidence Regions and Averaging for Trees

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier, France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

Joint Work with
Louis Billera (Mathematics, Cornell)
Persi Diaconis (Mathematics and Statistics, Stanford)
Karen Vogtmann (Mathematics, Cornell).
Confidence Regions and Averaging for Trees

Combinatorics and Geometry for tree space

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier, France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

Joint Work with
Louis Billera (Mathematics, Cornell)
Persi Diaconis (Mathematics and Statistics, Stanford)
Karen Vogtmann (Mathematics, Cornell).
Tree for all
Tree for all

Where do collections of trees come from?

Family trees or phylogenetic trees whose leaves are different evolutionary entities (species, genes, populations).
Clustering trees for multiple clustering arrangements obtained over time for the same entities, cells or genes in gene expression for instance.
Decision trees, such as Classification and Regression trees,
• whose leaves are the single observations before the trees have been pruned back.
Decision trees, such as Classification and Regression trees, whose leaves are the single observations before the trees have been pruned back.
Phylogenetic Trees
Confidence Statements for trees
## DNA Data for 12 species of primates

Mitochondria, 898 characters on 12 species.


Trees are built from DNA data such as the following:

<table>
<thead>
<tr>
<th>Species</th>
<th>DNA Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lemur_cat</strong></td>
<td>AAGCTTCATA</td>
</tr>
<tr>
<td><strong>Tarsius_s</strong></td>
<td>AAGTTTCATT</td>
</tr>
<tr>
<td><strong>Saimiri_s</strong></td>
<td>AAGCTTCACC</td>
</tr>
<tr>
<td><strong>Macaca_sy</strong></td>
<td>AAGCTTCTCC</td>
</tr>
<tr>
<td><strong>Macaca_fa</strong></td>
<td>AAGCTTCTCC</td>
</tr>
<tr>
<td><strong>Macaca_mu</strong></td>
<td>AAGCTTTTCTT</td>
</tr>
<tr>
<td>Hylobate</td>
<td>AAGCTTTTACA</td>
</tr>
<tr>
<td><strong>Pongo</strong></td>
<td>AAGCTTCACC</td>
</tr>
<tr>
<td><strong>Gorilla</strong></td>
<td>AAGCTTCACC</td>
</tr>
<tr>
<td><strong>Pan</strong></td>
<td>AAGCTTCACC</td>
</tr>
<tr>
<td><strong>Homosapie</strong></td>
<td>AAGCTTCACC</td>
</tr>
</tbody>
</table>

*Note: The DNA sequences are presented in uppercase for clarity.*
Color Coded version of the data, after alignment
Two Trees Built with this data by parsimony

---Homosapien
  +---11
    ! ! +---Pan
    +---9  +---10
    ! ! ! +---Gorilla
    +---------8 !
    ! ! ! +------Pongo
    ! ! !
    ! ! ! +------Hylobates
    +---7 !
    ! ! ! +---Macaca_fus
    ! ! !
    ! ! ! +---6
    ! ! ! +---5 +---Macaca_mul
    +---3 !
    ! ! ! +---4 +---Macaca_fas
    ! ! !
    +---2 ! +---Macaca_syl
    ! ! +---Saimiri_sc
    +---1 !
    ! +---Tarsius_sy
    +---Lemur_catt

---Homosapien
  +---11
    ! ! +---Pan
    +---10 +---Pan
    ! ! ! +---Gorilla
    +---------8 !
    ! ! ! +------Pongo
    ! ! !
    ! ! ! +------Hylobates
    +---7 !
    ! ! ! +---Macaca_fus
    ! ! !
    ! ! ! +---6
    ! ! ! +---5 +---Macaca_mul
    +---3 !
    ! ! ! +---4 +---Macaca_fas
    ! ! !
    +---2 ! +---Macaca_syl
    ! ! +---Saimiri_sc
    +---1 !
    ! +---Tarsius_sy
    +---Lemur_catt
Confidence Regions

From Efron, Halloran, Holmes, (1996)[0].
What is the curvature of the boundary?
How many neighbors does a region have?

From Efron, Halloran, Holmes, (1996)[0].
Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.
Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.
Classification And Regression Trees

Bootstrap for CART trees. 
*Average* Tree seems to be much better than one tree built with all the data. (Amit and Geman (1998))

What are the ways of averaging trees?
Consensus majority rule or strict consensus.
Clustering Trees

One hierarchical cluster per matrix.
Cube of matrices.
How do we flatten or follow in time the trees as they ‘change’, we want a space for following the pathways of trees.
Clustering Trees
For micro-arrays (Eisen et al, 1999)
Binary Trees
Other parameterizations for Tree Space


We would like to fill in the matching polytope.
Matchings and Binary Semi-labeled Trees This comes from Diaconis and Holmes (1998) A matching of $2(n-1)$ objects is a pairing off, without care for order within pairs or between pairs.

The Same matchings:

$(1, 4)(2, 5)(3, 6)$
$(6, 3)(4, 1)(2, 5)$
$(5, 2)(3, 6)(1, 4)$
Matchings and Binary Semi-labeled Trees This comes from Diaconis and Holmes (1998) A matching of $2(n-1)$ objects is a pairing off, without care for order within pairs or between pairs.

The Same matchings:

$$(1, 4)(2, 5)(3, 6)$$
$$(6, 3)(4, 1)(2, 5)$$
$$(5, 2)(3, 6)(1, 4)$$

Call $B_{n-1}$ the subgroup of $S_{2n-2}$ that fixes the pairs

$$\{1, 2\}\{3, 4\} \ldots \{2n - 3, 2n - 2\}$$

then

$$\mathcal{M}_{n-1} = S_{2n}/B_{n-1}$$

and

$$|\mathcal{M}_{n-1}| = \frac{(2n - 2)!}{2^{n-1}(n-1)!} = (2n - 3)!! = (2n - 3) \times (2n - 5) \times \cdots \times 3 \times 1$$
This formula for the number of trees was first proved using generating functions by Schroder (1873)\[0\].

\((S_{2n-2}, B_{n-1})\) form a Gelfand pair Diaconis and Shahshahani (1987) [?].

\[ L(M_{n-1}) = V_1 \oplus V_2 \oplus \ldots \oplus V_\lambda \]

A multiplicity free representation.

\[ L(M_{n-1}) = \bigoplus_{\lambda \vdash n} S^{2\lambda} \]

where the direct sum is over all partitions \(\lambda\) of \(m\), \(2\lambda = (2\lambda_1, 2\lambda_2, \ldots, 2\lambda_k)\) and \(S^{2\lambda}\) is associated irreducible representation of the symmetric group \(S_{2m}\).

Just to take the first few: for \(\lambda = n - 1\) \(S^\lambda\) are the constants, and this gives the sample size. for \(\lambda = (n-2, 1)\), \(S^\lambda\) are the number of times each pair appears. for \(\lambda = (n-3, 2)\), \(S^\lambda\) are the number of times partition of 4 appears in the tree. for \(\lambda = (n-3, 1, 1)\), \(S^\lambda\) are the number of times 2 pairs appear simultaneously. This decomposition is similar to what was done by Diaconis for permutation data.\[0\]
Matchings are useful

- For going through all trees systematically. (Gray code \([0]\) for Trees)
- Doing vigorous random walks on tree space.
- Doing Fourier Analysis on Tree Data.

But the matching distance is not satisfactory to the biologists.
The Matching Polytope
Mallow’s Model

\[ P(\tau) = \]
Mallow’s Model

\[ P(\tau) = Ke^{-\lambda d(\tau, \tau_0)}, \quad K \text{ is a normalizing constant} \]
Mallow’s Model

\[ P(\tau) = Ke^{-\lambda d(\tau, \tau_0)}, \quad K \text{ is a normalizing constant} \]

- Exponential family, it needs:
Mallow’s Model

\[ P(\tau) = Ke^{-\lambda d(\tau, \tau_0)}, \quad K \text{ is a normalizing constant} \]

- Exponential family, it needs:
  - A central tree \( \tau_0 \)
  - A distance between trees \( d(\tau, \tau_0) \)
Mallow’s Model

\[ P(\tau) = Ke^{-\lambda d(\tau, \tau_0)}, \quad K \text{ is a normalizing constant} \]

- Exponential family, it needs:
  - A central tree \( \tau_0 \)
  - A distance between trees \( d(\tau, \tau_0) \)

- It is possible to extend this to make a Bayesian model with a symmetric apriori distribution for \( \tau_0 \).
Mallow’s Model

\[ P(\tau) = Ke^{-\lambda d(\tau, \tau_0)}, \quad K \text{ is a normalizing constant} \]

- Exponential family, it needs:
  - A central tree \( \tau_0 \)
  - A distance between trees \( d(\tau, \tau_0) \)

- It is possible to extend this to make a Bayesian model with a symmetric apriori distribution for \( \tau_0 \).

Distances and centroids are essential
The permuto-associahedron
The permuto-associahedron

A book on polytopes. (Ziegler)

But the trees are extreme points
● A neighbor relation must be biologically acceptable (strong symmetry requirement)

● The trees must sit in the center of the regions.
• A neighbor relation must be biologically acceptable (strong symmetry requirement)

• The trees must sit in the center of the regions.

We could take the quotient of the polytope, but a direct construction is easier to visualize.
Rotation Moves
Rotation Moves
Rotation Moves

This is known to the biologists as the NNI moves.
Boundary for trees with 3 leaves.
Boundary for trees with 3 leaves

The geometric construction is joint work with Louis Billera and Karen Vogtmann [0]
The quadrant for one tree.
The quadrant for one tree
The quadrant for one tree.
The quadrant for one tree

(0,0)  (0,1)  (1,0)  (1,1)
The quadrant for one tree
The quadrant for one tree
The cube complex

A binary $n$-tree has the maximal possible number of interior edges $(n - 2)$. It determines the largest possible dimensional quadrant which is $n - 2$-dimensional. The quadrant corresponding to each tree which is not binary appears as a boundary face of at least three binary trees; in particular the origin of each quadrant corresponds to the (unique) tree with no interior edges.
The cube complex

A binary $n$-tree has the maximal possible number of interior edges $(n - 2)$. It determines the largest possible dimensional quadrant which is $n - 2$-dimensional. The quadrant corresponding to each tree which is not binary appears as a boundary face of at least three binary trees; in particular the origin of each quadrant corresponds to the (unique) tree with no interior edges. $T_n$ is built by taking one $n - 2$-dimensional quadrant for each of the $(2n - 3)!! = (2n - 3) \ast (2n - 5) \ast \cdots \ast 5 \ast 3 \ast 1$ possible binary trees, and gluing them together along their common faces.

For $n = 3$ there are three binary trees, each with 1 interior edge. Each tree thus determines a 1-dimensional “quadrant,” i.e. a ray from the origin. The three rays are identified at their origins. Figure for $n=3$. 
Three quadrants sharing a ray for \( n=4 \)

Note that the bottom boundary rays form a copy of \( T_3 \) embedded in \( T_4 \).
Three quadrants sharing a ray for $n=4$

Note that the bottom boundary rays form a copy of $T_3$ embedded in $T_4$. In general, $T_n$ contains many embedded copies of $T_k$ for $k < n$. For example, there is a copy of $T_k$ for each interior vertex $P$ such that the associated tree $T_P$ has $k$ leaves.
Link to the origin

All 15 quadrants for $n = 4$ share the same origin. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray; this graph is called the *link of the origin*. 
\[(0,1)\]

\[x + y = 1\]

\[(1,0)\]

\[(0,1)\]
Associahedron
given any three points $a, b$ and $c$ in $X$, with distances $d_1 = d(b, c), d_2 = d(a, c)$ and $d_3 = d(a, b)$, form a “comparison triangle” in the Euclidean plane with vertices $a', b'$ and $c'$ with side lengths $d_1 = d(b', c'), d_2 = d(a', c')$ and $d_3 = d(a', b')$. 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 $d(x', c') \leq d(x, c)$.

For further reference see Gromov [0].
Consequences

- We can define centroids. For two points, take their geodesic and then take the middle of it. \textbf{Theorem}

- We can make convex hulls. \rightarrow \text{Confidence regions.}

- We can use Mallow’s model.

- We know the number of neighbors of each tree.
Consequences

- We can define centroids. For two points, take their geodesic and then take the middle of it. **Theorem**

- We can make convex hulls. ⟷ Confidence regions.

- We can use Mallow’s model.

- We know the number of neighbors of each tree. \((2(n - 2))\)
Tree trajectories
Tree trajectories
How can abstract mathematics help?

• Decompositions that can be generalisable.

• Geometric Picture of Tree Space
  – A space for comparisons.
  – Ways of projecting.
  – Follow trees as they change, (paths of trees)
  – Centroids of trees
  – Neighborhoods (convex hulls of trees)....
  – Averages of trees

• Justification of commonsense, ground for generalizations.
Theorem In any CAT(0) space $X$,

1. Centroids exist for any finite set $Y \subset X$.

2. The centroid function is convex.

Proof: The proof is by induction on $n = |Y|$. The case $n = 2$ is in Bridson and Haefliger, 1999.

Suppose $n \geq 3$ and we have a convex centroid function $c(W)$ for $|W| = n - 1$. Let $Y = \{y_1, \ldots, y_n\}$ and $Y_i = Y \setminus \{y_i\}$. Suppose $Y$ has diameter $M$. Then by convexity for $(n - 1)$-sets, $d(c(Y_i), c(Y_j)) \leq \frac{1}{n-1} d(y_i, y_j) \leq \frac{1}{n-1} M$, and so $\text{diam } c^1(Y) \leq \frac{1}{n-1} M$. Thus the diameter of $c^k(Y)$ is bounded above by $\left(\frac{1}{n-1}\right)^k M$ and so goes to zero.

To show convergence, let $M_k$ denote the diameter of $c^k(Y)$, and consider the sequence $z_k \in c^k(Y)$, where $z_0 = y_1$, $z_1 = c(Y_1)$, $z_2 =$
c(\{c(Y_i) : i \neq 1\}), etc. It follows by convexity for \((n-1)\)-sets that 
\[ d(z_k, z_{k+1}) \leq \frac{1}{n-1} M_k. \]
Thus for \(l \geq k\),
\[ d(z_k, z_l) \leq M_k + \frac{1}{n-1} M_k + \left( \frac{1}{n-1} \right)^2 M_k + \cdots = \frac{n-1}{n-2} M_k, \]
showing that \(\{z_k\}\) is a Cauchy sequence. Thus, centroids exist for \(n\)-sets.

To show convexity of \(c(Y)\), \(|Y| = n\), suppose \(Y = \{y_1, \ldots, y_n\}\) and \(Y' = \{y'_1, \ldots, y'_n\}\). If \(Y_i = Y \setminus \{y_i\}\) and \(Y'_i = Y' \setminus \{y'_i\}\), then by convexity for \((n-1)\)-sets,
\[ d(c(Y_i), c(Y'_i)) \leq \frac{1}{n-1} \sum_{j \neq i} d(y_j, y'_j) \quad (1) \]
for each \(i\). Let \(\delta_i = d(y_i, y'_i)\) and \(d_0 = (\delta_1, \ldots, \delta_n)\). Then if \(d_k\) is the corresponding vector of distances between elements of \(c^k(Y)\) and \(c^k(Y')\), it follows from \(\text{[1]}\) that \(d_k \leq B^k_n d_0\), where \(B_n = \frac{1}{n-1}(J_n - I_n)\), \(J_n\) is the \(n \times n\) matrix of 1's and \(I_n\) is the \(n \times n\) identity matrix. Since \(B^k_n \rightarrow \frac{1}{n} J\) as \(k \rightarrow \infty\), it follows that \(d(c(Y), c(Y')) \leq \frac{1}{n} \sum d(y_i, y'_i)\) as desired. \(\text{back}\)


