MULTI-VARIATE ANALYSIS

PCA

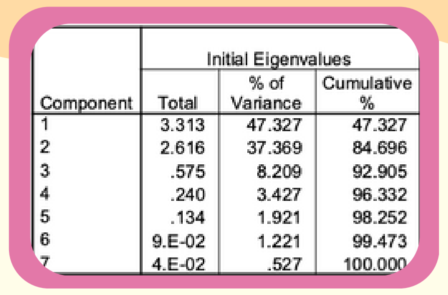
Discover and summarise joint or common variation among variables

Linear ordination, Euclidean distances
Continuous, normally distributed data

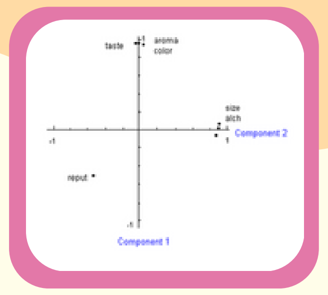
1. Compute covariance/correlation matrix
2. Extract Principle Components (p)
3. Retain Key Components/Latent Variables
   a. Kaiser's Criterion (eigenvalue >/= 1)
   b. Scree Plot (elbow)

FACTOR ANALYSIS

1. Plot and Rotate Latent Variables
2. Univariate Stats on key PC



Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.313	47.327	47.327
2	2.616	37.369	84.696
3	.575	8.209	92.905
4	.240	3.427	96.332
5	.134	1.921	98.252
6	9.E-02	1.221	99.473
7	4.E-02	.527	100.000



(DIS)SIMILARITY MATRIX

Quantify how similar or different each pair of samples is

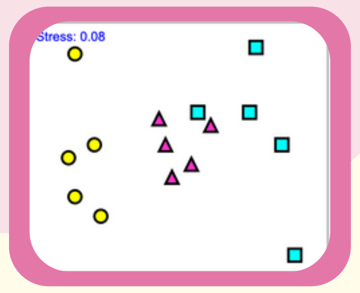
1. Choose a distance metric:
   a. Non-metric → Bray-Curtis (sum of absolute differences/ total abundances)
   b. Metric → Euclidean (Sq Root of SS differences)
2. Compute pairwise distances between samples.

NMDS

Distance between points reflects relative similarity

Non-parametric ordination based on rank similarities
More flexible, suited to ecological or non-normal data.

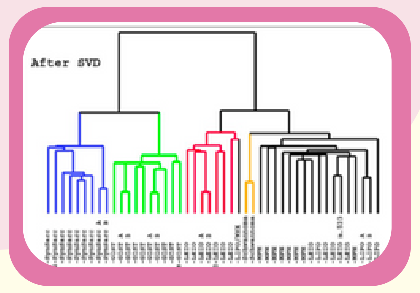
1. Iteratively arrange samples in few dimensions
2. Assess stress value:
   0.0 ~ perfect, 0.1 ~ decent,
   0.2 ~ ok, 0.3 ~ crap
1. Plot ordination.



CLUSTERING

Identification of clusters or groups of observations

1. Choose clustering approach
   a. Hierarchical Agglomerative vs Non-Hierarchical
   b. Choose linkage method (UPGMA)
2. Look for groupings



TEST GROUP DIFFERENCES

ANOSIM

Test whether samples are more similar within groups than between groups

Non-parametric test using rank similarities
Handles simple designs best (1 factor)

1. Compute the average rank distance
   a. Between groups:
   b. Within groups:
2. Calculate global R
   a. (rbetween - rwithin) /(standardising factor)
   b. (0 = no difference, 1 = complete separation)
3. Use permutations to assess significance

PERMANOVA

Test whether centroids of groups differ in multivariate space ~ which variables explain variation/interactions

Non-parametric test using permutations of distance measures
Handles multiple factors and continuous covariates

1. Convert to dissimilarity matrix to Euclidean space
2. Partition total variation into:
   a. Between-group variation
   b. Within-group variation
3. Permute samples among groups to build null distribution.
4. Report pseudo-F statistic and p-value.
   a. F= (SSwithin/dfwithin)/ (SSbetween/dfbetween)

MANOVA

Test of group main effects and interactions in multiple dependent variables.

Parametric, >/= 2 DV (continuous/ratio)
>/= 1 IV (categorical/ordinal)
Assumes multivariate normality and homogeneity of covariance.

1. Fit model (using R)
2. Examine Wilks' λ, or equivalent
   a. The ratio of error to effect plus error (|E| |H + E|)
3. Conduct post-hoc tests if significant.

SIMPER

Identify which variables contribute most to dissimilarity between groups.

1. Compute average between-group dissimilarities.
2. Rank variables by contribution to total difference.
3. Interpret main drivers

PRINCIPAL COMPONENTS & FACTOR ANALYSES

Goals:

Identify underlying dimensions/principal components of a distribution
Discover and summarise joint or common variation among variables

1. Matrix:

Units same -> covariance matrix
Units different ->correlation matrix
(standardises units)

2. Extract Principal Components:

p variables → p components
1st component ~ largest variance
Proportion of variance extracted
= eigenvalue/ p.
Each is orthogonal to the other

3. Retain Key Components:

Kaiser's Criterion: keep variables eigenvalues >/= 1

Scree Plot: keep variables above elbow

Plots/ Loadings and Rotations

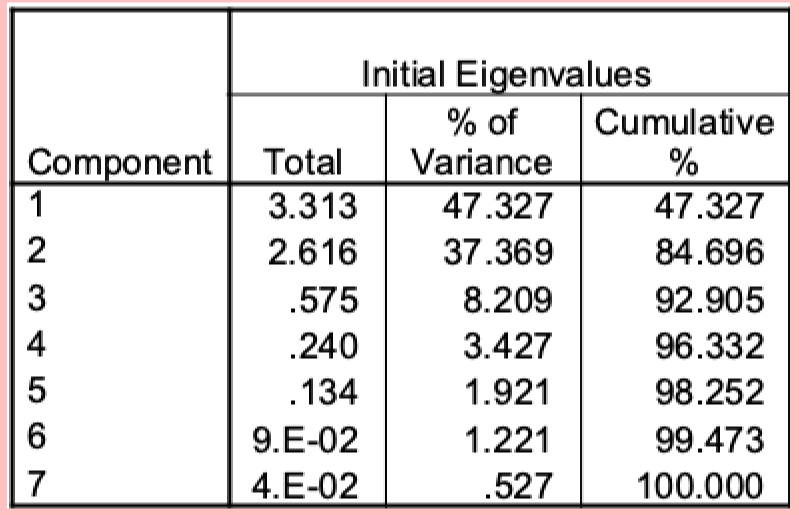
Loading Matrix = component matrix = rotations

Rotation helps discriminate among factors

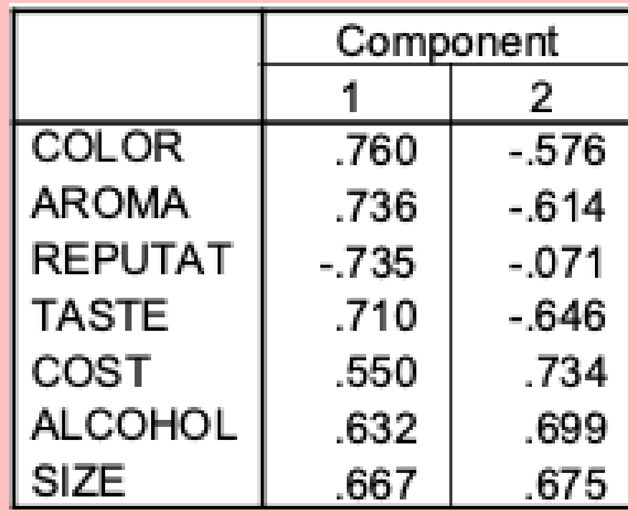
Varimax: orthogonal rotation that maximises dispersion of loadings

Rotate axes so two dimensions pass more neatly through the two major clusters

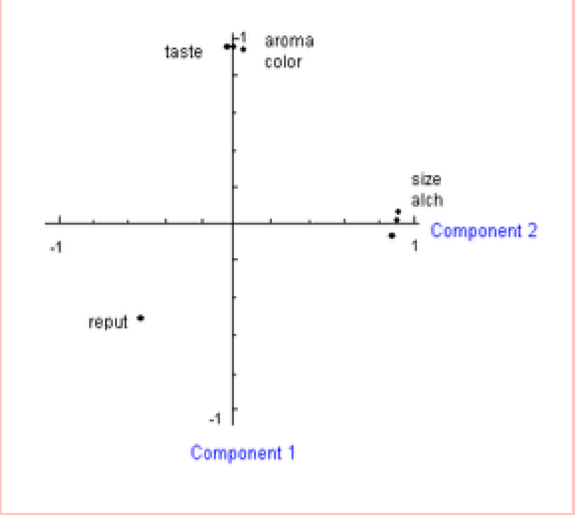
Can use PCs as response/predictor variables in additional stats/analyses

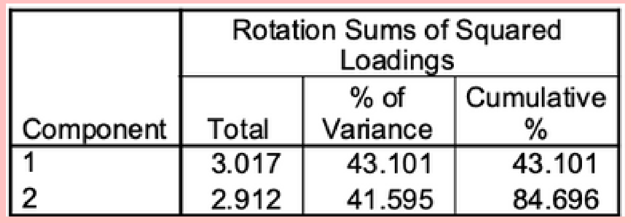


	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	3.313	47.327	47.327
2	2.616	37.369	84.696
3	.575	8.209	92.905
4	.240	3.427	96.332
5	.134	1.921	98.252
6	9.E-02	1.221	99.473
7	4.E-02	.527	100.000



	Component	
	1	2
COLOR	.760	-.576
AROMA	.736	-.614
REPUTAT	-.735	-.071
TASTE	.710	-.646
COST	.550	.734
ALCOHOL	.632	.699
SIZE	.667	.675





	Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %
1	3.017	43.101	43.101
2	2.912	41.595	84.696

Eigenvalues – amount of original variances explained by new derived variable

Eigenvectors – weights showing how much each original variable contributes to each newly derived variable and direction

Distance Matrices

How alike/different objects are

Dissimilarity measures (distance) ~ how "close" things in multivariate distance

Euclidean dissimilarity

Metric

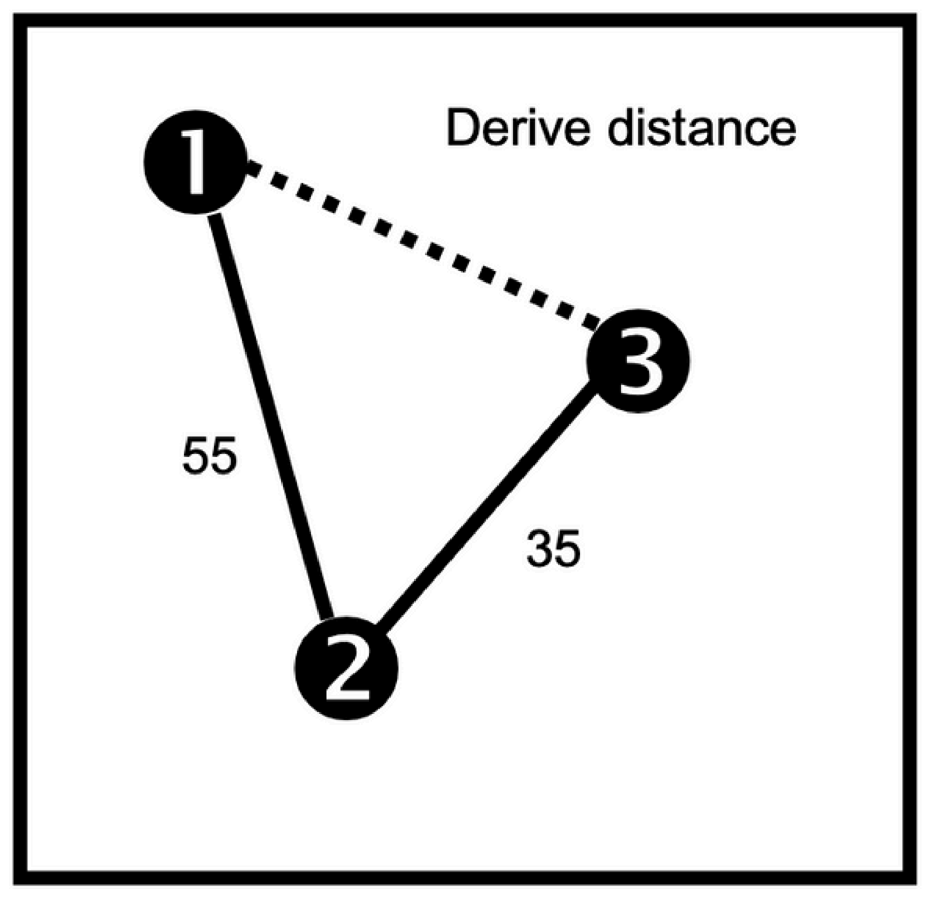
Sq Root of SS differences between each variable

Useful for "measurement" variables with true zeros

No upper limit

Includes joint zeros

Generally bounded by 0 and 100
0 = similar, 100 = dissimilar



Bray-Curtis dissimilarity

Percent dissimilarity
Ignores 0/0 (joint absences)

Well suited to species abundance data

Main determinants are variables with high values

Bray-Curtis = sum of absolute differences/ total abundances

Transformations

Absolute measures may over-emphasise the importance of prominent taxa

Down-weighting importance of dominant species by changing scale of measurement

Square root $\sqrt{y}$
Log (1+y)
4th root $\sqrt{\sqrt{y}}$ (preserves some abundance effects)
Presence/Absence

Standardisations

Adjust data so means and/or variances or totals are the same for each variable

Make each species equally important OR
Make each sample equally important

Express values as proportion of the largest value for that object

Change the fundamental interpretation, use in addition to "raw" data?

CLUSTERING

Goals:

Do samples associate with each other into groups?
Aims to find “natural groupings”

Steps:

1. Generate distance matrix
2. Choose clustering approach
3. Look for groupings

Hierarchical Agglomerative Cluster Analysis

Start with a pairwise similarity matrix among objects (individuals, sites, populations, taxa)

Most similar joined into a group

Similarities of new groups to all others is calculated

Two closest groups are combined repeatedly until one group remains

2-dimensional representation in dendrogram

Non-Hierarchical Clustering

Start with single object and cluster other objects that are similar to that one

Objects can be reassigned to clusters during clustering process

K-means clustering

Splits objects into pre-defined number (K) of clusters

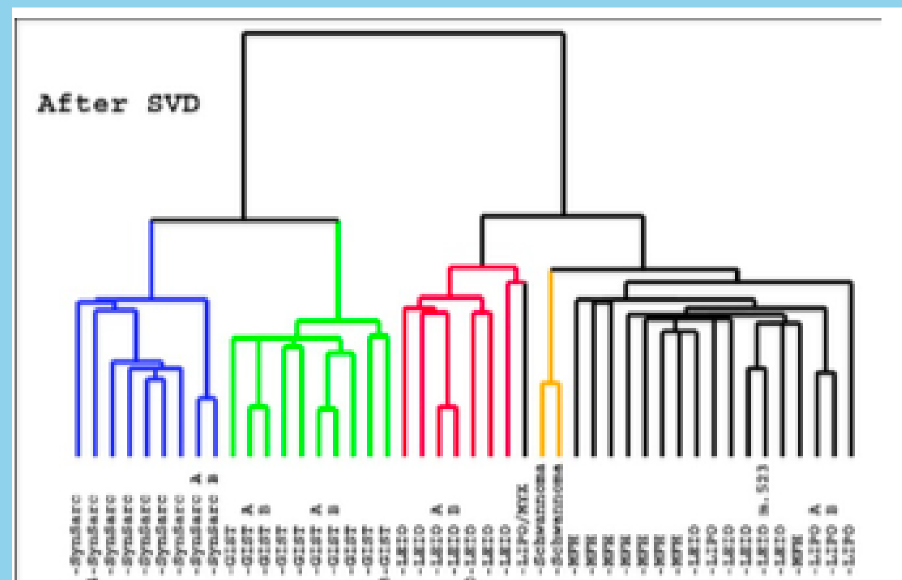
Cluster membership is iteratively re-evaluated by some criterion

Types of Hierarchical Agglomerative Cluster Analysis

Single Linkage (nearest neighbour)
Complete Linkage (furthest neighbour)
Average Linkage (group average or mean)

- **Unweighted pair-groups method using arithmetic averages (UPGMA)**
- Weighted pair-groups method using arithmetic averages (WPGMA)
- Unweighted pair-groups method using arithmetic centroids (UPGMC)

Ward’s Minimum Variance Clustering



NON-METRIC MULTIDIMENTIONAL SCALING

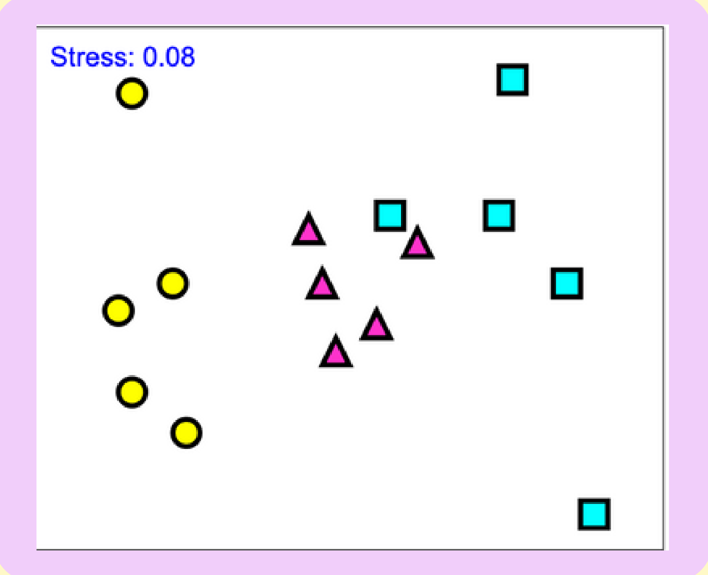
Uses the rank order of similarity relationships between samples

Places points in 2 (or 3) dimensional space to represent this ranked order

Distance between points reflects relative similarity

Stress Value = Quality/Accuracy

0.0 ~ perfect map
0.1 ~ decent map
0.2 ~ ok map
0.3 ~ don't bother



PARAMETRIC STATS

ANOSIM

Analysis of similarities

Handles simple designs best (1 factor)

Compares ranks among groups to ranks within group

Gives a robust, comparable and globally interpretable measure of magnitude of overall community change associated with each factor

Steps:

1. Calculate global R ~ reflects difference among sites to differences within sites
2. Recompute R under permutations of sample labels
3. Refer the observed value of R to permutation distribution to establish significance

$$R = \frac{(r_{between} - r_{within})}{\text{standardising factor}}$$

PERMANOVA

Works with any distance measure that is appropriate to the data, and uses permutations to make it distribution free

Says which variables explain variation and gives interactions

Steps:

1. Generate the distance matrix
2. Run the PERMANOVA using the same models as for MANOVA

Pseudo F based on $SS_{between}/SS_{within}$

SIMPER

Similarity Percentages

Calculates % contribution of each species to the dissimilarities between all pairs of sampling units in different groups

and % contribution to similarities between all pairs within groups

Output ~ a list of species in order of their percent contributions to dissimilarities between or similarities within groups

MANOVA

Extension of ANOVA ~ main effects and interactions are assessed on a combination of dependent variables
MANOVA tests whether mean differences among groups on a combination of dependent variables is likely to occur by chance
If MANOVA is significant, undertake separate ANOVA on each of the DVs

Basic requirements

>/= 2 dependent variables (continuous/ratio)
>/= 1 categorical independent variables (nominal/ordinal)

Key Questions

Which DVs are most important?

- For significant main effects or interactions, on which individual dependent variable is there the most difference, "caused" by the levels of the independent variable?
- Follow a significant MANOVA with individual ANOVAs in order to see extent of effects on dependent variables

Which levels of the IV are significantly different?

- If there are significant main effects on independent variables with more than two levels than you need to test which levels are different from each other
- If there are interactions the interactions need to be taken apart so that the specific causes of the interaction can be uncovered

Assumptions

Independence

Observations should be statistically independent

Linearity

Assumes linear relationships between all DVs

Random sampling

Data should be randomly sampled

Multivariate Normality

Assumes that the means of the various DVs in each cell and all linear combinations of them are normally distributed

If individual variables are normal, than they should have multivariate normality

Homogeneity of covariance matrices

The variance of the data within each group should be equal (homoscedasticity)

Levene's test: for each DV

Box's test: to test the variance-covaraince matrices

Wilk's Lambda

|E| |H + E|

The ratio of error to effect plus error

Bonferroni corrections

Corrections involve dividing alpha by n (or sequentially reducing n)

Use a p value other than 0.05 to claim a difference at alpha = 0.05