Statistical Statements for Phylogenetic Trees

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

Funded in part by a grant from NSF-DMS  
Collaborators: Karen Vogtmann, Persi Diaconis and Lou Billera, Aaron Staple, Henry Towsner.
• The statistical paradigm.
  ★ Estimates and confidence.
  ★ Statistical approaches to variability.
  ★ Robustness
  ★ Inspirations from ranked data.

• Building a treespace with a natural distance.

• The Bootstrap.

• Special cases: multivariate statistics and Treespace
Estimate $\hat{\tau}$ computed from the data:
Data can be:

- **Binary**
  
  ```
  Lemur_cat 00000000000001010100000
  Tarsius_s 10000010000000010000000
  Saimiri_s 10000010000001010000000
  Macaca_sy 00000000000000010000000
  Macaca_fa 10000010000000001000000
  ```

- **Aligned**


  ```
  12  60
  Lemur_cat AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC CTTACATCAT CCATATTATT
  Tarsius_s AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC CTCACCTCCT CCCTATTATT
  Saimiri_s AAGCTTCACC GGCGCAATGA TCCTAATAAT CGCTCACGGG TTTACTTCGT CTATGCTATT
  Macaca_sy AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA CTCACCTCCT CCCTATTATT
  Macaca_fa AAGCTTCTCC GGCACAAGCA CCCTTATAAT CGGCCACGGG CTCACCTCCT CCATGTATTT
  Macaca_mu AAGCTTTGCT GGCACAACCA TCCTCATGAT TGCTCACGGA CTCACCTCCT CCATATATTT
  ```
Gene order

These data sets usually come with their own metrics.
The parameter is a semi-labeled binary Tree

0 root

inner edges

1 3 2 4

leaves

inner edges

Inner Node
Statistical Paradigms

Classical Frequentist

- estimate the parameter,
  (either in a parametric (ML) way,
  semiparametric (Distance based methods),
  or nonparametric way (Parsimony))
- find the sampling Distribution of the estimator.

Bayesian

- Specify a Prior Distribution
- Update the prior using the Data
- Compute the Posterior Distribution

Difficulties arise as the estimators lie in a non Euclidean space.
Sampling Distribution for Trees
Data

1
2
3
Data

Treespace $T_n$
True Sampling Distribution
Bootstrap Sampling Distribution (non parametric)
Data

Bootstrap Sampling Distribution
(parametric)
How do we define distributions on Treespace

- Not the uniform distribution.
- By inspiration from ranked Data: Mallow’s model (1957)
  \[ P(\tau_i) = Ke^{-\lambda d(\tau_i, \tau_0)} \]
  - Uses a central tree \( \tau_0 \).
  - Uses a distance \( d \) in treespace.
  - But very symmetrical, maybe need a mixture model.
- Other distributions (see Aldous, 2001), one might want to include information about the estimation method used as this influences the shape of the tree.
Classical Statistical Summaries

- Expectation (center of the distribution).
  Open question: What distribution is the consensus such a center of?

- Median (multivariate median (Tukey, 1972)).

- Variance (second moment $E_{P_n} d^2(\hat{\tau}, \tau)$).

- Presence/Absence of a clade.

- Summaries of Multivariate Variabilities. (PCA, MDS,..).
Confidence Statements for trees
Confidence Statements in Statistics

Depend on local and global properties of a neighborhood.

From Efron, Halloran, Holmes, (1996)
Confidence Statements in Statistics

Depend on local and global properties of a neighborhood.

What is the curvature of the boundary? How many neighbors does a region have?

From Efron, Halloran, Holmes, (1996)
Simple confidence values

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

- Univariate.
- Multiple Testing.
- Composite Statements.
Do we care about confidence statements for phylogenetic trees?

Cetacees: recognising what is being sold as Whale meat in Japan?


Whale www.DNA.surveillance Earth Trust Press Release
Frequentist Confidence Regions

\[ P(\tau \in \mathcal{R}_\alpha) = 1 - \alpha \]

We will use the nonparametric approach of Tukey who proposed peeling convex hulls to construct successive ‘deeper’ confidence regions. But we need a geometrical space to build these regions in.
Aims

- Fill Tree Space and make meaningful boundaries.
- Define distances between trees.
- Define neighborhoods, meaningful measures.
- Principal directions of variations in tree space, summarizing: structure + noise.
- Confidence statements, convex hulls.
Rotation Moves
Rotation Moves

0

1 2 3 4

0

1 2 3 4
The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology this is called an ‘unresolved’ tree.
Boundary for trees with 3 leaves
The quadrant for one tree
The quadrant for one tree

The quadrant for one tree
The quadrant for one tree
The quadrant for one tree
The quadrant for one tree
The pendant edges are shared by all trees on the same $n$ leaves, so in fact we decompose treespace into a product space $R^n \times T_n$. 
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.
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. $\mathcal{T}_n$ is built by taking one $n - 2$-dimensional quadrant for each of the $(2n - 3)!! = (2n - 3) \times (2n - 5) \times \cdots \times 5 \times 3 \times 1$ 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.
Three quadrants sharing a ray for $n=4$
Three quadrants sharing a ray for n=4

Note that the bottom boundary rays form a copy of $T_3$ embedded in $T_4$. 
Note that the bottom boundary rays form a copy of $\mathcal{T}_3$ embedded in $\mathcal{T}_4$. In general, $\mathcal{T}_n$ contains many embedded copies of $\mathcal{T}_k$ for $k < n$. 
Link of 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.
consequence for time to convergence of MCMC chains of random walks based on NNI moves.
CAT(0) space, (Gromov)
CAT(0) space, (Gromov)

Triangles are thin.
Consequences

• Averaging works better than it should, (an argument against total evidence computation without decomposing??).

• We can build Bayesian priors based on distances.

• We can make a useful bootstrap statement.

• We can make convex hulls. $\longrightarrow$ Confidence regions.

• We know how many neighbors any tree has.

• We can make a useful bootstrap statement.
How many neighbors for a given tree? (W.H. Li, 1993)

We know the number of neighbors of each tree.
How many neighbors for a given tree? (W.H.Li, 1993)

We know the number of neighbors of each tree.
How many neighbors for a given tree? (W.H. Li, 1993)

We know the number of neighbors of each tree.

For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree: 15 neighbors. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have two contiguous “small edges” and all the other inner edges significantly bigger than 0.
This picture of treespace frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius $r$ around a given tree. All we have to do is check the sets of contiguous edges in the tree smaller than $r$, say there is only one set of size $k$, then the neighborhood will contain

$$(2k - 3)!! = (2k - 3) \times (2k - 5) \times \cdots \times 3$$

‘different’ trees.

If there are $m$ sets of sizes $(n_1, n_2, \ldots, n_m)$
In this case the number of trees within $r$ will be $15 \times 105 \times 3 = 4725$, in general:

$$(2n_1 - 3)!! \times (2n_2 - 3)!! \times (2n_3 - 3)!! \cdots \times (2n_m - 3)!!$$

A tree near the star tree at the origin will have an exponential number of neighbors.

This explosion of the volume of a neighborhood at the origin provides for interesting math problems.
These differing number of neighbors for different trees show that the bootstrap values cannot be compared from one tree to another.

This was implicitly understood by Hendy and Penny in their NN Bootstrap procedure.

Are there other ways of using the bootstrap than just counting clade appearances?
Beware the different number of neighbors matters if you think you are using a Monte Carlo method to estimate the distance to the boundary using the bootstrap.
Inferential Bootstrap

\[ \mathcal{X} \text{ original data} \rightarrow \hat{T} \text{ estimate.} \]

Call \( \mathcal{X}^* \) bootstrap samples consistent with the model used for estimating the tree:

- Non parametric multinomial resampling for a parsimony tree.
- Seqgen parametric type resampling with the same parameters for a ML.
- Bayesian GAMMA prior on rates and generation (Yang 2000) for random sequences according to \( \hat{T} \).
Can we use the distance for the bootstrap:

Classical Bootstrap Theory (open problem: to provide such a proof here)

\[ \text{Distribution}(d(\hat{T}, T)) = \]
Can we use the distance for the bootstrap:

**Classical Bootstrap Theory** (open problem: to provide such a proof here)

\[
\text{Distribution} \left( d(\hat{T}, T) \right) = \text{Distribution} \left( d(\hat{T}^*, \hat{T}) \right)
\]
Can we use the distance for the bootstrap:

Classical Bootstrap Theory (open problem: to provide such a proof here)

\[ \text{Distribution}(d(\hat{T}, T)) = \text{Distribution} (d(\hat{T}^*, \hat{T})) \]

A better pivot than \( d(\hat{T}^*, \hat{T}) \)?
(classically this would be of the form:

\[
\frac{d(\hat{T}^*, \hat{T})}{\sqrt{\text{var}(d(\hat{T}^*, \hat{T}))}}
\]

What is a variance estimate for trees? \( \sum_{i=1}^{n} d(\hat{T}^*, \hat{T})^2 \) ?

\[ \rightarrow \text{all open problems} \]
Bootstrap Distribution of Distances to $\hat{\tau}$
distances to original tree
Frequency
0.010 0.015 0.020 0.025 0.030 0.035 0.040
0 20 40 60 80 100 120

distances to original tree
Frequency
0.010 0.015 0.020 0.025 0.030 0.035 0.040
0 20 40 60 80 100 120
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions
Convex Hulls as Nonparametric Confidence Regions

(Tukey’s peeling, can also provide a multivariate median if we go all the way down to the inner peel)

There are multivariate versions of this: David Scott’s non parametric density estimation shells.
Multivariate Visualizations

Using the distances between trees as input to multidimensional scaling, find a useful ‘view’ of the trees variability.

Problems: The ‘meaning’ of the coordinates.
Example: John Endler’s bower birds:
There are many different sources of trees:

- DNA tree.
- Plumage Trees.
- Bower making trees.

(Interactive plots normally generated by xgvis, here are a few snapshots)
Maximum Likelihood Bootstrap
Robustness Bootstrap - stability

- Do small perturbations of my data make for changes in the tree?
- How close are the data to being treelike?
- Projection problem (residuals)
- Open problem: we need a notion of differential in treespace to study the influence functions.
How can mathematical statistics help?

Perspectives

- 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)
  - Aggregating trees, expectations for various measures.
  - Neighborhoods (convex hulls of trees)....

- How much does non-independence matter?

- Justification of commonsense, ground for generalizations.
Comments

-A tree is rounding the data, if everything else is noise, it purifies the picture, but if there in fact a mixture of 2 trees in the data?

- Data $\rightarrow$ Tree + Residuals.
Sometimes the residual is all we care about, so losing it is a loss.

- If we have two estimated trees and their sampling distributions $\hat{\tau}_1$ with $n$ leaves $\hat{\tau}_2$ with $m$ leaves, with an intersection of overlapping leaves that is $r$.

We can include more leaves than actually present by embedding $\mathcal{T}_n$ and $\mathcal{T}_m$ in $\mathcal{T}_k$, $k > n, m$, all the trees that represents $\hat{\tau}_1$ form a sub-complex of $\mathcal{T}_n$ and we can do the same for $\hat{\tau}_2$, the extra information that we can include is the sampling distributions for both trees. We can try and find an intersection in the support of the two distributions, this is more flexible than just looking for a consensus.

-Spaces of nonpositive curvature are generalisations of hyperbolic spaces (ie trees).
References


Proof by direct decomposition

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

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

then

$$M_{n-1} = S_{2n}/B_{n-1}$$

and

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

This formula for the number of trees was first proved using generating functions by Schroder (1873)[?].

$(S_{2n-2}, B_{n-1})$ 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.

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

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.