Geometry, Statistics and Tree Space.

Susan Holmes

Toulouse,
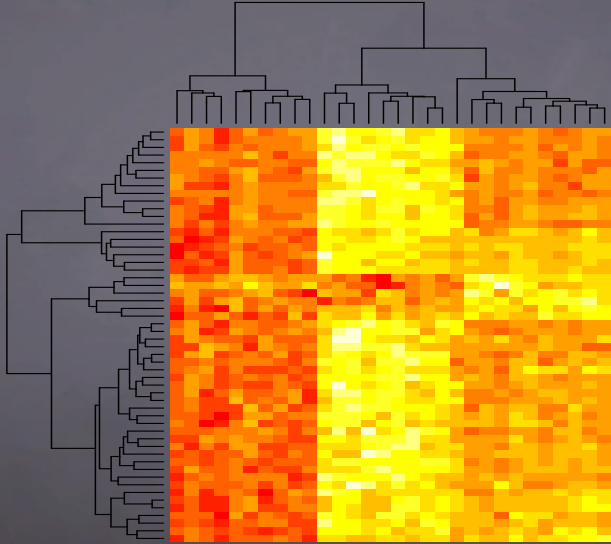
August 29, 2019

Motivation: Forests of Trees

- Different genes, same set of species.
- Bootstrapped Data by Multinomial Resampling, then estimating the tree.
- Bayesian Posterior Distributions on set of Trees.
- Simulated data according to certain evolutionary models (`seq-gen`).
- Data specimens in different conditions.
- Hierarchical Clustering Trees for (repeated) RNA-seq data (different time points, different space points, ...).

Hierarchical Clustering Trees

Some Methods for Generating Trees
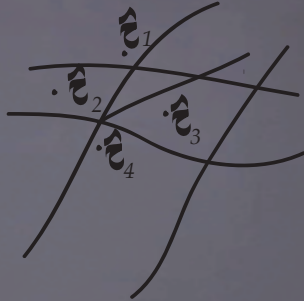
With advances in computational power we can use simulated data to evaluate clustering stability, either in a frequentist (Bootstrap) setting or by using a Bayesian paradigm where trees from a posterior distribution can be generated by MCMC (Monte Carlo Markov chain) methods.

We provide here a brief overview of the standard methods for generating distributions of trees. Different approaches to the problem of combining the trees are summarized. This combination of information on different trees is a non-standard statistical problem because trees do not lie in a Euclidean space ([1]).

Sampling Distribution for Trees

Data

t_1

t_2

t_3

Treespace T_n

Data

$\mathcal{F}$

$\cdot c_1$

$\cdot c_2$

$\cdot c_3$

$\cdot c_4$

True Sampling Distribution

$\hat{\theta}$

$\hat{\theta}_n$

Data

$\hat{\xi}^*_1$

$\hat{\xi}^*_2$

$\hat{\xi}^*_3$

$\hat{\xi}^*_4$

Bootstrap Sampling Distribution
(non parametric)

Bootstrap support for Phylogenies Taking as observations the *columns* of the matrix X of aligned sequences, the rows representing the species.

The sampling distribution of the estimated tree is estimated by resampling with replacement among the characters or columns of the data.

This provides a large set of plausible alternative data sets, each be used in the same way as the original data to give a separate tree (see [13] for a review).

Parametric Bootstrapping for Microarray Clusters

Bayesian posterior distributions for phylogenetic trees   - Prior distributions on the DNA mutation rates that occur during the evolutionary process and a uniform distribution on the original tree.

- Use of MCMC to generate instances of the posterior distribution.
- Implementations `MrBayes` [15] and `Beast` provide a sample of trees from the posterior distribution.
- The posterior distribution provides an estimate of variability.

Bayesian methods in hierarchical clustering Heller[23] provide a Bayesian nonparametric method for generating posterior distributions of hierarchical clustering trees.

Euclidean space (where through every point not on a line) is flat:
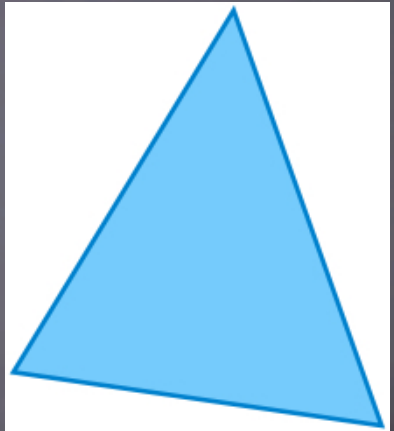
(sum of angles of a triangle is 180 ^{o}),

Hyperbolic space is ' negatively' curved:
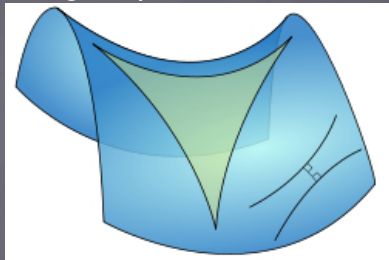
Euclid's parallel postulate is replaced.
In hyperbolic geometry there are at least two distinct lines through P which do not intersect l, so the parallel postulate is false.
A characteristic property of hyperbolic geometry is that the angles of a triangle add to less than 180 ^{o}.

Geodesic metric space:
If we have a distance defined between any two points of a space, we call it a metric space.
(The distance doesn't have to be defined through ordinary coordinates)
A geodesic metric space is a metric space where geodesics are defined to be the shortest path between points in the space.

δ-hyperbolic space is a geodesic metric space in which every geodesic triangle is δ-thin.

δ-thin: pick three points and draw geodesic lines between them to make a geodesic triangle. Then any point on any of the edges of the triangle is within a distance of δ from one of the other two sides.

For example, trees are 0-hyperbolic: a geodesic triangle in a tree is just a subtree, so any point on a geodesic triangle is actually on two edges.

Δacd

Normal Euclidean space is ∞-hyperbolic; i.e. not hyperbolic. Generally, the higher δ has to be, the less curved the space is.

Comparing Different Trees
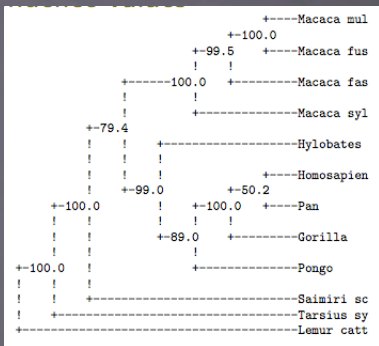
- Binomial Support Estimates (Consensus+support values).
- Split Differences, Visualization Programs .
- Distances.
- Recoding of Trees as binary columns.

Confidence Statements for trees
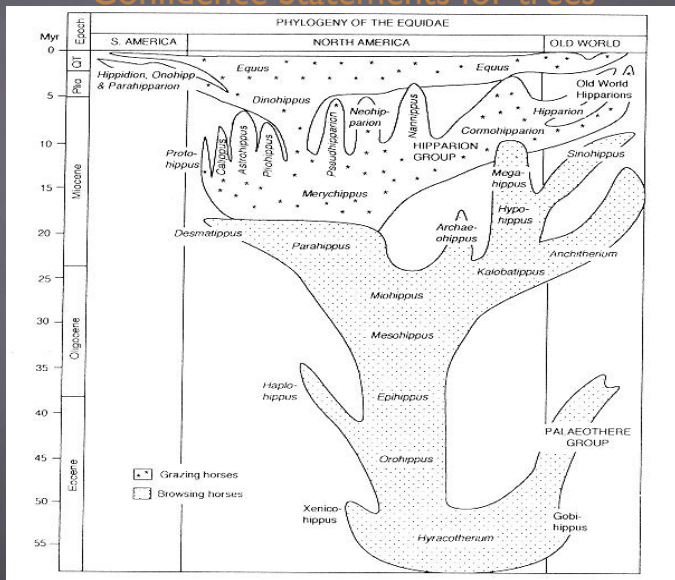
Confidence Statements in Statistics

Depend on local and global properties of a neighborhood.
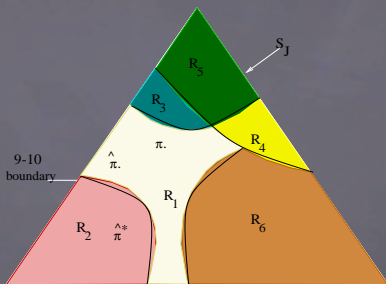
From Efron, Halloran, Holmes, (1996)

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

- Univariate.

- Multiple Testing.

- Composite Statements.

```
                              +----Macaca mul
                        +-100.0
               +-99.5   +----Macaca fus
               !   !
       +------100.0   +--------Macaca fas
       !   !
       !   !      +-------------Macaca syl
   +-79.4
       !   !   +------------------Hylobates
       !   !   !
       !   !   !         +----Homosapien
       !   +-99.0      +-50.2
   +-100.0   !   +-100.0  +----Pan
   !   !     !   !   !
   !   !     +-89.0   +--------Gorilla
   !   !           !
+-100.0  !         +-------------Pongo
!   !   !
!   !   +--------------------------Saimiri sc
!   +------------------------------Tarsius sy
+------------------------------------Lemur catt
```

Do we care about confidence statements for phylogenetic trees?

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

Steve

Phylogenetic Identification
of Whale and Dolphin Products

	KOREA	JAPAN	
	1994–97	1993–98*	1998-99
Pygmy Right			
Bowhead			
Right			
N. Minke	54	31	29
S.Minke	2	134	57
Bryde's	2	4	
Pygmy Bryde's	2		
Sei		1	4
Humpback		2	
Fin		21	1
Blue		2	
Gray			
Pygmy Sperm		1	
Sperm			1
Other Beaked whales		1	
Baird's		19	8
Cuvier's	1	3	
Porpoise		3	2
Killer whale	1	1	

Outgroups

Beaked
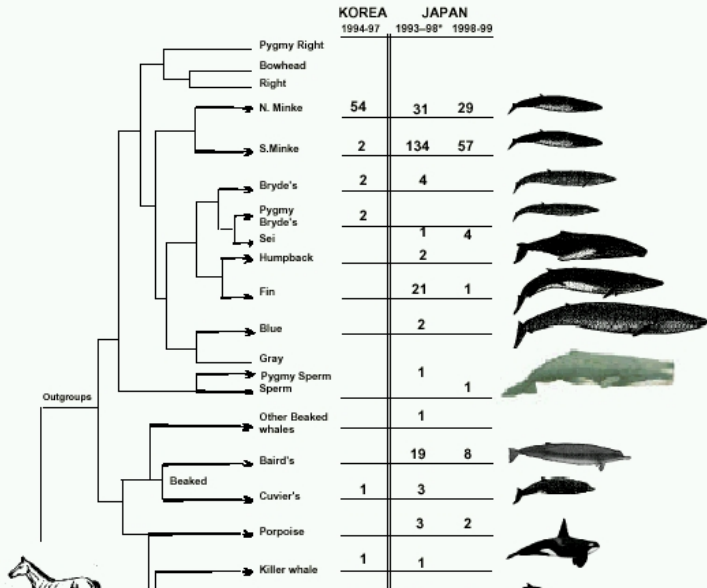
The River without a Paddle?

Human immunodeficiency virus: Phylogeny and the origin of HIV-1
The origin of human immunodeficiency virus type 1 (HIV-1) is
controversial.
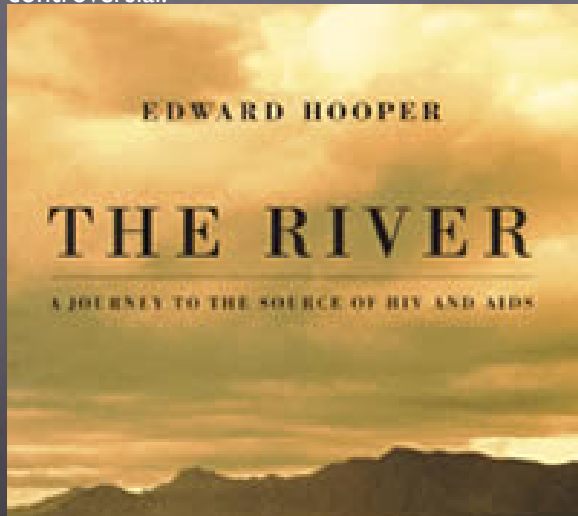
Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign, supporting a model of 'natural transfer' from chimpanzees to humans. If this timescale is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes of group M differentiated in chimpanzees before their transmission to humans.

Confidence Intervals ?

Korber and colleagues extrapolated the timing of the origin of HIV-1 group M back to a single viral ancestor in 1931, give or take about 12 years for 95% confidence limits.

Because this calendar of events obviously pre-dated the OPV trials, in the revised version of his book, Hooper suggested that group M first began to diverge in chimpanzees, and that there were then several independent transfers of virus to humans via OPV.

In that case, several OPV batches should bear evidence of their production in chimpanzee tissue, yet no such evidence has been found.

H

J

C

A

US B Clade
1954
(1941-1963)

M group
1930
(1911-1945)

B

B / D
1939
(1920-1940)

D
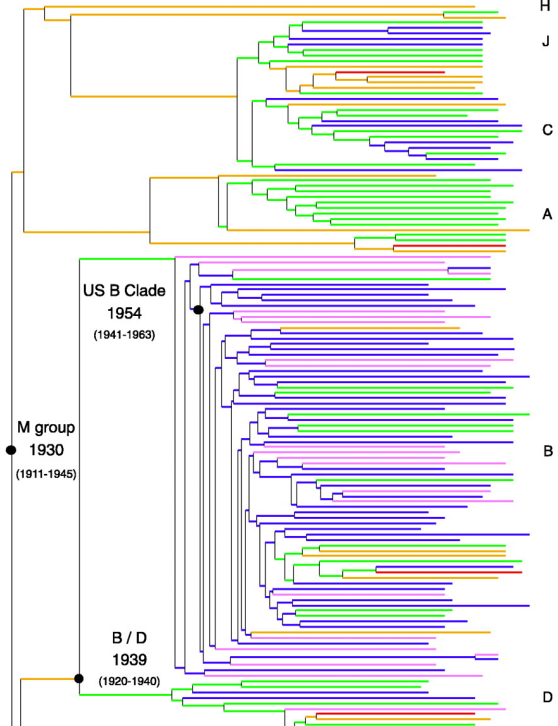
The OPV batch that Hooper considered to be under most suspicion, however, was CHAT 10A-11.
An original vial of the batch was found at Britain's National Institute for Biological Standards and Control, and the new tests show that it was prepared from rhesus-macaque cells.

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.

What does a neighborhood look like?

Need modern topology.
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.

Distances between Trees

- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI). Rotation Moves
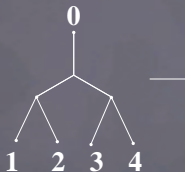
Distances between Trees

- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI). Rotation Moves

Distances between Trees

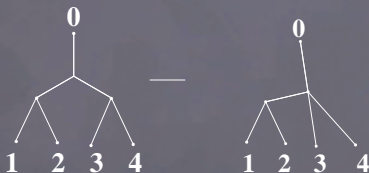- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI). Rotation Moves

Distances between Trees

- Robinson and Foulds, (bipartitions).
- Nearest Neighbor Interchange (NNI). Rotation Moves

- Subtree Prune Rebranch. (SPR)
- Fill-in of NNI moves: Billera, Holmes, Vogtmann (BHV). 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
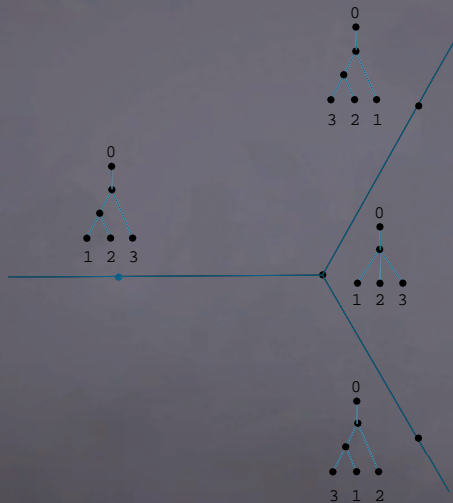
The quadrant for one tree
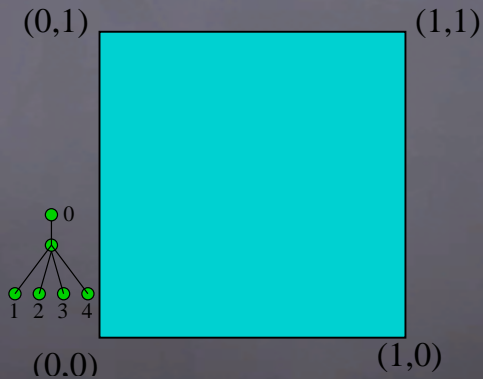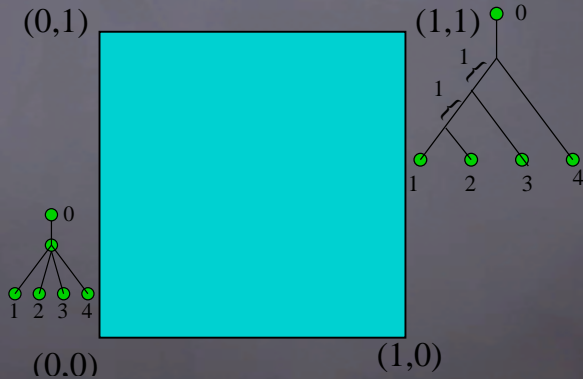
The quadrant for one tree
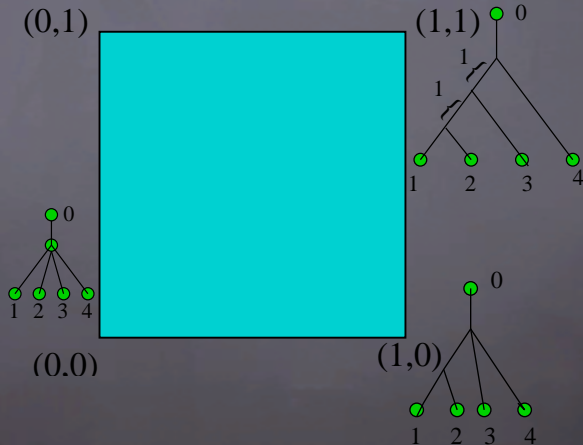
The quadrant for one tree
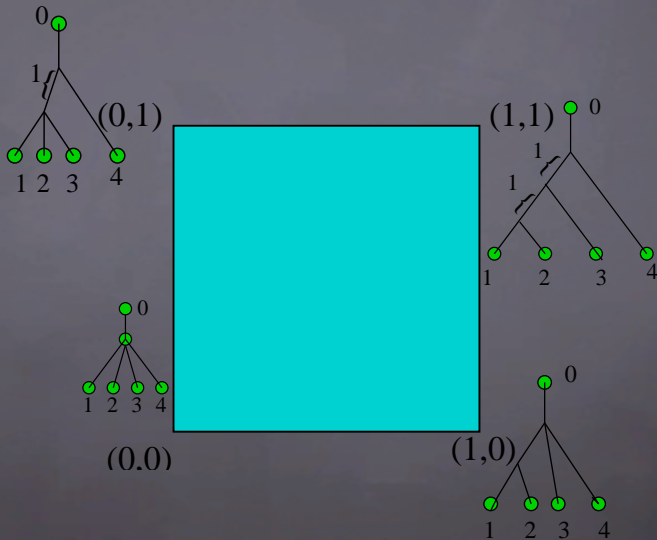
Link of the origin

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*.
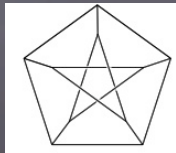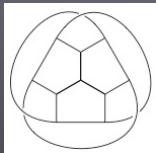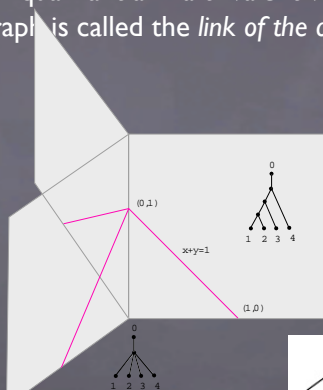
Cube complex of Euclidean Orthants
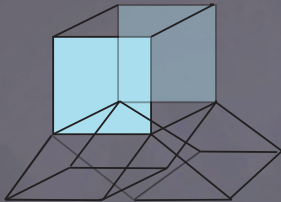
A path between two trees consists of line segments through a sequence of orthants. This sequence of orthants is the *path*.

A path is a *geodesic* when it has the smallest length of all paths between two points.

Cube complex of Euclidean Orthants
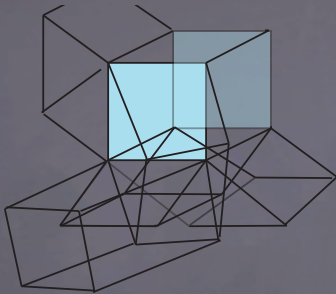
A path between two trees consists of line segments through a sequence of orthants. This sequence of orthants is the *path*.

A path is a *geodesic* when it has the smallest length of all paths between two points.

A Cone Path
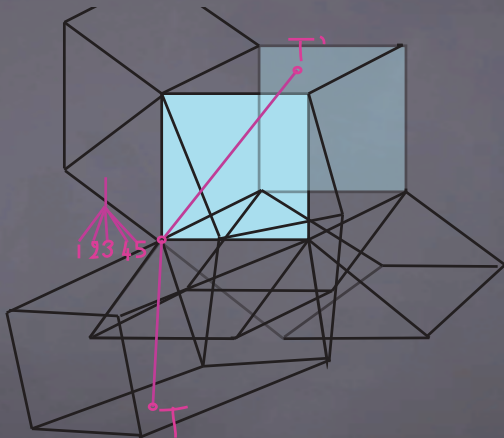
A path between two trees T and T' always exists. Since all orthants connect at the origin, any two trees T and T' can be connected by a two-segment path, this is called the cone-path.

Three orthants sharing a common boundary for $n = 4$ leaves.
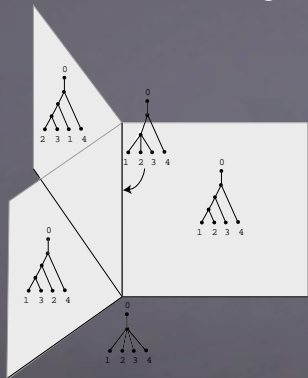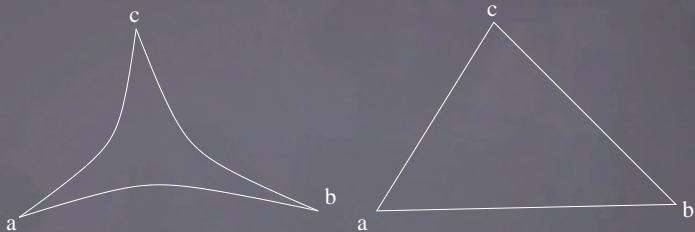
Theorem(Billera, Holmes, Vogtmann (BHV)): Tree space with BHV metric is a CAT(0) space, that is, it has non-positive curvature. This implies there are geodesic between any two trees (Gromov). It is not an Euclidean space.
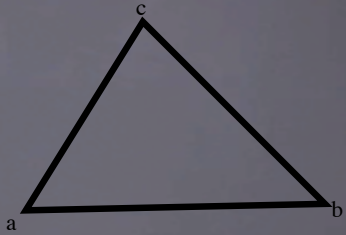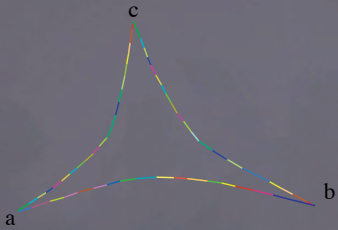
This has an effect on the existence of geodesics.
The speed at which MCMC methods work.
The size of the "variance".
The computation of the mean of a set of trees.
The number of neighbors of a tree.

We know that given a distance matrix we can give a treelike representation of the points with these distances by building a tree if the distances obey Buneman's four point condition (Buneman, 1974).

Buneman's four point condition

For any four points (u, v, w, x) :
The three sums: $d(u, v) + d(w, x), d(u, w) + d(v, x), d(u, x) + d(v, w)$ are equal, not less than the third.

We can see Gromov's definition the hyperbolicity contant δ as a relaxation of the above four-point condition:

Gromov's hyperbolicity contant

For any four points u,v,w,x, the two larger of the three sums $d(u, v) + d(w, x), d(u, w) + d(v, x), d(u, x) + d(v, w)$ differ by at most 2δ.

Can we embed trees in Euclidean space (approximately)

We can ask whether points are closer to a tree or to being embeddable in Euclidean space by using Gromov's δ.
Implementation:
`distory` is an R package written with John Chakerian[3] which both implements the geodesic BHV distance between trees using Owen and Provan (2009)'s algorithm and the computation of delta for any finite set of points.
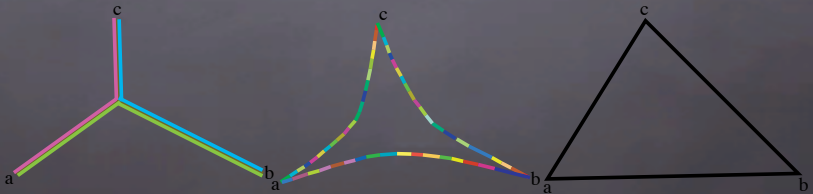
Multidimensional Scaling (MDS or PCoA)

Schoenberg's (1935) remarked that a symmetric matrix of positive entries with zeros on the diagonal is a Euclidean distance matrix between n points if and only if the matrix

$$-\frac{1}{2} H \Delta_2 H \text{ is semi-definite positive}$$

where $H = (I - \frac{1}{n} \mathbf{1} \mathbf{1}')$, and $\mathbf{1}' = (1, 1, 1 \ldots, 1)$

Approximating Non Euclidean Distances by Euclidean ones

Forward:Decomposition of Distances Suppose we did have an Euclidena space, variables measured in $\mathbb{R}^p$ that are not centered: Y, apply the centering matrix

$$X = HY, \qquad \text{with } H = (I - \frac{1}{n} \mathbf{1}\mathbf{1}'), \text{ and } \mathbf{1}' = (1, 1, 1 \ldots, 1)$$

Call $B = XX'$, if $D^{(2)}$ is the matrix of squared distances between rows of X in the euclidean coordinates,

$$d_{i,j} = \sqrt{(x_i^1 - x_j^1)^2 + \cdots + (x_i^p - x_j^p)^2}. \text{ and } -\frac{1}{2}HD^{(2)}H = B$$

Backward from D to X We can go backwards from a matrix D to X by taking the eigendecomposition of B in much the same way that PCA provides the best rank r approximation for data by taking the singular value decomposition of X, or the eigendecomposition of XX'.

$$X^{(r)} = US^{(r)}V' \text{ with } S^{(r)} = \begin{pmatrix} s_1 & 0 & 0 & 0 & ... \\ 0 & s_2 & 0 & 0 & ... \\ 0 & 0 & ... & ... & ... \\ 0 & 0 & ... & s_r & ... \\ ... & ... & ... & 0 & 0 \end{pmatrix}$$

This provides the best approximate representation in an Euclidean space of dimension r. The algorithm provides points in a Euclidean space that have approximately the same distances as those provided by D^2.

MDS Algorithm

In summary, given an $n \times n$ matrix of interpoint distances, one can solve for points achieving these distances by:

1. Double centering the interpoint distance squared matrix: $S = -\frac{1}{2} H D_2 H$.

2. Diagonalizing S: $S = U \Lambda U^T$.

3. Extracting $\tilde{X}$: $\tilde{X} = U \Lambda^{1/2}$.

Is it better to represent the distances by a tree or a Euclidean projection?

SPATIAL VERSUS TREE REPRESENTATIONS OF PROXIMITY DATA

SANDRA PRUZANSKY

BELL LABORATORIES

AMOS TVERSKY

STANFORD UNIVERSITY

J. DOUGLAS CARROLL

BELL LABORATORIES

In this paper we investigated two of the most common representations of proximities, two-dimensional euclidean planes and additive trees. Our purpose was to develop guidelines for comparing these representations, and to discover properties that could help diagnose which representation is more appropriate for a given set of data. In a simulation study, artificial data generated either by a plane or by a tree were scaled using procedures for fitting either a plane (KYST) or a tree (ADDTREE). As expected, the appropriate model fit the data better than the inappropriate model for all noise levels. Furthermore, the two models were roughly comparable: for all noise levels, KYST accounted for plane data about as well as ADDTREE accounted for tree data. Two properties of the data proved useful in distinguishing between the models: the skewness of the distribution of distances, and the proportion of elongated triangles, which measures departures from the ultrametric inequality, Applications of KYST and ADDTREE to some twenty sets of real data, collected by other investigators, showed that most of these data could be classified clearly as favoring either a tree or a two-dimensional representation.

Key words: multidimensional scaling, clustering, tree structures, additive trees.

Malaria Data as seen using ape
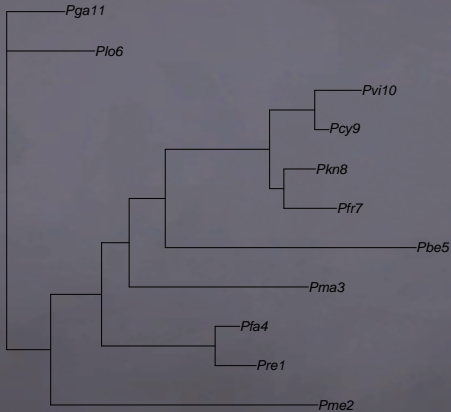
Eigenvalues of MDS for bootstrapped trees
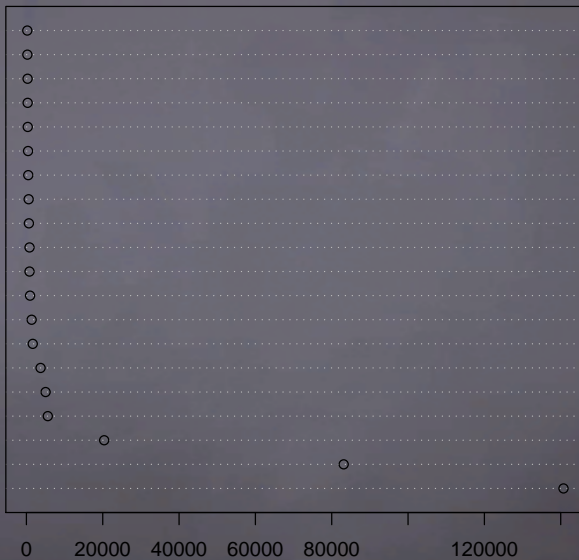
Bootstrapped trees
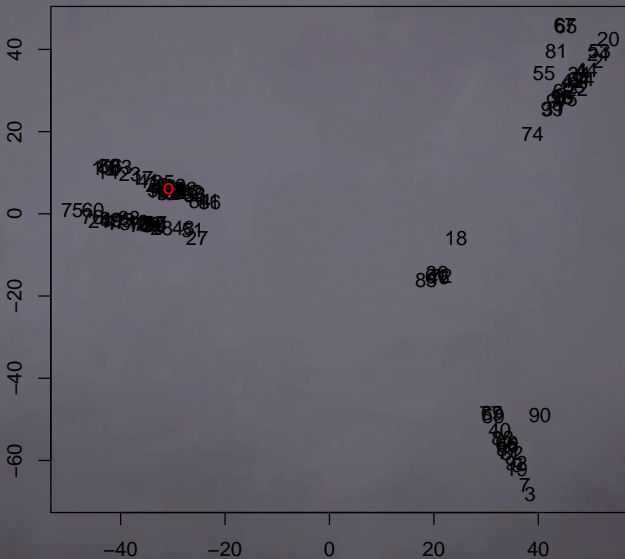
Probability Distributions on Tree Space

In Holmes (2005) I discuss the use of distances for making believable probability distributions on the space of trees, the simplest such model is

$$P(\tau_i) = Ke^{-\lambda d(\tau_i, \tau_0)}$$

This is really a Mallows[17] model for trees, and as such has possible extensions in similar ways than [10], [11] or those used for rankings developed in [4].
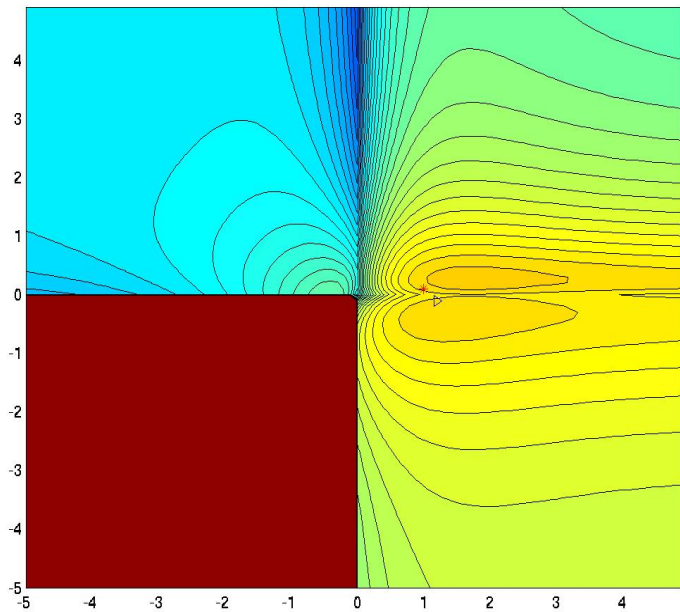
Empirical Evidence on Mixing on Bethe Lattice

Mossel noticed that one of the extreme points of tree space with regards to predicting the root was the Bethe Lattice:
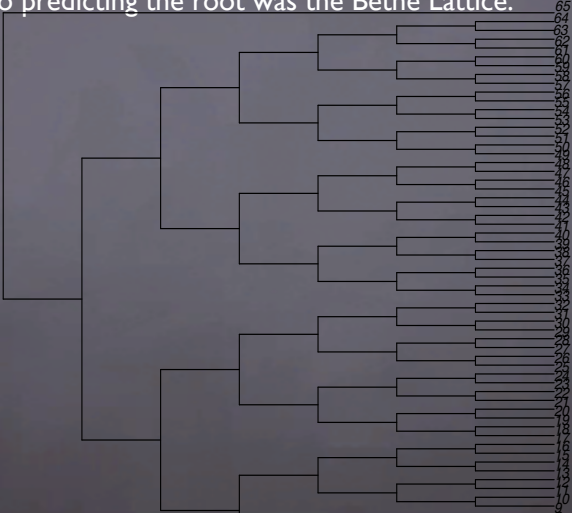
Can we hear the root?

For large enough independent sequences, say for k we can reconstruct the tree with probability $1 - \delta$

$$k > \frac{c \log n}{(1 - \theta_{m}ax)^2 \theta_{min}^d(T)}$$

However for large mutation rates, Mossel also proved the impossibility of estimating a tree if we only have short sequences and high mutation rates.

Distribution of Trees from seqgen Bethe Tree Data

$\alpha = 0.05, \ell = 1000$ MDS plot,
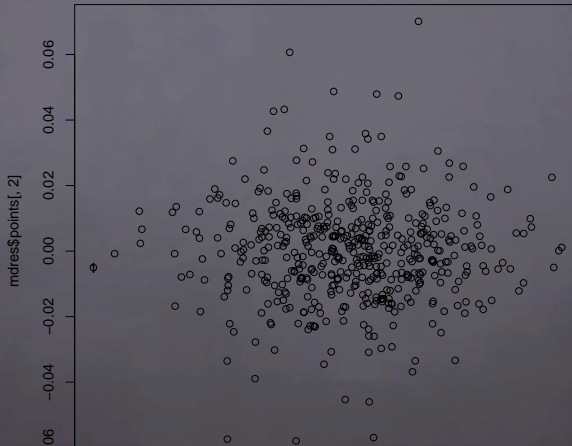
Distribution of Trees from seqgen Bethe Tree Data

$\alpha = 0.01, \ell = 1000$ MDS plot,

Seeing the Mutation Rate Gradient

We generated 9 sets of trees with mutation rates set from $\alpha = 0.01$ to $\alpha = 0.09$ and we generated the data according to the Bethe lattice tree.

Here are the results in the first plane of the MDS:
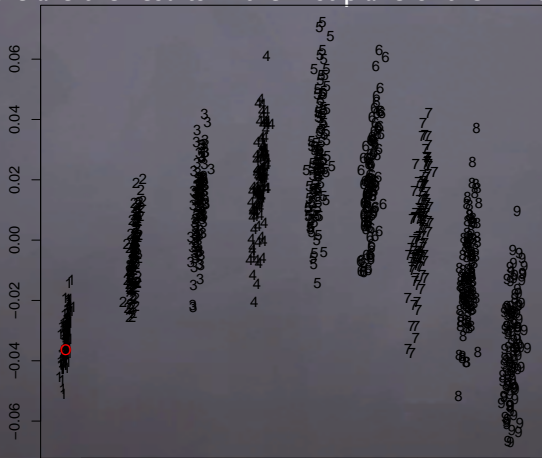
Tree of Trees

A tree is a complete CAT(0) space.
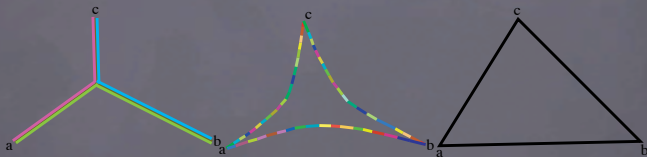
Since BHV,2001 [1] have shown that the space of trees is negatively curved (a **CAT(0)** space), the most natural representation of a collection of trees may be a tree.
Is this good for anything?

Mixture Detection

Mixtures pose problems when using MCMC methods in the Bayesian estimation context (Mossel, Vigoda 2005[20]). These authors note that MCMC methods in particular those used to compute Bayesian posterior distributions on trees can be misleading when the data are generated from a mixture of trees, because in the case of a 'well-balanced' mixture the algorithms are not guaranteed to converge.

They recommend separating the sequences according to coherent evolutionary processes.

Suppose the data come from the mixture of several different trees, we will see how the bootstrap and the various distances and representations can detect these situations.

Our procedure uses the bootstrap.

We use the distance between trees and then make a hierarchical clustering tree using single linkage (Similar to UPGMA) to provide a picture of the relationships between the trees.
In this simulated example we generate two sets of data of length $1,000$ from the two different trees represented:

Trees used to generate sequences of length 1000 each which are combined into one 2000 long aligned set ($\mathcal{X}_{12}$) and then bootstrapped.

A simulation experiment: we concatenate the data into one data set on which the standard phylogenetic estimation procedures are run. This provides the estimated tree for the data. We also generate 250 bootstrap resamples from the combined data. We then compute the distances between the 250 trees from each of the bootstrap resamples and make a hierarchical clustering single linkage from this distance matrix.
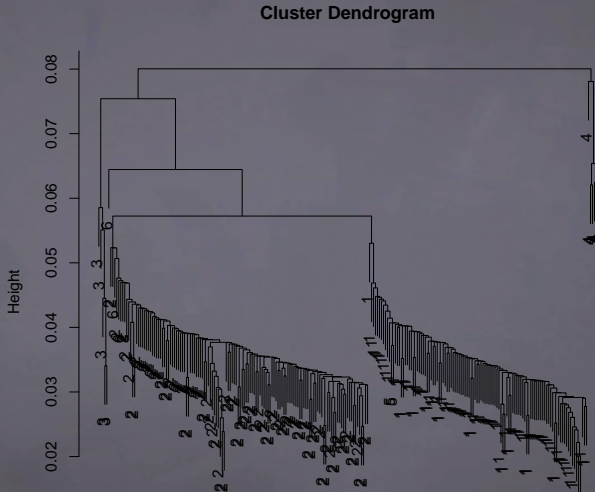
Figure: Hierarchical clustering of 250 trees resulting from a nonparametric bootstrap of the data generated by the double data set $\mathcal{X}_{12}$

Data	Distrib.	Dist	Max (sd)	Mean(sd)	δ
500	Unif	Manhat	13.8 (0.33)	8.33 (0.04)	7.
500	Unif	Euclid	3.04 (0.06)	2.03 (0.009)	1.
512	MVN	Manhat	49.14 (1.59)	28.22 (0.20)	21
512	MVN	Euclid	11.66 (0.41)	7.00 (0.05)	4.
512	Bethe	JC69	0.223 (0.008)	0.16 (0.003)	0.
512	Bethe	Raw	0.19 (0.006)	0.14 (0.002)	0.

Table: Different values of δ and the ratio $\delta/\max(d)$
for points generated both in bounded Euclidean
space and for points generated from trees. Each
value was estimated from 100 simulations, in the
Euclidean case the distances were computed from
points generated in 25 dimensions.

In particular, we used the δ/max statistic in the case of the bootstrapped trees represented by the MDS plot in the resulting ratio was 0.47, thus indicating given the calibration experiments in the above table that point configuration would be well approximated by a Euclidean MDS. The δ/max statistic is a rough approximation for scaling each triangle considered by its diameter; two other approximations, scaling by the perimeter and scaling by the max of the sums $A_{(1)}$ are implemented in the R package.

Statistical Uses for Distances

- Center of Cloud of Trees (equal weights): Find T_0 that minimizes either $\sum_{k=1}^{K} d^2(T_0, T_k)$ this is the (L^2) definition of the mean tree, or $\sum_{k=1}^{K} d(T_0, T_k)$ (L^1).
- Extend the above to cater for a measure on treespace.

$$P(T) = K exp(-\lambda d(T, T_0))$$

- Variability of the tree-points:
  Pseudovariance$= \frac{1}{K-1} \sum_{k=1}^{K} d^2(T_0, T_k) = \hat{s}^2$.
- Studentizing :

$$\frac{d(\hat{T}^*, \hat{T}_{obs})}{\hat{s}}$$

- Leverage of a position, as in leverage of an observation in regression.
- PCA with regards to Instrumental Variables- DPCOA. Explain a set of distances between trees by other distances between the same data.

Path between different tree topologies
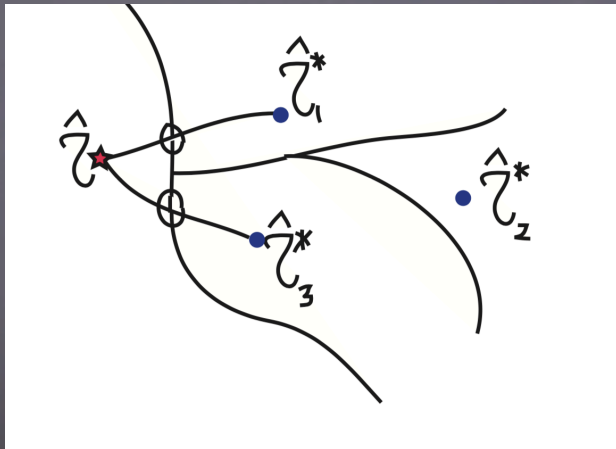
Finding the 'guilty characters'

Thinking like a Statistician....

Thinking like a Statistician....

and a geometer..

- How treelike are the data ?	Model Selection.
- Do we always need the tree,	Distances between Data.
- Are all the characters supporting the tree?	Leverage.
- Finding hidden gradients	Ordination of trees.
- Stability under perturbation	Evaluating the estimates.
- How variable are the trees?	Variance and Moments.

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.
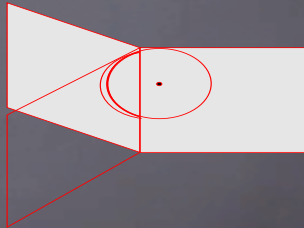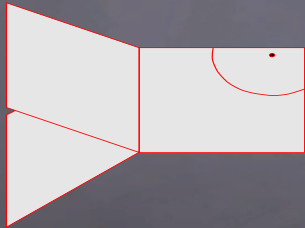
For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree: 15 neighbors. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have two contiguous "small 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

$$(2k - 3)!! = (2k - 3) \times (2k - 5) \times \cdots 3 \text{'different' trees.}$$

If there are m sets of sizes $(n_1, n_2, \ldots, n_m)$

1 2   3   4 5 6   7 8 9 10 11

1 2   3   4 5 6   7 8 9 10 11
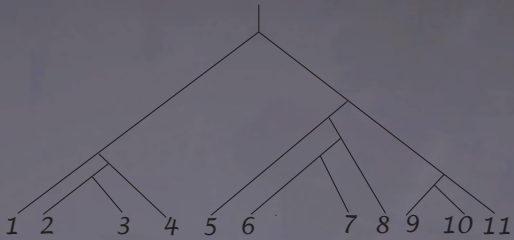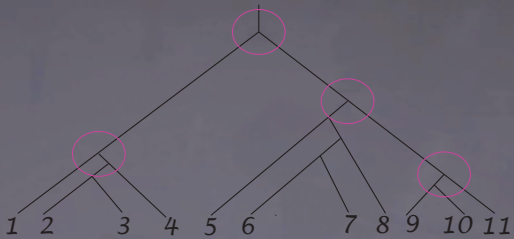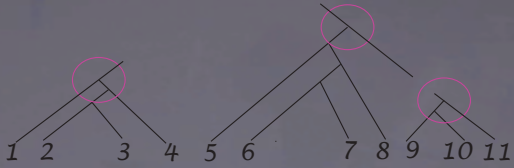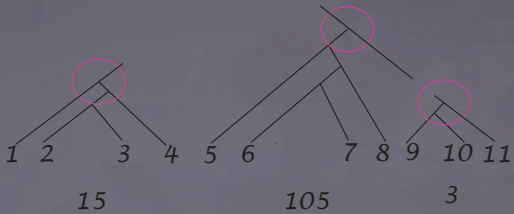
15   105   3

In this case the number of trees within r will be $15 * 105 * 3 = 4725$, 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?
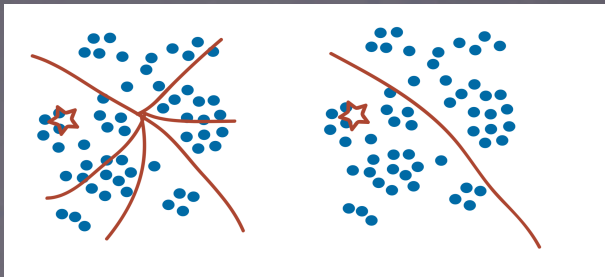
Beware the different number of neighbors matters if you think you are using a Monte Carlo method to estimate the distance to the boundary using the bootstrap.

Inferential Bootstrap
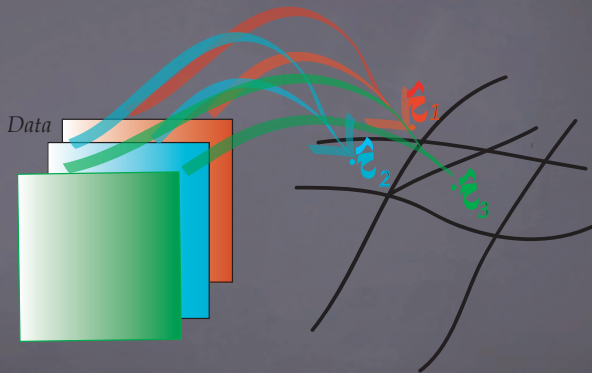
$\mathcal{X}$ original data $\longrightarrow \hat{\mathcal{T}}$ estimate.

How?
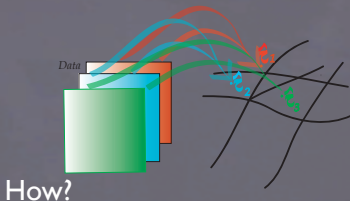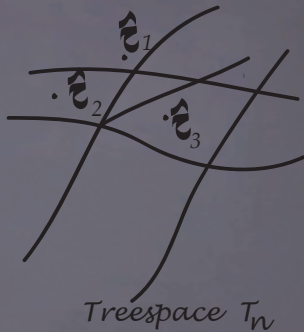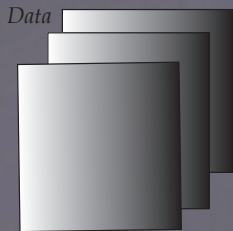
Call $\mathcal{X}^*$ bootstrap samples consistent with the model used for estimating the tree:

- 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 $\hat{\mathcal{T}}$

Sampling Distribution for Trees

Data

ξ_1

ξ_2

ξ_3

Treespace T_n

Data

$\mathcal{F}$

c_1
c_2
c_3
c_4

True Sampling Distribution

Bootstrap Sampling Distribution
(non parametric)

New resample D^* drawn by resampling rows (genes) from the original $D_{\text{nspecies} \times \text{nchar}}$ matrix.

- Are the characters (columns) independent?
  We actually have less information than we think?
  What is the unit of information?

- Block Bootstrap to generate dependent data.

Summarizing the bootstrap sampling distribution:
Why isn't enought to just count the branches in common?
Loss of all the multivariate information.

Tree Stability ?

Resample genes and compare the bootstrap tree to the original tree using a distance between trees (Billera, Holmes, Vogtman, 2001 for the distances and Holmes, Vogtmann, Staple, 2004 for the algorithm). Implemented in `ape`.

The bootstrap works (?)

**Conjecture:**
The bootstrap estimate of the sampling distribution of the distances $d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})$ is a good approximation to the true sampling distribution of $d(\hat{\mathcal{T}}, \mathcal{T})$.

Hypothesis Testing

As an additional element we have projected the star tree "S" (chosen with the lengths of the pendant edges closest to the original tree) to see whether it is in a small neighborhood, or credibility region of the bootstrapped trees.

This is analogous to seeing if 0 is in a confidence interval of differences between two random variables. If the star tree seems to be in central to a confidence region with a high probability coverage then we conclude that the data are not really treelike. In the figure , S appears to be on the outer convex hull of the projected points; we can conclude that the probability that the star tree belongs to the confidence region is low. To our knowledge, this is the first concrete implementation of the idea of using convex hulls to make confidence statements of this type [14] .

As an aside, note that the numbers in the Figure label the different types of branching patterns. We see that trees of the same topology are not necessarily closer to the original tree if we use the **BHV** with no modifications. In some cases we may want to give an extra weight to crossing orthants (ie changing branching pattern). We give examples of such modifications of the distance in the [**?**] vignette.
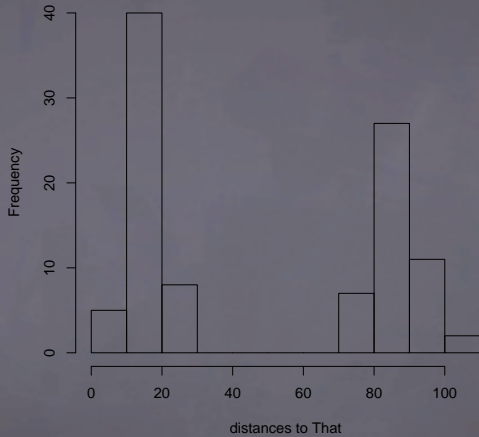
Eigenvalues of MDS for bootstrapped trees
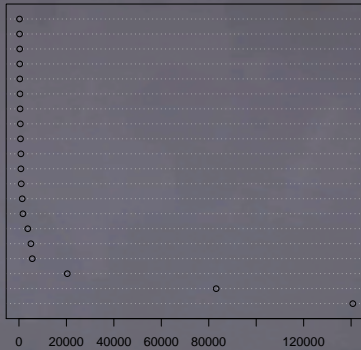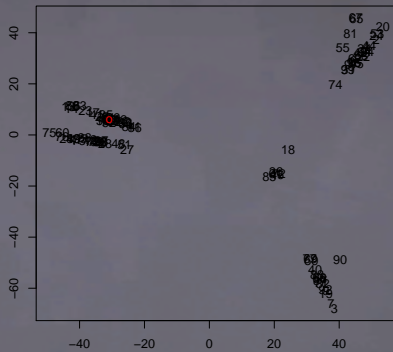
Bootstrapped trees

Who Cares?

Bacterial Species in the Gut: Example of a Metagenome.
Samples from IBS and healthy rats give abundance of about 1,000
species of bacteria.

Who Cares?

Bacterial Species in the Gut: Example of a Metagenome.
Samples from IBS and healthy rats give abundance of about 1,000
species of bacteria. To be continued...

References

[1] L. Billera, S. Holmes, and K. Vogtmann. The geometry of tree space. *Adv. Appl. Maths*, 771–801, 2001.

[2] J. Chakerian and S. Holmes. Computational methods for evaluating phylogenetic trees, 2010. arXiv.

[3] J. Chakerian and S. Holmes. distory:Distances between trees, 2010.

[4] Douglas E. Critchlow. *Metric methods for analyzing partially ranked data*. Springer-Verlag, Berlin, 1985.

[5] P. Diaconis, S. Goel, and S. Holmes. Horseshoes in multidimensional scaling and kernel methods. *Annals of Applied Statistics*, 2007.

[6] P. W. Diaconis and S. P. Holmes. Matchings and phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 95(25):14600–14602 (electronic), 1998.

[7] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[8] B. Efron, E. Halloran, and Susan P. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93:13429–34, 1996.

[9] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Boston, 2004.

[10] M. A Fligner and J. S Verducci. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369, 1986.

[11] Michael A Fligner and Joseph S Verducci. Multistage ranking models. *Journal of the American Statistical Association*, 83(403):892–901, 1988.

[12] M. Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, New York, 1987.

[13] S. Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2):241–255, 2003. Silver anniversary of the bootstrap.

[14] S. Holmes. Statistical approach to tests involving phylogenies. In *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford,UK, 2005.

[15] J. Huelsenbeck and F. Ronquist. MrBayes:Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–755, 2001.

[16] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[17] C. L. Mallows. Non-null ranking models. I. *Biometrika*, 44:114–130, 1957.

[18] K. Mardia, J. Kent, and J. Bibby. *Multiariate Analysis*. Academic Press, NY., 1979.

[19] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404 (electronic), 2004.

[20] E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–9, Sep 2005.

[21] M. Owen and J.S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 2–13, 2010.

[22] E. Paradis. Ape (analysis of phylogenetics and evolution) v1.8-2, 2006. http://cran.r-project.org/doc/packages/ape.pdf.

[23] R Savage, K Heller, Y Xu, and Z. Ghahramani. R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC*, Jan 2009.

[24] I.J. Schoenberg. Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'esp ace de Hilbert. *The Annals of Mathematics*, 36(3):724–732, July 1935.

[25] F. H. Sheldon and A. H. Bledsoe. Avian molecular systematics. *Annu. Rev. Ecol. Syst.*, 24:243–278, 1993.