Stats fest 2007

Multivariate analysis

murray.logan@sci.monash.edu.au

MONASH University

Multivariate analyses

- **Aims**
  - Data reduction
    - Reduce large numbers of variables into a smaller number – that adequately summarize the patterns
  - Reveal patterns in the data that cannot be found using isolated variables
    - Characterize things based on a large number of variables
      - Classify sites
      - Taxonomy

Multivariate analyses

- **Objects**
  - Things we wish to compare
    - Sampling or experimental units
    - E.g. sites, quadrats
- **Variables**
  - Characteristics measured from each object
    - Usually continuous variables
      - counts of many different species (species abundances)
      - Size of body parts (taxonomy)

Multivariate analyses

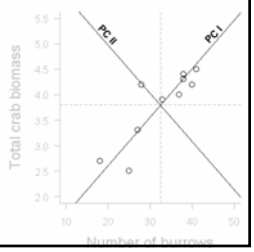
- **R-mode analyses**
  - Combine variables based on correlations
  - E.g. Principal components analysis (PCA)
- **Q-mode analyses**
  - Combine variables based on object dissimilarity
  - E.g. Multidimensional scaling (MDS)
  - Analysis of similarity (ANOSIM)
  - Autocorrelation
  - Cluster analysis

PCA

- **Aims**
  - Data reduction
  - Reveal patterns in the data that cannot be found using isolated variables
- **Data**
  - Many predictor variables measured from the same sampling units

PCA

- **If there are variables that are correlated to one-another**
  - Combine them together
- **Axis rotation**
  - First component
    - Surface of best fit
    - Explains most variation
  - Next component
    - Perpendicular
    - Explains next most
  - Next ….

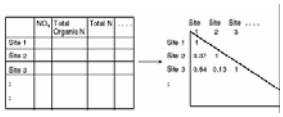


PCA

- **Difficult to visualize when more than 3 variables**
- **Eigenanalysis**
  - Matrix algebra used to do axes rotation in multidimensional space
  - Start with *p* original variables
  - End with *p* new completely uncorrelated variables (principal components)

PCA

- **Eigenanalysis**
  - Calculate correlation matrix between all *p* variables
  - Calculate new principal components (PC)
  - Eigenvalues (latent roots)
    - Amount or original variation explained by each new principal component
    - Adds up to the number of original variables
  - Component loadings
    - contribution of each original variable to each of the new PC
  - Factor scores

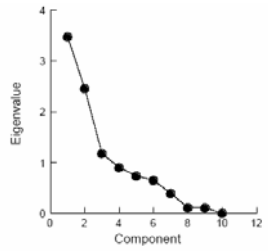


PCA

- **How many components to keep**
  - Eigenvalue > 1 rule
    - The sum of the eigenvalues is always equal to the number of original variables
    - Any PC > 1 must be explaining more than its share of the variation
      - Retain
    - Any PC < 1 not explaining much
      - Do not retain

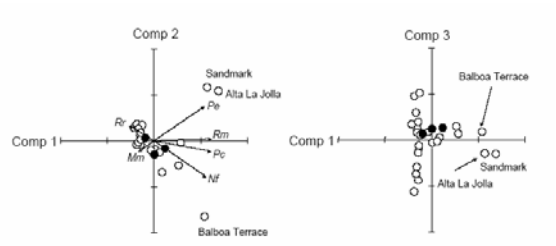
PCA

- **How many components to keep**
  - Scree test
    - Obvious ‘elbow’



PCA

- **Ordination plot**



PCA

- **Assumptions**
  - Because it is based on correlations
    - Assumes linearity

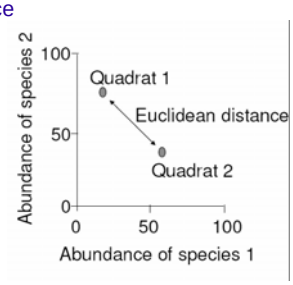
Q-mode analyses

- **Distance (dissimilarity) measures**
  - Measure of the degree of difference between each pair of objects based on a set of variables
    - How different sites are with respect to species composition
    - How different organisms are with respect to a suit of morphological and/or genetic characteristics
  - Smaller dissimilarities represent higher degree of similarity

Dissimilarity

- **Euclidean distance**
  - Geometric distance between objects (j and k) in multidimensional space
  - 0 when two objects identical
  - No maximum value
  - Joint absences considered similar

$$\text{dist} = \sqrt{\sum (y_{ij} - y_{ik})^2}$$



Dissimilarity

- **Bray-Curtis dissimilarity**
  - 0 when two objects are identical
  - Reaches a maximum of 1 when two objects have no variables in common
  - Joint absences ignored

$$\text{dist} = 1 - \frac{2\sum (y_{ij}, y_{ik})}{\sum (y_{ij} + y_{ik})}$$

Quadrat	Sp1	Sp2	Sp3
1	0	3	0
2	2	0	4

Euclidean = 5.4
Bray-Curtis = 1

Dissimilarity

- **Distance matrix**

Site	Sp1	Sp2	Sp3
A	0	14	68
B	0	29	49
C	0	1	0
D	0	4	11
E	1	0	24

Distance matrix

Euclidean distances

	A	B	C	D	E
A	0.00	24.2	69.2	57.9	46.2
B	24.2	0.00	57.9	45.5	28.3
C	69.2	56.4	0.00	11.4	24.0
D	57.9	45.5	11.4	0.00	13.6
E	46.2	38.3	24.0	13.6	0.00

Bray-Curtis distances

	A	B	C	D	E
A	0.00	0.21	0.96	0.69	0.55
B	0.21	0.00	0.97	0.68	0.53
C	0.96	0.97	0.00	0.87	1.00
D	0.69	0.68	0.87	0.00	0.45
E	0.55	0.53	1.00	0.45	0.00

Dissimilarity which is best?

- **Species abundance data**
  - Zeros common
  - Max value when quadrats have no species in common
  - Bray-Curtis preferred

```
> library(vegan)
> *.bc <- vegdist(variables, "bray")
```

- **Measurement/morphological data**
  - Zeros rare
  - Euclidean distance OK

```
> library(vegan)
> *.euc <- vegdist(variables, "euc")
```

Dissimilarity

- **Other distances**
  - Genetic distances from gene frequencies
    - Nei's distance
    - Edward's (Angular) distance
    - Coancestrality coefficient (Reynolds') distance
    - Classical Euclidean (Rogers') distance
    - Absolute genetics (Provesti's) distance

Standardizations

- **Aim**
  - To allow all variables to have an equal influence on patterns
  - Avoids overweighting by highly abundant species
  - Allows rare species to contribute
  - Different environmental variables measured on different scales
- **Scale each variable**
  - Divide all observations by max for that variable
  - Scale to a mean of 0 and sd of 1

Standardizations

- **Scale to maximums**

Raw data

Site	Sp1	Sp2	Sp3
A	0	14	68
B	0	29	49
C	0	1	0
D	0	4	11
E	1	0	24

Standardized (max) data

Site	Sp1	Sp2	Sp3
A	0.00	0.48	1.00
B	0.00	1.00	0.72
C	0.00	0.02	0.00
D	0.00	0.14	0.16
E	1.00	0.00	0.35

```
> library(vegan)
> *.stnd <- decostand(*[,2:4], "max")
```

Standardizations

- **Scale to mean of 0, sd of 1**

Raw data

Site	Sp1	Sp2	Sp3
A	0	14	68
B	0	29	49
C	0	1	0
D	0	4	11
E	1	0	24

Standardized (max) data

Site	Sp1	Sp2	Sp3
A	-0.447	0.361	1.350
B	-0.447	1.593	0.668
C	-0.447	-0.706	-1.092
D	-0.447	-0.460	-0.697
E	1.789	-0.789	-0.230

```
> *.stnd <- scale(*[,2:4])
```

Multidimensional scaling (MDS)

- **Aims**
  - Graphical representation of dissimilarity between objects in as few dimensions (axes) as possible
  - Axes are new variables

MDS

1. Setup data

- Objects (sites) in rows
- Variables (species) in columns

Site	Sp1	Sp2	Sp3	Sp4	Sp5
1	54	0	0	5	0
2	37	1	0	4	0
3	68	2	0	2	0
4	60	0	0	0	1
5	47	0	0	2	0
6	60	0	0	0	0

MDS

2. Calculate dissimilarity (Bray-Curtis)

```
> library(vegan)
> *.bc <- vegdist(variables, “bray”)
```

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
Site 1	0.00					
Site 2	0.20	0.00				
Site 3	0.67	0.65	0.00			
Site 4	0.22	0.33	0.76	0.00		
Site 5	0.33	0.41	0.80	0.19	0.00	
Site 6	0.34	0.43	0.80	0.18	0.05	0.00

MDS

3. Decide on the number of dimensions for ordination

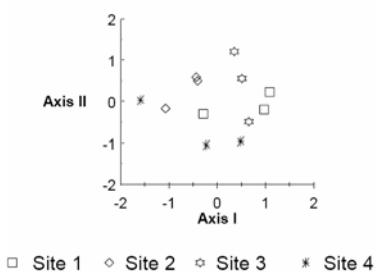
- Suspected number of major underlying ecological gradients
- Minimize new axes (variables) but still retain information
- Usually between 2 and four dimensions

```
> library(MASS)
> *.mds <- isoMDS(*.dist, k=2)
```

MDS

4. Arrange objects on ordination plot

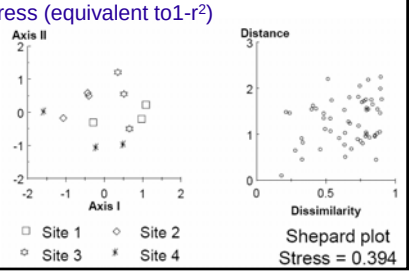
- Starting configuration
- Usually random



MDS

5. Compare distances on ordination plot with dissimilarity distances

- Strength of the relationship between ordination distances and dissimilarity distances
- Kruskal's stress (equivalent to1-r^2)



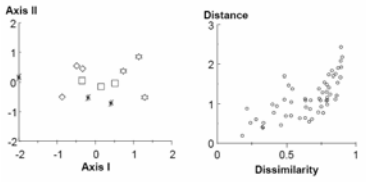
MDS

28

6. Move objects on ordination iteratively

- Each move improves the match between dissimilarities and ordination distances
- Lower stress value

After 20 iterations



Stress = 0.119

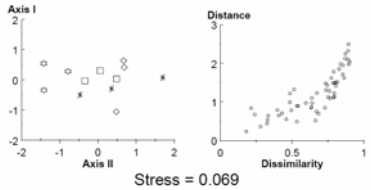
MDS

29

7. Final configuration

- When further moving of objects no longer improves the match
- Stress low as possible

Final configuration - 50 iterations



Stress = 0.069

MDS

30

- **Non-metric MDS**
  - Rank-based regression
    - Similar to rank based correlation
  - Better for ecological data
- **How low should stress be?**
  - > 0.20 (20%) – basically random
  - <0.15 (15%) is good match
  - <0.1 (10%) is ideal
    - Ordination configuration is close to actual dissimilarities
    - A small number of new variables explain most of the patterns contained in all the original variables

Hypothesis testing?

- **Is there are difference between the habitats with respect to species composition?**
  - Can we use the new axes scores in ANOVA?
  - NO! (why?)
  - BUT…

	Site	Sp1	Sp2	Sp3	Sp4	Sp5
Habitat 1	1	54	0	0	5	0
	2	37	1	0	4	0
	3	68	2	0	2	0
Habitat 2	4	60	0	0	0	1
	5	47	0	0	2	0
	6	60	0	0	0	0

Analysis of Similarities (ANOSIM)

- **Aim**
  - To compare groups based on similarities of objects
  - Uses dissimilarity matrices
- **Data**
  - Categorical variable
  - Multiple continuous response variables
    - Dissimilarity matrix
- **H_0:**
  - Average rank dissimilarities between objects within groups = Average rank dissimilarities between objects between groups
    - No difference in species composition between groups

Analysis of Similarities (ANOSIM)

Habitat	Site	Sp1	Sp2	Sp3	Sp4	Sp5
A	1	54	0	0	5	0
A	2	37	1	0	4	0
A	3	68	2	0	2	0
B	4	60	0	0	0	1
B	5	47	0	0	2	0
B	6	60	0	0	0	0

		Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
A	Site 1	0.00					
A	Site 2	0.20	0.00				
A	Site 3	0.67	0.65	0.00			
B	Site 4	0.22	0.33	0.76	0.00		
B	Site 5	0.33	0.41	0.80	0.19	0.00	
B	Site 6	0.34	0.43	0.80	0.18	0.05	0.00

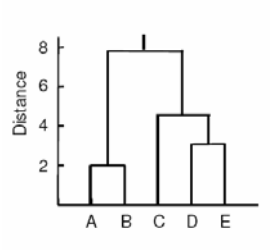
$R = r_b - r_w$ / number of sites

Analysis of Similarities (ANOSIM)

- **Dissimilarities not normally distributed**
  - Based on ranks
- **Dissimilarities not independent**
  - Uses randomization procedures to construct a probability distribution
- **Generates own test statistic (called R)**

Cluster analysis

- **Aims**
  - Combines similar objects together into clusters which are displayed as a dendogram



Cluster analysis

- **Single linkage (Nearest neighbour) method**

1. **Calculate dissimilarity matrix**
2. **First cluster is formed between two objects with smallest dissimilarity**

Distance Matrix

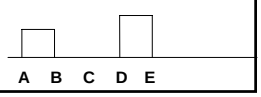
	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-

Cluster analysis

3. **Next cluster is between the next most similar pairs and so on**
   - Length of linkage reflects dissimilarity

Distance Matrix

	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-

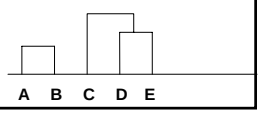


Cluster analysis

3. **Next cluster is between the next most similar pairs and so on**

Distance Matrix

	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-



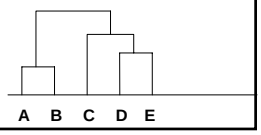
Cluster analysis

3. **Next cluster is between the next most similar pairs and so on**

4. **Procedure continues until all objects are linked in clusters**

Distance Matrix

	A	B	C	D	E
A	-				
B	2	-			
C	6	5	-		
D	10	9	4	-	
E	9	8	5	3	-

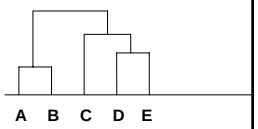


Cluster analysis

- **Other linkage methods**
  - Average linkage
    - Unweighted Pair-Group Method of Arithmetic Averaging (UPGMA)
    - Average neighbour
  - Complete linkage (Furthest neighbour)
    - Distance between clusters determined by most dissimilar objects in their groups

Clustering

- **How well do the cluster groups match the dissimilarity patterns**
  - cophenetic correlation
    - 0.823



Dissimilarity distances

	A	B	C	D	E
A	-	-	-	-	-
B	2	-	-	-	-
C	6	5	-	-	-
D	10	9	4	-	-
E	9	8	5	3	-

Cluster distances

	A	B	C	D	E
A	-	-	-	-	-
B	2	-	-	-	-
C	5	5	-	-	-
D	5	5	4	-	-
E	5	5	4	3	-

Minimum spanning trees

- **Mapped over ordination plots**

1. **Find smallest dissimilarity**
2. **Join these objects with a line**
3. **Find the next lowest dissimilarity and join objects**
4. **Repeat until all points joined**
5. **Short lines represent within clusters, long lines between clusters**

Autocorrelation- Mantal test

43

- **Aim**
  - Investigate association between two distance matrices
- **H_0:**
  - No association between two matrices
    - E.g. no correlation between species abundances and environmental characteristics

1. **Calculate correlation (*r*) between matrices**
2. **Since dissimilarities not independent or normal, use randomization test to reshuffle one of the matrices repeatedly**

Autocorrelation- Mantal test

44

Matrix A

	1	2	3	4	5
1	0				
2	20	0			
3	41	39	0		
4	12	25	53	0	
5	13	14	45	17	0

Matrix B

	1	2	3	4	5
1	0				
2	84	0			
3	26	51	0		
4	10	17	45	0	
5	22	35	28	32	0

Random permutation rows (and columns) of Matrix A

	2	1	5	4	3
2	0				
1	20	0			
5	14	13	0		
4	25	12	17	0	
3	39	41	45	53	0

Matrix B unchanged

- **Calculate probability**