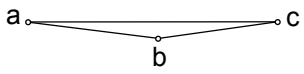
Additive vs ultrametric: triangle inequality

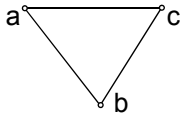
- Pairwise distance: any numeric variable that:
  - Expresses the difference between two objects *a,b*.
  - Has at least the properties of being metric:
    - (1) d(*a,a*)=0, d(*b,b*)=0, and d(*a,b*)≥0
      - Identical objects are indistinguishable.
      - Nonidentical objects might or might not be distinguishable.
    - (2) If *a*≠*b*, then d(*a,b*)>0
      - If the objects differ in properties being measured, then their distance must be greater than zero.
    - (3) d(*a,b*)=d(*b,a*)
      - Symmetry.

Additive vs ultrametric: triangle inequality

- (4) d(*a,c*) ≤ d(*a,b*) + d(*b,c*)
  - Triangle inequality.



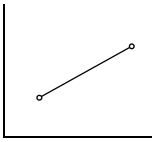
- Any set of distances having the 4 metric properties will produce an additive tree.
- Properties can be superimposed on distances, even if they don't possess the properties.
  - 'Forces' the distances into an additive tree.



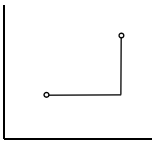
- 4th property can be relaxed:
  - (4) d(*a,c*) ≤ max{d(*a,b*), d(*b,c*)}
    - Any set of distances having these properties will produce an ultrametric tree.
    - Any distances can be 'forced' into an ultrametric tree.
    - Ensures that clustering monotonically increases with distance.

Some commonly used distances

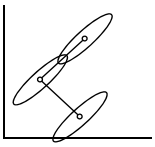
- Euclidean distance:



- Manhattan distance:

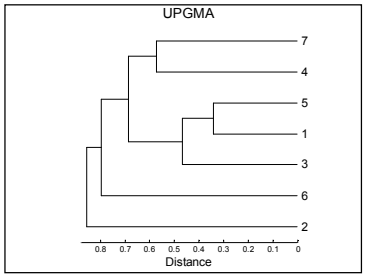


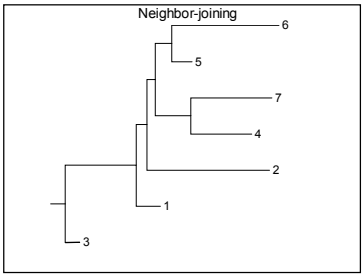
- Mahalanobis distance:



- Hamming distance between two strings of equal length: number of positions in which symbols are different (→ percent dissimilarity).

Ultrametric and additive trees for same data





Monotonically
Increasing clustering

General procedure for hierarchical cluster analysis

(1) Begin with N x N symmetric distance matrix.

(2) Each object (taxon) initially considered to be a separate cluster.

(3) Find two objects *i* and *j* separated by smallest distance.

(4) Combine objects *i* and *j* into a new cluster, *k*.

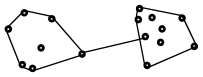
(5) Calculate distance between the new cluster and all other existing clusters.

- Reduces size of the distance matrix.

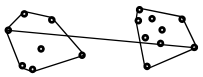
(6) Go to step 3. Continue until all objects have been merged into a single cluster.

Cluster analysis

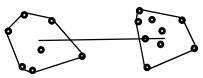
- Methods differ according to (5) how distances are calculated between new cluster and all other existing clusters:
  - Single linkage:



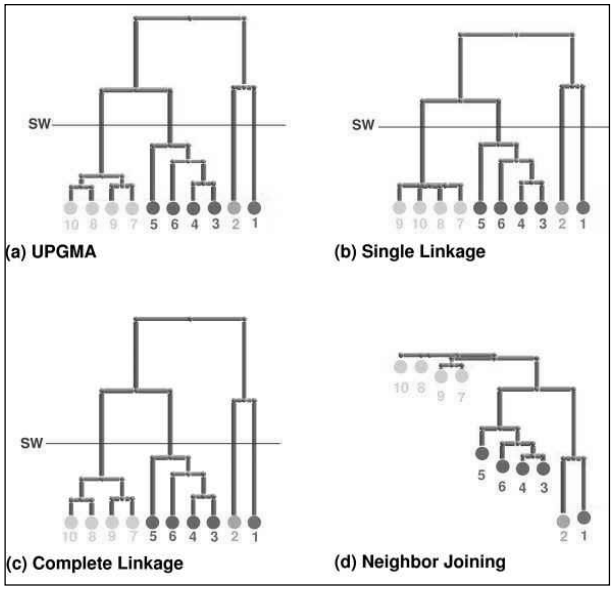
  - Complete linkage:



  - UPGMA, WPGMA:



Cluster analyses



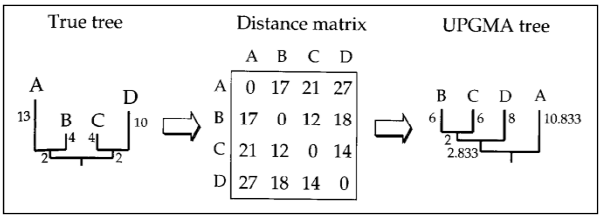
http://www.biomedcentral.com/1471-2105/9/90/figure/F5

Distances are always estimates

- Pairwise distance values are never exact:
  - Complete phylogenetic record of all genetic/phenotypic events would constitute set of distances that are completely:
    - Additive.
    - Mutually consistent across all taxa.
  - Observed distances are approximations:
    - Comprise ‘true’ distances, plus error.
    - Don’t display complete additivity and mutual consistency.
  - Additivity (metricity) and ultrametricity are properties of a distance matrix.
    - Can be tested statistically.
  - Q: to what extent do methods recover ‘true’ distances, in presence of error.

UPGMA

- UPGMA and other ultrametric methods assume that evolutionary rates have been constant:
  - Simulations show will work if rates are variable, but constant on average (clocklike, =stationary).
  - But if rates are not clocklike, can give misleading results.



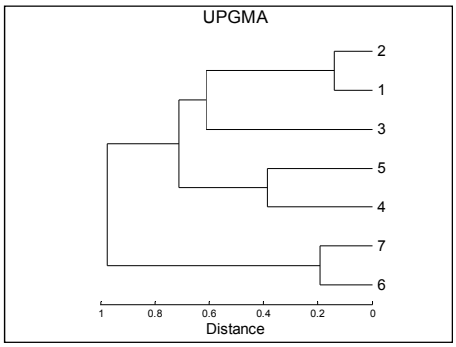
  - Estimation error can mimic effects of nonstationarity.

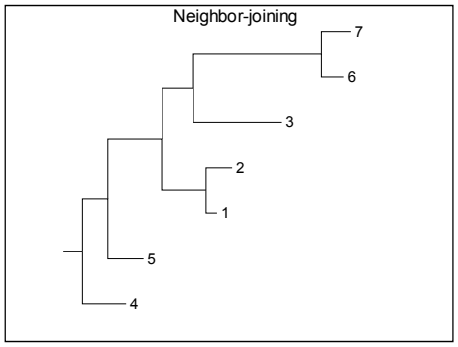
Additive trees

- Calculated in iterative fashion.
  - Several algorithms, most giving similar results.
  - Update placement of nodes rather than formation of clusters.
- Rate uniformity not assumed.
  - Corrects original distances for unequal divergence among branches.
- Least-squares and neighbor-joining trees guaranteed to recover true tree if distance matrix is an exact reflection of a tree.

Phenetic vs patristic distances

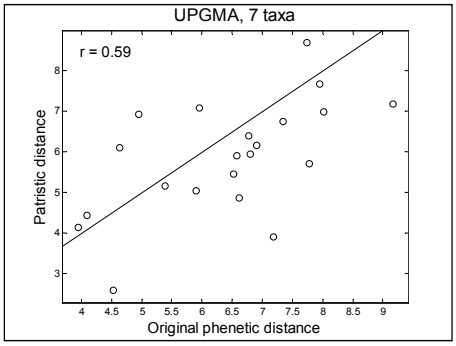
- Assessing agreement between tree and original distance matrix:
  - Patristic distance: predicted between two taxa by tree.
    - UPGMA: 2 x distance to nearest common node.
    - NJ: sum of horizontal branch lengths between taxa.

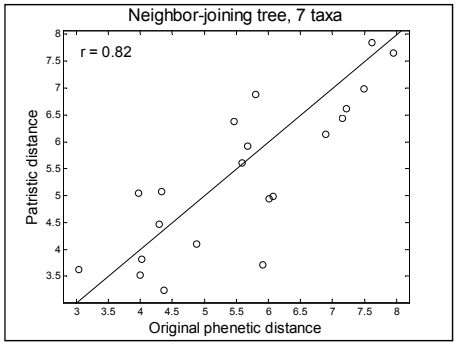




Correlations between phenetic and patristic distances

- *Ex:* phenetic distances calculated from clocklike tree:
  - Add noise.
  - Calculate cophenetic correlation between phenetic and patristic distances.





Assessing confidence in trees

- Two basic kinds of questions:
  (1) Is the overall tree "significant", or significantly better than another tree?
    - Permutation tests.
    - Frequency distributions of tree lengths.
      - E.g., g_1 statistic.
    - Likelihood and Bayesian scores.

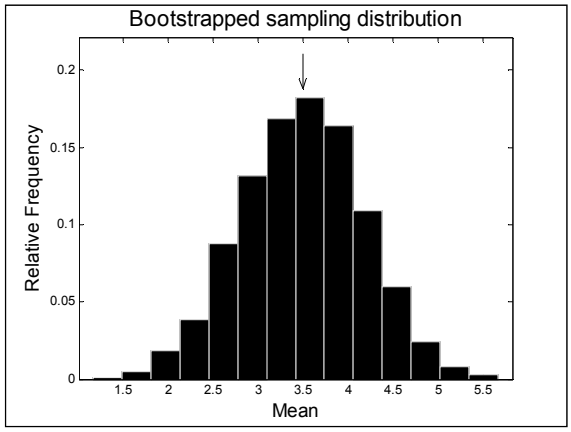
  (2) Are particular parts of the tree "significant"?
    - Bootstrapping.
    - Bayesian posterior probabilities.
    - Bremer support.

Bootstrap

- Bootstrap: general randomization procedure for estimating reliability of statistics:
  - Used primarily for estimating sampling distributions and associated confidence intervals.
  - Bootstrap (or bootstrapped) sample from a sample of *N* observations is a resample of those *N* observations *with replacement*.
    - Any particular observation might be sampled once, twice, or more often, or might be missed altogether.
    - Number of unique bootstrap samples: N^N.
    - In practice, use a *random set* of bootstrap samples, sufficient to stabilize the values we're trying to estimate.

Original sample		Bootstrapped resamples (of 46656 possible)									
1		1	2	6	5	4	5	6	3	6	4
2		3	3	3	6	5	6	5	1	5	1
3		4	1	2	3	1	6	4	5	1	5
4		2	5	1	1	4	6	1	4	3	3
5		4	1	4	6	1	5	4	5	6	2
6		4	3	4	2	1	3	5	5	6	2

Bootstrap



- Rationale and strong assumption: all observations are equivalent (i.e., independent replicates) and are representative off all possible observations that might have been sampled from the population.

Bootstrapping trees

- Very important topic in the theory and practice of phylogenetic inference.
- Originally proposed for trees estimated by "parsimony" with discrete data (Felsenstein 1985).
- Can be extended to all trees and networks:
  - Cladograms, dendrograms, phenograms, etc.

Bootstrapping trees

- Requires data matrix from which distance matrix is calculated:
  - Sample characters, with replacement, maintaining number of characters.
    - Bootstrapped sample, = pseudoreplicate.
    - Acts as a proxy for a true replicate from nature.
  - Convert pseudoreplicate data matrix to distance matrix.
    - For ultrametric and additive trees.
  - Construct tree, keep track of appearance of clusters (groups, clades, etc.).
    - Proportion of bootstrapped trees in which observed clusters occur.
- Gives bootstrap support frequency (BSF).

Bootstrapping trees

Original data matrix

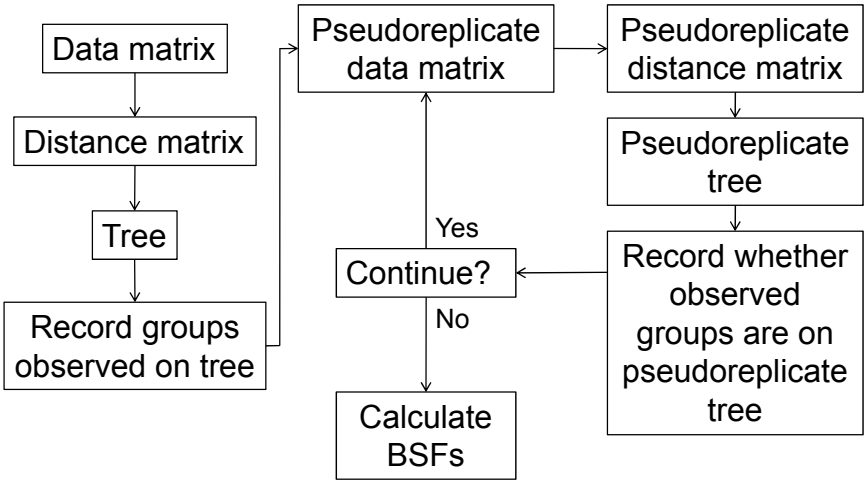
	Characters			
Taxon	a	b	c	d
A	1	3	2	1
B	2	4	5	1
C	1	2	2	1
D	2	1	3	3
E	2	1	2	3

Bootstrapped data matrix

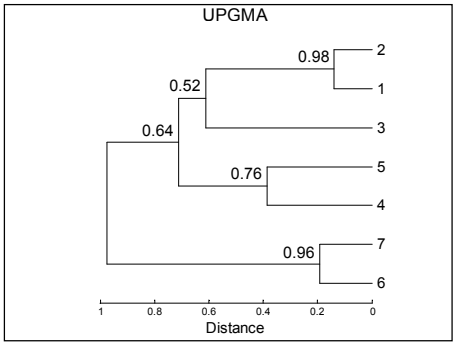
	Characters			
Taxon	a	a	c	c
A	1	1	2	2
B	2	2	5	5
C	1	1	2	2
D	2	2	3	3
E	2	2	2	2

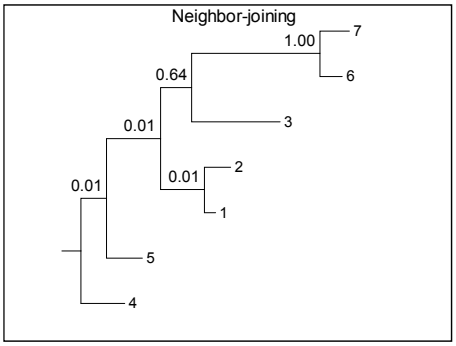
- Assumes that characters are:
  - Independent (uncorrelated).
  - Equivalent (equally weighted).
  - Representative of an underlying pool of characters that might have been sampled.

Bootstrapping trees



Bootstrapped UPGMA and NJ trees from same random data





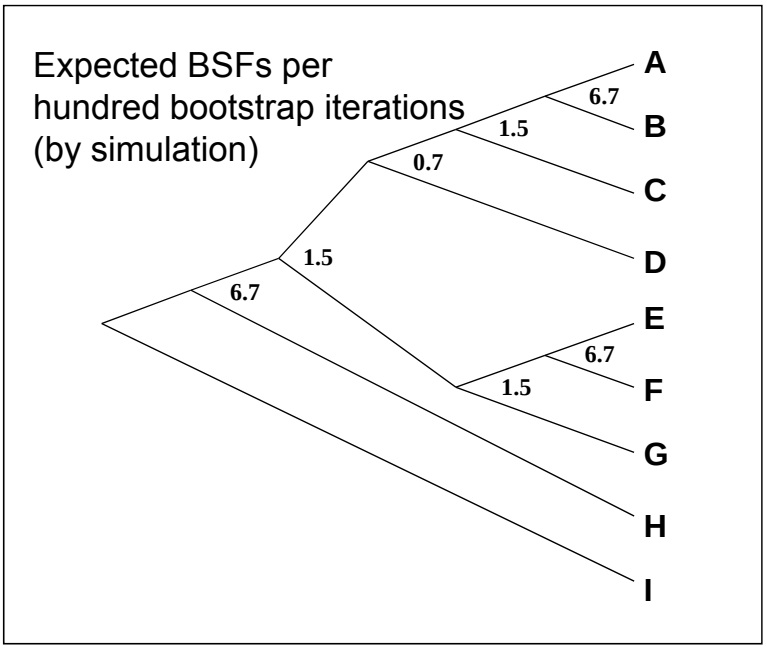
Common interpretations of bootstrap support frequencies

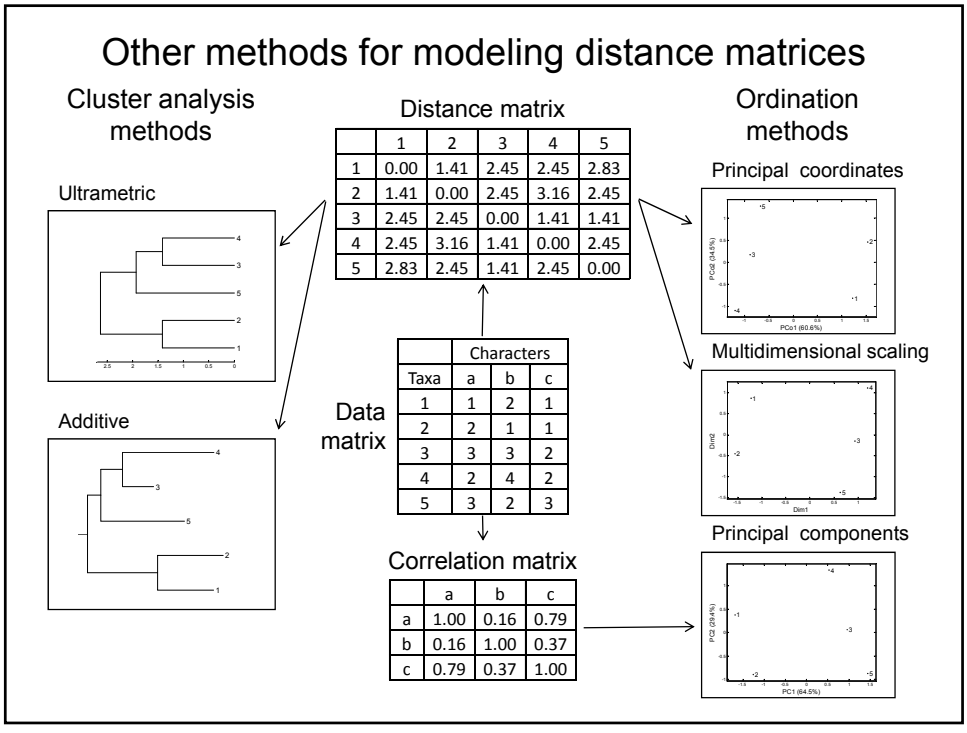
- Probability that the cluster (clade) will be observed on repeated sampling of many characters from the underlying pool of characters.
- Confidence limit on a cluster.
- Level of “significance” of a cluster.
- Probability that a given cluster is a “real” group.
- General measure of support for a given cluster.

Assessing the “significance” of a bootstrap support frequency

- Minimum-50% (majority consensus) rule.
- Gestalt 70% rule (Hillis & Bull 1993).
- 1-*P* rule (Felsenstein & Kishino 1993).

- However:
  - None of these rules is adequate.
  - All can be misleading.





	1	2	3	4	5
1	0.00	1.41	2.45	2.45	2.83
2	1.41	0.00	2.45	3.16	2.45
3	2.45	2.45	0.00	1.41	1.41
4	2.45	3.16	1.41	0.00	2.45
5	2.83	2.45	1.41	2.45	0.00

	Characters		
Taxa	a	b	c
1	1	2	1
2	2	1	1
3	3	3	2
4	2	4	2
5	3	2	3

	a	b	c
a	1.00	0.16	0.79
b	0.16	1.00	0.37
c	0.79	0.37	1.00

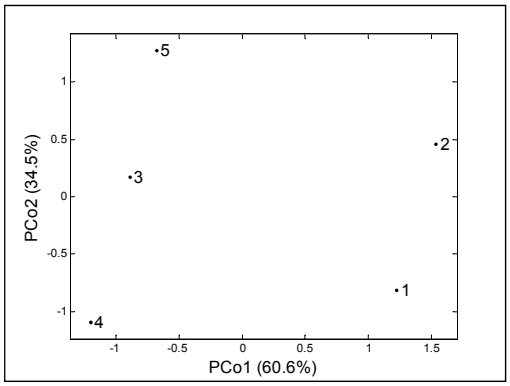
Principal coordinates analysis (PCoA)

- Eigenanalysis method:
  - Decomposes information in the distance matrix into a set of orthogonal axes:
    - Axes = principal coordinates.
    - Orthogonal = statistically independent.
  - Principal coordinates:
    - PCo1 accounts for the maximum information in the distance matrix.
    - PCo2 accounts for the maximum residual information in the distance matrix, independent of PCo1.
    - PCo3 accounts for the maximum residual information in the distance matrix, independent of both PCo1 and PCo2.
    - Etc.
  - For an *N*×*N* distance matrix, there are *N* principal coordinates.
    - The full set of *N* principal coordinates accounts for all of the information in the distance matrix.
  - Observations (taxa) are projected onto the axes to give projection scores, which can be plotted with scattergrams.

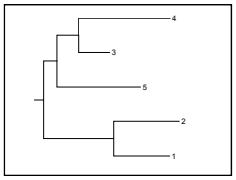
Principal coordinates analysis (PCoA)

Distance matrix

	1	2	3	4	5
1	0.00	1.41	2.45	2.45	2.83
2	1.41	0.00	2.45	3.16	2.45
3	2.45	2.45	0.00	1.41	1.41
4	2.45	3.16	1.41	0.00	2.45
5	2.83	2.45	1.41	2.45	0.00



Additive



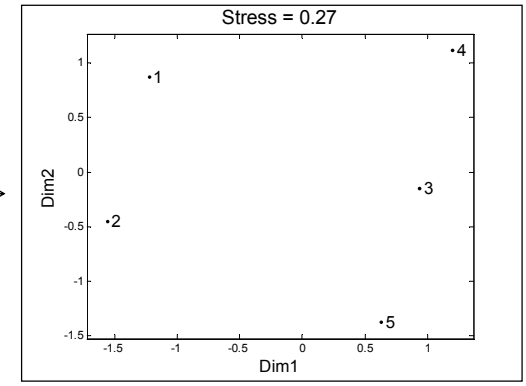
Multidimensional scaling (MDS)

- Not an eigenanalysis method:
  - Information in distance matrix is not decomposed.
- Points representing observations (taxa) are 'squeezed' into 1D, 2D, or 3D… space.
  - Placed into space so that interpoint distances reconstruct original distances as much as possible.
    - Metric and nonmetric (monotonic) versions.
  - Axes are arbitrary, and points can be arbitrarily rotated.
  - Amount of distortion measured as 'stress'.
- Observations (taxa) are projected onto the axes to give projection scores, which can be plotted with scattergrams.

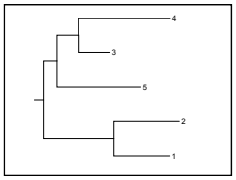
Multidimensional scaling (MDS)

Distance matrix

	1	2	3	4	5
1	0.00	1.41	2.45	2.45	2.83
2	1.41	0.00	2.45	3.16	2.45
3	2.45	2.45	0.00	1.41	1.41
4	2.45	3.16	1.41	0.00	2.45
5	2.83	2.45	1.41	2.45	0.00



Additive



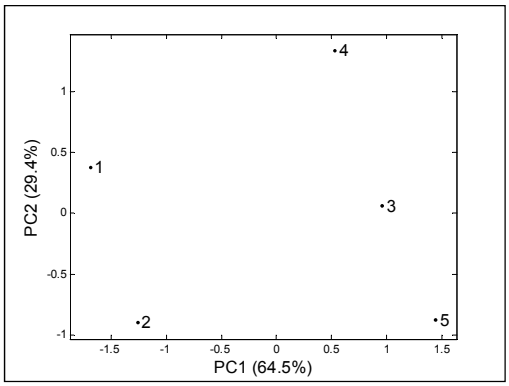
Principal components analysis (PCA)

- Eigenanalysis method, but based on correlation or covariance matrix:
  - Decomposes information in the correlation matrix into a set of orthogonal axes:
    - Axes = principal components.
    - Orthogonal = statistically independent.
  - Principal components:
    - PC1 accounts for the maximum information in the distance matrix.
    - PC2 accounts for the maximum residual information in the distance matrix, independent of PC1.
    - PC3 accounts for the maximum residual information in the distance matrix, independent of both PC1 and PC2.
    - Etc.
  - For an *N*×*N* correlation matrix, there are *N* principal components.
    - The full set of *N* principal components accounts for all of the information in the distance matrix.
  - Observations (taxa) are projected onto the axes to give projection scores, which can be plotted with scattergrams.

Principal components analysis (PCA)

Correlation matrix

	a	b	c
a	1.00	0.16	0.79
b	0.16	1.00	0.37
c	0.79	0.37	1.00



Additive

