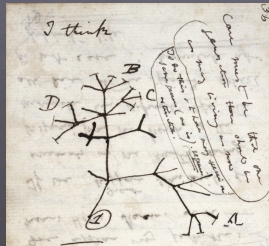


Molecular Evolution: Phylogenetic Tree Building

Lecture I: Geometry and Statistics, Susan Holmes

Bio-X and Statistics, Stanford University

August 29, 2019



Background foundations of Phylogeny

1. Statistics versus mathematics.
2. What is a Tree?
3. Gene Tree.
4. Model for Molecular Evolution.
5. Mutation Rates and Edge Lengths.
6. Examples of estimation methods for trees: parsimony.
7. ML estimation.
8. Parametric Bootstrap for ML.
9. Bayesian Approach.
10. Distance based tree building.
11. Hierarchical Clustering Trees.

Mathematical Logic

$$(A \rightarrow B) \iff (\neg B \rightarrow \neg A)$$

Observation: Non B = $\neg B$.

Conclusion: Observing $\neg B$, allows us to say: A is not true.

Statistical Logic

$$(A \rightarrow B) \iff (\neg B \rightarrow \neg A)$$

Observation: X

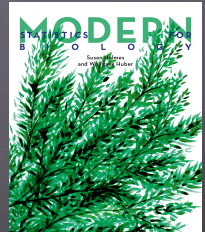
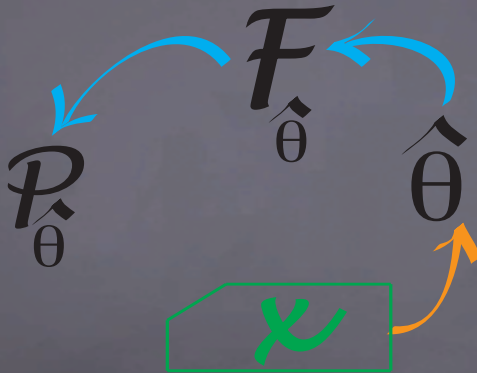
If the observed X makes $P(B)$ very small, then we infer A is unlikely.

Statistical Logic: induction

$$(H_0 \rightarrow E) \iff (\neg E \rightarrow \neg H_0)$$

If the observed X makes $P(E)$ is very small, then we infer H_0 is unlikely.

Statistics: separate the model from the data



See a complete book:

<http://bios221.stanford.edu/book/>

Phylogenetic Trees

11 1620

```
Pre1      GTACTTGTGA GGCCTTATAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pme2      GTATCTGTGA AGCCTTATAA AAAGATAGT- T-TAAATTAA AGGAATTATA
Pma3      GTATTTGTGA AGCCTTATAA GAGAAAAGTA TATTAACCTA AGGA-TTATA
Pfa4      GTATTTGTGA GGCCTTATAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pbe5      GTATTTGTGA AGCCTTATAA GAAAAA--T- TTTTAATTAA AGGAATTATA
Plo6      GTATTTGTGA AGCCTTATAA GAAAAAAGT- TACTAACTAA AGGAATTATA
Pfr7      GTACTTGTGA AGCCTTATAA GAAAGAAGT- TATTAACCTA AGGAATTATA
Pkm8      GTACTTGTGA AGCCTTATAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pcy9      GTACTCGTGA AGCCTTTTAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pvi10     GTACTTGTGA AGCCTTTTAA GAAAAAAGT- TATTAACCTA AGGAATTATA
Pga11     GTATTTGTGA AGCCTTATAA GAAAAAAGT- TATTAATTAA AGGAATTATA
```



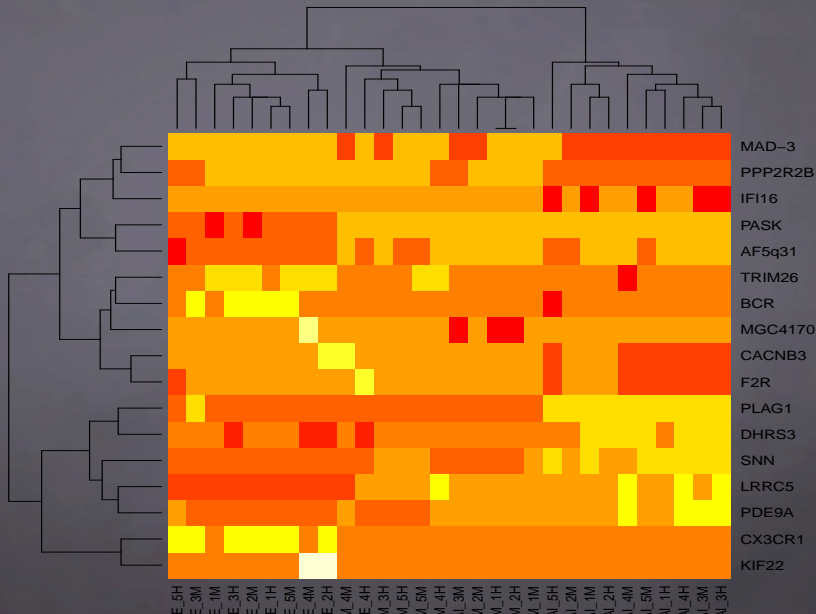
```
ACAAAGAAGT AACACGTAAT AA--ATTAT TTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTAAT AA--ATTATA TTTATTA--- -AGTGTGTAT
ACAAAGAAGT AACACATAAT AAA-TTTCGA -ATATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTAAT AA--ATTAT TTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACATAAT AT--ATTAC TATATTT--- -AGTGTGTAT
ACAAAGAAGT AACACATAAT AAAGCTCGCT CTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTGAA ATGGATTAACT TCCATTTTTT TAGTGTGTAT
ACAAAGAAGT AACACGTAAT --GGATTCT- TCCATTTT-- TAGTGTGTAT
ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
ACAAAGAAGT AACACATAAT AAAACTTTGT TTTATTT--- -AGTGTGTAT
```

Phylogenetic tree is the unknown parameter

Estimated in different ways from DNA/AA data:

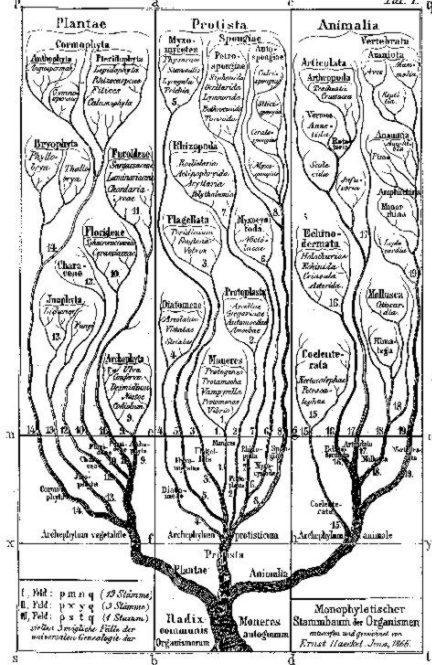
- Parametric: ML estimation, PAML, Phylml, FastML, RaxML,...
- Distance based methods: Neighbor Joining, UPGMA,..
- Parsimony: Steiner tree problem: nonparametric.
- Bayesian estimation, Mr Bayes by MCMC, from posterior sampling distribution.

Hierarchical Clustering Trees



An introduction to Phylogeny

Representation of biological families by trees predates Darwin's theory of evolution, although the latter gave such representations a true explanatory justification. For biologists, at each branch of the tree are situated separation events that split orders or families or genera or species. An early example is the classification made by Haeckel, 1870.

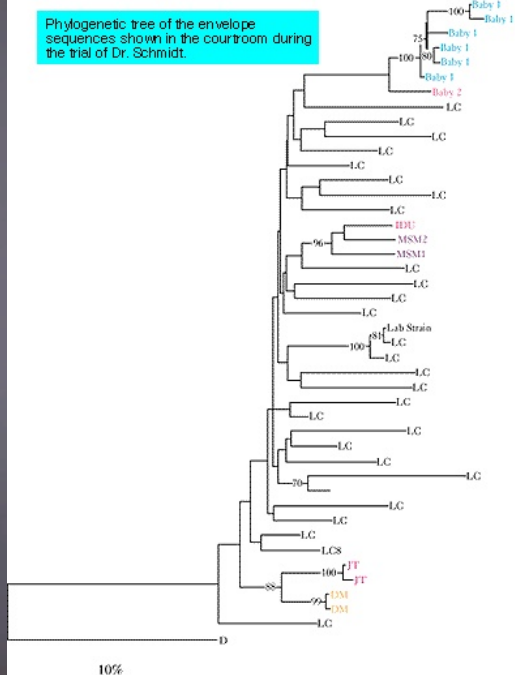


I. Feld: p m n q (D. Steiner)
 II. Feld: p x y q (J. Steiner)
 III. Feld: p a t q (F. Steiner)
 Das folgende Bild der waterenheit kinologisch dar

Monophyletischer
 Stammbaum der Organismen
 tierischer und pflanzenlicher
 Ernst Haeckel, Jena, 1866.

Radix organismorum
 Moneras autogenum

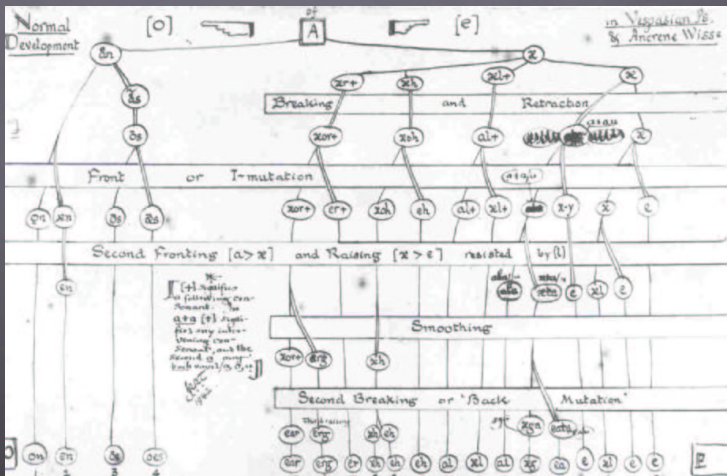
Phylogenetic tree of the envelope sequences shown in the courtroom during the trial of Dr. Schmidt.



10%

Less symmetrical Phylogenies

Linguistics use trees to map out the history of language. Linguists use trees, but they have an ancient form and a novel form. So their trees do not have symmetry between siblings.



Number of trees ?

Felsenstein, 1978 published the number of phylogenetic trees

$$(2n - 3)!! = (2n - 3) \times (2n - 5) \times \dots 5 \times 3$$

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

Coding Trees as Perfect Matchings

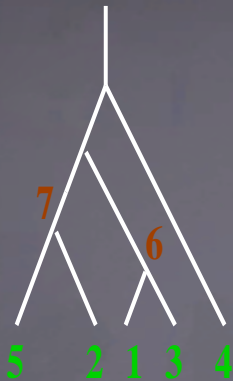
A perfect matching on $2n$ points is a partition of $1, 2, \dots, 2n$ into n two-element subsets. It is well known that there are $(2n)!/2^n n!$ distinct perfect matchings. When $n = 2$, the three perfect matchings are

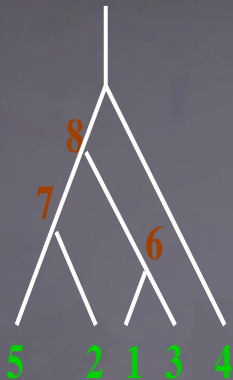
$$\{1, 2\}\{3, 4\}; \{1, 3\}\{2, 4\}; \{1, 4\}\{2, 3\}$$

From Trees to Matchings





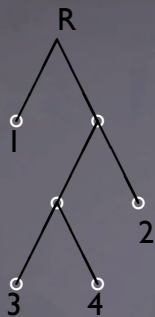




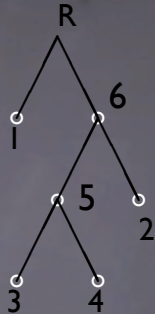
Put down the sibling pairs:

$$(1, 3)(2, 5)(6, 7)(8, 4)$$

We briefly describe the correspondence between matchings and trees. Begin with a tree with ℓ labeled leaves. Label the internal vertices sequentially with $\ell + 1, \ell + 2, \dots, 2(\ell - 1)$ choosing at each stage the ancestor which has both children labeled and who has the descendent lowest possible available label (youngest child). Thus the tree



is labeled



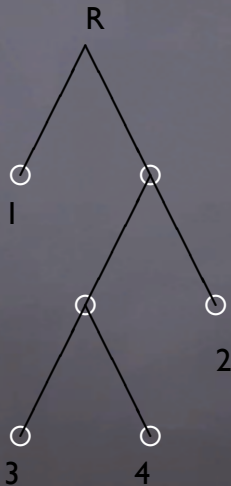
When all nodes are labeled, create a matching on $2n = 2(\ell - 1)$ vertices by grouping siblings. In the example above, this yields

$$\{3, 4\}\{2, 5\}\{1, 6\}.$$

From matchings to trees

To go backward, given a perfect matching of $2n$ points, note that at least one matched pair has both entries from $\{1, 2, 3, \dots, n+1\}$. All such labels are leaves; if there are several leaf-labeled pairs, choose the pair with the smallest label. Give the next available label ($n+2 = \ell+1$) to their parent node. There are then a new set of available labeled pairs. Choose again the pair with the smallest label to take the next available label for its parent, and so on.

For example, $\{3, 4\}\{2, 5\}\{1, 6\}$ has $2n = 6$ and $\{3, 4\}$ has both entries from $\{1, 2, 3, 4\}$. The parent of these is labeled 5 and thus matched with 2 and then the parent of $\{2, 5\}$ is matched with 1, yielding



Matchings and Decompositions

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 \mathcal{B}_{n-1} the subgroup of \mathcal{S}_{2n-2} that fixes the pairs

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

then

$$\mathcal{M}_{n-1} = \mathcal{S}_{2n} / \mathcal{B}_{n-1}$$

and

$$|\mathcal{M}_{n-1}| = \frac{(2n-2)!}{2^{n-1}(n-1)!} = (2n-3)!! = (2n-3) \times (2n-5) \times \dots \times 3 \times 1$$

$(\mathcal{S}_{2n-2}, \mathcal{B}_{n-1})$ form a Gelfand pair Diaconis and Shahshahani (1987)

$$L(\mathcal{M}_{n-1}) = V_1 \oplus V_2 \oplus \dots \oplus V_\lambda$$

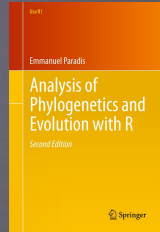
A multiplicity free representation.

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

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

Just to take the first few: for $\lambda = n - 1$ \mathcal{S}^λ are the constants, and this gives the sample size. for $\lambda = (n - 2, 1)$, \mathcal{S}^λ are the number of times each pair appears. for $\lambda = (n - 3, 2)$, \mathcal{S}^λ are the number of times partition of 4 appears in the tree. for $\lambda = (n - 3, 1, 1)$, \mathcal{S}^λ are the number of times 2 pairs appear simultaneously.

Matchings are useful

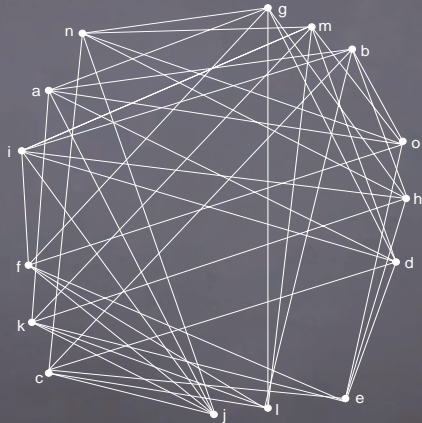


- For going through all trees systematically. (Gray code for Trees)
- Doing vigorous random walks on tree space.
- Doing Fourier Analysis on Tree Data.

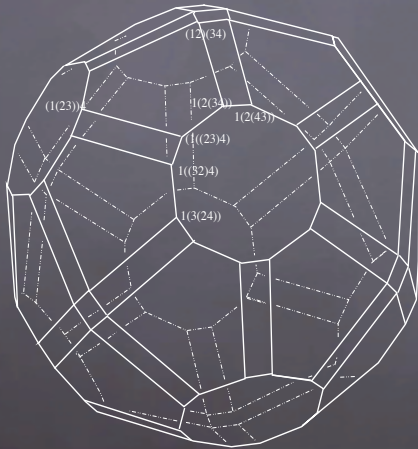
But the matching distance is not satisfactory to the biologists.

The Matching Polytope

o



Cornell, 1997: The permuto-associahedron



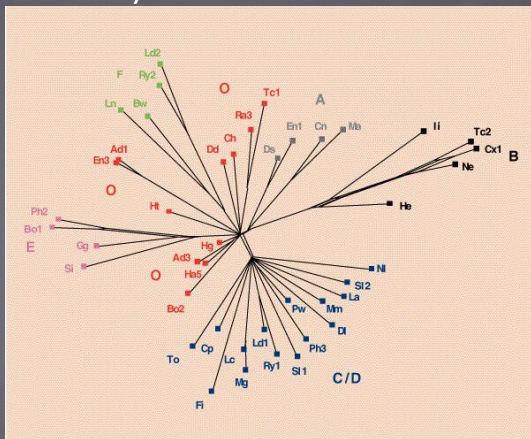
A book on polytopes. (Ziegler)
But the trees are extreme points
Quotients (?/!)

One tree from one gene : many gene trees

A gene sequence might be about 2000 base pairs long. One of the problems that has occurred in the last 20 years is that biologists believe that the way evolution works is that there would only be one species tree.

Different genes have different histories, so you get different gene trees. Putting them together is also a statistical problem: trying to find out what the average of the different genes are. We're going to study the evolutionary process as one of our models for trying to understand what happens over time and how these mutations occur. What we see with the data is some columns with changes. We're going to try to make a model for how these substitutions occur and use that model in various ways to try to make up the tree. The models we use are all Markovian. If you write them in discrete time, we have probability of a change occurring as the transition probability.

Copying Model not only for DNA



Chaucer

Continuous time Markov chains

Memoryless Property $P(Y(u+t) = j | Y(t) = i)$ doesn't depend on time before t

Time homogeneity $P(Y(h+t) = j | Y(t) = i)$ doesn't depend on t , only depends on h , time between the events.

Instantaneous transition rate

$$P_{ij}(h) = q_{ij}h + o(h), j \neq i.$$

$$P_{ii}(h) = 1 - q_i(h) + o(h), \quad q_i = \sum_{j \neq i} q_{ij}$$

q_{ij} is known as the instantaneous transition rate.

Times between changes are exponential

$$P(T \geq t + h) = P(T \geq t)P(T \geq t + h | T \geq t)$$

$$P(T \geq t + h) = P(T \geq t)P(T \geq h)$$

$$= P(T \geq t)(1 - q_i h + \dots)$$

$$\frac{P(T \geq t + h) - P(T \geq t)}{h} = -q_j P(T \geq t)$$

$$\frac{dP(T \geq t)}{dt} = -q_i P(T \geq t)$$

$$P(T \geq 0) = 1$$

gives solution

$$P(T \geq t) = e^{-q_i t}$$

$$P(T \leq t) = 1 - e^{-q_i t}$$

$$f(t) = q_i e^{-q_i t} \sim \text{Exp}(q_i)$$

Derivative of P

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = -q_j P_{ij}(t) + \sum_{k \neq j} q_{kj} P_{ik}(t)$$

as $h \rightarrow 0$,

$$\frac{dP_{ij}(t)}{dt} = -q_j P_{ij}(t) + \sum_{k \neq j} q_{kj} P_{ik}(t)$$

The simplest possible model we'll study, the mutations are all equally likely. This model, called a Jukes-Cantor model is a one parameter model. We suppose that every transition is reversible and that the probability is that they're all equal.

Particular case of Jukes-Cantor: $q_j = 3\alpha$ and $q_{ij} = \alpha, i \neq j$.

$$\begin{aligned}\frac{dP_{ij}(t)}{dt} &= -3\alpha P_{ij}(t) + \alpha \sum_{k \neq j} P_{ik}(t) \\ &= -3\alpha P_{ij}(t) + \alpha(1 - P_{ij}(t)) \\ &= \alpha - 4\alpha P_{ij}(t) \\ P_{ii}(0) &= 1 \text{ and } P_{ij}(0) = 0\end{aligned}$$

gives solutions

$$\begin{aligned}P_{ii}(t) &= \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \\ P_{ij}(t) &= \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\end{aligned}$$

The rate matrix Q is of the form:

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ T & \alpha & -3\alpha & \alpha & \alpha \\ C & \alpha & \alpha & -3\alpha & \alpha \\ G & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

The Kimura two parameter model is:

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & -\alpha - 2\beta & \beta & \beta & \alpha \\ T & \beta & -\alpha - 2\beta & \alpha & \beta \\ C & \beta & \alpha & -\alpha - 2\beta & \beta \\ G & \alpha & \beta & \beta & -\alpha - 2\beta \end{array}$$

The 12 parameter model is of the form

$$Q = \begin{array}{c|cccc} & A & T & C & G \\ \hline A & - & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ T & \alpha_{2,1} & - & \alpha_{2,3} & \alpha_{2,4} \\ C & \alpha_{3,1} & \alpha_{3,2} & - & \alpha_{3,4} \\ G & \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & - \end{array}$$

The substitution matrix gives the probability of the change of a nucleotide during a time t as the continuous Markov chain with infinitesimal generator Q .

In the case of the amino acids we would have bigger matrices (20×20 instead of 4×4), but most of the other computations carry through. The best reference about these subjects are the books by W. H Li and WH Li and D. Graur. See also Page and E. Holmes on Molecular Evolution: A phylogenetic approach.

Estimating the rates

- Call λ the amino acid replacement rate per year,

$$\lambda = \frac{K}{2t} = \frac{\#substit.}{2 \times divergence\ time}$$

- Probability that a site stays unchanged through t intervals is $(1 - \lambda)^{2t}$
- The probability D_t of one or more replacements occurring in t units of time is

$$1 - (1 - \lambda)^{2t}$$

-

$$\begin{aligned}1 - D_t &= (1 - \lambda)^{2t} \\ \log(1 - D_t) &= 2t \log(1 - \lambda) \\ \log(1 - D_t) &= \frac{K}{\lambda} \log(1 - \lambda) \simeq -K\end{aligned}$$

Expected proportion of differences between sequences at time t .

Example : β globin molecule in primates

contains 146 amino acids, the estimates of the number of differences

	Time of div. (millions of years)	Average # of amino acid changes	average \hat{D} differ.	$-\log(1 - \hat{D})$
	85	25.5	25.5/146	.192
are:	60	24	24/146	.180
	42	6.25	6.25/146	.044
	40	6.0	6.0/146	.042
	30	2.5	2.5/146	.018
	15	1.5	1.5/146	.007

The slope is around $a = .002$, and the evolution rate is half of this, so: 10^{-3} per million years or 10^{-9} per year.

Human	MVHLTPEEKSAVTALWGKVNVEVGGGALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Gorilla	MVHLTPEEKSAVTALWGKVNVEVGGGALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Rabbit	MVHLSSEEEKSAVTALWGKVNVEVGGGALGRLLVVYPWTQRFFESFGDLSSANAVMNNPK
Cow	M..LTAEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPK
Goat	M..LTAEKAAVTGFWGKVKVDEVGAEALGRLLVVYPWTQRFFEHEFGDLSSADAVMNNAK
Mouse	MVHLTDAEKAAVSCLWGKVNSEVGGGALGRLLVVYPWTQRFFESFGDLSSASAIMGNNAK
Chicken	MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMM
Carp	MVEWTDAAERSAIIIGLWGKLNDELGPQALARCLIVYPWTQRFFASFGNLSSPAAIMGNPK

61

120

Human	VKAHGKKVLGAFSDGLAHLDNLKGTFAATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Gorilla	VKAHGKKVLGAFSDGLAHLDNLKGTFAATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
Rabbit	VKAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVIVLSHHFG
Cow	VKAHGKKVLDSFNSGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFG
Goat	VKAHGKKVLDSFNSGMKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVVVLARHHG
Mouse	VKAHGKKVITAFNDGLNHLDSLKGTFAALSELHCDKLHVDPENFRLLGNMIVIVLGHHLG
Chicken	VRAHGKKVLTSGDAVNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIIVLAAHFS
Carp	VAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLVHVPDNPFRLLADICITVCAAMKFG

121

148

Human	.KEFTPPVQAAYQKVVAGVANALAHKYH
Gorilla	.K.....
Rabbit	.KEFTPQVQAAYQKVVAGVANALAHKYH
Cow	.KEFTPVLQADFQKVVAGVANALAHRYH
Goat	.SEFTPLLQAQEFQKVVAGVANALAHRYH
Mouse	.KDFTPAAQAAFQKVVAGVATALAHKYH
Chicken	.KDFTPQCQAQWQKLVVVVAHALARKYH
Carp	PSGFSPNVQEAQKFLSVVVSALCRQYH

Human beta-globin vs. Gorilla beta-globin

Percent Similarity: 100

Percent Identity: 99

```

      .           .           .           .
Human   1 MVHLTPEEKSAVTALWGKVNVDENVGGEALGRLLVVYPWTQR
      ||||||||||||||||||||||||||||||||||||||||||||
Gorilla 1 MVHLTPEEKSAVTALWGKVNVDENVGGEALGRLLVVYPWTQR
      .           .           .           .
51  TPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHC
      ||||||||||||||||||||||||||||||||||||||||||||
51  TPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHC
      .           .
101  PENFRL LGNVLVCVLAH HFGK 121
      |||:||||||||||||||
101  PENFKL LGNVLVCVLAH HFGK 121
```

We're going to separate out two problems, which in today's age of computing, should be mixed together: alignment and trees.

I'm going to suppose we have sequences either of amino acids or nucleotides which we have aligned. This is an example data set I did in my first phylogeny paper I wrote was with Brad Efron in which we analyzed malaria data. These are malaria sequences from 11 different species of malaria. Two of the species of malaria are human malaria. The others are from different animals. The question in trying to find out information from the families has a lot of influence on designing vaccines.

Malaria Data

11 1620

```
Pre1      G T A C T T G T T A   G G C C T T A T A A   G A A A A A A A G T -   T A T T A A C T T A   A G G A A T T A T A
Pme2      G T A T C T G T T A   A G C C T T A T A A   A A A G A T A G T -   T - T A A A T T A A   A G G A A T T A T A
Pma3      G T A T T T G T T A   A G C C T T A T A A   G A G A A A A G T A   T A T T A A C T T A   A G G A - T T A T A
Pfa4      G T A T T T G T T A   G G C C T T A T A A   G A A A A A A A G T -   T A T T A A C T T A   A G G A A T T A T A
Pbe5      G T A T T T G T T A   A G C C T T A T A A   G A A A A A - - T -   T T T T A A T T A A   A G G A A T T A T A
Plo6      G T A T T T G T T A   A G C C T T A T A A   G A A A A A A A G T -   T A C T A A C T A A   A G G A A T T A T A
Pfr7      G T A C T T G T T A   A G C C T T A T A A   G A A A G A A G T -   T A T T A A C T T A   A G G A A T T A T A
Pkn8      G T A C T T G T T A   A G C C T T A T A A   G A A A A G A G T -   T A T T A A C T T A   A G G A A T T A T A
Pcy9      G T A C T C G T T A   A G C C T T T T A A   G A A A A A A A G T -   T A T T A A C T T A   A G G A A T T A T A
Pvi10     G T A C T T G T T A   A G C C T T T T A A   G A A A A A A A G T -   T A T T A A C T T A   A G G A A T T A T A
Pgal1     G T A T T T G T T A   A G C C T T A T A A   G A A A A A A A G T -   T A T T A A T T T A   A G G A A T T A T A
```

```
ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTAAT AA--ATTATA TTTATTA--- -AGTGTGTAT
ACAAAGAAGT AACACATAAT AAA-TTTCGA -ATATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTAAT AA--ATTTAT TTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACATAAT AT--ATTTAC TATATTT--- -AGTGTGTAT
ACAAAGAAGC AACACATAAT AAAGCTGCGT CTTATTT--- -AGTGTGTAT
ACAAAGAAGT AACACGTGAA ATGGATTAAC TCCATTTTTT TAGTGTGTAT
ACAAAGAAGT AACACGTAAT --GGATTC TCCATTTT-- TAGTGTGTAT
ACAAAGAAGT AACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
ACAAAGAAGC GACACGTAAT --GGATCCG- TCCATTTT-- TAGTGTGTAT
ACAAAGAAGC AACACATAAT AAAACTTTGT TTTATTT--- -AGTGTGTAT
```

Transitions and Transversions

The probability of changing from a purine to a pyrimidine is called a transversion. If you think about coding sequences, the amino acids you don't code the amino acid if you have a transition. We make the two parameter model is the most used in the study of evolution. We don't have discrete time, that's just a simplification.

Model 0: Jukes Cantor

This model is not a completely realistic model.

All mutations, transversions and translations are equally likely.

The probability of it not changing is $1 - 3\alpha$. This is discrete time markov chain matrix.

You can look at it stationary distribution because you have a perfect symmetry, the left eigenvector is $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$.

This stationary distribution of $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$.

If for a long time you have sequences evolving over time and you're lost track of time and you pull a nucleotide at random it has equal probability of being any of those.

Transitions and Transversions

The probability of changing from a purine to a pyrimidine is called a transversion. If you think about coding sequences, the amino acids you don't code the amino acid if you have a transition. We make the two parameter model is the most used in the study of evolution. We don't have discrete time, that's just a simplification.

Distance based methods Variants of hierarchical cluster analysis.

The aim is to reconstruct the distances as computed between the two sequences of the two species x and y by distances along the edges of the tree forming a path between x and y .

First a distance matrix is constructed between the N units in some way. These distances d_{xy} are supposed to estimate the unknown 'true evolutionary' distances between x and y as they would be measured along the unknown true tree \mathcal{T} .

For the Jukes-Cantor model which assumes equal rates of substitution between all base pairs provides the estimate of distances between sequences x and y as:

$$d_{xy} = -\frac{3}{4} \log\left(1 - \frac{4}{3}\left(1 - \left(\frac{\#AA}{k} + \frac{\#CC}{k} + \frac{\#GG}{k} + \frac{\#TT}{k}\right)\right)\right)$$

where k denotes the number of characters (columns) in the data matrix, and $\#AA$ denotes the number of times there is an A in x matched with an A in y .

Once the distances are decided upon, the parametric model is left behind and a clustering technique such as hierarchical clustering with average groups is used to find the tree from the distances.

Remarks:

If we knew the true evolutionary distances between species, we could build an additive tree that reproduced the distances along the tree in a unique way.

The existence of an additive tree reproducing the distances faithfully is not always ensured, a sufficient condition for this to be possible is called the **four point condition**(for all quadruples):

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC}).$$

This means that one of the two sums is minimum and the other two are equal. Notice that this is not the same as the ultrametric property which says that for any three points: A, B, C:

$$d_{AC} \leq \max(d_{AB}, d_{BC})$$

If the distances obey the ultrametric property the distances can be fit to a binary tree with leaves equally distant from the root.

Unfortunately distances computed from real data never obey this property.

Additivity is destroyed by:

- Homoplasy (reversal, parallelism and convergence) which is caused by superimposed changes.
- An uneven distribution of change rates.
- Measurement error.
- **Paralogous** sequences.

We concentrate on distances that are computed from substitution models such as Jukes and Cantor's one-parameter model, Kimura's two-parameter model, or even the complex 12-parameter model for the substitution matrices. These models provide estimates of differences between sequences computed from the frequencies of various changes in the sequences.

Parsimony method

Nonparametric procedures. Farris (1983), has a justification for parsimony : “minimizes requirements of ad hoc hypotheses of homoplasy”.

Analogy is made between homoplasies and residuals, (part of the data that the tree does not explain), minimizing homoplasies is akin to minimizing residuals in regression.

Roughly this method can be seen as based on the assumption that “evolution is parsimonious” which means that there should be no more evolutionary steps than necessary.

Thus the best trees are the ones that minimize the number of changes between ancestors and descendants. Under independence of each of the characters, this has a clear combinatorial translation.

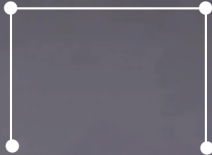
The parsimony tree as a combinatorial problem

Unrooted parsimony trees.

Recall that the Hamming distance between two units is the number of changes needed to bring one to the other. This assumes that all changes in a categorical character are counted as one step.

$$d_H(\text{AACTGGG}, \text{AACTGGC}) = d_H(\text{AACTGGG}, \text{AACTGGA}) = 1$$

Here, given N points in a metric space, the Steiner problem is that of finding the shortest tree connecting the N points where one is allowed to add extra vertices. Thus, with 4 points arranged at the vertices of a unit square, one would add a fifth point in the center to form the Steiner tree.



The minimum spanning tree and the Steiner tree of the 4 vertices of a rectangle.

Although statisticians are not familiar with minimal Steiner trees, they may have encountered minimal spanning trees as used by Friedman and Rafsky (1985).

The relation between the two is well explained in Gardner's wonderful chapter on Steiner trees (Chapter 22, Gardner (1997)). He explains how minimal spanning trees are good "starting points" since in the plane for instance they can only be 13% longer than Steiner trees.

As a combinatorial problem, the maximum parsimony tree is the problem of finding the Steiner points or Steiner tree for Hamming distance between the units, under the constraint that the tree be binary.

The problem of finding a minimal Steiner tree is that of finding the Steiner points (representing ancestors) that minimize the complete length of the tree. Steiner points are points that are added to a graph so that its minimal spanning tree becomes shorter.

Computation issues

The minimal Steiner tree problem is NP-hard, meaning that no algorithm is known that will compute an optimal tree in polynomial time in the number of species N .

Much work has been done to implement good heuristic algorithms for finding approximately optimum trees. Swofford's PAUP, Felsenstein's `PhyLip`, and Goloboff's `NONA` all contain clever use of branch and bound techniques and branch swapping to find acceptable answers.

#species=1500 can now be done routinely.

Parsimony as a statistical procedure

Felsenstein (1983) lists parsimony in a section entitled a section on parsimony as “non-statistical approaches”. Farris says (1983) says the “statistical approach to phylogenetic inference was wrong from the start, for it rests on the idea that to study phylogeny at all one must first know *in great detail* how evolution has proceeded”. Both these authors identify statistics with parametric modeling.

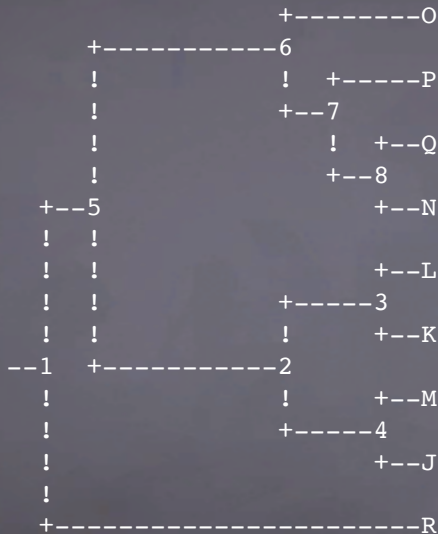
In fact parsimony is just a nonparametric method of estimating the tree parameter.

Simple Example

T7 data experimentally generated phylogeny, Hillis et al. (1992) for which the parsimony program will be seen to produce the correct answer. Here is the part of the data set (in `phylip` form) composed of the informative sites:

```
9 21
R   C C G C C G G C C G G C C A G C G G G G T
J   C C C C G T A C C G G T C A A C G G G G T
K   T C C C G C A C C G A T C A A T G G G G G
L   T C C C G C A C C G A T C A A T G G G G G
M   C T C C G T A C C G G T C A A C G G G G T
N   C C T T A C G T T A G C T G G C A A A A T
O   C T C C G C G C T G G C C G G C A G A A T
P   C C C C A C G C T G G C C G G C A G A A T
Q   C C T T A C G T T A G C T G G C A A A A T
```

One most parsimonious tree found:



remember: this is an unrooted tree!

requires a total of 25.000

steps in each site:

Output: the Newick notation

The output file called `treefile` contains the following line (the tree in parentheses format):

```
((O,(P,(Q,N))),((L,K),(M,J))),R);
```

Rooting the Tree

At least one of the taxonomic units has a special function. For a statistician it would be seen as a simple outlier: the biologists voluntarily include what they call an **outgroup** to locate the root of the tree. The root is situated by creating an unrooted tree and the edge that joins the outgroup to the other species will be the support for the root.

This is a clever use of prior information that simplifies the problem considerably, (by a factor of $(2N - 3)$). What is less obvious to the outsider is why, once the root's position is decided upon, the biologists keep the outgroup in the data set - it seems to distort the image of the closer group (called the **ingroup**), in fact outgroups also provide information on the root's characters, and so on the ancestral states of the character.

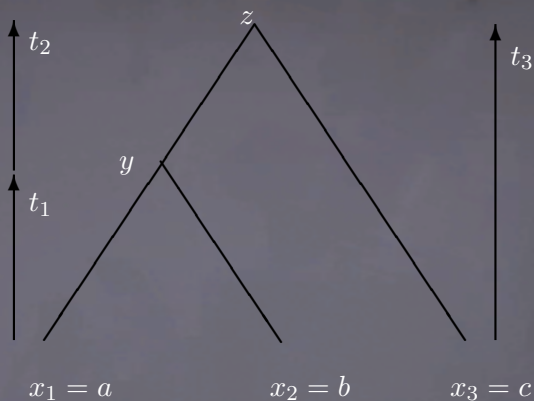
Maximum likelihood trees

For a statistician this is the easiest of the methods to understand. A parametric model (θ, T) is postulated, θ is a η -dimensional vector that we explain below and T is the tree's topology. Under this model the likelihood for each possible tree T is separately computed for each character, the independence of characters then allows the total likelihood of the tree for all data to be computed by taking the product.

The first part of the vector of parameters θ comes from the Markovian substitution model as explained before.

The number of other parameters that have to be specified depends on the complexity of the model. If a molecular clock¹ is postulated, speciation times $\{t_1, t_2, \dots, t_{N-2}\}$ (splitting events) are the other parameters. Otherwise both the branch lengths $\{v_1, v_2, \dots, v_{N-2}\}$ and the different rates along those branches have to be parametrized.

¹branch lengths in evolutionary change depend linearly on time



The substitution parameters are estimated from the data. A complete model including distributions of separation events is postulated and the likelihood can be computed for each possible tree by computing the likelihood of the tree for each site $X_{.j}$:

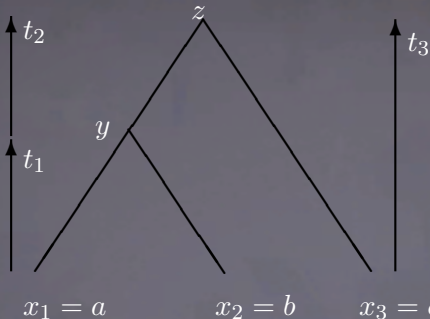
$$f(X_{.j} | \theta_1, \theta_2, \dots, \theta_\eta, \mathcal{T}).$$

This actually requires computing the likelihood of all the subtrees, so the method is recursive.

$$\mathcal{L}(\theta_1, \theta_2, \dots, \theta_\eta | X_{.1}, X_{.2}, \dots, X_{.k}, \mathcal{T}) = \prod_{j=1}^k f(X_{.j} | \theta, \mathcal{T})$$

The essential assumptions:

1. Each site in the sequence evolves independently.
2. Different lineages evolve independently.
3. Each site undergoes substitution at an expected rate (can be extended to a series of rates with a given distribution).



$$x_1 = a \quad x_2 = b \quad x_3 = c$$

Likelihood: $P(\text{data} | \text{Tree}, t\text{'s}, \text{ancestors}, \text{mutation rates})$. Based on the probabilities computed given the tree and for potential ancestors ($t_3 = t_1 + t_2$)

$$P(a, b, c, y, z | T, t) = P(a|y, t_1)P(b|y, t_1)P(c|z, t_3)P(y|z, t_2)P(z)$$

$$P(a, b, c, | T, t) = \sum_z \pi_z P_{zc}(t_3) \sum_y P_{zy}(t_2) P_{ya}(t_1) P_{yb}(t_1)$$

This is a function of t_1, t_2 whose values are estimated as the maximum for a given tree topology, then for the ml estimate is made for each T.

The T with the maximum value is the maximum likelihood estimate. We can consider the likelihood computation, one character at a time. Starting from the root, or starting from the leaves, Felsenstein's transversal method starts from the leaves, we abbreviate the character we are interested from x_{ij} to x_i . For two leaves with the residue a at their common ancestor (the root here):

$$P(x_1, x_2, a | \mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \pi_a P(x_1 | a, \theta_1) P(x_2 | a, \theta_2)$$

The root is an unknown nuisance parameter that we integrate out:

$$P(x_1, x_2 | \mathcal{T}, \theta_1 = t_1, \theta_2 = t_2) = \sum_a \pi_a P(x_1 | a, \theta_1) P(x_2 | a, \theta_2)$$

Call $m[i]$ the direct parent of i , and $P(L_i|a)$ denote the probability of all nodes below i given that the node i is a . We number the inner nodes from $(n+1)$ to $(2n-2)$, these ancestral nodes are all unknown, so we have to sum the probabilities of all their possible assignments to compute the complete likelihood of the tree, given its edge lengths $(\theta_1, \theta_2, \dots, \theta_{2n-2})$.

The algorithm is similar to the forward algorithm in HMM.

Sum over possible paths, working upwards from the leaves.

Compute $P(L_j|e), P(L_k|f)$ for all e and f at daughter nodes j, k of i

$$P(L_i|a) = \sum_{b,c} P(b|a, t_j) * P(L_j|b) * P(c|a, t_k) * P(L_k|c)$$

We can write down the complete probability as a sum.

We denote the alphabet of possible residuals \mathcal{A} ,

$$P(x^1, x^2, \dots, x^{(2n-2)} | \mathcal{T}, \theta) \\ = \sum_{(a^{n+1}, \dots, a^{2n-1}) \in \mathcal{A}^{n-2}} \pi_{a^{2n-1}} \prod_{n+1}^{2n-2} P(a^i | a^{m[i]}, \theta_i) \prod_1^n P(x^i | a^{m[i]}, \theta_i)$$

the computational algorithm evaluates $P(L_i | a)$ for the children j and k such that $m[j] = m[k] = i$, we compute $P(L_j | b)$ and $P(L_k | c)$ for all possible b and c .

These instructions allow us to compute the likelihood of any tree, given its branching order (sometimes called topology) and its branch lengths.

For the maximum likelihood computation, we need to compute the tree that maximizes the likelihood, first for a given branching order, find the branch lengths that maximize the likelihood. This can be done by taking the derivative $\frac{\partial P(x^j | x^{m[j]}, \theta_k)}{\partial \theta_j}$ in order to use the conjugate gradient method for optimising the edge lengths, or we can take an EM approach as Felsenstein, 1981 suggests and implemented in his `phylip` program.

Complexity: Hard

Finding the likelihood of one tree is an NP complete problem

Remark :There is no known polynomial time algorithm that finds the tree with maximum likelihood.

Thus as we need to look at all the topologies, of which there are exponentially many; we see the exact computation becomes quickly intractable as the number of leaves increases.

Nice implementations:

phylip, RaXML, FastML, PhyML, (see wikipedia)...

From R: phangorn, phyml.

Maximum likelihood trees: Output from phylip program
dnaml:

Nucleic acid sequence Max. Likelihood, vers. 3.572c

Empirical Base Frequencies:

A 0.27778 G 0.22685

C 0.22325 T(U)0.27212

Transition/transversion ratio = 2.000000

(Transition/transversion parameter = 1.519971)


```

+J
!
!           +R
!       +--1
!       !   !   +N
!       !   +--4
!       !       !   +O
!   +--5       +--3
!   !   !           !   +P
!   !   !           +--2
--7--6   !           +Q
!   !   !
!   !   +L
!   !
!   +M
!
+K

```

Ln Likelihood = -344.10331

Examined 95 trees

Between	And	Length	Approx.Conf.Limits
-----	---	-----	-----
7	J	0.00006	(zero, infinity)

7		6	0.00003	(zero, infinity)	
6		5	0.00006	(zero, infinity)	
5		1	0.00936	(zero, 0.02236)	**
1	R		0.00466	(zero, 0.01384)	**
1		4	0.00469	(zero, 0.01389)	**
4	N		0.00462	(zero, 0.01369)	**
4		3	0.00003	(zero, infinity)	
3	O		0.00462	(zero, 0.01369)	**
3		2	0.00003	(zero, infinity)	
2	P		0.00462	(zero, 0.01369)	**
2	Q		0.00003	(zero, infinity)	
5	L		0.00006	(zero, infinity)	
6	M		0.00003	(zero, infinity)	
7	K		0.00003	(zero, infinity)	

* = significantly positive, $P < 0.05$

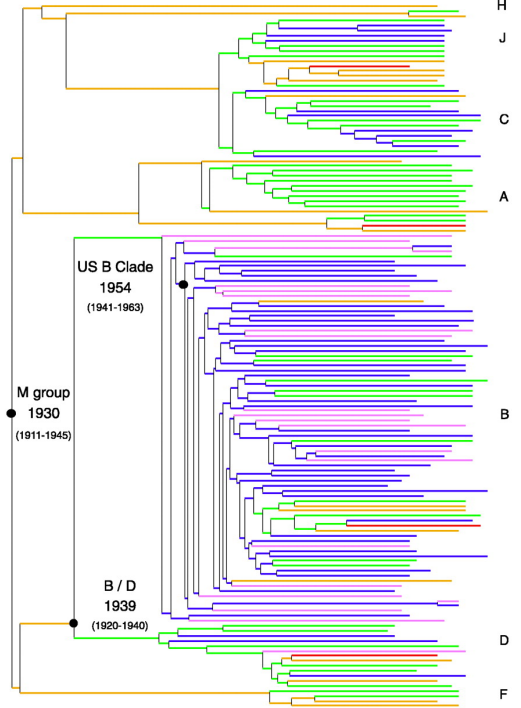
** = significantly positive, $P < 0.01$

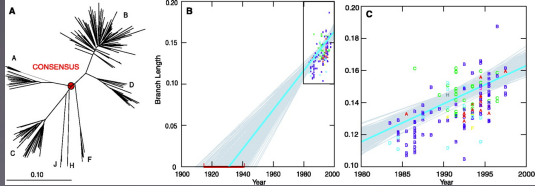
ML Estimate Application: Origins of HIV

The article by Korber et al. provides an estimate of a most recent ancestor. When you see two sequences, how much time went by until the most recent common ancestor.

The English author, Hooper, hypothesis that HIV was spread by dispensaries who were giving the polio vaccination in East Africa. They were supposed to be responsible for diffusing AIDS because the vaccination was grown in monkey tissue. The idea was to try to disprove this occurred at the time of the vaccination program in 1957 and this study was trying to make a confidence interval of the time of the most recent ancestor using as many sequences as they had to make up the whole tree.

One of the reasons this data seemed interested is that this data is freely available on Los Alamos National Laboratories.





The idea is that of the models we are using for molecular evolution, they have this molecular clock.

You have a homogenous process, the number of mutations will be proportionate to time.

There hasn't been much progress in disproving or in proving this molecular clock hypothesis, so the way it's justified is the average amount of mutation that occurs over time.

Parametric bootstrap generation of sequences

Suppose we had the treefile from a previous phymlip output, the generation of sequences is done using Seq-gen (Rambaut and Grassly, 1997) by :

```
seq-gen -mHKY -t3.0 -l27 -n100 < treefile > example
```

For which the output looks like:

Sequence Generator - seq-gen, Version 1.04
(c) Copyright, 1996 Andrew Rambaut and Nick Grassly
Department of Zoology, University of Oxford
South Parks Road, Oxford OX1 3PS, U.K.
Simulating 11 taxa, 27 bases
for 1 tree(s) with 100 dataset(s) per tree
Branch lengths assumed to be number of substitutions
per site
Rate homogeneity of sites.
Model=HKY
transition/transversion ratio = 3 (kappa=6)
frequencies = A:0.25 C:0.25 G:0.25 T:0.25
0% | _____ | 100%
[.....]
Time taken: 0.12 seconds

The data file example.T7 generated looks like this:

```
11 27
R      CCGACCTCCAAGATTCGCTATGACAAT
P      CCGACCTCCAAGATTCGCTATGACAAT
Q      CCGACCTCCAAGATTCGCTATGACAAT
L      CCGACCTCCAAGATTCGCTATGACAAT
M      CCGACCTCCAAGATT.....etc
..
11 27
R      ATGGTAGCGGATAACTGACTTCATCGA
P      ATGGTAGCGGATAACTGACTTCATCGA
Q      ATGGTAGCGGATAACTGACTTCATCGA
L      ATGGTAGCGGATAACTGACTTCATCGA
M      ATGGTAGCGGATAACTGACTTCATCGA
.....      ATGGTAGCGGATAA.....etc
```


This file example. T7 was then submitted to the `phylip` program `dnapsars` with the option `multiple` data sets indicating that there were 100 data sets to analyze, the first part of the output from this looked like this:

```
((R,((((M,K),L),N),Q),(J,P)),O)[0.0100];
((R,((((M,K),L),N),(J,Q)),P),O)[0.0100];
((R,((((M,K),L),(J,N)),Q),P),O)[0.0100];
((R,((((M,K),(J,L)),N),Q),P),O)[0.0100];
((R,((((M,(J,K)),L),N),Q),P),O)[0.0100];
((((((J,M),(R,K)),L),N),Q),P),O)[0.0100];
((((((J,(R,M)),K),L),N),Q),P),O)[0.0100];
((((((R,J),M),K),L),N),Q),P),O)[0.0100];
((R,((((J,M),K),L),N),Q),P),O)[0.0100];
((((((R,(J,M)),K),L),N),Q),P),O)[0.0100];
((R,J),((((M,K),L),N),Q),P),O)[0.0100];
((J,(R,((((M,K),L),N),Q),P)),O)[0.0100];
((R,(J,((((M,K),L),N),Q),P)),O)[0.0100];
((R,(J,(((M,K),L),N),Q)),P),O)[0.0100];
((R,(((J,((M,K),L),N),Q),P)),O)[0.0100];
((R,(((J,((M,K),L)),N),Q),P),O)[0.0100];
((R,((((J,(M,K)),L),N),Q),P),O)[0.0100];
(((J,(R,M)),(((K,L),N),Q),P),O)[0.0100];
(((R,J),M),(((K,L),N),Q),P),O)[0.0100];
(((R,(J,M)),(((K,L),N),Q),P),O)[0.0100];
((M,((R,J),(((K,L),N),Q),P)),O)[0.0100];
(((R,J),M,(((K,L),N),Q),P)),O)[0.0100];
(((R,J),((M,((K,L),N),Q)),P),O)[0.0100];
```

Notice at the end of each tree is associated a weight.

Molecular Clock

Says that the probability of changes along the edges of the tree are proportional to edgelengths:

More believable models of Evolution:

The likelihood was computed as:

$$\mathcal{L}(\theta_1, \theta_2, \dots, \theta_\eta | x_{.1}, x_{.2}, \dots, x_{.k}, \mathcal{T}) = \prod_{j=1}^k f(x_{.j} | \theta, \mathcal{T})$$

Variation of rates of substitution among sites.

Variable sites models for the rates considers the sites to have different rates. The new likelihood takes the different rates into account:

$$P(x|T, t, r_K) = \prod_{k=1}^K P(x_k|T, r_k t)$$

We do not have enough information about the sites to know what these rates should be, so we integrate out the variation by integrating out over all values of r using a prior for the rates. Yang proposes to use a gamma $g(r, \alpha, \alpha)$ prior which has mean 1 and variance $1/\alpha$ for the rates.

The likelihood now becomes:

$$P(x|T, t, \alpha) = \prod_{k=1}^K \int_0^{\infty} P(x_k|T, rt)g(r, \alpha, \alpha)dr$$

For each T, this is maximised with respect to t and α .

Actually better by far to use α from other data.

In practice a discrete sum approximation is sufficient.

Similar approach is to use a hidden Markov model for the states (Felsenstein and Churchill)

$$P(x|T, t, \alpha_s) = \prod_{k=1}^K \sum_{l=1}^m a_{kl} P(x_k|T, r_l)g(r, \alpha, \alpha)$$

Different areas can thus be defined:

- Surface sites of proteins may be exposed to more substitutions.
- Loops with exposed sites.
- Beta sheets have an alternance of buried and exposed sites.

Full Bayesian Method

- Prior distribution on all tree branching patterns.
- Gamma distribution for the rates.
- Compute posterior distribution using MCMC.

Implementations: MrBayes , Beast

Open Questions:

- Prior probability model for trees , open question. Uniform distribution on all trees poses big problem:
 $2n - 3!!$ different binary rooted semi-labeled trees with n leaves.
With 10, you have more than a million trees.
- How long to run the MCMC? (Diaconis and Holmes, EJP cannot touch the real case)
Negative results by Mossel and Vigoda on problems with mixtures.
- Using the output from MCMC runs ...we will talk about this.

Distance Based Methods

In phylogenetics, neighbor joining is very similar to the algorithms used for hierarchical clustering.

The aim is to reconstruct the distances as computed between the two sequences of the two species x and y by distances along the edges of the tree forming a path between x and y .

First a distance matrix is constructed between the N units in some way. These distances d_{xy} are supposed to estimate the unknown 'true evolutionary' distances between x and y as they would be measured along the unknown true tree \mathcal{T} .

For the Jukes-Cantor model which assumes equal rates of substitution between all base pairs provides the estimate of distances between sequences x and y as:

$$d_{xy} = -\frac{3}{4} \log\left(1 - \frac{4}{3}\left(1 - \left(\frac{\#AA}{k} + \frac{\#CC}{k} + \frac{\#GG}{k} + \frac{\#TT}{k}\right)\right)\right)$$

where k denotes the number of characters (columns) in the data matrix, and $\#AA$ denotes the number of times there is an A in x matched with an A in y .

Iterative Agglomeration: Bottom Up heuristic

Once the distances are decided upon, the parametric model is left behind and a clustering technique such as hierarchical clustering with average groups is used to find the tree from the distances.

Remarks:

If we knew the true evolutionary distances between species, we could build an additive tree that reproduced the distances along the tree in a unique way.

The existence of an additive tree reproducing the distances faithfully is not always ensured, a sufficient condition for this to be possible is called the **four point condition**(for all quadruples):

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC}).$$

This means that one of the two sums is minimum and the other two are equal. Notice that this is not the same as the ultrametric property which says that for any three points: A, B, C:

$$d_{AC} \leq \max(d_{AB}, d_{BC})$$

$$d_{AC} \leq \max(d_{AB}, d_{BC})$$

If the distances obey the ultrametric property the distances can be fit to a binary tree with leaves equally distant from the root.

Unfortunately distances computed from real data never obey this property.

This can be destroyed by:

- Homoplasy (reversal, parallelism and convergence) which is caused by superimposed changes.
- An uneven distribution of change rates.
- Measurement error.
- **Paralogous** sequences.

Hierarchical clustering trees

Built from distances or dissimilarities between the rows of the data matrix [7].

Common examples include computations of dissimilarities in gene expression or in occurrence of words in texts or webpages.

The resulting hierarchical clustering tree has the advantage over simple partitioning methods that one can look at the output in order to make an informed decision as to the relevant number of clusters for a particular data set.

Microarray studies have popularized the use of a double hierarchical clustering or bi-clustering trees where both the rows and columns of the data are clustered. This is the most popular method for visualizing both relations between genes and patient groups in gene expression studies [1, 5].

Many implementations are available; the illustration in Figure in the introduction was made with `heatmap` function in R [9].

Consequences for statistics on treespace

- The uniform distribution on tree is irrelevant.
- Statistical inference involving phylogenetic trees require more sophisticated probabilities on treespace.
- Would benefit from a notion of neighborhood for trees.

References

- [1] D.B. Carr, R. Somogyi, and G. Michaels. Templates for looking at gene expression clustering. *Statistical Computing & Statistical Graphics Newsletter*, 7:20–29, 1997.
- [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] P. W. Diaconis and S. P. Holmes. Matchings and phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 95(25):14600–14602 (electronic), 1998.
- [5] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863, 1998.
- [6] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Boston, 2004.
- [7] J Hartigan. Representation of similarity matrices by trees. *Journal of the American Statistical Association*, Jan 1967.

- [8] S. Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2):241–255, 2003. Silver anniversary of the bootstrap.
- [9] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.
- [10] E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–9, Sep 2005.
- [11] E. Paradis. Ape (analysis of phylogenetics and evolution) v1.8-2, 2006. <http://cran.r-project.org/doc/packages/ape.pdf>.

Geometry, Statistics and Tree Space.

Susan Holmes

Toulouse,

August 29, 2019



Motivation: Forests of Trees

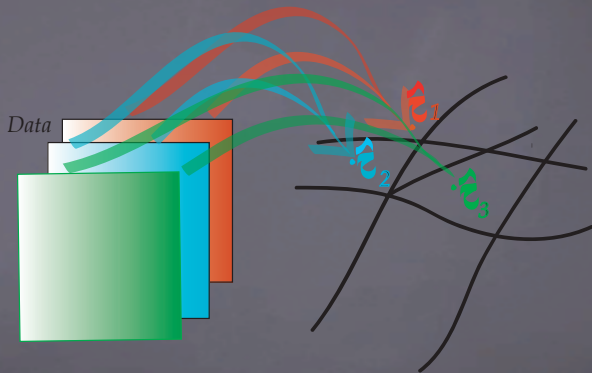
- Different genes, same set of species.
- Bootstrapped Data by Multinomial Resampling, then estimating the tree.
- Bayesian Posterior Distributions on set of Trees.
- Simulated data according to certain evolutionary models (seq-gen).
- Data specimens in different conditions.
- Hierarchical Clustering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

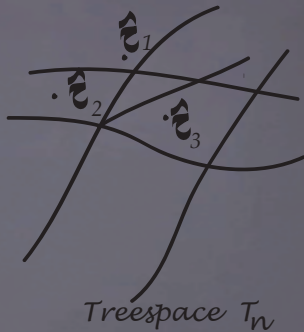
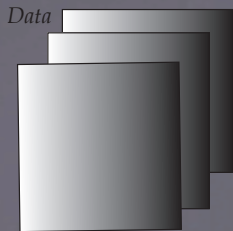
Some Methods for Generating Trees

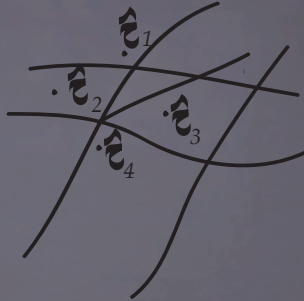
With advances in computational power we can use simulated data to evaluate clustering stability, either in a frequentist (Bootstrap) setting or by using a Bayesian paradigm where trees from a posterior distribution can be generated by MCMC (Monte Carlo Markov chain) methods.

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







True Sampling Distribution



*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.

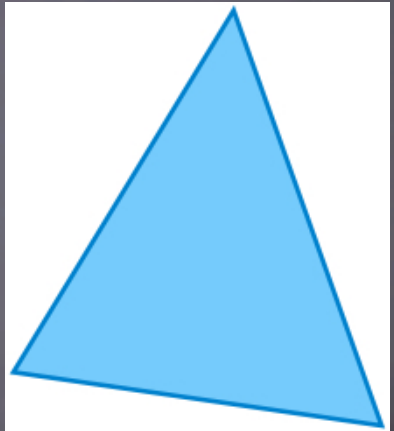
This provides a large set of plausible alternative data sets, each be used in the same way as the original data to give a separate tree (see [13] for a review).

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.

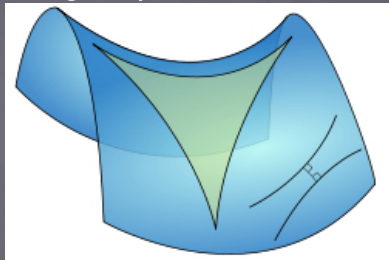
Bayesian methods in hierarchical clustering Heller[23] provide a Bayesian nonparametric method for generating posterior distributions of hierarchical clustering trees.

Euclidean space (where through every point not on a line) is flat:



(sum of angles of a triangle is 180°),

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° .

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.

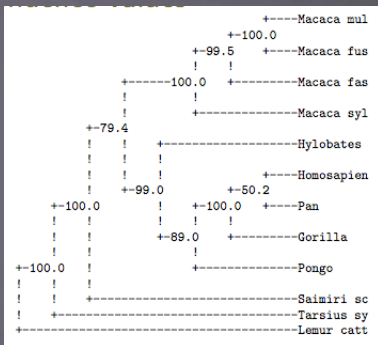
δ -thin: pick three points and draw geodesic lines between them to make a geodesic triangle. Then any point on any of the edges of the triangle is within a distance of δ from one of the other two sides.

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.



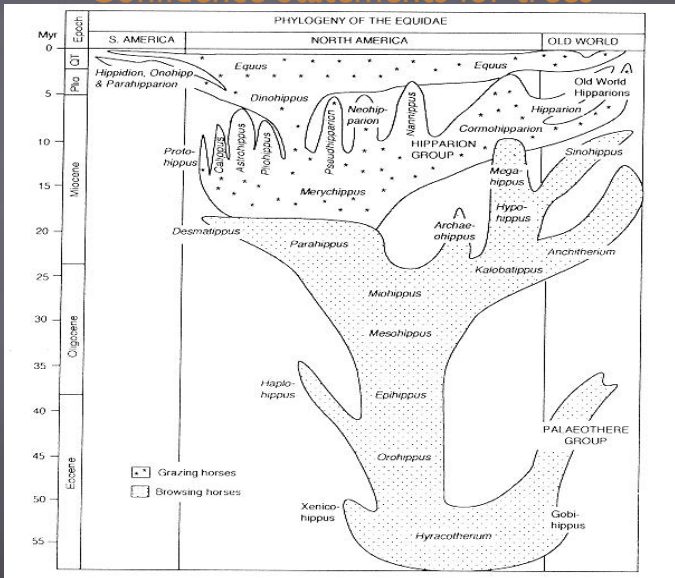
Normal Euclidean space is ∞ -hyperbolic; i.e. not hyperbolic.
Generally, the higher δ has to be, the less curved the space is.

Comparing Different Trees



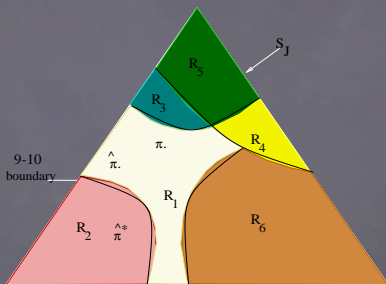
- Binomial Support Estimates (Consensus+support values).
- Split Differences, Visualization Programs .
- Distances.
- Recoding of Trees as binary columns.

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?

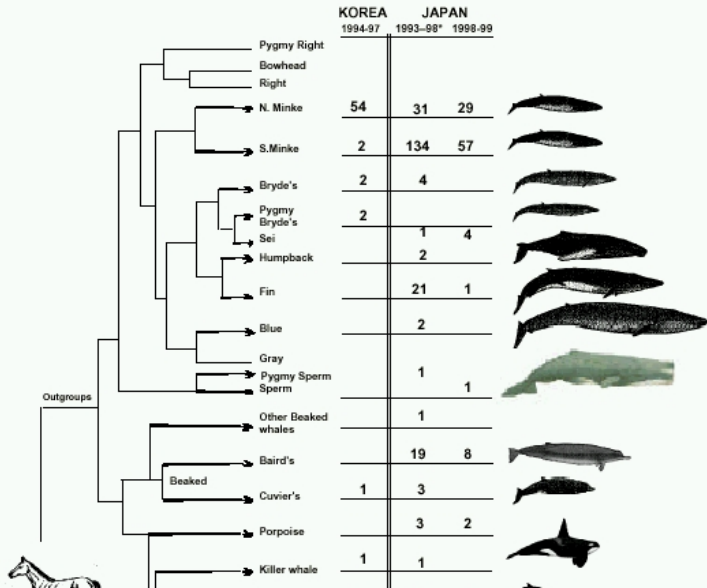
Do we care about confidence statements for phylogenetic trees?

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



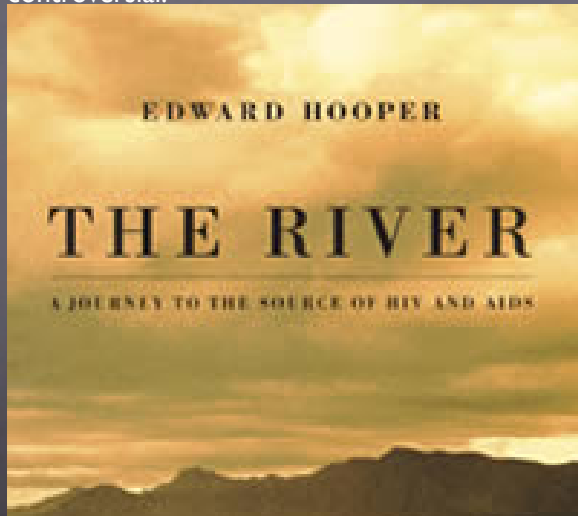


Phylogenetic Identification of Whale and Dolphin Products



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.



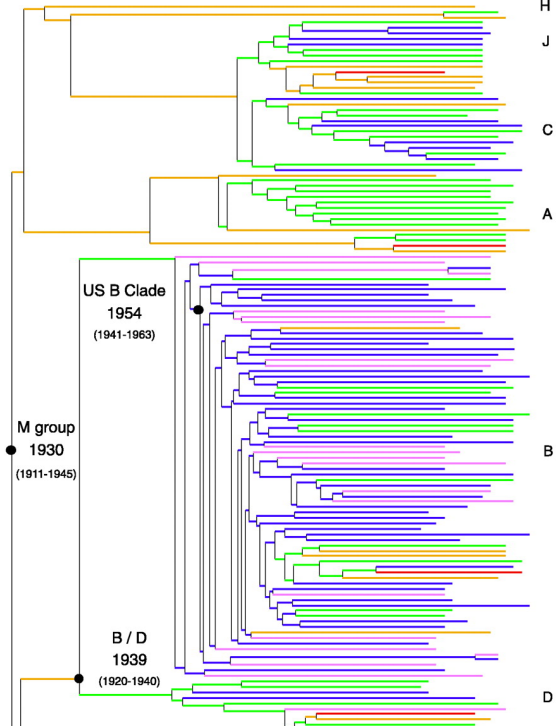
Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign, supporting a model of 'natural transfer' from chimpanzees to humans. If this timescale is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes 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.



Closure: Polio vaccines exonerated
Nature 410, 1035 - 1036 (2001)



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?

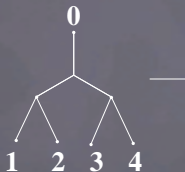
Need modern topology.

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.

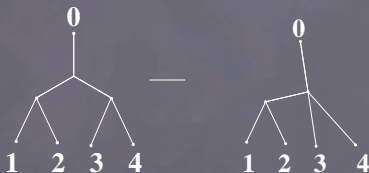
Distances between Trees

- 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**



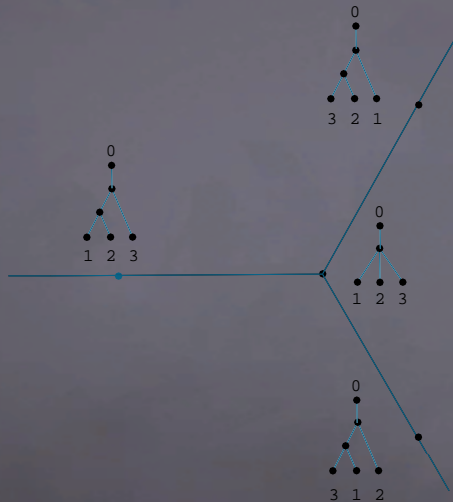
Distances between Trees

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

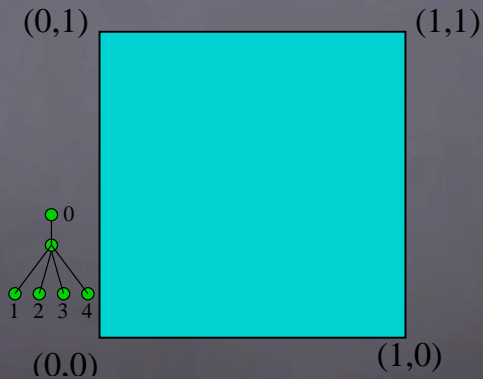


- Subtree Prune Rebranch. (SPR)
- Fill-in of NNI moves: Billera, Holmes, Vogtmann (BHV).
The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology this is called an 'unresolved' tree.

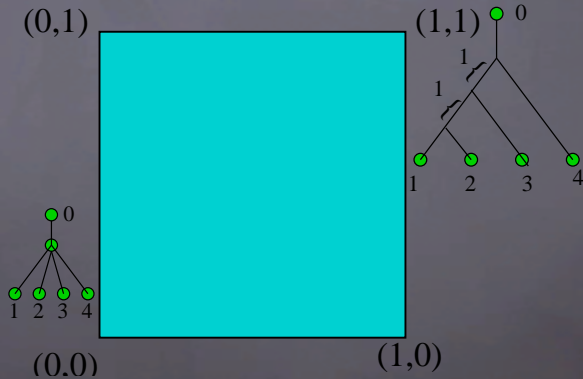
Boundary for trees with 3 leaves



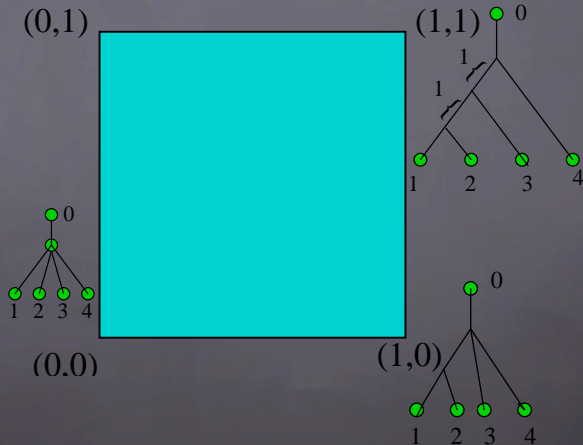
The quadrant for one tree



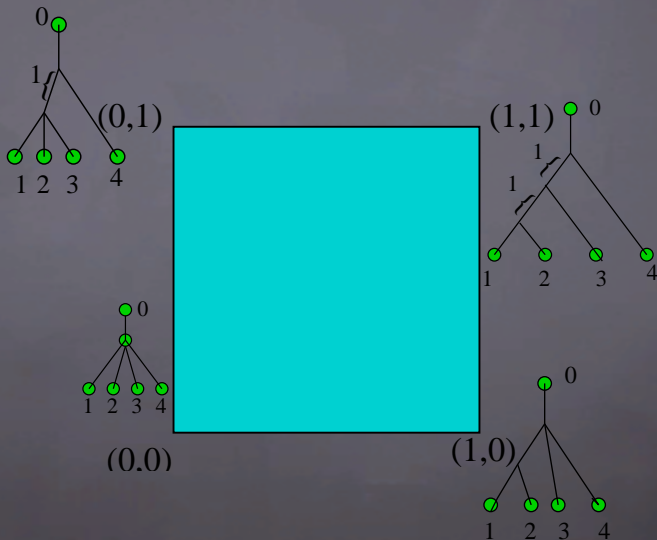
The quadrant for one tree



The quadrant for one tree

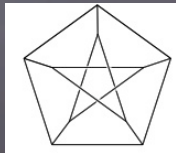
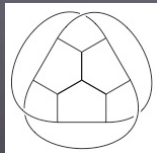
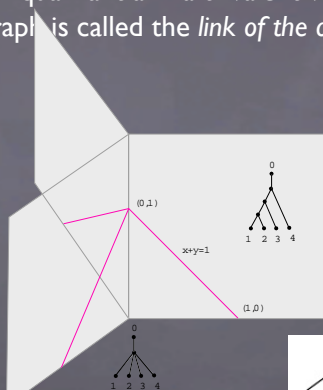


The quadrant for one tree

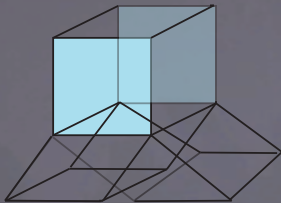


Link of the origin

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*.



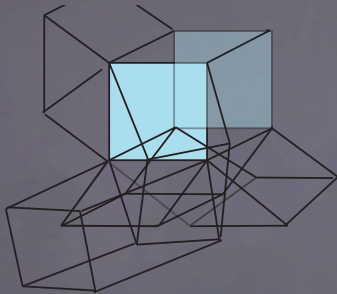
Cube complex of Euclidean Orthants



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.

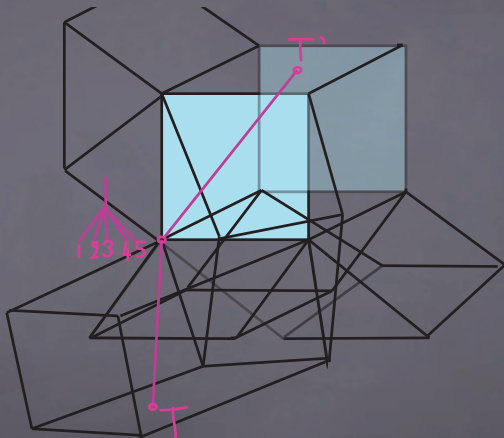
Cube complex of Euclidean Orthants



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

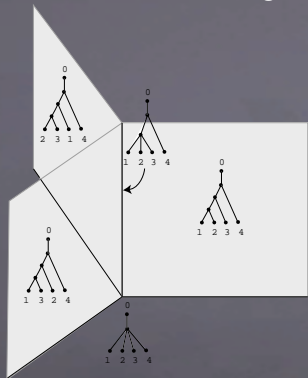
A path is a *geodesic* when it has the smallest length of all paths between two points.

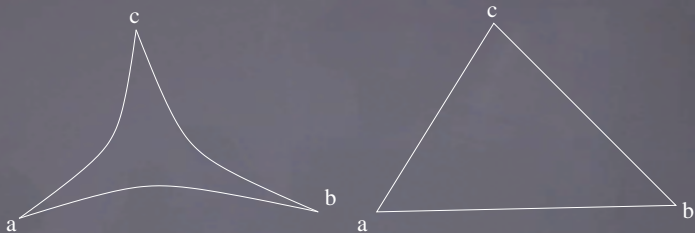
A Cone Path



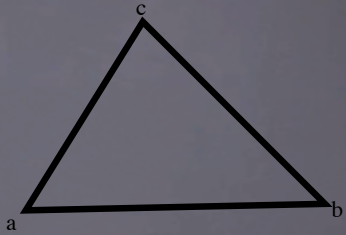
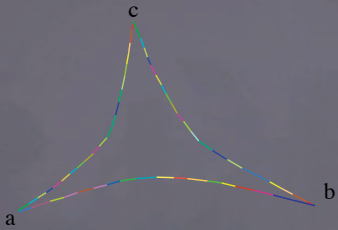
A path between two trees T and T' always exists. Since all orthants connect at the origin, any two trees T and T' can be connected by a two-segment path, this is called the cone-path.

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 constant δ as a relaxation of the above four-point condition:

Gromov's hyperbolicity constant

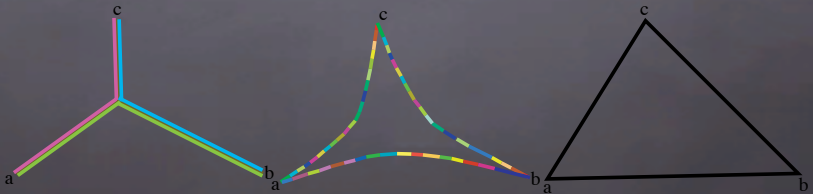
For any four points u,v,w,x , the two larger of the three sums $d(u,v) + d(w,x)$, $d(u,w) + d(v,x)$, $d(u,x) + d(v,w)$ differ by at most 2δ .

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

$$-\frac{1}{2}H\Delta_2H \text{ is semi-definite positive}$$

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

Approximating Non Euclidean Distances by Euclidean ones

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

$$X = HY, \quad \text{with } H = \left(I - \frac{1}{n} \mathbf{1} \mathbf{1}'\right), \text{ and } \mathbf{1}' = (1, 1, 1, \dots, 1)$$

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

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

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' .

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

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:

$$S = -\frac{1}{2}HD_2H.$$

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?

PSYCHOMETRIKA—VOL. 47 NO. 1.
MARCH 1982

SPATIAL VERSUS TREE REPRESENTATIONS OF PROXIMITY DATA

SANDRA PRUZANSKY

BELL LABORATORIES

AMOS TVERSKY

STANFORD UNIVERSITY

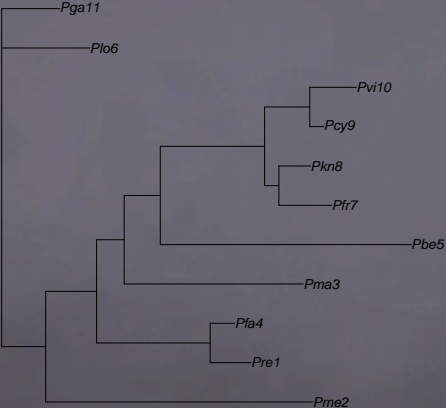
J. DOUGLAS CARROLL

BELL LABORATORIES

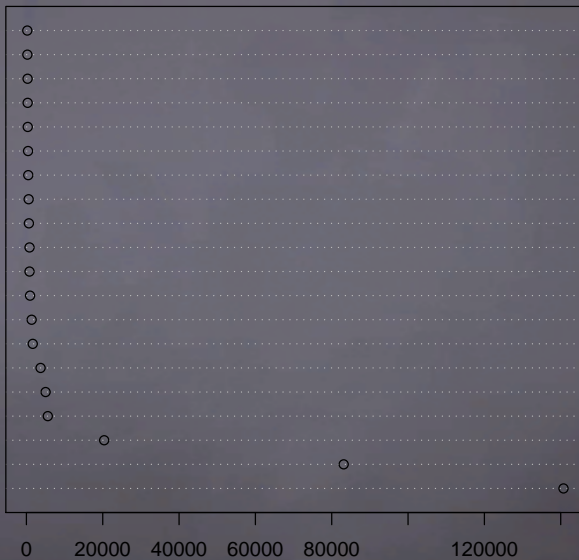
In this paper we investigated two of the most common representations of proximities, two-dimensional euclidean planes and additive trees. Our purpose was to develop guidelines for comparing these representations, and to discover properties that could help diagnose which representation is more appropriate 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.

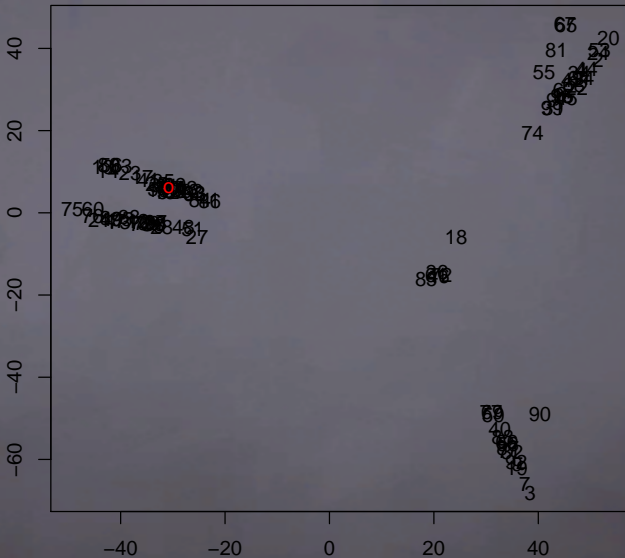
Malaria Data as seen using ape



Eigenvalues of MDS for bootstrapped trees



Bootstrapped trees



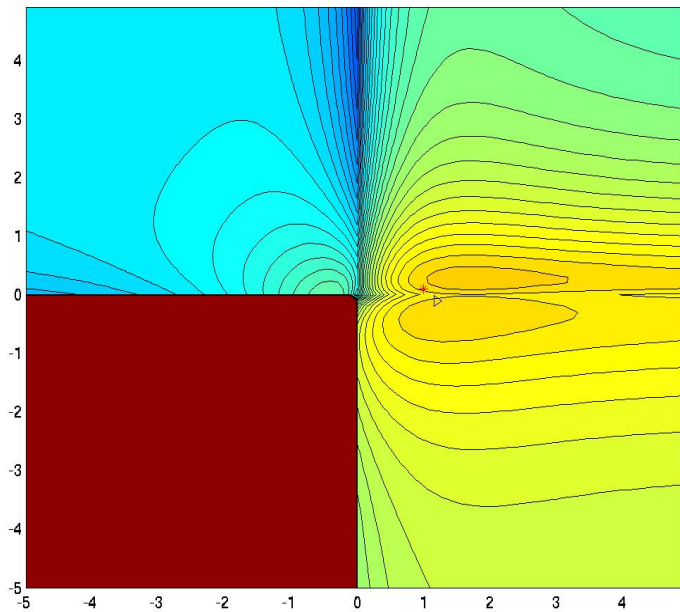
Probability Distributions on Tree Space

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

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

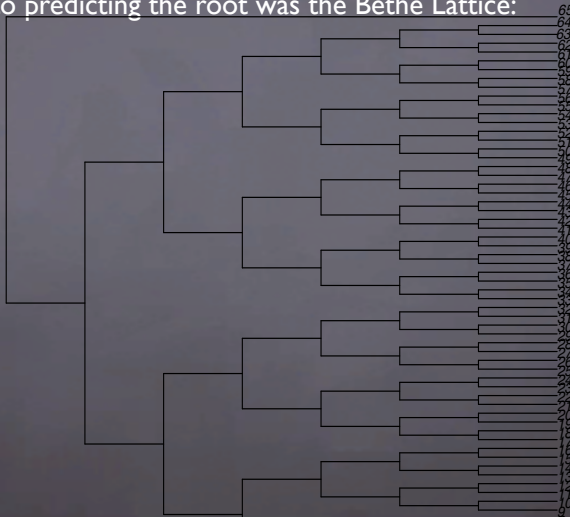
This is really a Mallows[17] model for trees, and as such has possible extensions in similar ways than [10], [11] or those used for rankings developed in [4].

Maximum Likelihood Bootstrap



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?



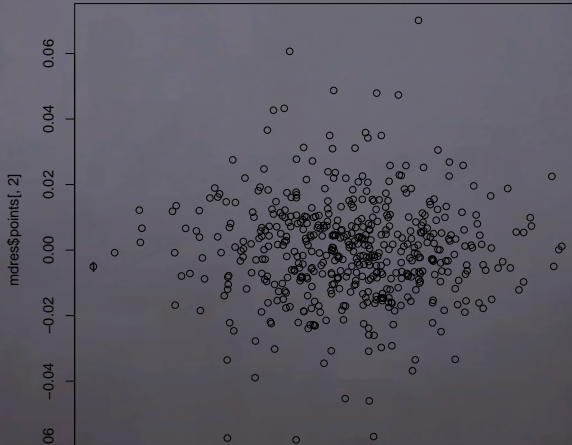
For large enough independent sequences, say for k we can reconstruct the tree with probability $1 - \delta$

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

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

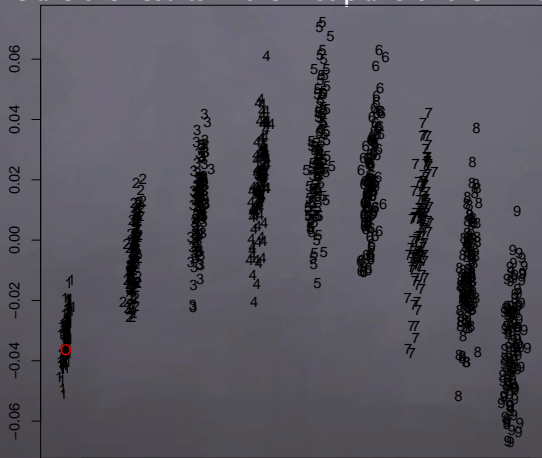
$\alpha = 0.01, \ell = 1000$ MDS plot,



Seeing the Mutation Rate Gradient

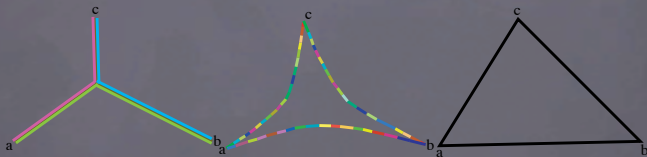
We generated 9 sets of trees with mutation rates set from $\alpha = 0.01$ to $\alpha = 0.09$ and we generated the data according to the Bethe lattice tree.

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.

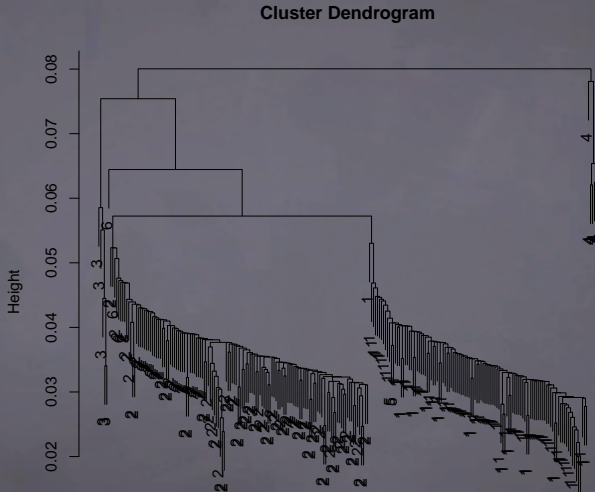


Figure: Hierarchical clustering of 250 trees resulting from a nonparametric bootstrap of the data generated by the double data set \mathcal{X}_{12}

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.

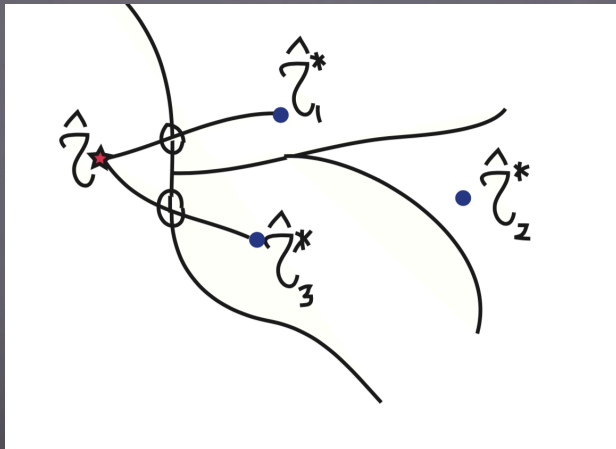
$$P(T) = K \exp(-\lambda d(T, T_0))$$

- 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'



Pfa	A	C	G	T	A	G	C
Pme	A	C	T	G	A	G	C
Pre	A	C	T	T	A	G	C
Pga	A	A	G	T	C	G	A
Pcy	A	A	T	T	C	T	G
Pfr	A	A	T	T	C	T	G

Thinking like a Statistician...

Thinking like a Statistician....

and a geometer..

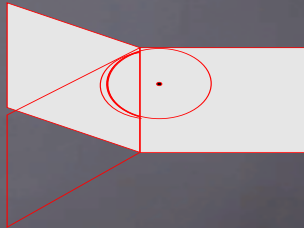
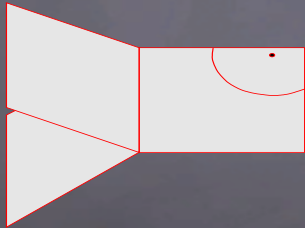
- How treelike are the data ? Model Selection.
- Do we always need the tree, Distances between Data.
- Are all the characters supporting the tree? Leverage.
- Finding hidden gradients Ordination of trees.
- Stability under perturbation Evaluating the estimates.
- How variable are the trees? Variance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

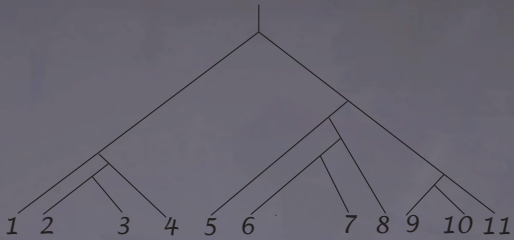


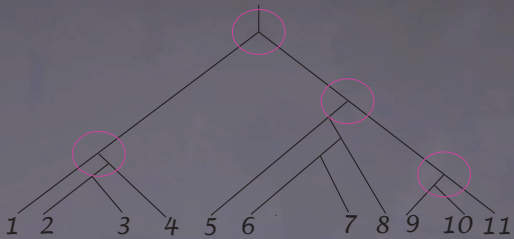
For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree:
15 neighbors. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have two contiguous “small edges” and all the other inner edges significantly bigger than 0.

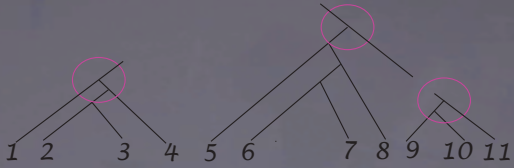
This picture of treespace frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius r around a given tree. All we have to do is check the sets of contiguous edges in the tree smaller than r , say there is only one set of size k , then the neighborhood will contain

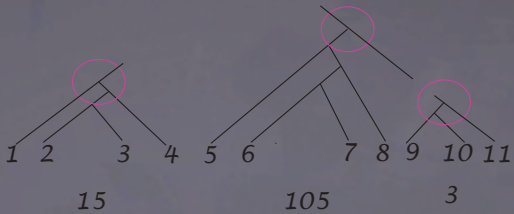
$$(2k - 3)!! = (2k - 3) \times (2k - 5) \times \cdots \times 3 \text{ 'different' trees.}$$

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









In this case the number of trees within r will be $15 * 105 * 3 = 4725$,
in general:

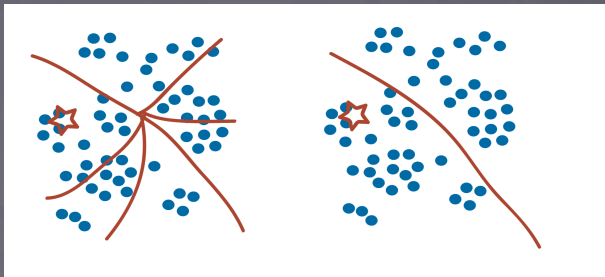
$$(2n_1 - 3)!! \times (2n_2 - 3)!! \times (2n_3 - 3)!! \cdots \times (2n_m - 3)!!$$

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.

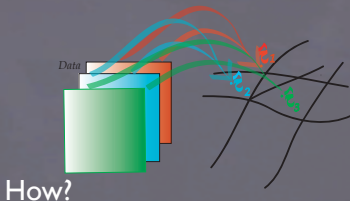
Are there other ways of using the bootstrap than just counting clade appearances?



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

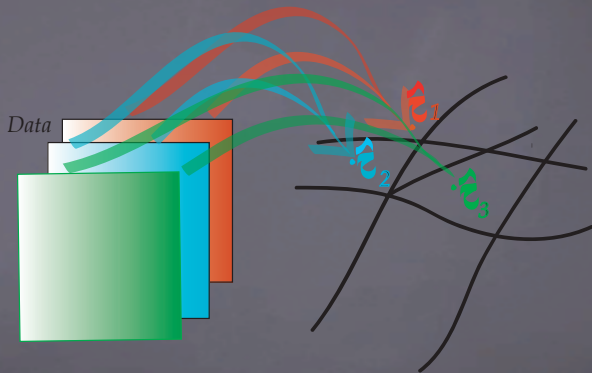
\mathcal{X} original data $\longrightarrow \hat{\mathcal{T}}$ estimate.

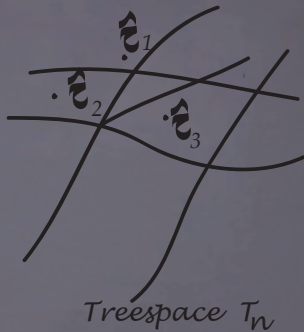
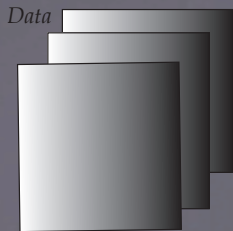


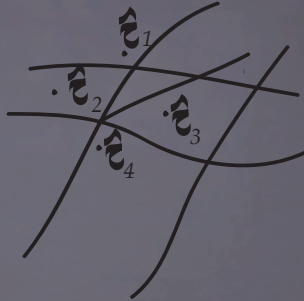
Call \mathcal{X}^* bootstrap samples consistent with the model used for estimating the tree:

- 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







True Sampling Distribution



*Bootstrap Sampling Distribution
(non parametric)*

New resample D^* drawn by resampling rows (genes) from the original $D_{n_{\text{species}} \times n_{\text{char}}}$ 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 enough 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:

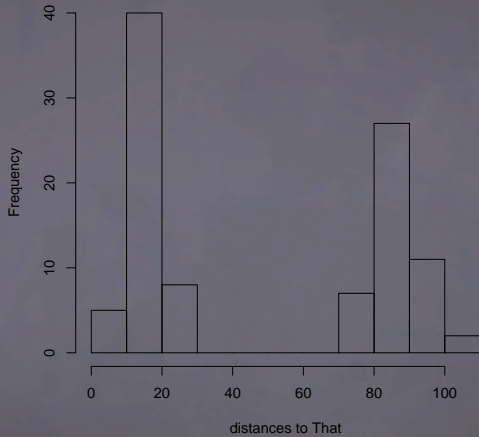
The bootstrap estimate of the sampling distribution of the distances $d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})$ is a good approximation to the true sampling distribution of $d(\hat{\mathcal{T}}, \mathcal{T})$.

Hypothesis Testing

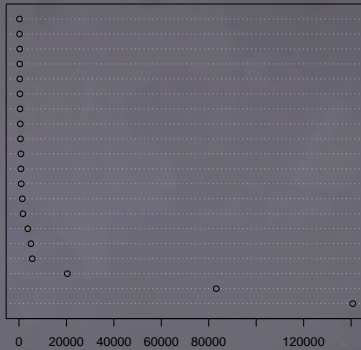
As an additional element we have projected the star tree “ S ” (chosen with the lengths of the pendant edges closest to the original tree) to see whether it is in a small neighborhood, or credibility region of the bootstrapped trees.

This is analogous to seeing if θ 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] .

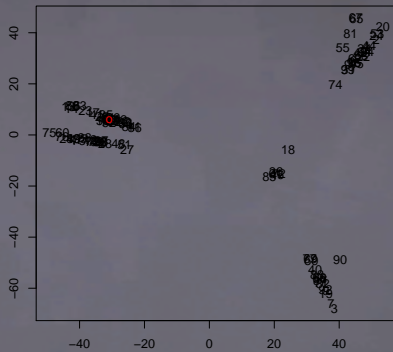
As an aside, note that the numbers in the Figure label the different types of branching patterns. We see that trees of the same topology are not necessarily closer to the original tree if we use the **BHV** with no modifications. In some cases we may want to give an extra weight to crossing orthants (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. *Multivariate 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'espace 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.

Using distances for heterogeneous data analyses.

Susan Holmes

<http://www-stat.stanford.edu/~susan/>

Bio-X and Statistics, Stanford University

September 1, 2019

Part I

Heterogeneity

'Homogeneous data are all alike;

all heterogeneous data are heterogeneous

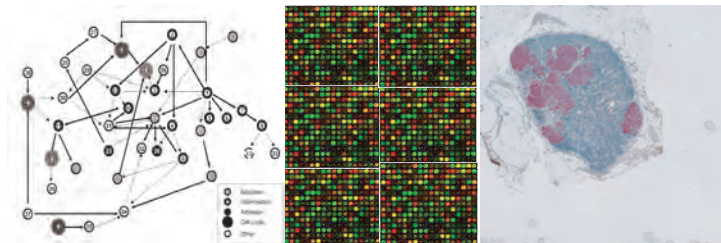
in their own way.'

Heterogeneity of Data

- ▶ Status : response/ explanatory.
- ▶ Hidden (latent)/measured.
- ▶ Types :
 - ▶ Continuous
 - ▶ Binary, categorical
 - ▶ Graphs/ Trees
 - ▶ Images
 - ▶ Maps/ Spatial Information
 - ▶ Rankings
- ▶ Amounts of dependency: independent/time series/spatial.
- ▶ Different technologies used (454, Illumina, PacBio, MassSpec, RNA-seq, Cytof).

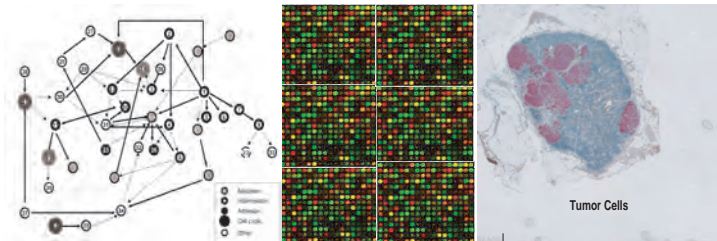
Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration



Goals in Modern Biology: Systems Approach

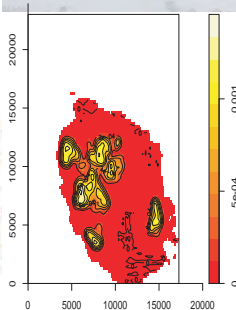
Look at the data/ all the data: data integration



$$\begin{pmatrix} 0110-11000-1 \\ 0110000001 \\ 01-10-10000-1 \\ 0110000101 \\ 01100-11011 \end{pmatrix}$$

$$X_{Blood} = \begin{pmatrix} 0.5 & 1.1 & 1.6 & 1.2 & \dots \\ 0.3 & 1.9 & 2.2 & 1.1 & \dots \\ 1.1 & 0 & 3.2 & 0.4 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 2.7 & 2.3 & 1.2 & 1.1 & \dots \end{pmatrix}$$

$$X_{LN} = \begin{pmatrix} 0.45 & 0.13 & 1.06 & 1.2 & \dots \\ 0.53 & 0.95 & 2.26 & 5.12 & \dots \\ 0.11 & 0 & 3.2 & 1.24 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0.27 & 0.33 & 4.2 & 1.1 & \dots \end{pmatrix}$$



What do statisticians do?

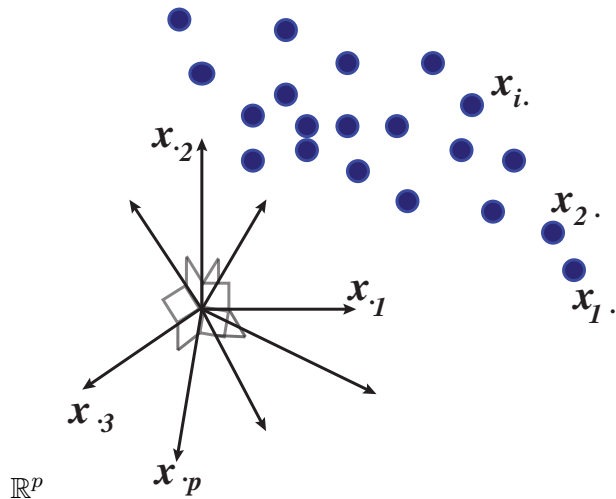
- ▶ Design new experiments to test scientific hypotheses.
- ▶ Visualize and summarize data in ways that account for uncertainties.
- ▶ Look for meaningful differences or structure in high dimensional noisy data.
- ▶ Predict the class of new observations given previously observed ones.
- ▶ Predict the value of a response variable given a whole set of other explanatory variables.
- ▶ Combine different sources of data to understand complex interactions.

Today's challenge

- ▶ Data are not uniformly distributed from some manifold.
- ▶ Data are not an identically distributed random sample.
- ▶ Data are not independent.
- ▶ Data may be combined from different source types (multiway).



Data can often be seen as points in a state space



Distances in Statistics

- ▶ Euclidean Distances, spatial distances.
- ▶ Weighted Euclidean distances: Mahalanobis distance for discriminant analysis.
- ▶ Chisquare distances for contingency tables and discrete data.
- ▶ Jaccard distances for presence absence is one of 50 distances used in Ecology.
- ▶ Earth Mover's distance **on** trees or graphs.
- ▶ Distances **between** aligned graphs or trees.
- ▶ Biologically meaningful distances (DNA, haplotype, Proteins).

What do statisticians use distances for?

- ▶ Summaries through Fréchet Means and Medians and pseudo variances.
 - ▶ Center of Cloud of Objects T_k (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 Fréchet mean object,
 - ▶ or $\sum_{k=1}^K d(T_0, T_k)$ (L^1 or Geometric Median).
 - ▶ Pseudovariance = $\frac{1}{K-1} \sum_{k=1}^K d^2(T_0, T_k) = \hat{s}^2$.

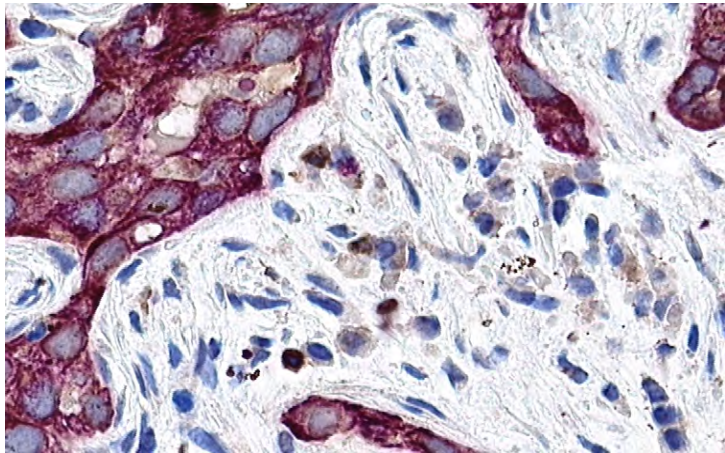
What do statisticians use distances for?

- ▶ Summaries through Fréchet Means and Medians and pseudo variances.
- ▶ Dimension reduction and visualization.
- ▶ Nearest Neighbor Methods.
- ▶ Clustering.
- ▶ Make network edges from close points.
- ▶ Prediction by minimizing weighted residual distances.
- ▶ Cross-products: correlations, autocorrelations.
- ▶ Generalizations of analysis of variance.

Finding the right distance usually solves the statistical problem.

Part II

The Geometries of Data



First example: cell segmentation

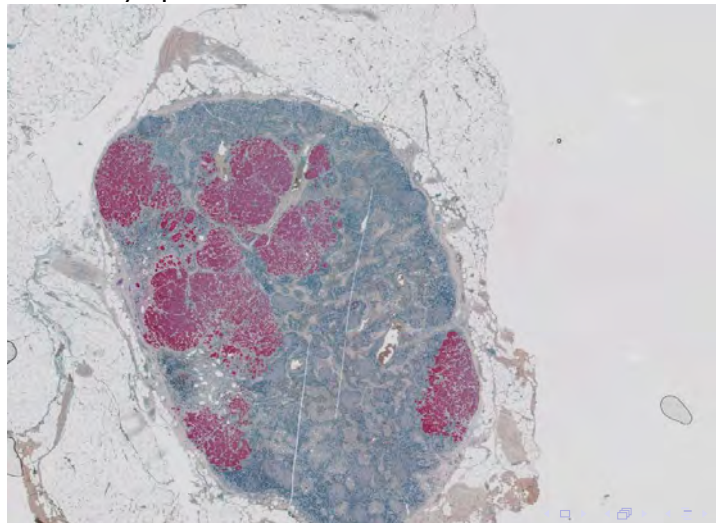
Joint work with Adam Kapelner and PP Lee.

Stained biopsy slides.
levels/wavelengths).

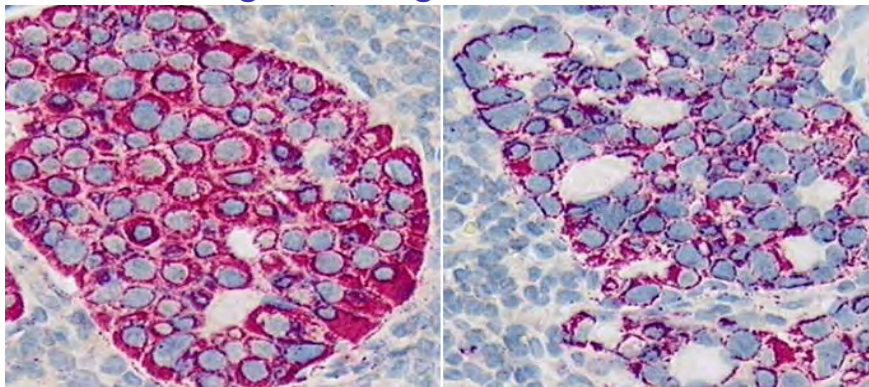
Stained Lymph Node

Multispectral imaging (8

Aim to identify cell.



Problem : Staining is heterogeneous



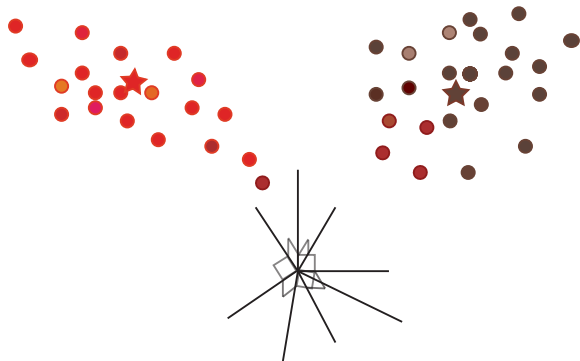
Both images are from the same image set. The stained cells are cancer cells stained with Fast Red red.

Some regions of the tissue stain like the image on the left and other regions stain as the left.

This shows the level of heterogeneity. These are two “subclasses” of the same phenotype (the left is named subclass “A,” the right, subclass “B”).

Problem : Staining is heterogeneous

Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.

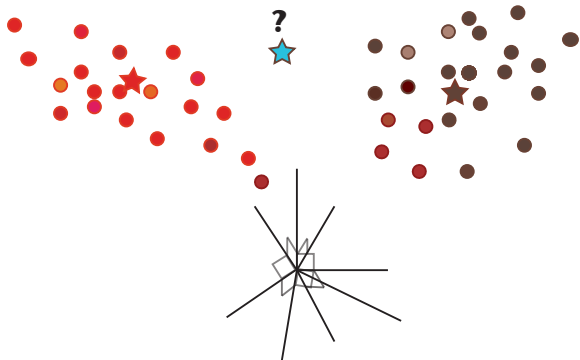


Simple nearest neighbor approach:

- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor

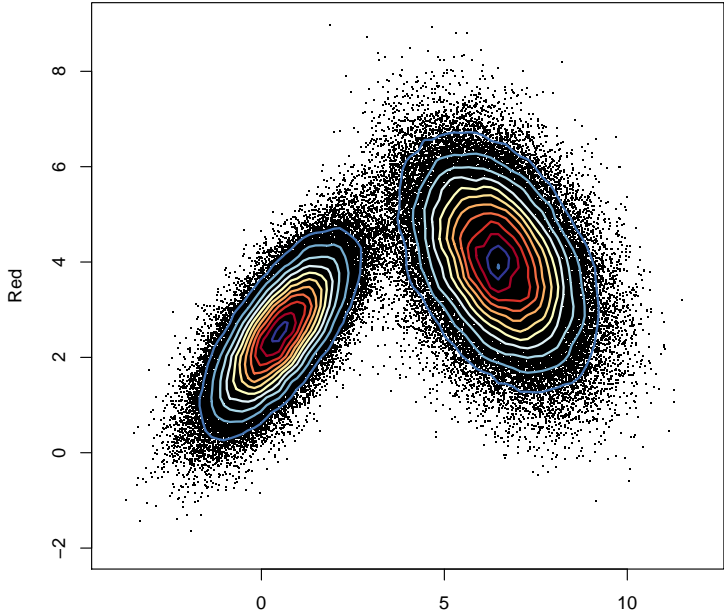
Problem : Staining is heterogeneous

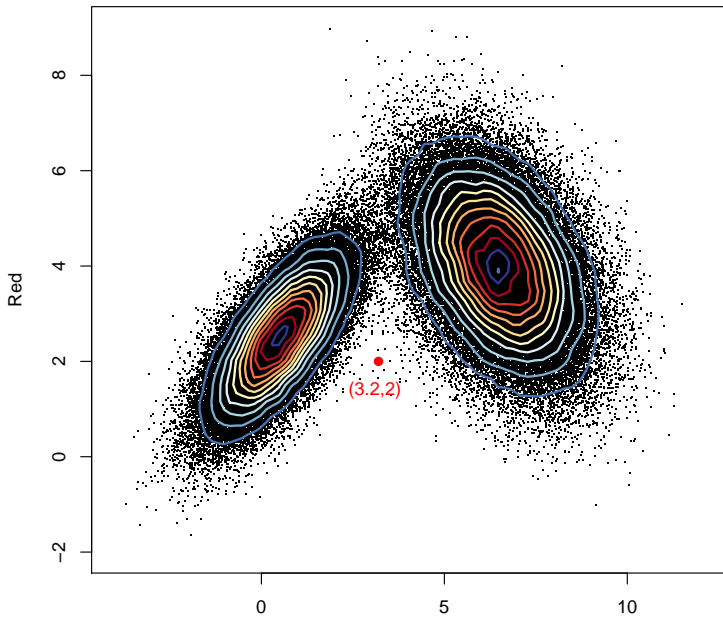
Extreme variability in the image colors/intensity/contrast.
Pixels from a same cell not independent and identically distributed across the different slides or across different cell types.

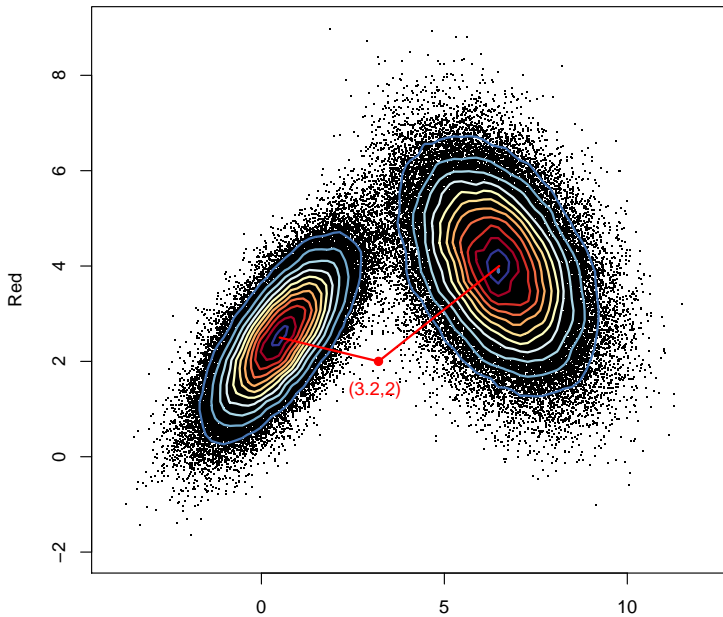


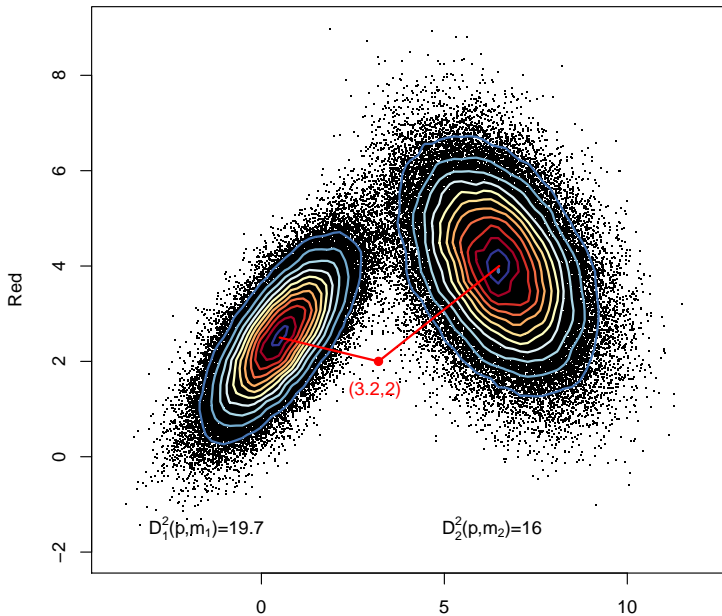
Simple nearest neighbor approach:

- Take 8 dimensional pixels points.
- Assigning the point to the closest neighbor









Multivariate Normal Data

Mahalanobis Transformation.

Several different clusters with different variance-covariance matrices and different means.

(μ_1, Σ_1) (μ_2, Σ_2)

$$D_1^2(x, \mu_1) = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$

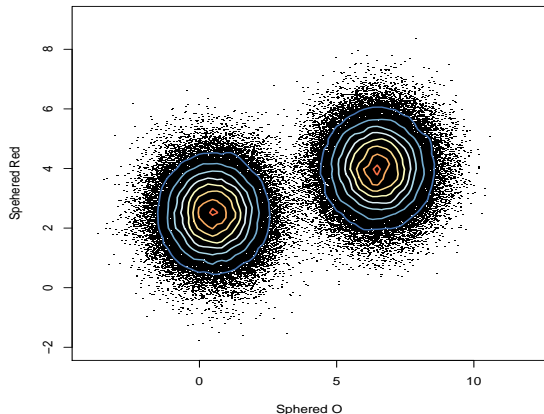
$$D_2^2(x, \mu_2) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$$

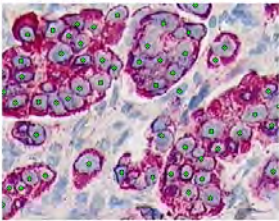
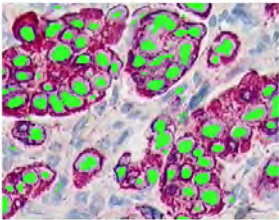
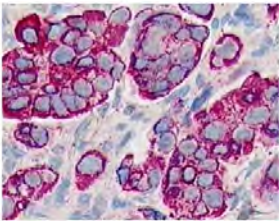
Corresponding Data Transformation

$$H = I - 1D_n1^T, \quad S = X'HD_nHX$$

$$z_{i.} = S^{-\frac{1}{2}}(x_{i.} - \bar{x})$$

This is sometimes called 'data sphering'.





Add Phen **Delete Phen**

Color:

rMin: 1

rMax: 9

Pix/Cent:

Name: tic cells 26

Color:

rMin: 1

rMax: 9

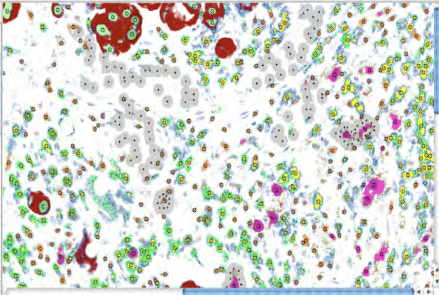
Pix/Cent:

Name: Tumor 111


Color:

rMin: 1

rMax: 10



Zoom Level:



Settings:

Visualize:

See points:

See Typel Errors (1)

Image Colors:

T_cells identified pixels visibility

Denritic cells identified pixels visibility

Tumor identified pixels visibility

other_cells identified pixels visibility

T_cells centroids visibility

Denritic cells centroids visibility

Tumor centroids visibility

other_cells centroids visibility


Mark / Delete Training Point

Location: (962,319)

Boost on classified images


<none> Image 0/239

0-7-0-0-0




stage 0232

185-18-7-31-12




stage 0093

258-14-6-35-13




stage 0147

112-10-5-11-9




stage 0181

165-12-11-10-10




stage 0096

0-0-0-2-0




stage 0278

237-4-3-20-12




stage 0318

0-0-4-0-0



stage 0255

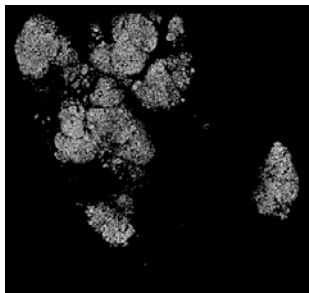
0-0-0-2-0



stage 0126

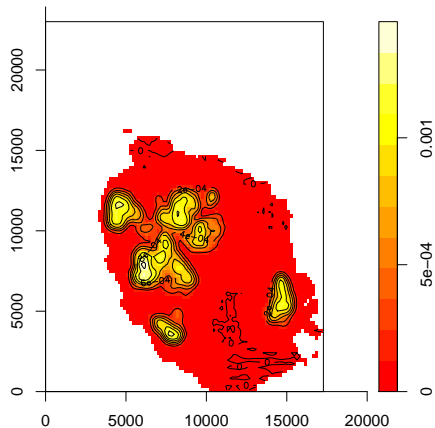
Output Data

Tumor



Number of Tumor cells: 27,822

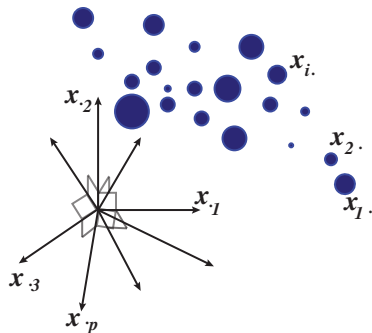
Tumor Cells



We can add information through choice of distances

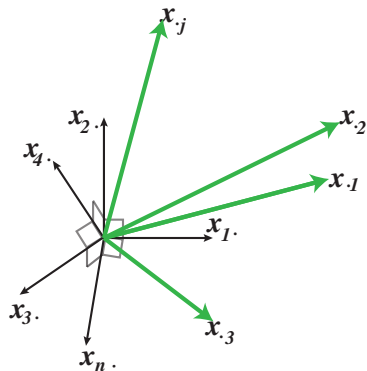
Sample data can often be seen as points in a state space.

\mathbb{R}^p



Variables are 'vectors' in data point space

\mathbb{R}^n



$$x^t Q y = \langle x, y \rangle_Q$$

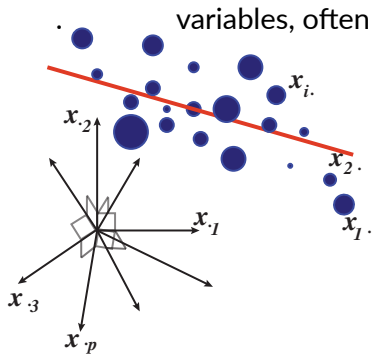
Duality : Transposable data.

$$x^t D y = \langle x, y \rangle_D$$

Data Analysis: Geometrical Approach

- i. The data are p variables measured on n observations.
- ii. X with n rows (the observations) and p columns (the variables).
- iii. D is an $n \times n$ matrix of weights on the “observations”, which is most often diagonal but not always.
- iv Symmetric definite positive matrix Q , weights on

$$Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \dots \\ 0 & 0 & \ddots & 0 & \dots \\ \vdots & \dots & \dots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$$



Euclidean Space and dimension reduction

These three matrices form the essential “triplet” (\mathbf{X} , \mathbf{Q} , \mathbf{D}) defining a multivariate data analysis.

Q and D define geometries or inner products in \mathbb{R}^p and \mathbb{R}^n , respectively, through

$$\begin{aligned}x^t Q y &= \langle x, y \rangle_Q & x, y &\in \mathbb{R}^p \\x^t D y &= \langle x, y \rangle_D & x, y &\in \mathbb{R}^n.\end{aligned}$$

This can be extended to more inner products giving what is known as **Kernel** methods.

Principal Component Analysis: Dimension Reduction

PCA seeks to replace the original (centered) matrix X by a matrix of lower rank, this can be solved using the singular value decomposition of X :

$$X = USV', \text{ with } U'DU = I_n \text{ and } V'QV = I_p \text{ and } S \text{ diagonal}$$

$$XX' = US^2U', \text{ with } U'DU = I_n \text{ and } S^2 = \Lambda$$

PCA is a linear nonparametric multivariate method for dimension reduction. D and Q are the relevant metrics on the dual row and column spaces of n samples and p variables.

A Commutative Diagram Approach

Caillez and Pages, 1976. Escoufier, 1977.

Statisticians search for approximations with certain properties, for the case of PCA for instance, we rephrase the problem as follows:

- ▶ Q can be seen as a linear function from \mathbb{R}^p to $\mathbb{R}^{p^*} = \mathcal{L}(\mathbb{R}^p)$, the space of scalar linear functions on \mathbb{R}^p .
- ▶ D can be seen as a linear function from \mathbb{R}^n to $\mathbb{R}^{n^*} = \mathcal{L}(\mathbb{R}^n)$.
- ▶

$$\begin{array}{ccccc} \mathbb{R}^{p^*} & \xrightarrow{\quad X \quad} & \mathbb{R}^n & & \\ & & & & \\ V = X^t D X & Q \uparrow & \downarrow V & D \downarrow & \uparrow W & W = X Q X^t \\ & & \mathbb{R}^p & \xleftarrow{\quad X^t \quad} & \mathbb{R}^{n^*} & \end{array}$$

This duality gives 'transposable' data.

Properties of the Diagram

Rank of the diagram:

X, X^t, VQ and WD all have the same rank.

For Q and D symmetric matrices, VQ and WD are diagonalisable and have the same eigenvalues.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_r \geq 0 \geq \dots \geq 0.$$

Eigendecomposition of the diagram: VQ is Q symmetric, thus we can find Z such that

$$VQZ = Z\Lambda, Z^t QZ = \mathcal{I}_p, \text{ where } \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p). \quad (1)$$

Modern extensions to this approach include Kernel methods in Machine Learning.

Comparing Two Diagrams: the RV coefficient

Many problems can be rephrased in terms of comparison of two “duality diagrams” or put more simply, two characterizing operators, built from two “triplets”, usually with one of the triplets being a response or having constraints imposed on it. Most often what is done is to compare two such diagrams, and try to get one to match the other in some optimal way. ($O = WD$)

To compare two symmetric operators, there is either a vector covariance as inner product

$covV(O_1, O_2) = Tr(O_1^t O_2) = \langle O_1, O_2 \rangle$ or a vector correlation (Escoufier, 1977)

$$RV(O_1, O_2) = \frac{Tr(O_1^t O_2)}{\sqrt{Tr(O_1^t O_1) tr(O_2^t O_2)}}.$$

If we were to compare the two triplets $(X_{n \times 1}, 1, \frac{1}{n} I_n)$ and $(Y_{n \times 1}, 1, \frac{1}{n} I_n)$ we would have $RV = \rho^2$.

PCA: Approximating one diagram by another

PCA can be seen as finding the matrix Y which maximizes the *RV* coefficient between characterizing operators, that is, between $(X_{n \times p}, Q, D)$ and $(Y_{n \times q}, I, D)$, under the constraint that Y be of rank $q < p$.

$$RV(XQX^tD, YY^tD) = \frac{Tr(XQX^tDYY^tD)}{\sqrt{Tr(XQX^tD)^2 Tr(YY^tD)^2}}.$$

This maximum is attained where Y is chosen as the first q eigenvectors of XQX^tD normed so that $Y^tDY = \Lambda_q$. The maximum RV is

$$RV_{max} = \frac{\sum_{i=1}^q \lambda_i^2}{\sum_{i=1}^p \lambda_i^2}.$$

Of course, classical PCA has $D = \frac{1}{n}\mathcal{I}$, $Q = \mathcal{I}$, but the extra flexibility is often useful. We define the distance between triplets (X, Q, D) and (Z, Q, M) where Z is also $n \times p$, as the distance deduced from the RV inner product between operators XQX^tD and ZMZ^tD .

Discriminant Analysis as a duality diagram

Case of a categorical response variable (group labels).

Let A be the $g \times p$ matrix of group means in each of the p variables. This satisfies

$$Y^t D X = \Delta_Y A \quad \text{where } \Delta_Y = Y^t D Y = \text{diag}(w_1, w_2, \dots, w_g),$$

and $w_k = \sum_{i:y_{ik}=1} d_i$, the w_k 's are the group weights, as they are the sums of the weights as defined by D for all the elements in that group.

Call T the matrix $T = X^t D X$, in the standard case with all diagonal elements of D equal to $\frac{1}{n}$ this is just the standard variance-covariance, otherwise it is a generalization thereof.

The generalized between group variance-covariance is $B = A^t \Delta_Y A$ and call the between group variance covariance the matrix $W = (X - Y A)^t D (X - Y A)$.

A generalized Huyghens' formula:

$$T = B + W$$

Proof: Expanding W gives

$$\begin{aligned} W &= X^t DX - X^t DYA - A^t Y^t DX + A^t Y^t DYA \\ &= T - A' \Delta_Y A - A' \Delta_Y A + A' \Delta_Y A = T - B \end{aligned}$$



Duality Diagram for LDA

The duality diagram for linear discriminant analysis is

$$\begin{array}{ccc} \mathbb{R}^{p^*} & \xrightarrow{\quad A \quad} & \mathbb{R}^g \\ T^{-1} \uparrow & & \downarrow \Delta_Y \\ \mathbb{R}^p & \xleftarrow{\quad A^t \quad} & \mathbb{R}^{g^*} \\ & & \uparrow AT^{-1}A^t \end{array}$$

This corresponds to the triple (A, T^{-1}, Δ_Y) , because

$$(X^t D Y) \Delta_Y^{-1} (Y^t D X) = A^t \Delta_Y A$$

and gives equivalent results to the triple $(Y^t D X, T^{-1}, \Delta_Y^{-1})$.

The discriminating variables are the eigenvectors of the operator

$$A^t \Delta_Y A T^{-1}.$$

Part III

Combine and Compare Trees,
Graphs and Contingent Count Data
for the Human Microbiome

Layers of Data in the Microbiome

Joshua Lederberg: 'the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease'

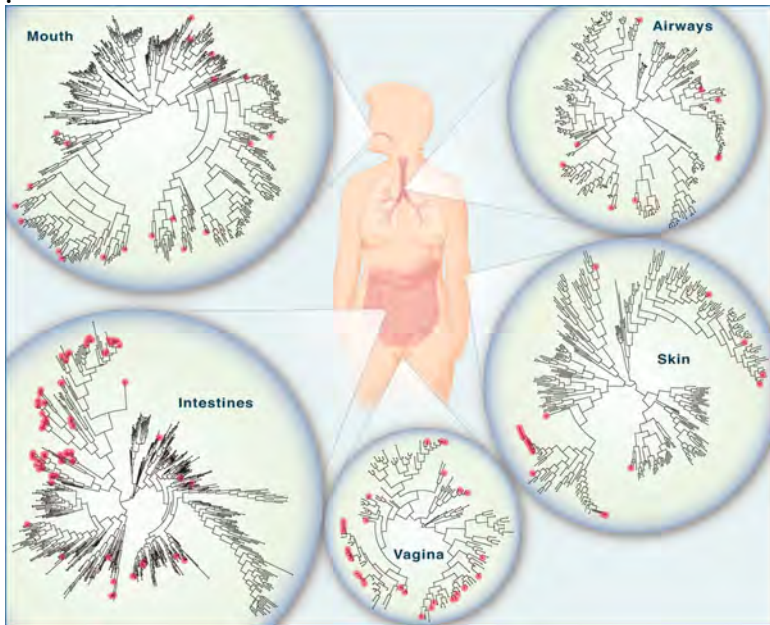
Microbiome Complete collection of genes contained in the genomes of microbes living in a given environment.

Numbers Humans shelter 100 trillion microbes (10^{14}), (we are made of 10×10^{12} cells).

Metagenome Composition of all genes present in an environment (soil, gut, seawater), regardless of species.

Transcriptome These are the mRNA transcripts in the cell, it reflects the genes that are being actively expressed at any given time.

Metabolome The metabolites (small molecules) nucleic or fatty acids, sugars,... present in the sample either endogenous or exogenous (medication, pollution).



Source: YK Lee and SK Mazmanian Science, 2010.

Bacteria etc... and Us

The human microbiome or human microbiota is the assemblage of microorganisms that reside on the surface and in deep layers of skin, in the saliva and oral mucosa, in the conjunctiva, and in the gastrointestinal tracts.

- ▶ They include bacteria, fungi, and archaea.
- ▶ Some of these organisms perform tasks that are useful for the human host. (live in symbiosis)
- ▶ Majority have no known beneficial or harmful effect.

Human Microbiome: What are the data?

DNA The Genomic material present (16sRNA-gene especially, but also shotgun).

RNA What genes are being turned on (gene expression), transcriptomics.

Mass Spec Specific signatures of chemical compounds present (LC/MS, GC/MS).

Clinical Multivariate information about patients' clinical status, medication, weight.

Environmental Location, nutrition, drugs, chemicals, temperature, time.

Domain Knowledge Metabolic networks, phylogenetic trees, gene ontologies.

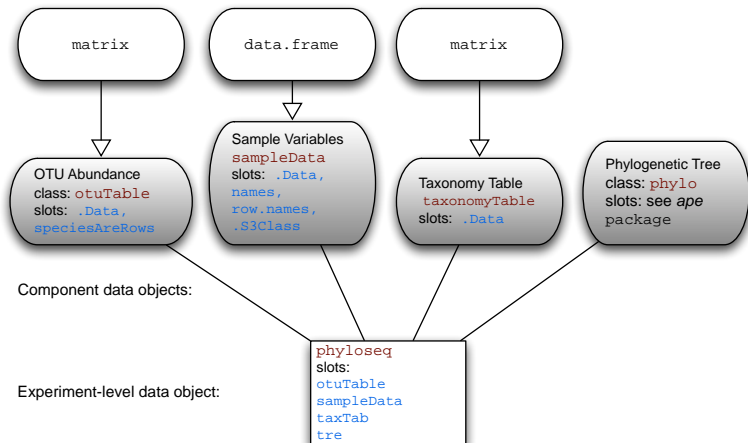
An example of taxa/specimen table.

ASV/OTU	Ctrl1	Ctrl2	Ctrl3	Ctrl4	Ctrl5	IBD1	IBD2	IBD3
Bacteroides	1822	913	147	2988	4616	172	3516	612
Bifidobacterium	0	162	0	0	84	0	85	19
Collinsella	1359	0	0	206	0	327	0	0
Enterococcus	621	0	0	3	40	0	0	0
Streptococcus	75	139	2161	110	97	1820	85	5

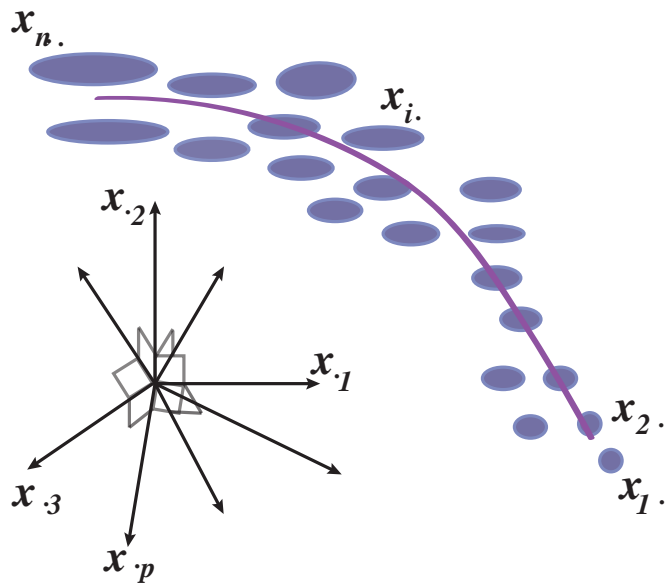
Heterogeneous Data Objects

Object oriented input and data manipulation with phyloseq
(McMurdie and Holmes, 2013, Plos ONE)

Object oriented data in R:

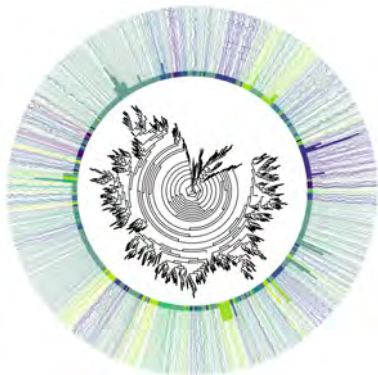


Points are measured with unequal variance



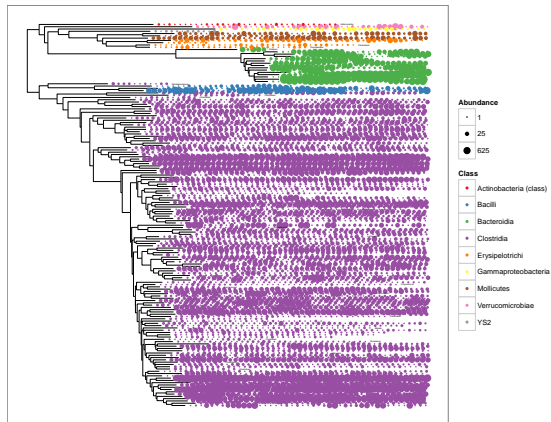
Part IV

Combining a phylogenetic tree with
the count data



A distance on the known tree

Monge-Kantorovich earth mover's distance on the tree.
Used to compare two samples or body sites for instance.
Incorporate taxa abundances and phylogenetic tree



Duality diagram methods that can use any dependency structure.

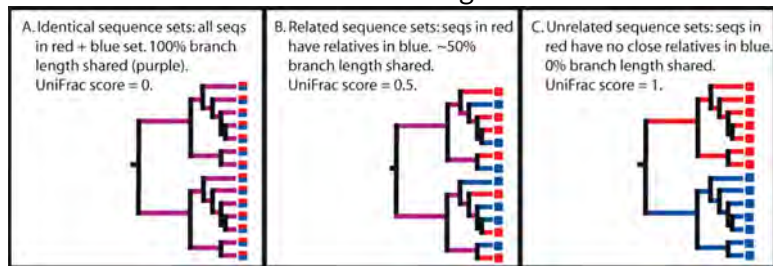
UniFrac Distance (Lozupone and Knight, 2005)

is a distance between groups of organisms that are related to each other by a tree.

Suppose we have the OTUs present in sample 1 (blue) and in sample 2 (red).

Question: Do the two samples differ phylogenetically?

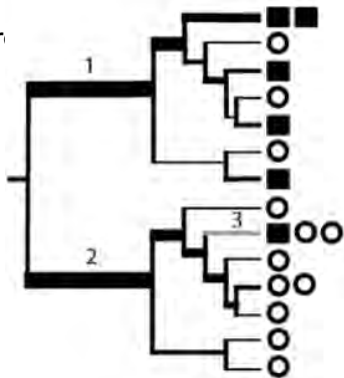
It is defined as the ratio of the sum of the lengths of the branches leading to members of group A or members of group B but not both to the total branch length of the tree.



Weighted UniFrac distance A modification of UniFrac, weighted UniFrac is defined in (Lozupone et al., 2007) as

$$\sum_{i=1}^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- ▶ n = number of branches in the tree
- ▶ b_i = length of the i th branch
- ▶ A_i = number of descendants of i th branch in group A
- ▶ A_T = total number of sequences in group A

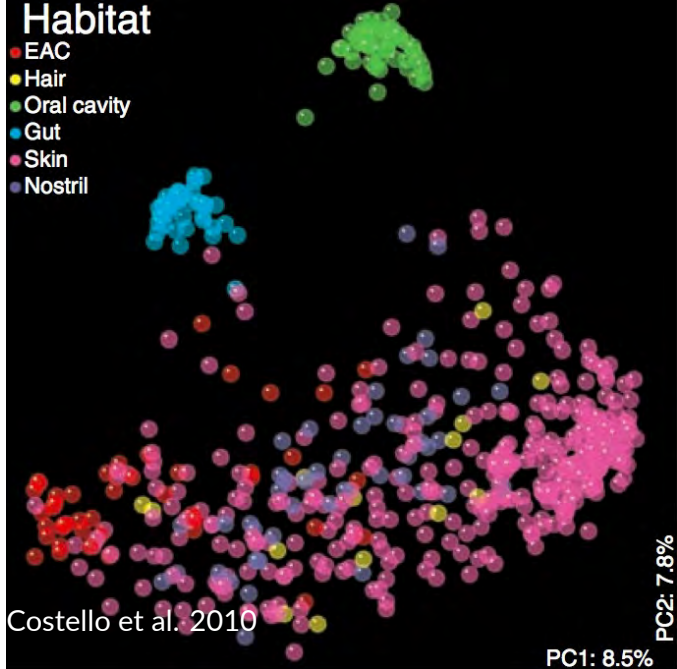


[6].

[7].

Habitat

- EAC
- Hair
- Oral cavity
- Gut
- Skin
- Nostril



Rao's Distance

We start with a distance between individuals.

The heterogeneity of a population (H_i) is the average distance between members of that population.

The heterogeneity between two populations (H_{ij}) is the average distance between a member of population i and a member of population j .

The distance between two populations is

$$D_{ij} = H_{ij} - \frac{1}{2}(H_i + H_j)$$

Decomposition of Diversity

If we have populations $1, \dots, k$ with frequencies π_1, \dots, π_k , then the diversity of all the populations together is

$$H_0 = \sum_{i=1}^k \pi_i H_i + \sum_i \sum_j \pi_i \pi_j D_{ij} = H(w) + D(b)$$

Double Principal Coordinate Analysis

Pavoine, Dufour and Chessel (2004), Purdom (2010) and Fukuyama et al. (2011). .

Suppose we have n species in p locations and a (euclidean) matrix Δ giving the squares of the pairwise distances between the species. Then we can

- ▶ Use the distances between species to find an embedding in $n - 1$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in Δ .
- ▶ Place each of the p locations at the barycenter of its species profile. The euclidean distances between the locations will be the same as the square root of the Rao dissimilarity between them.
- ▶ Use PCA to find a lower-dimensional representation of the locations.

Give the species and communities coordinates such that the inertia decomposes the same way the diversity does.

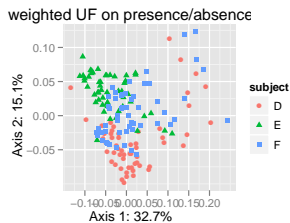
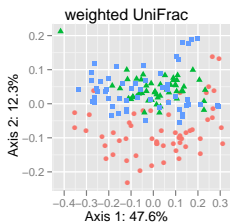
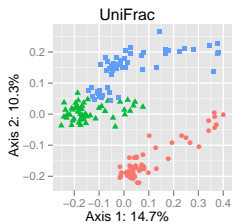
Fukuyama and Holmes, PSB, 2012.

Method	Original description	New formula	Properties
DPCoA	square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i (A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
wUniFrac	$\sum_i b_i A_i/A_T - B_i/B_T $	$\sum_i b_i A_i/A_T - B_i/B_T $	Less sensitive to outliers/more sensitive to noise than DPCoA
UniFrac	fraction of branches leading to exactly one group	$\sum_i b_i \mathbf{1}\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

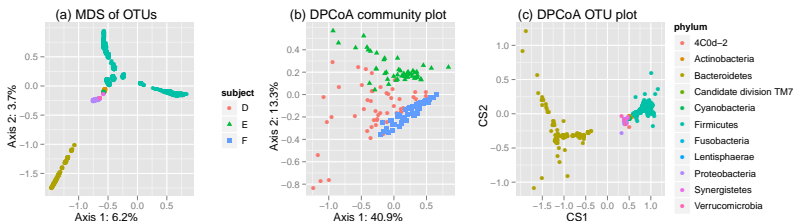
Summary of the methods under consideration. "Outliers" refers to highly abundant taxa, and noise refers to noise in detecting low-abundance taxa.

Antibiotic Time Course Data

Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F)
Each patient sampled ~ 50 times during the course of treatment with ciprofloxacin (an antibiotic).
Times categorized as Pre Cp, 1st Cp, 1st WPC (week post cipro), Interim, 2nd Cp, 2nd WPC, and Post Cp.



Comparing the UniFrac variants. From left to right: PCoA/MDS with unweighted UniFrac, with weighted UniFrac, and with weighted UniFrac performed on presence/absence data extracted from the abundance data used in the other two plots

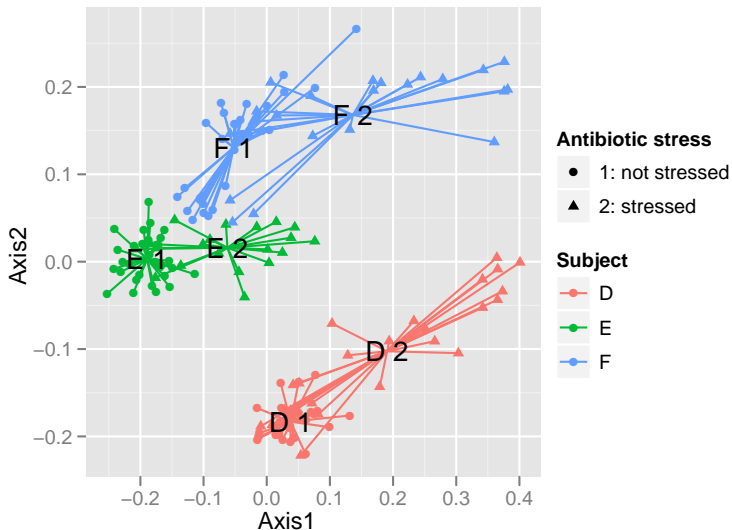


(a)

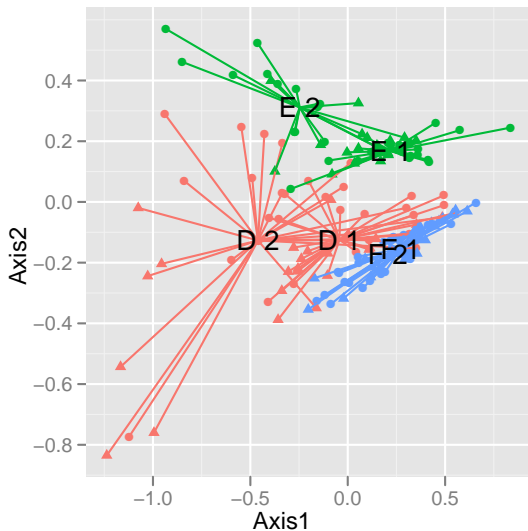
PCoA/MDS of the OTUs based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

Antibiotic Stress

We next want to visualize the effect of the antibiotic. Ordinations of the communities due to DPCoA and UniFrac with information about the whether the community was stressed or not stressed (pre cipro, interim, and post cipro were considered “not stressed”, while first cipro, first week post cipro, second cipro, and second week post cipro were considered “stressed”). We see that for UniFrac, the first axis seems to separate the stressed communities from the not stressed communities. DPCoA also seems to separate the out the stressed communities along the first axis (in the direction associated with *Bacteroidetes*), although only for subjects D and E.



PCoA/MDS with unweighted UniFrac. The labels represent subject plus antibiotic condition.



Community points as represented by DPCoA. The labels represent subject plus antibiotic condition.

Conclusions for Antibiotic Stress

Since UniFrac emphasizes shallow differences on the tree and since PCoA/MDS with UniFrac seems to separate the subjects from each other better than the other two methods, we can conclude that the differences between subjects are mainly shallow ones.

However, DPCoA also separates the subjects and the stressed versus non-stressed communities, and examining the community and OTU ordinations can tell us about the differences in the compositions of these communities.

Modulating the tree-based distances

We would like the axes to be both smooth on the tree and for which the projections of the samples have a large variance.

We can design an inner product on the rows which will pull out axes with these properties.

One extreme will be PCA without a tree, the other is DPCoA.

We create a family of gPCAs interpolating between DPCoA and standard PCA or as giving us a tunable parameter controlling how smooth we want the principal axes to be.

Adaptive gPCA



Fukuyama, Julia (2019), Ann. of Appl. Statistics.

We want to incorporate the prior (tree-like) information about the structure of the variables.

The intuition is that the variables which are similar to each other should behave in similar ways (in the case of microbiome data the idea is that species close together on the tree will behave similarly).

Perform generalized PCA on the posterior estimate of each sample given the data, taking into account the variance structure of the posterior.

Varying the scalings of the prior and noise variances gives a one-dimensional family of generalized PCAs which favor progressively smoother solutions according to the structure of the variables.

Data

Suppose we have a positive definite similarity matrix $Q \in \mathbb{R}^{p \times p}$ (a kernel matrix) between the variables.

To prevent scaling issues, assume that $\text{tr}(Q) = p$.

Note that since Q is positive definite, it is also a covariance matrix, and a random vector with covariance Q will have stronger positive correlations between variables which are more similar to each other.

Special case of the phylogenetic tree

Q is the matrix where Q_{ij} represents the amount of shared ancestral branch length between species i and j .

This is the kernel implicit in DPCoA; it is also related to the covariance of a Brownian motion run along the branches of the tree.

With this in mind, consider the following model for our data matrix X :

$$\mathbf{x}_i \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma_2^2 I), \quad i = 1, \dots, n \quad (2)$$

$$\mu_i \stackrel{\text{iid}}{\sim} N(0, \sigma_1^2 Q), \quad i = 1, \dots, n \quad (3)$$

We are simply including prior knowledge into our model. The prior incorporates information about the structure in our variables: since the μ_i 's have covariance equal to a scalar multiple of Q , inference using this prior will allow us to regularize towards this structure, or to smooth the data towards our expectation that similar variables will behave in similar ways.

PCA on Bayes estimates

We are interested in the “true” values given in μ_i and not the observed data \mathbf{x}_i , and so the appropriate next step is to compute the posterior distribution of the the μ_i 's and then perform PCA on these posteriors. We can compute the posterior distribution $\mu_i | \mathbf{x}_i$ using Bayes' rule, which is

$$\mu_i | \mathbf{x}_i = x \sim N(\sigma_2^{-2} Sx, S) \quad (4)$$

with

$$S = (\sigma_1^{-2} Q^{-1} + \sigma_2^{-2} I)^{-1} \quad (5)$$

Now we want to perform PCA on the posterior estimates of the μ_i 's. We need to take into account the fact that the posterior distributions for each μ_i have non-spherical variance, and so we need to use gPCA instead of standard PCA.

Theorem

The row scores from gPCA on the posterior estimates $\mu_i \mid \mathbf{x}_i$ from the model are the same, up to a scaling factor, to the row scores from gPCA on (X, S, I_n) . The principal axes from gPCA on the posterior estimates are the same, up to a scaling factor, as the principal axes from gPCA on (X, S, I_n) pre-multiplied by S .

From this theorem, we see that when we perform gPCA on the posterior estimates obtained from the model, different scalings of the prior and the noise variances simply lead to gPCAs with different row inner product matrices.

A family of gPCAs

Now we can explore the family of inner product matrices which our model gives rise to. Up to a scaling factor, the matrix $S = (\sigma_1^{-2}Q^{-1} + \sigma_2^{-2}I)^{-1}$ depends only on the relative sizes of σ_1 and σ_2 , the scalings for the prior and the noise. We therefore have a one-dimensional family of gPCAs determined by the relative sizes of σ_1 and σ_2 . To get some insight into this family, we can first consider the endpoints.

As $\sigma_1/\sigma_2 \rightarrow 0$, that is, as the noise becomes very small compared to the prior structure, S becomes more and more like a scalar multiple of the identity, and so we approach a scalar multiple of gPCA on the triple (X, I, I) , or standard PCA. At the other end, as $\sigma_2/\sigma_1 \rightarrow 0$, we approach a scalar multiple of gPCA on the triple (X, Q, I) . The gPCA on (X, Q, I) turns out to be very closely related to double principal coordinates analysis (DPCoA), which is another method for incorporating information about the variables into the analysis.

Automatic selection of family member

If we do not want to assume σ_1 and σ_2 are known, we can estimate the values σ_1 and σ_2 from the data itself by maximum marginal likelihood. To be more concrete, according to our data model we have

$$\mathbf{x}_i \stackrel{\text{iid}}{\sim} N(0, \sigma_1^2 Q + \sigma_2^2 I) \quad (6)$$

The overall log likelihood of the data is therefore (up to a constant factor)

$$\ell(X; \sigma_1, \sigma_2) = -\frac{n}{2} \log |\sigma_1^2 Q + \sigma_2^2 I| - \sum_{i=1}^n \frac{1}{2} \mathbf{x}_i^T (\sigma_1^2 Q + \sigma_2^2 I)^{-1} \mathbf{x}_i \quad (7)$$

Maximizing this likelihood is not a convex problem: we transform it into a one parameter problem over the unit interval. Let $r = \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$, and let $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Let $Q = V\Lambda V^T$ be the eigendecomposition of Q where V is an orthogonal matrix and Λ is diagonal containing the eigenvalues $\lambda_1, \dots, \lambda_p$. Finally, let $\mathbf{x}_i = V^T \mathbf{x}_i$ and \tilde{x}_{ij} be the j th element of $\tilde{\mathbf{x}}_i$. The log likelihood in the new parameterization is

$$\ell(X; r, \sigma) = -\frac{np}{2}\sigma^2 \log |rQ + (1-r)I| - \sigma^{-2} \sum_{i=1}^n \frac{1}{2} \mathbf{x}_i^T (rQ + (1-r)I) \mathbf{x}_i \quad (8)$$

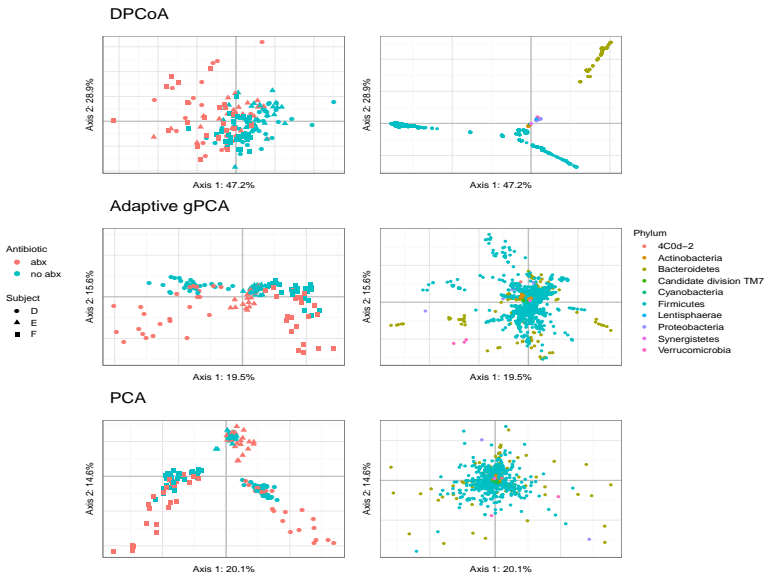
$$= -\frac{np}{2}\sigma^2 \sum_{j=1}^p \log(r\lambda_j + 1 - r) - \sigma^{-2} \sum_{i=1}^n \sum_{j=1}^p \frac{1}{2} \frac{\tilde{x}_{ij}^2}{r\lambda_j + 1 - r} \quad (9)$$

Based on the expression above, we can find a closed-form solution for the maximizing value of σ^2 for any fixed r .

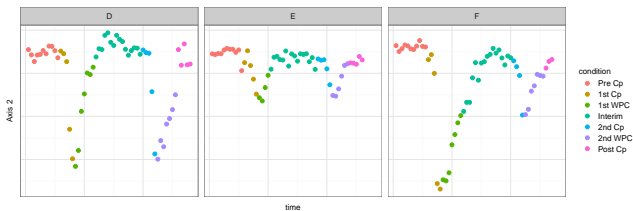
This gives us

$$\sigma^{2*}(r) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \tilde{x}_{ij}^2 / (r\lambda_i + 1 - r) \quad (10)$$

We re-write the likelihood as a function of r only. This is still not convex but only has one parameter which lies on the unit interval, the optimization can be performed numerically.



Sample (left) and species (right) plots for DPCoA (top), adaptive gPCA (middle), and standard PCA (bottom). Colors in the sample plots represent a binning of the sample points into abx (either when the subject was on antibiotics or the week immediately following)



A plot of the scores along the second axis from adaptive gPCA by time, plotted for each of the three individuals. We see very clearly that this axis is capturing species that change during the administration of the antibiotic but which are stable otherwise. The corresponding plots for PCA and DPCoA are much less compelling.

Alternatives

We could add a ridge penalty to Q , resulting in gPCA on $(X, Q + \lambda I, I)$. This family has the same endpoints as the family we have described: when $\lambda = 0$ we have gPCA on (X, Q, I) , and as $\lambda \rightarrow \infty$ we get standard PCA.

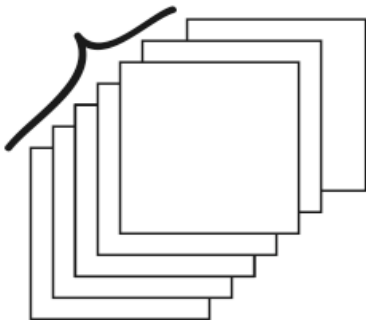
Very roughly, when we add a ridge penalty to Q , the main effect is to increase the small eigenvalues, but when we add a ridge penalty to Q^{-1} we make the large eigenvalues more similar to each other.

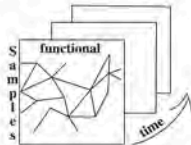
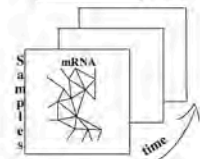
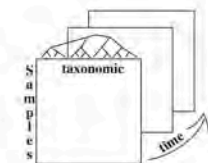
Small eigenvalues of Q correspond to eigenvectors that are very rough, while the large eigenvalues correspond to eigenvectors that are smooth.

When we do structured dimensionality reduction, we want to dampen any variance along rough eigenvectors, but we don't necessarily prefer variance in the direction of an extremely smooth eigenvector over variance in the direction of a mostly-smooth eigenvector. When we use $Q + \lambda I$, we remove the dampening on the rough directions, but when we use $S = (\sigma_1^1 Q^{-1} + \sigma_2^{-2} I)^{-1}$ we keep the eigenvalues of the rough directions small and decrease the difference between eigenvalues of smooth eigenvectors.

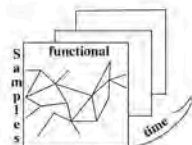
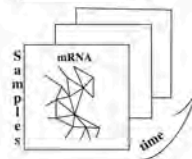
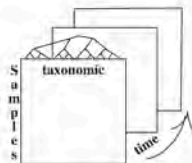
Part V

Multitable methods for
heterogeneous data





Unperturbed



Perturbed

Multi-table methods: use Inertia/Co-Inertia

Generalize variance and covariance \rightarrow moments of inertia.
weighted (p_i) sum of distances.

Abundance data in a contingency table \rightarrow weighted sum of the squares
weighted frequencies (chisquare).

Co-Inertia

When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the *covariance*.

$$\text{sum}(x1 * y1 + x2 * y2 + x3 * y3)$$

if x and y co-vary -in the same direction this will be big.

A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA).

Co-inertia analysis (CIA) is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples or the same time points.

That is the rows or columns of the matrix have to be weighted similarly and thus must be matchable.

RV coefficient

The global measure of similarity of two data tables as opposed to two vectors can be done by a generalization of covariance provided by an inner product between tables that gives the RV coefficient, a number between 0 and 1, like a correlation coefficient, but for tables.

$$RV(A, B) = \frac{Tr(A'B)}{\sqrt{Tr(A'A)}\sqrt{Tr(B'B)}}$$

Survey on RV: Josse, Holmes (2015) Statistics Surveys, [arXiv link](#).

Example

Combining different types of data (antibiotic study).

Taxa Read counts (3 patients taking cipro: two time courses) : .

Mass-Spec Positive and Negative ion Mass Spec features and their intensities: .

RNA-seq Metagenomic data on genes :.

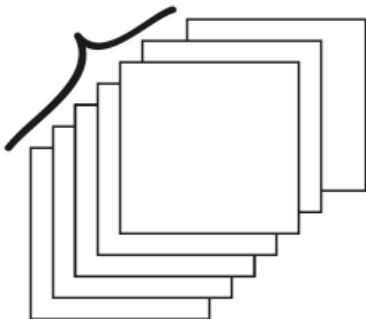
Here is the RV table of the three array types:

```
> fourtable$RV
```

	Taxa	Kegg	MassSpec+	MassSpec-
Taxa	1	0.565	0.561	0.670
Kegg	0.565	1	0.686	0.644
MassSpec+	0.561	0.686	1	0.568
MassSpec-	0.670	0.644	0.568	1

Part VI

Distances between "aligned" graphs



Bacteria 'sharing' between mice

Using the Jaccard index that measures the co-occurrence or co-occurrence of species between mice.

$$\text{Jaccard Similarity} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

$$\text{Jaccard Disimilarity} = \frac{f_{01} + f_{10}}{f_{01} + f_{10} + f_{11}}$$

```
mouse1
```

```
0 0 0 1 0 1 0 1 0 0 0 0 0 0 1
```

```
mouse4
```

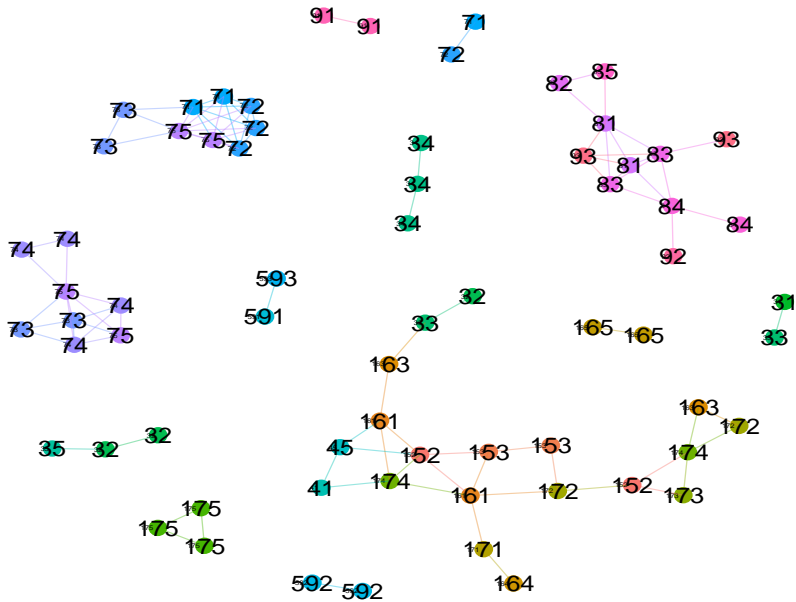
```
1 0 0 0 0 0 0 0 0 0 0 0 0 0 1
```

```
vegdist(rbind(mouse1,mouse4),method="jaccard")
```

```
0.8
```

Bacteria 'sharing' between mice as a network

```
netbaseline=make_network(phy_pifn_glom)
p=plot_network(netbaseline,phy_pifn_glom,
color="mousenames",label="mousenames",point_size=7)
+geom_text(aes(label=mousenames),size=7)
p+scale_colour_hue(guide="none")
```



Does the network relate to 'communities'?

Friedman and Rafsky (1979) devised a nonparametric test for multivariate data using the minimum spanning tree with any metric.

Then compute the number of 'pure' edging connecting labels from the same groups compared to the mixed edges connecting labels from different groups, call F_o the observed statistic.

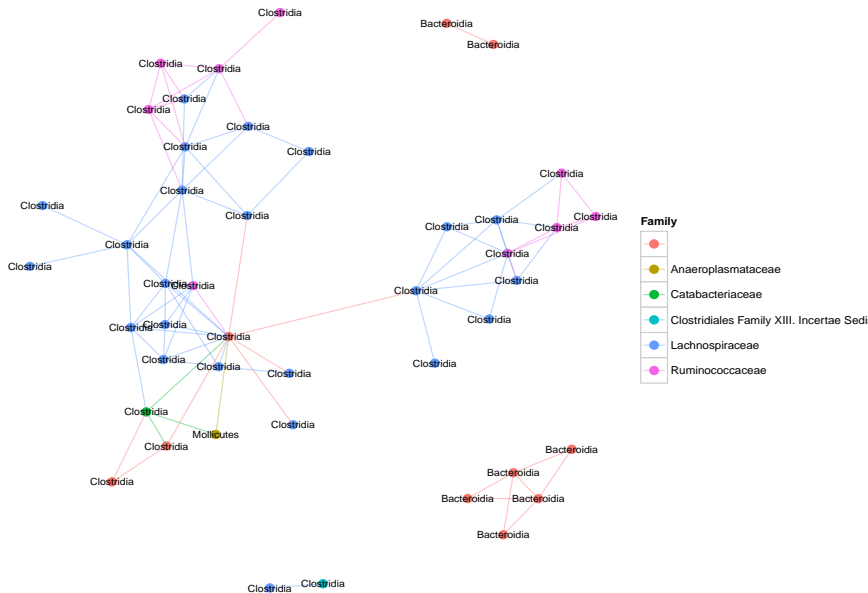
In our example: $F_o = 82$

Keeping the graph fixed, permute the labels and recompute the number of pure edges.

All 1000 simulated values had $F_s < 82$ so $p < 0.001$.

Co-occurrence networks for taxa of the baseline mice

```
p=plot_network(netbasetaxa,phy_pifn_glom,color="Family",  
type="taxa",label=NULL)  
p+geom_text(aes(label=Class),size=3)
```



Changes of the network over time?

OTU Network Plot

Type: Taxa

Mouse 71 (infected)
 Mouse 72 (infected)
 Mouse 51 (uninfected)
 Mouse 52 (uninfected)
 Mouse 62 (uninfected)
 Mouse 64 (uninfected)

Color Attribute: Genus

Distance: jaccard

Time Interval G1: 0 2 27

Time Interval G2: 0 10 27

Decimal: 0 0.5 1

Coloring taxa based on Genus



Genus

- Adierocytia
- Akkermansia
- Anaerostipes
- Bifida
- Butyrivibrio
- Clostridium
- Coprococcus
- Epulopiscium
- Escherichia
- Eubacterium
- Lachnospira
- Lachnobacterium
- Lactobacillus
- Moryella
- Oscillopsira
- Roseburia
- Ruminococcus
- Tributella
- Weissella



Genus

- Ad
- Ac
- Bu
- Cl
- Co
- De
- Ep
- Es
- Eu
- La
- La
- La
- M
- O
- Ru
- Ru
- Tr

Differences between two graphs?

Edges Added

Edges Removed



Family

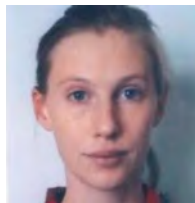
- Catabacteriaceae
- Clostridiaceae
- Clostridiales Family XIII. Incertae Sedis
- Conobacteriaceae
- Dehalobacteriaceae
- Enterobacteriaceae
- Erysipelotrichaceae
- Lactosporiaceae
- Lactobacillaceae
- Ruminococcaceae



Family

- Catabacteriaceae
- Enterobacteriaceae
- Erysipelotrichaceae
- Lactobacillaceae
- Lactosporiaceae
- Leuconostocaceae
- Ruminococcaceae

Distances between (node-identified graphs)



Claire Donnat, SH, Ann. of Applied Stat., 2018.

Example:

Each graph corresponds to a cuisine (French, American, Greek, etc...).

Each of 1,530 ingredients constitutes a node in the graph and each of the 49 cuisines is assigned to a weighted graph.

The weight on the edge is the frequency of co-occurrence of the two ingredients for that particular cuisine. Some graphs includes a collection of disconnected nodes (ingredients that never co-occur in a single recipe) and a weighted connected component.

Graphs with identified vertices

$G = (\mathcal{V}, \mathcal{E})$ the graph with vertices \mathcal{V} and edges \mathcal{E} . $N = |\mathcal{V}|$, $i \sim j$ if nodes i and j are neighbors. A refers to the adjacency matrix of the graph, and D to its degree matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \quad \text{and } D = \text{Diag}(d_i)_{i=1 \dots N} \text{ s.t. } d_i = \sum_{j=1}^N A_{ij}$$

Restricting ourselves to undirected graphs, the matrix A is symmetric: $A^T = A$.

Hamming distance

It measures the number of edge deletions and insertions necessary to transform one graph into another.

$$d_H(G, \tilde{G}) = \sum_{i,j} \frac{|A_{ij} - \tilde{A}_{ij}|}{N(N-1)} = \frac{1}{N(N-1)} \|A - \tilde{A}\|_1 \quad (11)$$

This defines a metric between graphs, since it is a scaled version of the L_1 norm between the adjacency matrices A and \tilde{A} . It defines a distance bounded between 0 and 1 over all graphs of size N .

The Jaccard distance

$$d_{\text{Jaccard}}(G, \tilde{G}) = \frac{|G \cup \tilde{G}| - |G \cap \tilde{G}|}{|G \cup \tilde{G}|} = \frac{\sum_{i,j} |A_{ij} - \tilde{A}_{ij}|}{\sum_{i,j} \max(A_{i,j}, \tilde{A}_{i,j})} = \frac{\|A - \tilde{A}\|_1}{\|A + \tilde{A}\|_*} \quad (12)$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix.

Eq. 12 is known to define a proper distance between the graphs. A straightforward way to see this is to use the Steinhaus Transform: for (X, d) a metric and c a fixed point, the transformation $\delta(x, y) = \frac{2d(x,y)}{d(x,c)+d(y,c)+d(x,y)}$ produces a metric. Apply this transformation, with d the Hamming distance and c the empty graph, to see:

$$\begin{aligned} \delta(G, \tilde{G}) &= \frac{2\|A - \tilde{A}\|_1}{\|A\|_1 + \|\tilde{A}\|_1 + \|A - \tilde{A}\|_1} = \frac{2(|G \cup \tilde{G}| - |G \cap \tilde{G}|)}{2|G \cup \tilde{G}|} \quad (*) \\ &= d_{\text{Jaccard}}(G, \tilde{G}). \end{aligned}$$

The recipes graphs

Each cuisine-graph has nodes that represent ingredients; edges are co-occurrence frequencies.

Cuisines can be better characterized by typical associations of ingredients.

For instance, the Japanese cuisine might be characterized by a higher associativity of ingredients such as “rice” and “nori” than Greek cuisine.

We use the co-occurrence counts of 1,530 different ingredients for 49 different cuisines (Chinese, American, French, etc.) Each cuisine is then characterized by its own co-occurrence graph.

The weight on the edge is the frequency of co-occurrence of the two ingredients in a given cuisine. The final graph for a given cuisine thus consists in a collection of disconnected nodes (ingredients that never appear in a single recipe for that cuisine) and a weighted connected component.

Ingredient comparisons (heat-wavelet based distances)		
Cuisine	Neighbor	top changes (char. distance)
Middle Eastern	Indian	mustard, dill, bread, thyme, oregano, feta cheese, walnut sesame seed, coconut, olive
	Moroccan	chive, nut, red wine, feta cheese, cane molasses, yogurt, rose, oregano, fennel, walnut
	Spanish	apricot, lentil, mint, zucchini walnut, pork sausage, feta cheese, sesame seed, lamb, yogurt
Chinese	Asian	black bean, oyster, turmeric, cumin, lime juice, nira, coconut, basil, beef broth, lime
	Japanese	lemon, oyster, salmon, buckwheat enokidake, tuna, radish, barley, kelp, katsuobushi
	Thai	peanut butter, mint, roasted peanut, fenugreek, turmeric, lime juice, cumin, coconut, basil, lime

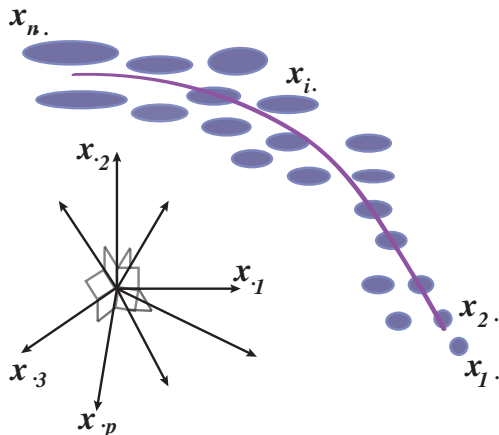
Identification of the ingredients that change the most from one graph to another

Distances enable statisticians to....

- ▶ Summarize data with medians, means and principal directions.
- ▶ Encode some variations in uncertainty.
- ▶ Make comparisons of heterogeneous sources of information.
- ▶ Integrate network and tree information.
- ▶ Measure diversity, inertia and generalize the notion of variance.

Questions for mathematicians

- ▶ How to build distances between images that account for unequal measurement errors, even locally?



Work by Adler, Taylor and Worsley (2003,2005,2007) using Random Fields.

Questions for mathematicians






- ▶ How well can the Euclidean embedding approximations do compared to the inherent noise?
- ▶ Are there better ways of approximating the commutative diagrams?

This is also an important point of contact with the use of Stein's method in probability theory.

Questions for mathematicians

- ▶ How to distinguish between the effect of the curvature of a state space and the effect of the unequal sampling?

References

-  L. Billera, S. Holmes, and K. Vogtmann.
The geometry of tree space.
Adv. Appl. Maths, 771–801, 2001.
-  J. Chakerian and S. Holmes.
distory:Distances between trees, 2010.
-  Daniel Chessel, Anne Dufour, and Jean Thioulouse.
The ade4 package - i: One-table methods.
R News, 4(1):5–10, 2004.
-  P. Diaconis, S. Goel, and S. Holmes.
Horseshoes in multidimensional scaling and kernel methods.
Annals of Applied Statistics, 2007.
-  Y. Escoufier.
Operators related to a data matrix.
In J.R. et al. Barra, editor, *Recent developments in Statistics.*,
pages 125–131. North Holland,, 1977.



Steven N Evans and Frederick A Matsen.

The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples.

arXiv, q-bio.PE, Jan 2010.



M Hamady, C Lozupone, and R Knight.

Fast unifracs: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data.

The ISME Journal, Jan 2009.



Susan Holmes.

Multivariate analysis: The French way.

In D. Nolan and T. P. Speed, editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 56 of *IMS Lecture Notes–Monograph Series*. IMS, Beachwood, OH, 2006.



Ross Ihaka and Robert Gentleman.

R: A language for data analysis and graphics.

Journal of Computational and Graphical Statistics,
5(3):299–314, 1996.



K. Mardia, J. Kent, and J. Bibby.
Multivariate Analysis.
Academic Press, NY., 1979.



P. J. McMurdie and S. Holmes.
Phyloseq: Reproducible research platform for bacterial
census data.
PlosONE, 2013.
April 22,.



Serban Nacu, Rebecca Critchley-Thorne, Peter Lee, and
Susan Holmes.
Gene expression network analysis and applications to
immunology.
Bioinformatics, 23(7):850–8, Apr 2007.



Sandrine Pavoine, Anne-Béatrice Dufour, and Daniel
Chessel.

From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis.

Journal of Theoretical Biology, 228(4):523–537, 2004.



Elizabeth Purdom.

Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree.

Annals of Applied Statistics, Jul 2010.



C. R. Rao.

The use and interpretation of principal component analysis in applied research.

Sankhya A, 26:329–359., 1964.

Part VIII

Dimension Reduction: the
Euclidean embedding workhorse:
MDS

Metric Multidimensional Scaling

Schoenberg (1935)

ANNALS OF MATHEMATICS
Vol. 36, No. 3, July, 1935

REMARKS TO MAURICE FRÉCHET'S ARTICLE "SUR LA DÉFINITION AXIOMATIQUE D'UNE CLASSE D'ESPACE DISTANCIÉS VECTORIELLEMENT APPLICABLE SUR L'ESPACE DE HILBERT"

BY I. J. SCHOENBERG

(Received April 16, 1935)

1. Fréchet's developments in the last section of his article suggest an elegant solution of the following problem.

Let

$$a_{ik} = a_{ki} \quad (i \neq k; i, k = 0, 1, \dots, n)$$

be $\frac{1}{2}n(n+1)$ given positive quantities. What are the necessary and sufficient conditions that they be the lengths of the edges of a n -simplex $A_0A_1 \dots A_n$? More general, what are the conditions that they be the lengths of the edges of a n -"simplex" $A_0A_1 \dots A_n$ lying in a euclidean space R_r ($1 \leq r \leq n$) but not in a R_{r-1} ?

This problem is fundamental in K. Menger's metric investigation of euclidean spaces ([6] and [7], particularly his third fundamental theorem in [7], pp. 737-743). It was solved by Menger by means of equations and inequalities involving certain determinants. Theorem 1 below furnishes a complete and independent solution of this problem. Theorem 2 solves the similar problem for spherical spaces previously treated by Menger's methods by L. M. Blumenthal and G. A. Garrett ([1]) and Laura Klanfer ([5]); it may be conveniently applied (Theorems 3 and 3') to prove and extend a theorem of K. Gödel ([4]). The method of Theorem 1 is finally applied to solve the corresponding problem for spaces with indefinite line element recently considered by A. Wald ([8]) and H. S. M. Coxeter and J. A. Todd ([2]).

From Coordinates to Distances and Back

If we started with original data in \mathbb{R}^p that are not centered: Y , apply the centering matrix

$$X = HY, \quad \text{with } H = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right), \text{ and } \mathbf{1}' = (1, 1, 1, \dots, 1)$$

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

$$-\frac{1}{2}HD^{(2)}H = B$$

Schoenberg's result: exact Euclidean distance If B is positive semi-definite then D can be seen as a distance between points in a Euclidean space.

Reverse engineering an Euclidean embedding

We can go backwards from a matrix D to X by taking the eigendecomposition of $B = -\frac{1}{2}HD^{(2)}H$ 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' .

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

Multidimensional Scaling (MDS)

Simple classical multidimensional scaling.

- ▶ Square D elementwise $D^{(2)} = D_2$.
- ▶ Compute $\frac{-1}{2}HD_2H = B$.
- ▶ Diagonalize B to find the principal coordinates SV' .
- ▶ Choose a number of dimensions by inspecting the eigenvalue's screeplot.

The advantage is that the original distances don't have to be Euclidean.

Taking Categorical Data and Making it into a Continuum

Horseshoe Example: Joint with Persi Diaconis and Sharad Goel (Annals of Applied Stats, 2005). Data from 2005 U.S. House of Representatives roll call votes. We further restricted our analysis to the 401 Representatives that voted on at least 90% of the roll calls (220 Republicans, 180 Democrats and 1 Independent) leading to a 401×669 matrix of voting data.

The Data

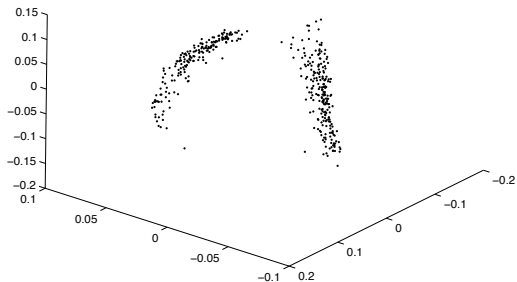
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
R1	-1	-1	1	-1	0	1	1	1	1	1	...
R2	-1	-1	1	-1	0	1	1	1	1	1	...
R3	1	1	-1	1	-1	1	1	-1	-1	-1	...
R4	1	1	-1	1	-1	1	1	-1	-1	-1	...
R5	1	1	-1	1	-1	1	1	-1	-1	-1	...
R6	-1	-1	1	-1	0	1	1	1	1	1	...
R7	-1	-1	1	-1	-1	1	1	1	1	1	...
R8	-1	-1	1	-1	0	1	1	1	1	1	...
R9	1	1	-1	1	-1	1	1	-1	-1	-1	...
R10	-1	-1	1	-1	0	1	1	0	0	0	...

L_1 distance

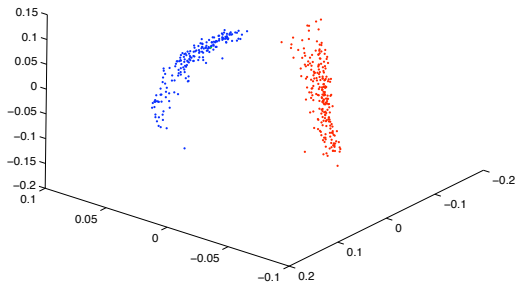
We define a distance between legislators as

$$\hat{d}(l_i, l_j) = \frac{1}{669} \sum_{k=1}^{669} |v_{ik} - v_{jk}|.$$

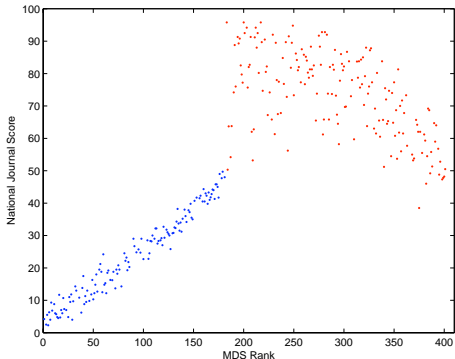
Roughly, $\hat{d}(l_i, l_j)$ is the percentage of roll calls on which legislators l_i and l_j disagreed.



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. We used dissimilarity indices $1 - \exp(-\lambda d(R_1, R_2))$



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.



Comparison of the MDS derived rank for Representatives with the National Journal's liberal score

Uncertainty propagation with heterogenous data.

Susan Holmes

<http://www-stat.stanford.edu/~susan/>

Bio-X and Statistics, Stanford University

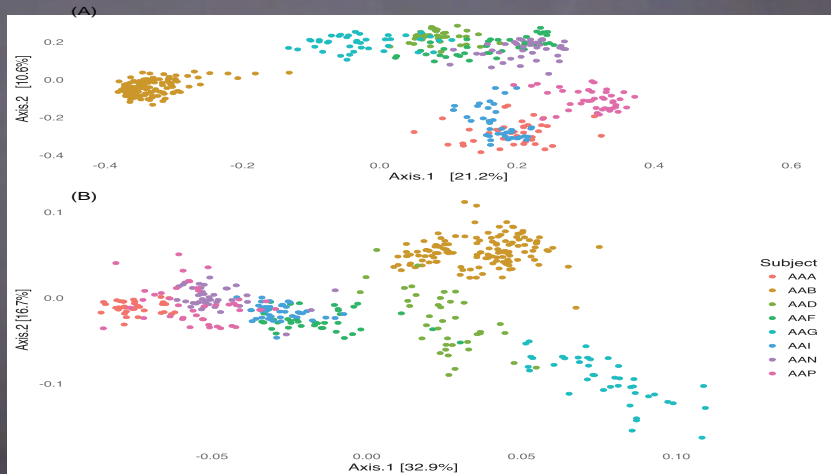
Toulouse, September 2, 2019

Part I

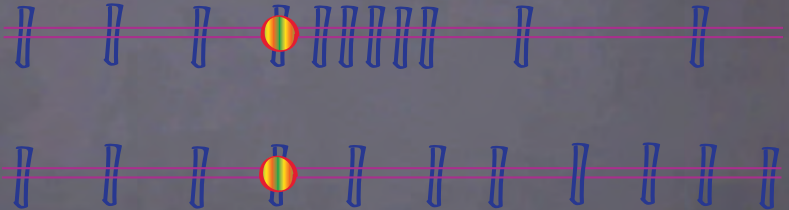
Experimental Design and data

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0 D		791	0	79	108	11
1 D		1616	0	1413	192	31
2 D		1323	0	915	165	23
3 D		1846	0	1366	170	31
4 D		2314	0	689	135	26
5 D		2244	0	776	310	175
6 D		1652	0	609	235	181

Subject to Subject variation is largest source of variation



Not equally distant time points.



Between point variation should be equal.

See Peter Diggle's text : Analysis of Longitudinal Data, 2002.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p}_{\text{tt}} = (p_1, p_2, \dots, p_J) \quad \text{For } J \text{ ASV's}$$

$$\mathbf{p}_{\text{ctl}} = (p_1, p_2, \dots, p_J) \quad \Delta = \text{diff}(\mathbf{p}_{\text{tt}} - \mathbf{p}_{\text{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p}_{\text{tt}} = (p_1, p_2, \dots, p_J) \quad \text{For } J \text{ ASV's}$$

$$\mathbf{p}_{\text{ctl}} = (p_1, p_2, \dots, p_J) \quad \Delta = \text{diff}(\mathbf{p}_{\text{tt}} - \mathbf{p}_{\text{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors. We need to quantify the uncertainty we have on the parameters.

Example in microbiome: unknown parameters?

The relative abundances of bacteria and their differences.

Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017)

$$\mathbf{p}_{\text{tt}} = (p_1, p_2, \dots, p_J) \quad \text{For } J \text{ ASV's}$$

$$\mathbf{p}_{\text{ctl}} = (p_1, p_2, \dots, p_J) \quad \Delta = \text{diff}(\mathbf{p}_{\text{tt}} - \mathbf{p}_{\text{ctl}})$$

We estimate these by accounting for different sequencing depths and provide estimates of the standard errors.

Models for noise: hierarchical Gamma-Poisson: we know how to transform the data to stabilize the variance (Delta-method).

McMurdie and Holmes (2014) "Waste Not, Want Not: Why

Read data are counts, the data are not compositional.

We do not summarize them to ratios or “relative abundance”.

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (decontam).
- Estimating depth bias requires read numbers.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Transform the data to equalize the variance.

Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



hidden variables.

Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.

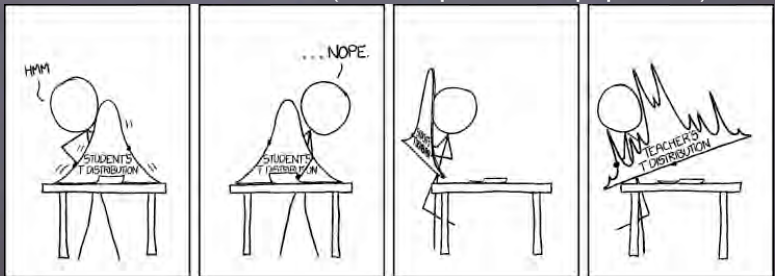


hidden variables.



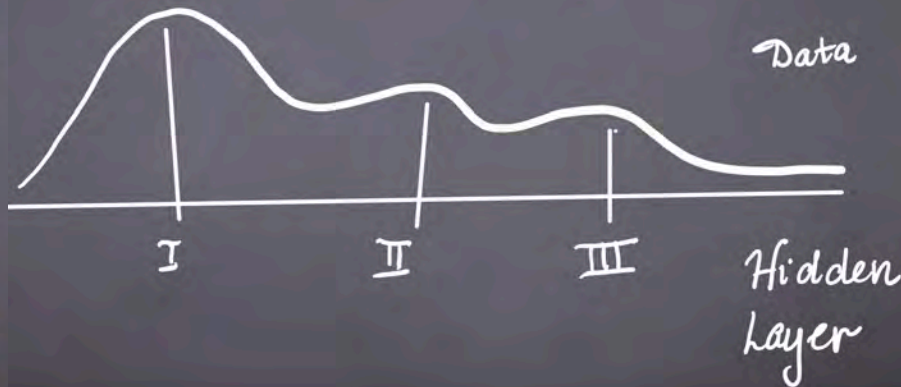
Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).



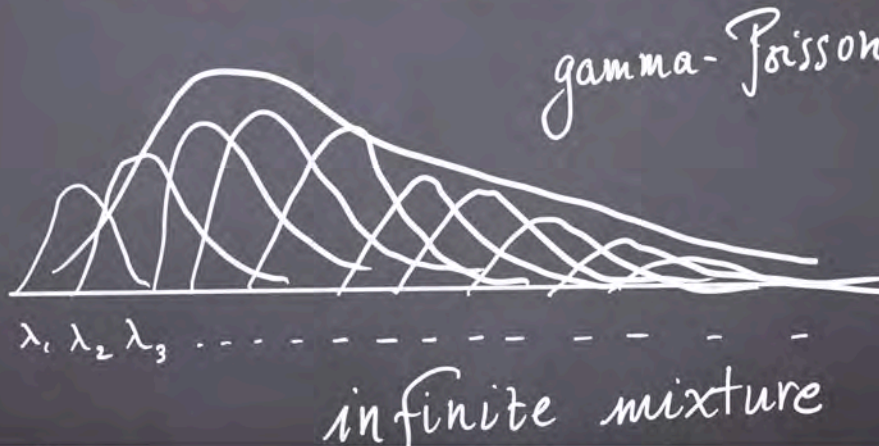
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation



Part II

*Models for Microbial Communities
over time.*

Pregnancy data: perturbation, stability and preterm birth

A case-control study of 49 pregnant women:

- 15 delivered preterm.
- From 40 of these women: 3,766 specimens collected weekly during gestation, and monthly after delivery.
- Sites:vagina, distal gut, saliva, and tooth/gum.
- 9 women: validation set collected after the first study was complete.

Methods used: variance stabilization through negative binomial, testing perturbations through linear mixed-effects modeling. Preterm prediction through medoid-based clustering and simple Markov chain. Provided: Simple community temporal trends, community structure, and vaginal community state transitions.

Attention to detail

- Careful noise models (dada2) and variance stabilization (DESeq2, vsn, voom).
- Random effects, mixed models.
- Finite State Markov chains.
- Differential abundance testing provides biomarkers for preterm birth.

DiGiulio DB, Callahan BJ, McMurdie PJ, ... & Holmes, SP and Relman, DA

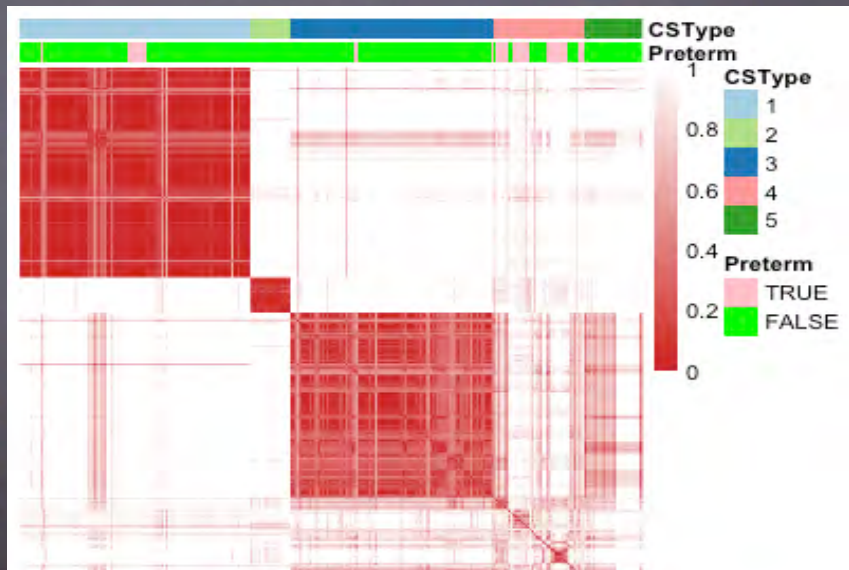
Temporal and spatial variation of the human microbiota during pregnancy. PNAS, 2015, 112(35):11060-5.

Co-occurrence networks

Dual networks:

- Edges are created between taxa if in more than a certain proportion of samples share that taxa.
This can be seen as a geometric graph with the distance being the Jaccard distance.
- Edges are created between samples if they share more than a certain proportion of taxa in common.

Communities of bacteria organize into 5 different types

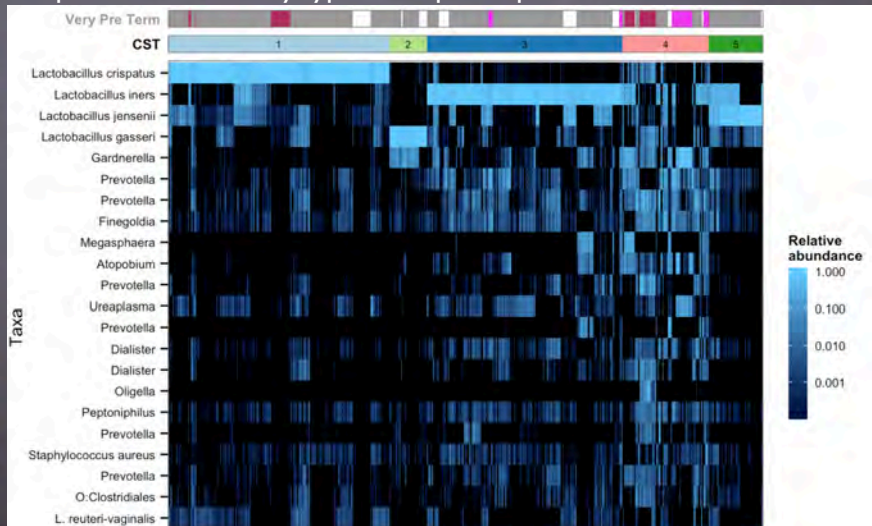


Questions asked?

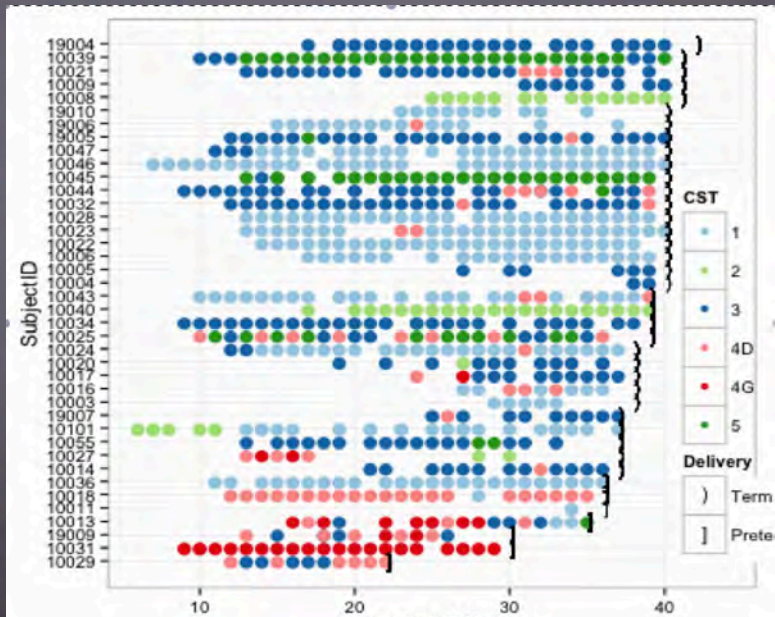
- Are the community state types the same as seen in previous studies?
- How stable are the communities within each individual during pregnancy?
- What alterations of the vaginal microbiome predict preterm birth?
- How early do these alterations occur?

Previously known Microbial Community State Types: Latent categorical variable.

Samples into community types and species patterns associated.

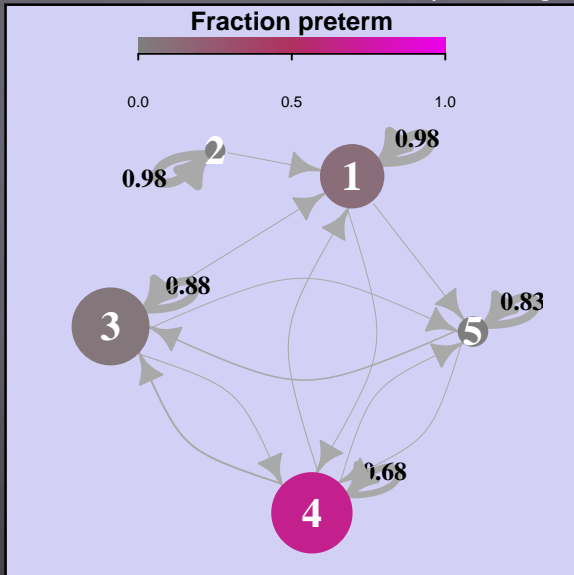


Longitudinal Analyses



Markov Chain Model

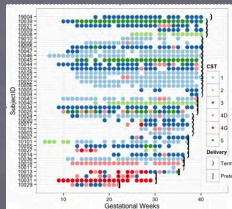
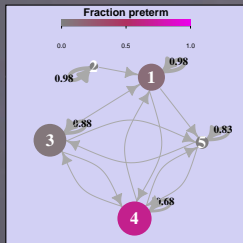
Transitions between states, as in simple ecological models.



Conclusions for this study

- Microbiota community and diversity stable during pregnancy.
- Prevalence of a Lactobacillus-poor vaginal community state type (CST 4) was inversely correlated with gestational age at delivery ($p=0.0039$).
Risk for preterm birth was more pronounced for subjects with CST 4 accompanied by elevated Gardnerella or Ureaplasma abundances.
- Finding validated with a separate diagnostic set of 246 vaginal specimens from nine women (four of whom delivered preterm).

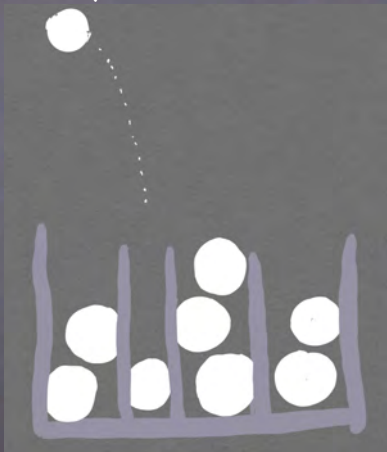
Illustration through Analyses



- Delivery Perturbation
- Preterm Prediction
- Stability

Part III

The Dirichlet for the multinomial



General Ideas about the multinomial

- Balls in boxes, not necessarily the same size.
- The number of balls is the number of reads, the boxes are the ASVs.
- Multinomial model gives the probability of seeing say (4,2,3,1) if the probabilities of the four boxes are $p_1 = 0.3, p_2 = 0.2, p_3 = 0.4, p_4 = 0.1$ this number is:

```
> dmultinom(c(4,2,3,1),prob=c(0.3,0.2,0.4,0.1))  
[1] 0.02612736
```
- Apart from the fact that if a lot of balls fall in the first box there will be less balls for the other boxes, the boxes' contents are independent: that is BAD.

Dirichlet

Make the p 's vary randomly.

Hierarchical Model:

$ps \sim \text{Dirichlet}(\alpha, \alpha, \alpha, \alpha)$

Uniform on the simplex (four cornered pyramid).

```
x <- round(gtools::rdirichlet(5, c(1,1,1,1) ),2)
```

```
> x
```

	[,1]	[,2]	[,3]	[,4]
[1,]	0.06	0.50	0.08	0.36
[2,]	0.20	0.57	0.18	0.05
[3,]	0.07	0.20	0.55	0.18
[4,]	0.57	0.04	0.00	0.39
[5,]	0.02	0.16	0.27	0.55

Multinomial needs to be modified

Multivariate dependencies in bacterial communities

Data depart from a multinomial distribution within each row:

- Some taxa are quasi-exclusive (*Lactobacillus crispatus* and *Gardnerella*).
- Co-occurrence through syntrophy, in which a molecular hydrogen-consuming species (typically a methanogen, like *Methanobrevibacter smithii* in the human gut) enhances the growth of a molecular hydrogen-producing species (any of a number of secondary fermenters in the gut).
- In the mouth (subgingival crevice), where in cases of moderate to severe periodontitis, a methanogen (*Methanobrevibacter oralis*) is always found with a syntrophic partner.
- There are not a finite number of taxa a priori, taxa evolve, some are sample-specific.

Part IV

Interpretability: Latent
variables and topic analysis

Discrete/disconnected Community state types are rare

Each sample is assigned to only one type of community.

Need a more nuanced model: mixtures.



Mixture models

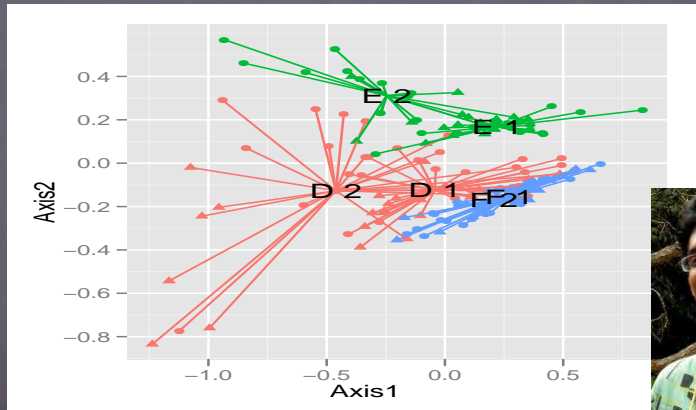
- In clustering and hidden discrete categorical variables, every sample belonged to a community state type.
- In a topic mixture model, every sample can be composed of several topics.

Most useful parallel: natural language processing.

Generative model

- Pick topics at random among a certain number of topics.
- Each topic corresponds to a probability distribution for many words.
- Pick a word at random according to the chosen topic.

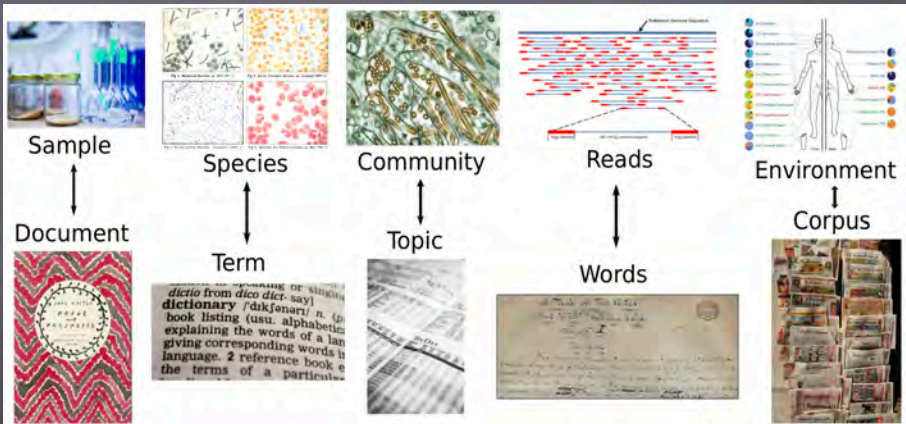
How to understand the the taxa involved in the perturbation?



Kris Sankaran

Biostatistics, 2018,
Latent Variable Modeling for the Microbiome.
[Kris Sankaran's Topic Page](#)

Parallel between topic and community analyses



Credit: Kris Sankaran

Parallel between topic and community analyses

index	book	elizabeth	darcy	bennet	miss	jane	bingley	time
0	P & P	0	0	4	0	1	3	0
1	P & P	1	0	5	0	1	4	0
2	P & P	0	0	6	0	0	5	1
3	P & P	1	4	5	1	0	9	1
4	P & P	3	3	5	4	4	5	3
5	P & P	3	0	0	2	1	6	1
6	P & P	0	6	6	7	1	5	1

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0	D	791	0	79	108	11
1	D	1616	0	1413	192	31
2	D	1323	0	915	165	23
3	D	1846	0	1366	170	31
4	D	2314	0	689	135	26
5	D	2244	0	776	310	175
6	D	1652	0	609	235	181

Statistical Model

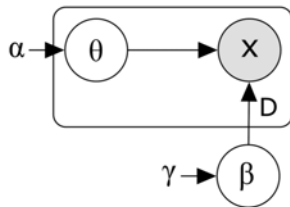
Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling.

It assumes samples have mixed memberships across topics.
(See Pritchard et. al 2000, Blei et. al. 2003)

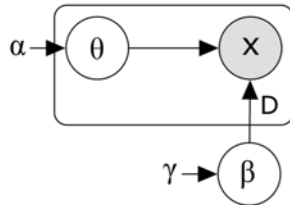
Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling.

Observed microbiomes \sim mixtures of underlying community types.

Statistical Model



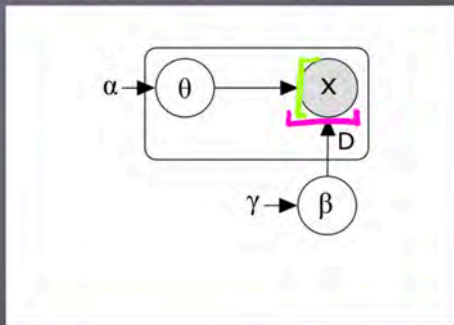
Statistical Model



Statistical Model

samples
layer
observa. $\left\{ \begin{array}{l} \text{sample 1} \dots \\ \vdots \\ \text{sample n} \dots \end{array} \right.$ rows (X)

↑
hidden layer for
taxa



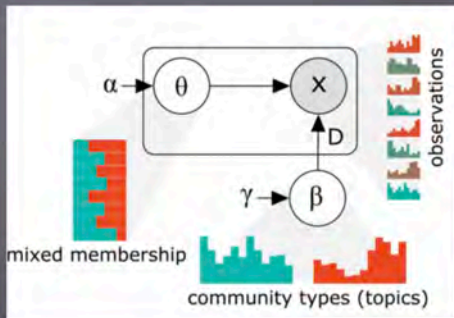
D Documents
K communities or topics

Statistical Model

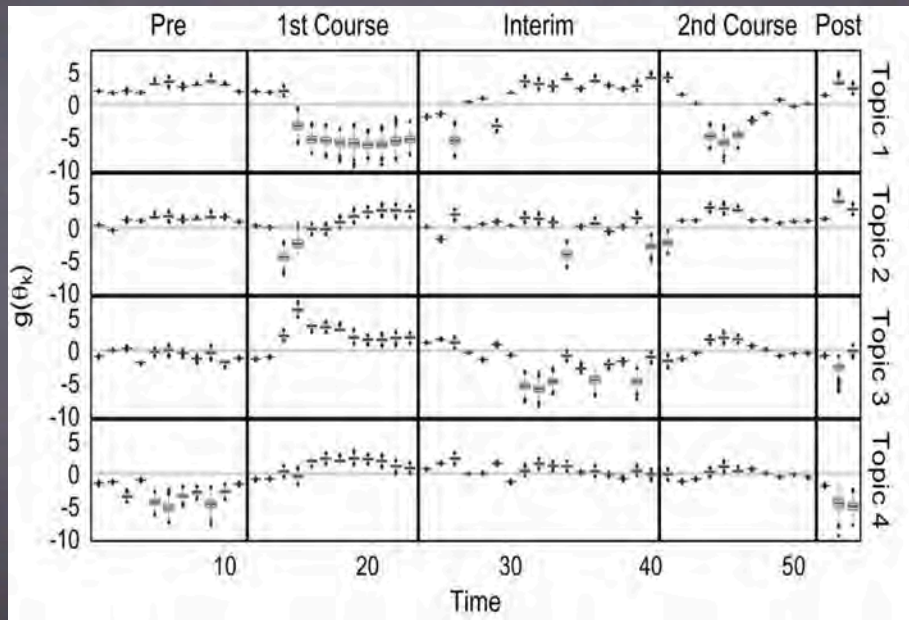
$$x_d \mid \beta \sim \text{Mult}(N_d, B\theta_d)$$

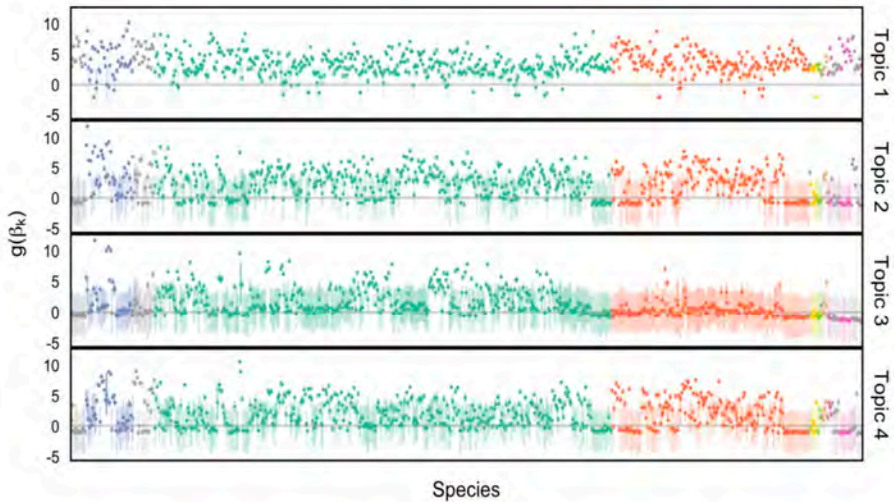
$$\theta_d \sim \text{Dir}(\alpha)$$

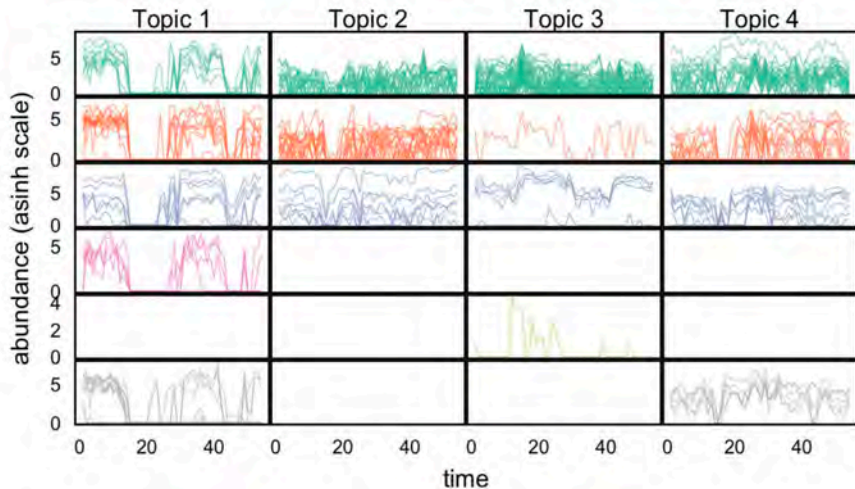
$d=1 \dots D$



$$\beta_k \sim \text{Dir}(\gamma), k=1 \dots K$$







Family

■ Lachnospiraceae	■ Bacteroidaceae	■ Streptococcaceae
■ Ruminococcaceae	■ uncultured	■ other

Part V

Distances cannot provide
all the information



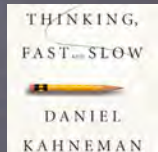
THE
UNDOING
PROJECT

A Friendship that Changed Our Minds

**But distances are not
everything...remember the baseline**



Amos Tversky and Danny Kahneman



Heuristics and Biases, more particularly the representativeness heuristic.

Heuristics are described as "judgmental shortcuts that generally get us where we need to go - and quickly - but at the cost of occasionally sending us off course."

Heuristics are useful because they use effort-reduction and simplification in decision-making.

For representativeness of an event, similarity or a small distance is not enough, the baseline frequencies (ie probability) are essential to conclude.

We need careful realistic probability models for treespace, no real data has ever been uniform, no multivariate data is ever multivariate normal.

Diversities in the microbiome depend on the number of taxa

- α -diversity: Number of 'species'-taxa in a biological sample (from one location).
- β -diversity: Differentiation in diversity among different samples from different locations.

Extremely sensitive to noise.

Fake species:

Microbial diversity in the deep sea and the underexplored "rare biosphere"

Mitchell L. Sogin^{*†}, Hilary G. Morrison^{*}, Julie A. Huber^{*}, David Mark Welch^{*}, Susan M. Huse^{*}, Phillip R. Neal^{*}, Jesus M. Arrieta^{†5}, and Gerhard J. Herndl[‡]

^{*}Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and [†]Royal Netherlands Institute Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

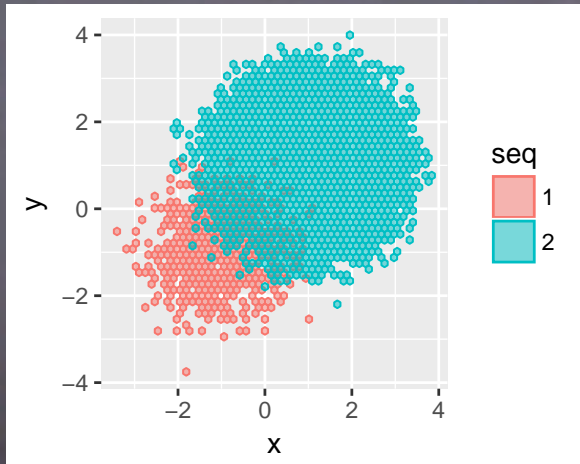
Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

The evolution of marine microbes over billions of years predicts Gene sequences, most commonly those encoding

How many words does Professor D. know?

- Maybe 15,000, 20,000?
- Start sampling..... banana, bannana, bannanna, orange, orange, muscle, musel, muscel, foreign, forene, forane,.....
- How many real words does Prof D. know?
- Use more information than the spelling...

From reads to Operational Taxonomic Units

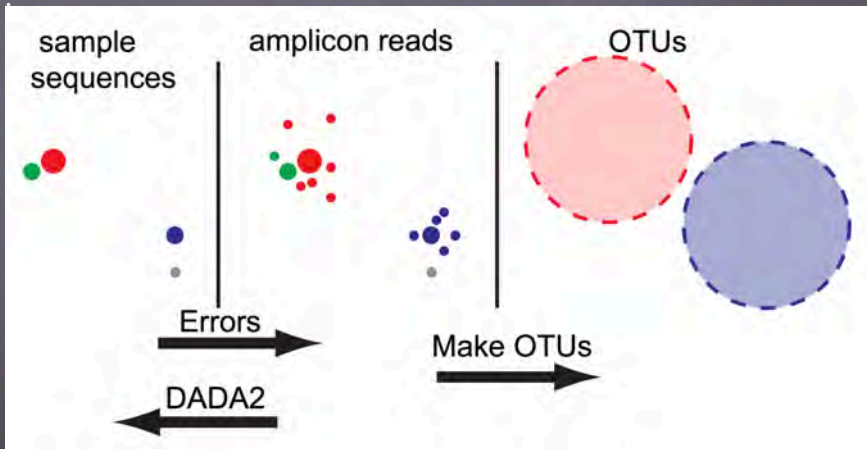


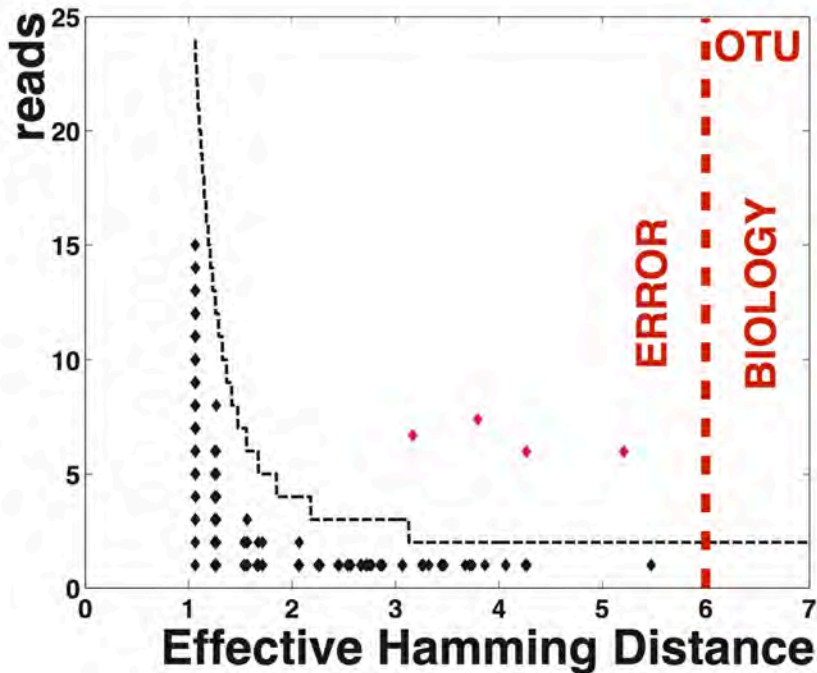
From reads to Operational Taxonomic Units

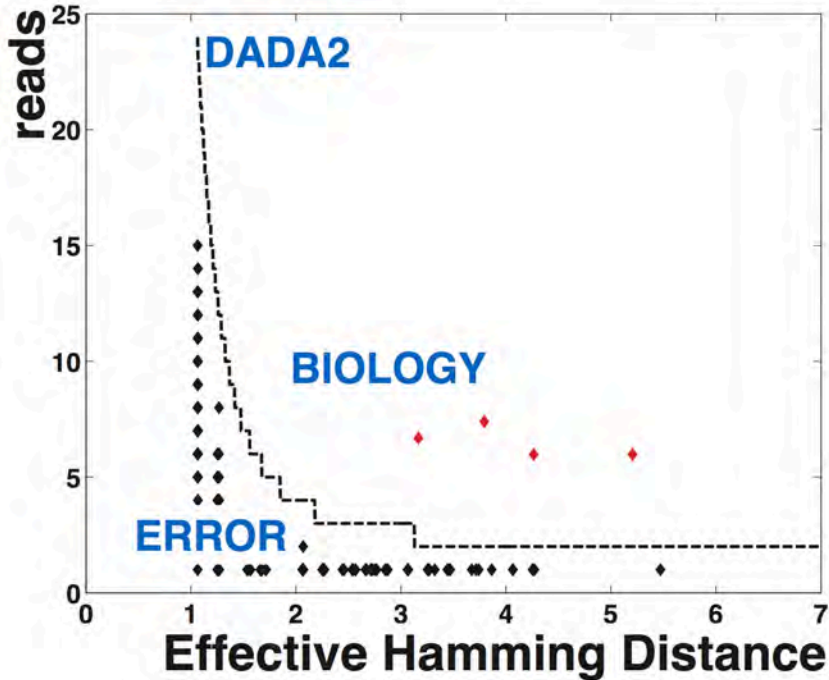


Current practice (qiime, mothur, rdp, ...): 97% similarity

Probabilistic Model

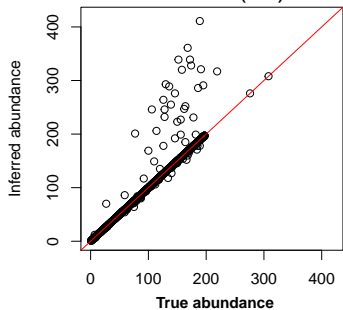






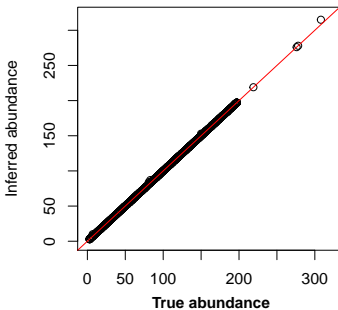
Accuracy: Simulated data

mothur (an)



TP: 978
FP: 272
FN: 77
cor: 0.935

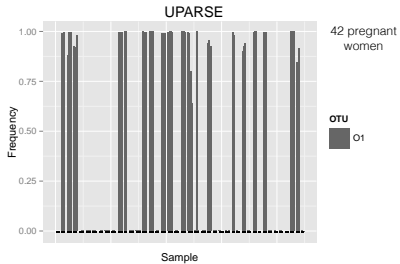
DADA2



TP: 1042
FP: 0
FN: 13
cor: 0.999

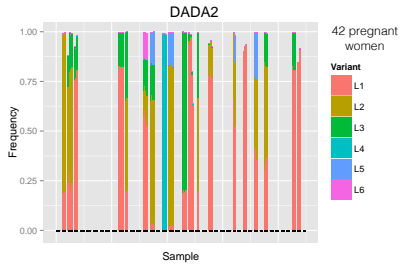
Data: Kopylova, et al. mSystems, 2016.

Resolution: *L. crispatus*



Data: MacIntyre et al. Scientific Reports, 2015.

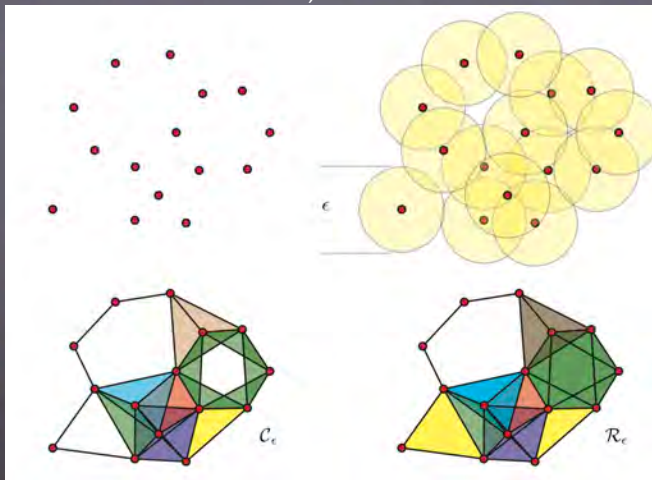
Resolution: *L. crispatus*



Data: MacIntyre et al. Scientific Reports, 2015.

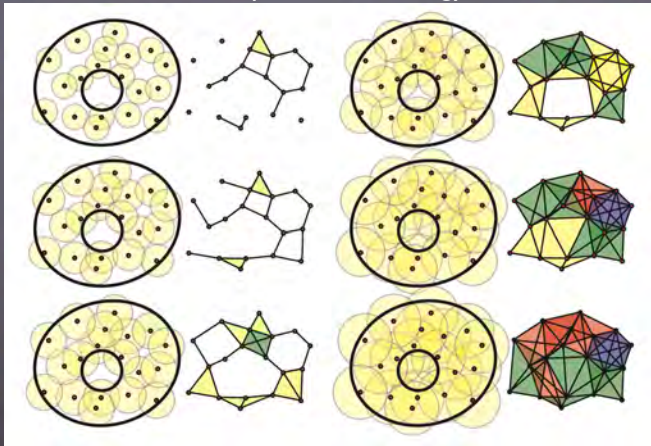
Mathematicians also have baseline measure problems

Examples in Topological Data Analysis (Edelsbrunner, Carlsson, Zomorodian, Ghrist et al.).



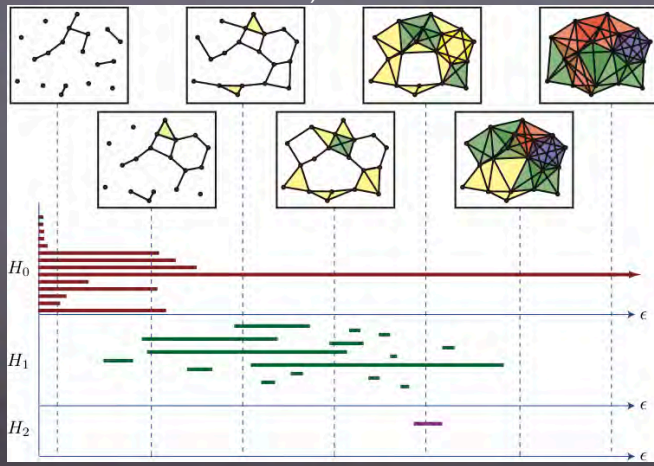
Mathematicians also have baseline measure problems

Ghrist, R. Barcodes: persistent toology of data, AMS, 2008



Mathematicians also have baseline measure problems

Examples in Topological Data Analysis (Edelsbrunner, Carlsson, Zomorodian, Ghrist et al.).

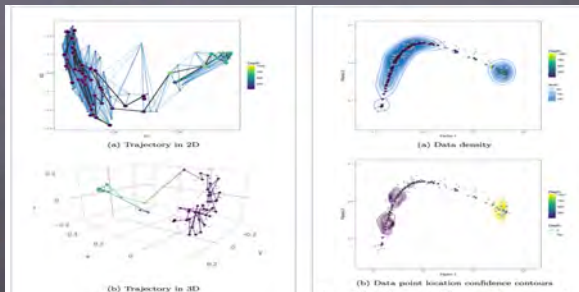


(Ghrist)

Open Question: How to make a method designed for uniformly

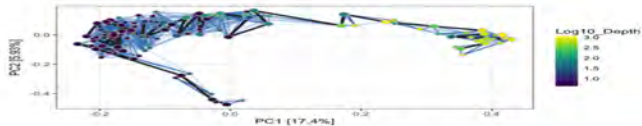
Part VI

Uncertainty quantification for Latent gradients

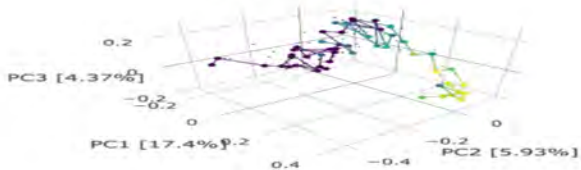


Uncertainty Quantification for rankings and gradients

Bayesian Unidimensional Scaling (Lan Huong Nguyen and Susan Holmes, 2017, BMC Bioinformatics).



(a) Trajectory in 2D



(b) Trajectory in 3D

Bayesian model for distances

$$d_{ij} | \delta_{ij} \sim \text{Gamma}[\mu_{ij} = \delta_{ij}, \sigma_{ij}^2 = s_{ij}^2 \sigma_\epsilon^2], \quad (1)$$

$$\delta_{ij} = |\tau_i - \tau_j|,$$

$$\tau_i | \alpha_\tau, \beta_\tau \sim \text{Beta}(\alpha_\tau, \beta_\tau),$$

$$\alpha_\tau \sim \text{Cauchy}^+(1, \gamma_\tau),$$

$$\beta_\tau \sim \text{Cauchy}^+(1, \gamma_\tau),$$

$$\sigma_\epsilon \sim \text{Cauchy}^+(0, \gamma_\epsilon),$$

Modeling the heteroscedasticity

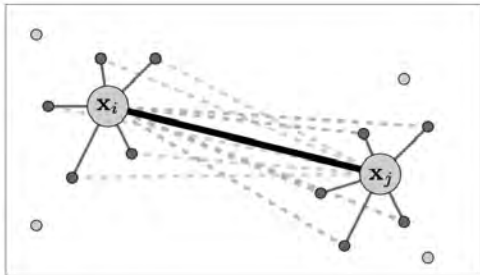
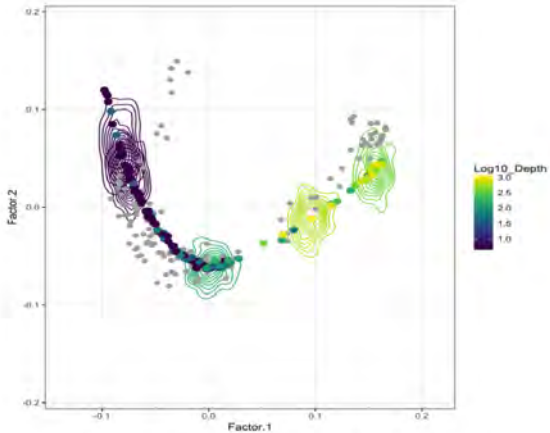


Figure 1 Graphical representation of points x_i and x_j together with their neighbors. The set of (dashed) distances from x_i to the K -nearest-neighbors of x_j , and from x_i to the K -nearest-neighbors of x_i is used to compute $\hat{s}^2(d_{ij})$, the estimate of the variance of d_{ij} . Here we chose $K = 5$.

$$s(\hat{d}_{ij}) = \frac{1}{|D_{ij}^K|} \sum_{d \in D_{ij}^K} (d - \bar{d}_{ij}^K)^2$$

Scale parameter for the error term: $s_{ij}^2 = s(\hat{d}_{ij})/s(\bar{d}_{ij})$.



(b) Datapoint location confidence contours

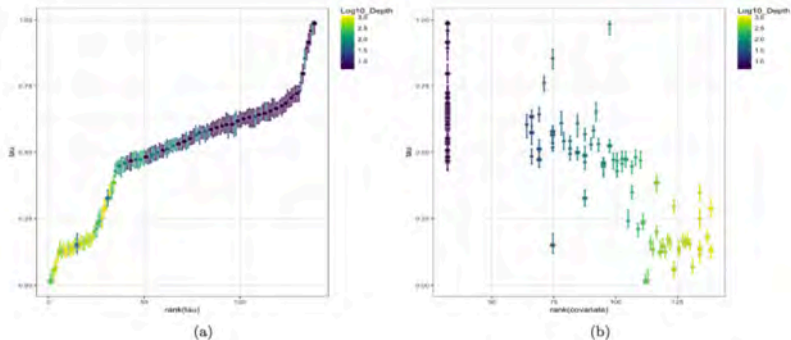


Fig. 4

Latent ordering in TARA Oceans dataset shown with uncertainties. The differences in the slope of plot (a) indicate varying data coverage along the underlying gradient. Correlation between the water depth and the latent ordering in microbial composition data is shown in (b). Coloring corresponds to log10 of the water depth (in meters) at which the ocean sample was collected

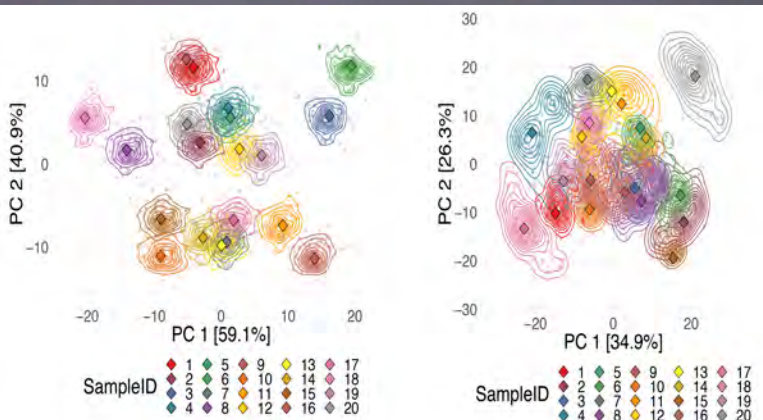
Code using stan

```
fit_buds <- function(D, K = NULL,
                    method = c("vb", "mcmc"),
                    hyperparams = list(
                      "gamma_tau" = 2.5,
                      "gamma_epsilon" = 2.5,
                      "gamma_bias" = 2.5,
                      "gamma_rho" = 2.5,
                      "min_sigma" = 0.03),
                    init_from = c("random", "principal_cu
                    seed = 1234, max_trials = 20, ...) {
```

buds package on github.

Part VII

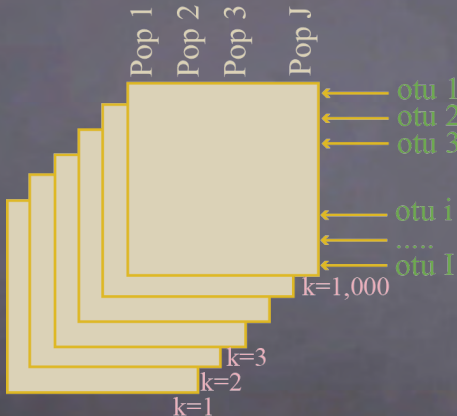
Uncertainty quantification for Latent factors



Full Bayesian nonparametric model

- We do not know the number of OTUs.
- We suppose underlying low dimensional latent variables for the sample P^j 's.
- We use dependent microbial distributions, marginal priors of discrete distributions are built using manipulation of a Gaussian process and then extending this to multiple correlated distributions.

Generalization: Bayesian posterior uncertainty measures



Parameters for samples $\mathbf{Y}^j, j \in \mathcal{J} = \{1, \dots, J\}$

Define a joint prior on these factors through the Gram matrix $(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$

The parameters \mathbf{Y}^j can be interpreted as key characteristics of the biological samples that affect the relative abundance of ASVs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j},$$

$\epsilon_{i,j}$ iid Normal

Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren, Bacallado, Favaro, Holmes, Trippa (2017, JASA).

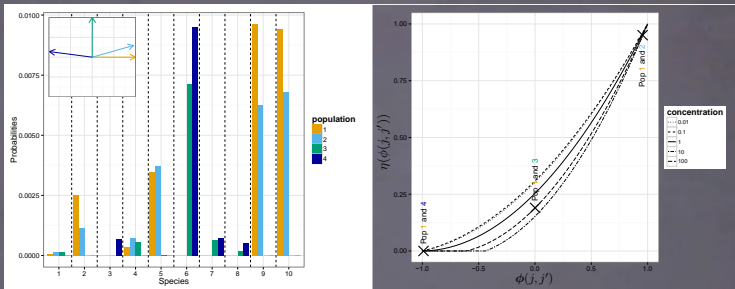


Figure: **Left panel:** realization of 4 microbial distributions from a dependent Dirichlet processes with 10 OTUs **Right panel:** correlation of two random probability measures when the cosine $\phi(j, j')$ between \mathbf{Y}^j and $\mathbf{Y}^{j'}$ varies from -1 to 1 . (Ren et al, JASA, 2017).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \dots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

The parameters \mathbf{Y}^j can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs.

$$Q_{i,j} = \langle \mathbf{x}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \quad (1)$$

where the $\epsilon_{i,j}$ are independent Normal variables.

The degree of similarity between the discrete distributions $\{P^j; j \in \mathcal{J}\}$ is summarized by the Gram matrix $(\phi(j, j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle; j, j' \in \mathcal{J})$.

The dependent Dirichlet processes is defined by setting

$$P^j(A) = \frac{\sum_i \mathbb{I}(Z_i \in A) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}, \quad \forall j \in \mathcal{J}, \quad (2)$$

for every $A \in \mathcal{F}$. Here the sequence (Z_1, Z_2, \dots) and the array $(\mathbf{X}_1, \mathbf{X}_2, \dots)$, contain independent and identically distributed random variables, while σ is a Poisson process on the unit interval defined by using a prior on $\sigma = (\sigma_1, \sigma_2, \dots)$, the distribution of ordered points $(\sigma_i > \sigma_{i+1})$ in a Poisson process on $(0, 1)$ with intensity

$$\nu(\sigma) = \alpha \sigma^{-1} (1 - \sigma)^{-1/2}, \quad (3)$$

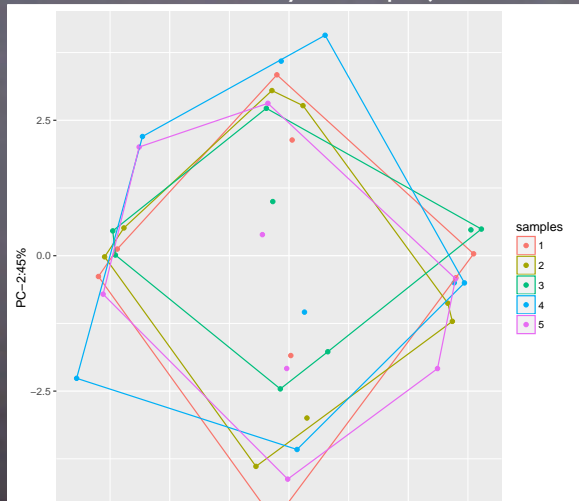
where $\alpha > 0$ is a concentration parameter.

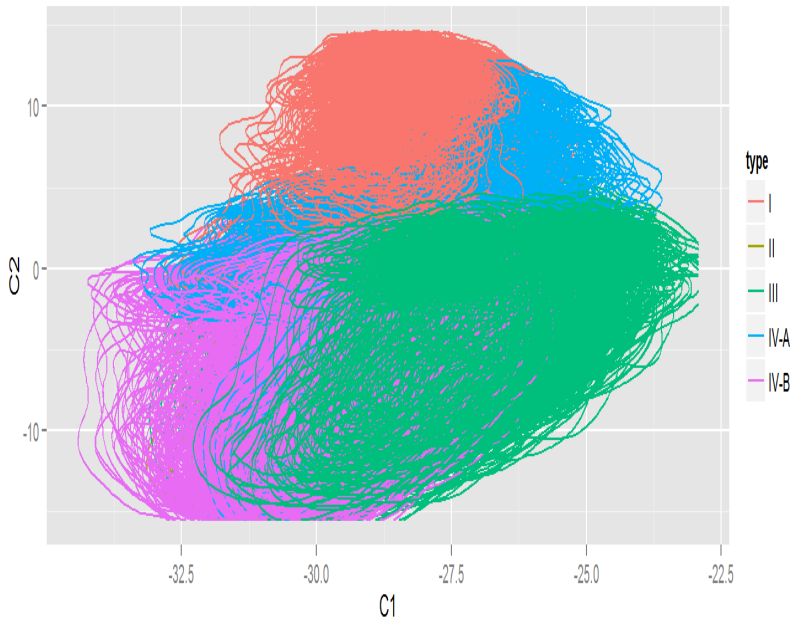
We will use the notation $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$.

The methods that we consider here are all related to PCA and use the normalized Gram matrix \mathbf{S} between biological samples. \mathbf{S} is the correlation matrix of $(Q_{i,1}, \dots, Q_{i,J})$. Based on a single posterior instance of \mathbf{S} , we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.

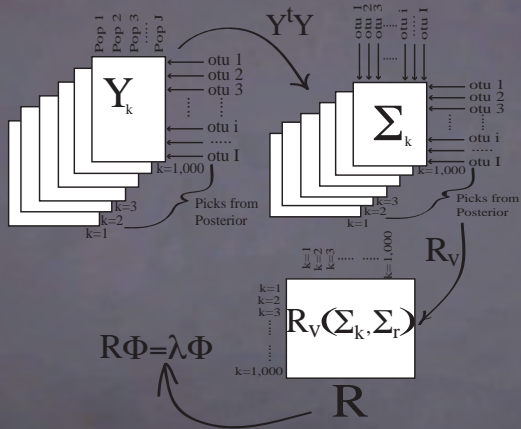
A projection approach

Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of \mathbf{S} on the same plot *could* show the variability of the projections.





Alternatively



We identify a consensus lower dimensional space for all posterior samples using STATIS (Escoufier, 1980, see Holmes, 2005). We list the three main steps used to visualize the variability of \mathbf{S} .

Registration: Find \mathbf{S}_0



Identify a Gram matrix \mathbf{S}_0 that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \dots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.

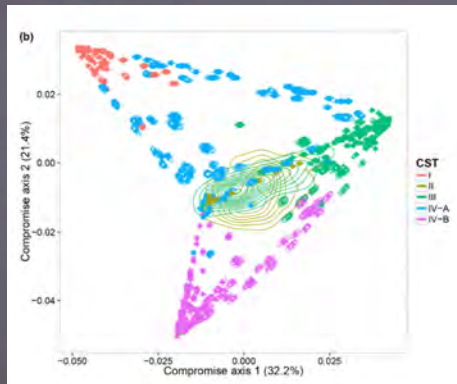
We prefer to choose \mathbf{S}_0 , the Gram matrix that maximizes similarity with $\mathbf{S}_1, \dots, \mathbf{S}_K$.

We use the **RV** similarity metric between two symmetric square matrices **A** and **B**

$$RV(\mathbf{A}, \mathbf{B}) = \text{Tr}(\mathbf{AB}) / \sqrt{\text{Tr}(\mathbf{AA})\text{Tr}(\mathbf{BB})}$$

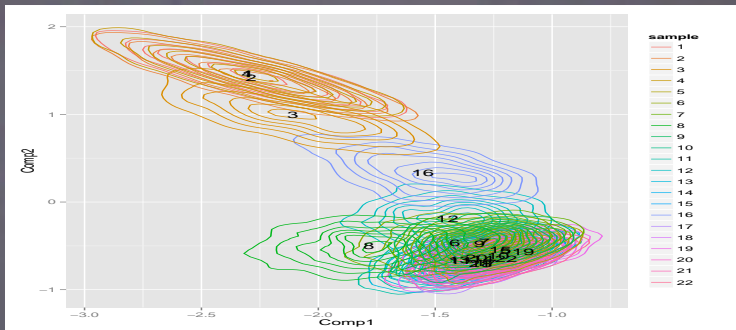
We diagonalize the **RV** matrix to obtain \mathbf{S}_0 .

We can see the uncertainties



Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2017 (JASA).

A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space \mathcal{V} .



A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space V .

R packages and resources

phyloseq: <http://bioconductor.org/packages/stats/bioc/phyloseq/>

dada2: <http://bioconductor.org/packages/stats/bioc/dada2/>

treelapse: <https://krisrs1128.github.io/treelapse/>

treelapse antibiotics <http://statweb.stanford.edu/~kriss1/antibiotic.html>

microbiome_pvlm: https://github.com/krisrs1128/microbiome_plvm

decontam: <https://github.com/benjjneb/decontam/>

adaptiveGPCA: <https://cran.r-project.org/web/packages/adaptiveGPCA/index.html>

bootLong: <https://github.com/PratheepaJ/bootLong/blob/master/vignettes/Workflow.Rmd>

Modern Statistics for Modern Biology

<http://bios221.stanford.edu/book/>

Solutions for microbiome analyses: respect the data.

- Poor data quality, information → quality scores & probability.
- Maintain all information → sequences are names.
- Interpretation → latent variables (gradients or clusters).
- Nonlinearity: gradients → t-sne and buds for manifold estimation.
- Reproducibility → complete code source.
- Heterogeneity → multicomponent objects: phyloseq.
- Training and collaboration → Rmd and html.
- Find the right "statistic" to bootstrap or compute posterior distribution for.

Benefitting from the tools and schools of Statisticians.....

Thanks to the R and Bioconductor community and to co-authors.



Wolfgang Huber, Joey McMurdie, Ben Callahan, JJ Allaire and Rob Gentleman.

Lab Group and David Relman



Postdoctoral Fellows Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Pratheepa Jeganathan, Nina Miolane. **Students:** John Cherian, Diana Proctor, Daniel Sprockett, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran, Claire Donnat. **Funding from** NIH TR01 and NSF-DMS.

phyloseq



Joey McMurdie (joey711 on github).

Available in Bioconductor.

How can I (my students, my postdocs...) learn more?

Ask me.

<http://www-stat.stanford.edu/~susan/>