

Classification and Correlation Techniques

Cluster Analysis and Mantel Test

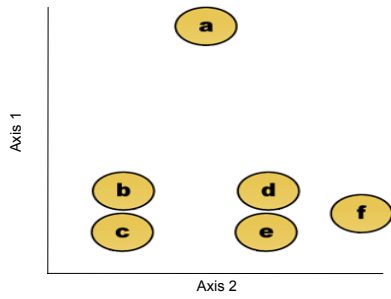
Cluster Analysis

- seeks to identify homogeneous subgroups of cases in a population.
- Used when the researcher does not know the number of groups in advance but wishes to establish groups and then analyze group membership.
- Contrasts to DFA which analyzes group membership for known groups pre-specified by the researcher.
- Cluster analysis seeks to identify a set of groups which both minimize within-group variation and maximize between-group variation.

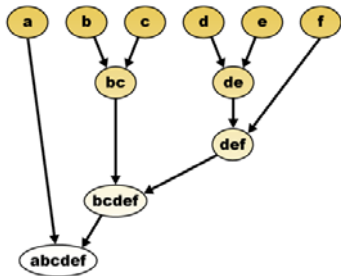
Cluster Analysis

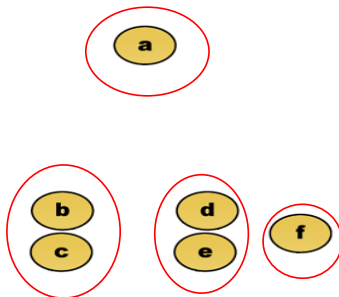
- The first step in cluster analysis is to select a distance measure to determine how the similarity of each data point is measured
- eg. Euclidean, block distance, Bray-Curtis, etc.

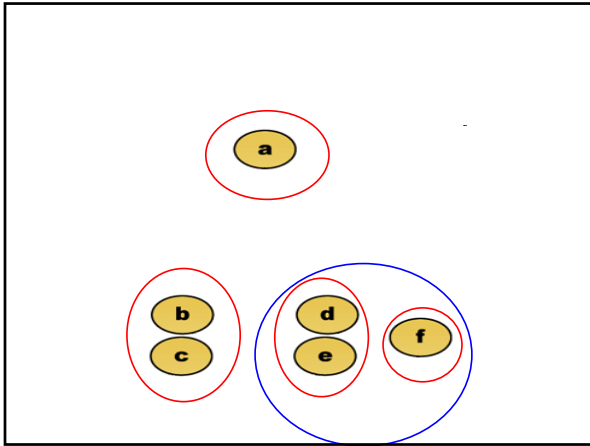
Samples are ordinated in multidimensional space

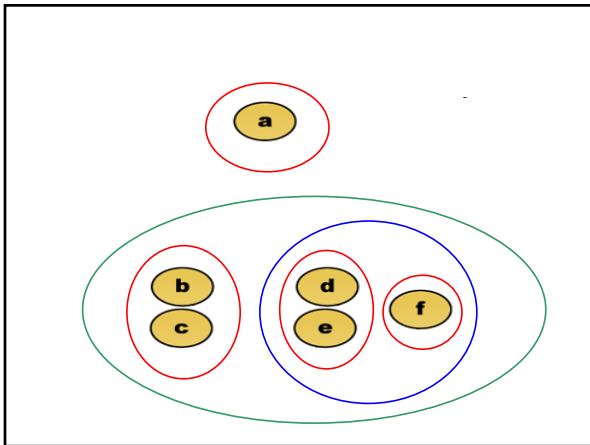


Clustering can be either agglomerative or divisive



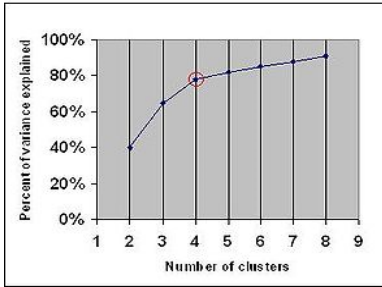






• one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Choose the minimum number of clusters that explain the most variance



Cluster Analysis

- Three general approaches to cluster analysis:
 - *Hierarchical clustering* allows users to select a definition of distance, then select a linking method for forming clusters, then determine how many clusters best suit the data. Hierarchical clustering generates representation of clusters in dendograms.
 - *K-means clustering* has the researcher specify the number of clusters in advance, then the algorithm calculates how to assign cases to the K clusters. K-means clustering is much less computer-intensive and is therefore sometimes preferred when datasets are large (ex., > 1,000). K-means clustering generates an ANOVA table showing mean-square error.
 - *Two-step clustering* creates pre-clusters, then it clusters the pre-clusters using hierarchical methods. Two step clustering handles very large datasets, is the method chosen when data are categorical (it supports continuous variables also).

Hierarchical Clustering

- Hierarchical clustering is appropriate for smaller samples (typically < 250). When n is large, the solution gets very computationally intensive
- User defines the distance measure and whether clustering is agglomerative or divisive
- Once clusters are grouped together, they stay together

K-means Clustering

- Researcher specifies the number of groups
- Better for larger groups because all pairwise distance comparisons are not calculated
- Samples may be shifted from one cluster to another during the iterative process of converging on a solution
- the algorithm seeks to minimize within-cluster variance and maximize variability between clusters

Two-step clustering

- Used for very large data sets
- Handles categorical (three or more levels) as well as continuous data
- Identifies pre-clusters in a first step, then treats these as single cases in a second step which uses hierarchical clustering

The utility of clusters must be assessed by three criteria

- *Size*. All clusters should have enough cases to be meaningful. One or more very small clusters indicates the researcher has requested too many clusters. Analysis resulting in a very large, dominant cluster may indicate too few clusters have been requested
- *Meaningfulness*. Ideally the meaning of each cluster should be readily construed from the variables used to create the clusters
- *Criterion validity*. The number of clusters should be consistent with other variables known from theory or prior research to correlate with the concept which clustering is supposed to reflect

- Failure to meet these criteria may indicate the researcher has requested too many or too few clusters, or possibly that an inappropriate distance measure has been selected.
- It is also possible that the hypothesized conceptual basis for clustering does not exist, resulting in arbitrary clusters.

Applications

- Can be used to compare assemblages of taxa from a heterogeneous environment or to determine how heterogeneous an environment is from the perspective of the study organism
- Can be used to create phylogenies based on shared attributes

Mantel Test

- The Mantel test is used to test the correlation between two distance matrices
- Originally developed to evaluate spatial and temporal clustering of diseases like leukemia (Mantel 1967)
- It was later introduced to the fields of systematics and biogeography (Sokal 1979)

Mantel Test

- Each matrix is calculated from a different set of variables measured on the same sample units

	sp1	sp2	sp3
tree 1	4	1	3
tree 2	3	4	3
tree 3	1	5	3

	height	Stem growth	BTD
tree 1	5	3	3
tree 2	2	5	5
tree 3	2	3	4

Mantel Test

- The Mantel test is an alternative to regressing one set of variables against another
- Because the cells of distance matrices are not independent of each other we cannot accept the p-values from standard techniques that assume independence of the observations

Mantel Test

- The Mantel test is used to evaluate the congruence between two distance matrices of the same dimensions
- The two matrices must have the same set of sample units in the same order

	sp1	sp2	sp3
tree 1	4	1	3
tree2	3	4	3
tree 3	1	5	3

	height	Stem growth	BTD
tree 1	5	3	3
tree 2	2	5	5
tree3	2	3	4

Mantel Test

- Seek linear relationships between two matrices
- The ability to construct the matrices from any distance measure, similarity measure, or design variable leads to high power and flexibility

How It Works

- Tests the significance of the correlation between matrices by evaluating results from repeated randomization
 - How often does randomization of one matrix result in a correlation that's as strong or stronger than the observed correlation?
- Strong correlation structure between matrices will rarely be preserved or enhanced if one matrix is shuffled

- A test statistic (Z) is calculated for each run
- A p-value is calculated from the number of randomizations that yield a test statistic equal to or more extreme than the observed value
- The standardized Mantel statistic (r) is calculated as the Pearson correlation coefficient between the two matrices

Mantel Test

- Are trees that are more similar in their arthropod communities also more similar in their growth?

	sp1	sp2	sp3
tree 1	4	1	3
tree2	1	4	2
tree 3	3	1	3

	height	Stem growth	BTD
tree 1	5	3	3
tree 2	2	5	7
tree 3	5	3	4
