Multivariate analyses II – classification and ordination

Chapter 17-18 in the course book

Göran Arnqvist, Animal Ecology

Data

- Many data sets contains large amounts of data, for example surveys, inventories and genetic (e.g. microarray) data.
- Typically many samples with data on, for example, the abundance or prescence/absence of things in each sample... (many variables/species = multivariate)
- Previous lecture: grouping factor/s known! But what of no groups are known *a priori*?
- Need multivariate methods to help us see groups and patterns in data – analyses are ”blind” to group belonging.
- Rather ”heavy” statistics...
- Two distinct aims: multivariate classification and ordination

Objectives

- Classification: placing samples into different subsets, or clusters, so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.
- Ordination: mainly for visualization - serves to summarize multivariate data (such as species abundance data) by producing a low-dimensional "ordination space" in which similar samples are plotted close together, and dissimilar samples are placed far apart.

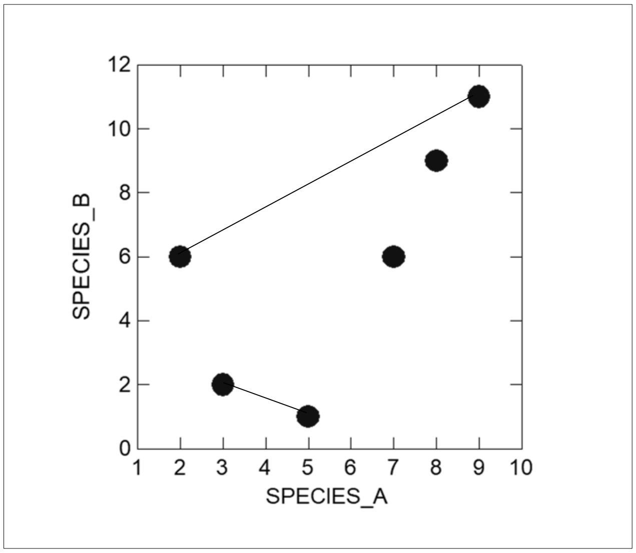
→ Classification is the placement of sample units into **discrete** groups and ordination is the arrangement or 'ordering' of sample units along **continuous** gradients.

First, a key question: how similar are two samples?

	SpeciesA	Species B	Species C	...	Species Z
1	32	5	9	.	6
2	76	22	19	.	23
3	34	11	12	.	17
4	0	2	22	.	2
...	.	.	.	.	.
N	5	8	1	.	33

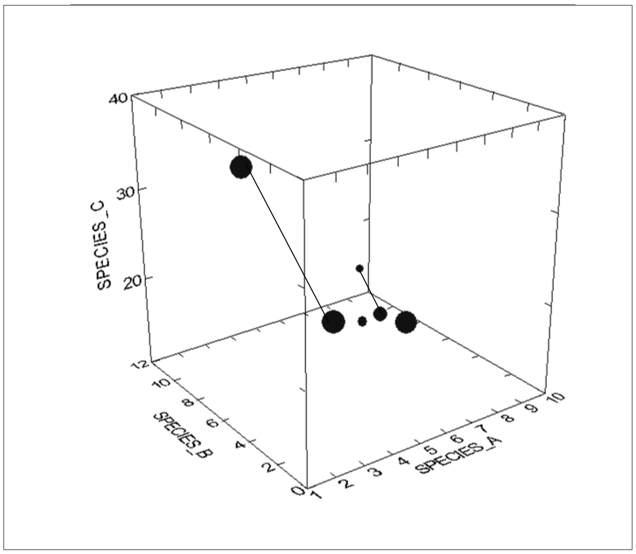
Multivariate measures of similarity/dissimilarity

One common measure: Euclidean distance – the geometric distance between two points (e.g., samples)



Multivariate measures of similarity/dissimilarity

One common measure: Euclidean distance – the geometric distance between two points (e.g., samples) in n dimensions



Multivariate measures of similarity/dissimilarity

There are many, slightly different, measures/indicies (see book!)

Common for abundance (i.e., continuous) data:

- Euclidean distance
- Mahalanobis distance
- Bray – Curtis (Czekanowski)
- Percentage similarity (PS)

Common for presence/absence (i.e., 0 or 1) data:

- Sørensen index
- Jaccard index

Results *will* differ somewhat depending on which index is used…

Similarity/dissimililarity/distance matrix

E.g.; how similar / dissimilar are sample 1 and 2? Can contain either of the different kinds of measures

	1	2	3	4	5
1	-				
2	4.2	-			
3	1.9	15.8	-		
4	10.3	3.4	7.3	-	
5	3.6	2.8	3.3	1.1	-

Two important points

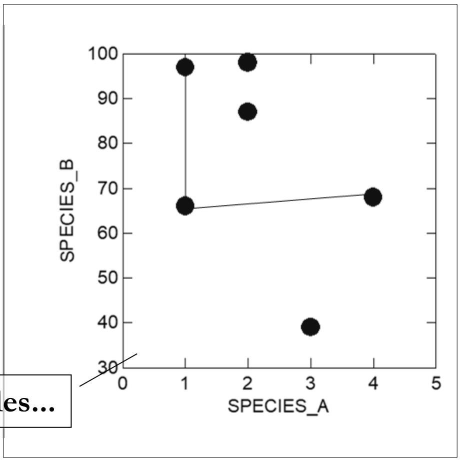
1. Abundant variables (eg species) will often ”dominate” the analysis!

Often ok!

If not ok, either

A) Analyse presence / absense data...*or better...*

B) Standardize data per variable prior to analysis – use standardized data



Cf scales...

Two important points

1. Abundant variables will often ”dominate” the analysis!

Standardization puts all varables on the same scale – gives all species the same ”importance”, e.g.:

a) $X' = \frac{X - \bar{X}}{S} + c$ where *c* is a common constant for all species such that no X´ values are negative (typically, *c* = 2 - 3).

b) $X' = \frac{X}{X_{\max}}$ where X_{max} is the largest value observed for that species.

Two important points

1. Abundant variables will often ”dominate” the analysis!

Standardized

Relativized

X1	X2		X´1(a)	X´2(a)		X´1(b)	X´2(b)
1	97		0.6	3.2		0.25	1
4	57		2.9	1.6		1	0.59
2	58		1.4	1.6		0.5	0.6
4	88		2.9	2.8		1	0.91
3	38		2.1	0.8		0.75	0.39

Two important points

2. The ”size”/effort of the sampling units must be the same in most cases – affects the ”size” of numbers...

Rarely ok if they differ!

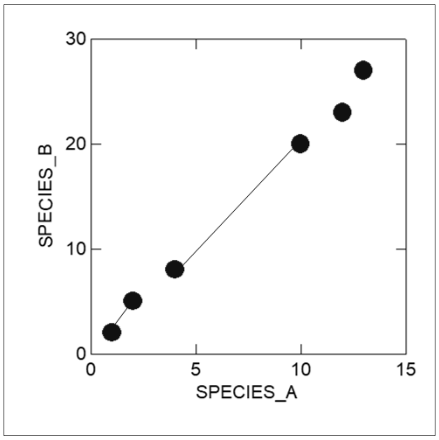
If not ok,

then standardize all data

per sample prior to

analysis (use

standardization (a) above).



Two important points

2. The ”size” of the sampling units must be the same in most cases!

X1	X2	X3		X´1(a)	X´2(a)	X´3(a)
1	2	5		1.2	1.7	3.1
4	8	3		1.6	3.1	1.2
2	5	7		0.9	2.1	2.9
43	94	111		0.9	2.3	2.8
39	77	134		1.1	1.9	3.1

Important note I: for everything said from here on, you can (should?) run analyses on both ”raw” data and on standardized data per variable – results will often differ somewhat...as will the interpretations...

Important note II: Here, I focus on classifying/ordinating samples (how similar are samples with regards to the variables) but one can also do the reverse (how similar are the variables with regards to the samples)...

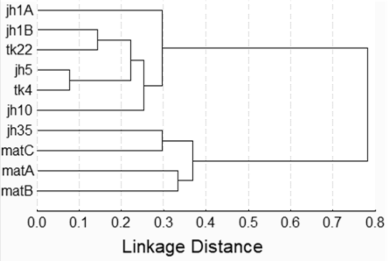
Important note III: The variables can be almost any variables. In biology, often different genes, transcripts, proteins, peptides, species, etc, etc...

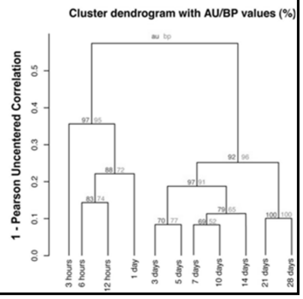
Classification

Typical aim: grouping samples into discrete clusters or classes, so that the data in each cluster share some similarity. These methods are ”guided” by or based on some dis/similarity matrix. Several more or less related methods, but two kinds useful and common....

Classification

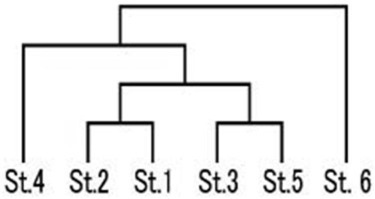
A) **Cluster Analysis** – uses a dissimilarity matrix to produce dendrograms (tree plots) with hierarchical clusters that are built ”bottom-up” (by **UPGMA** - (Unweighted Pair Group Method with Arithmetic mean**)**. Several different options and different measures of dissimilarity can be used. Branch lengths are informative! Common in many fields! Empirical support for bifurcations (i.e., nodes) can be tested with resampling tests.

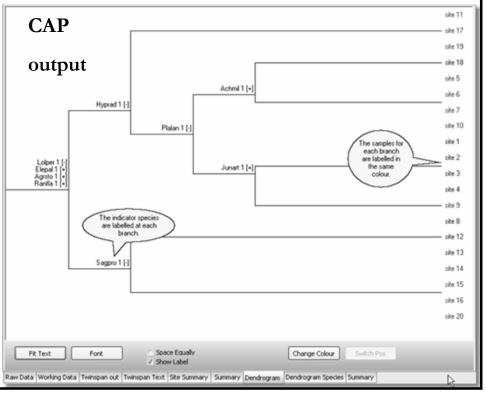




Classification

B) **TWINSPAN**– uses a different method to place samples in classes by building ”top-down”. Very popular in plant ecology, but not so much outside of this domain. Useful for identifying indicator species. Branch lengths not informative.





CAP

output

□ Space Equally
☑ Show Label

Ordination

Typical aim: to (i) reduce multivariate data (for example, many species) into fewer dimensions that can (ii) then be used to plot samples (typically in 2D [=bivariate] space) in order to visualize and seek patterns. Similar samples are close together in such ordination plots, and dissimilar samples are placed far apart!

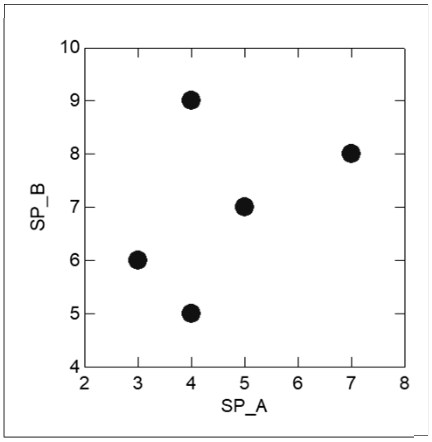
A very large number of more or less different methods available (>10) – only the common and most useful mentioned here...

Ordination

First: reducing data into fewer dimensions – how?

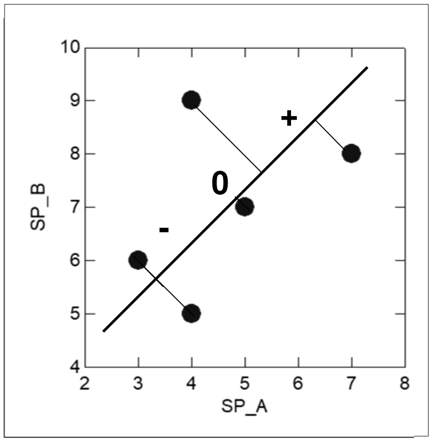
New derived variables (Z_{ij}), that we call principal components or axes or dimensions, are created. Methods differ somewhat in how these derived/latent variables are constructed, but all are ”made up” by our data in one way or another. ”Guided” by either a dissimilarity matrix or the variance-covariance matrix.

Ordination



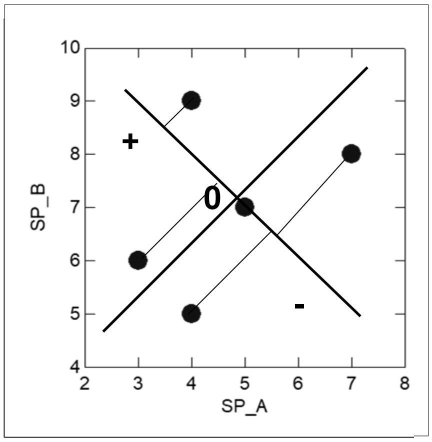
	Sp.A	Sp.B			
A	3	6			
B	5	7			
C	4	5			
D	7	8			
E	4	9			

Ordination



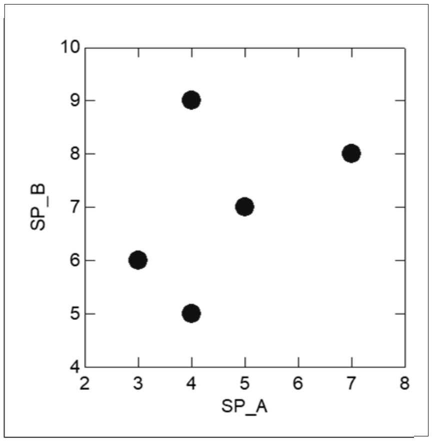
	Sp.A	Sp.B		PC1	
A	3	6		-2.1	
B	5	7		0.1	
C	4	5		-2.1	
D	7	8		1.9	
E	4	9		0.6	

Ordination



	Sp.A	Sp.B		PC1	PC2
A	3	6		-2.1	0.4
B	5	7		0.1	0
C	4	5		-2.1	-0.7
D	7	8		1.9	-0.8
E	4	9		0.6	1.9

Ordination

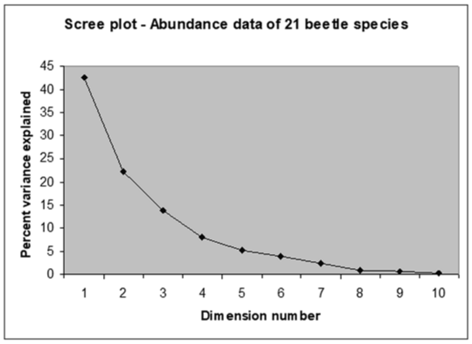


These two new dimensions/axes explain 100% of the abundance of Species A and B in our data...but we need to *reduce* the number of variables!

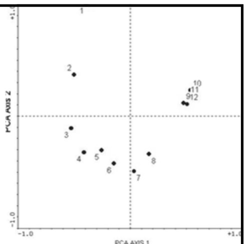
	Sp.A	Sp.B		PC1	PC2
A	3	6		-2.1	0.4
B	5	7		0.1	0
C	4	5		-2.1	-0.7
D	7	8		1.9	-0.8
E	4	9		0.6	1.9

Ordination

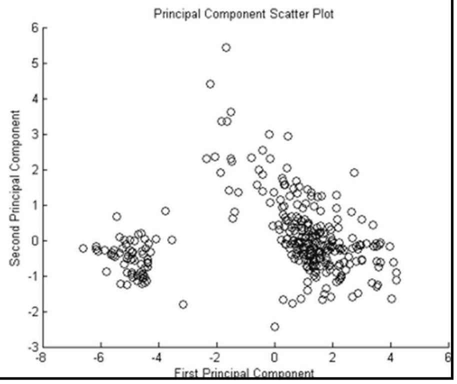
- With multivariate biological data, we can capture most variation in our data (for example species abundance data) with much fewer dimensions/axes – a reduction of dimensions. An example of a ”scree plot”:



Ordination



- PCA – Principal component analysis Is used for many purposes, one being ordination. Dimensions/axes are called ”principal components”. The linear nature of PCA sometime a problem – causes a ”horseshoe” effect...not always useful for ordination...



Ordination

- PCoA - Principal Coordinates Analysis (also called metric multidimensional scaling). As for NMDS, maximizes the correlation between distance measures in matrix and distance in the ordination.

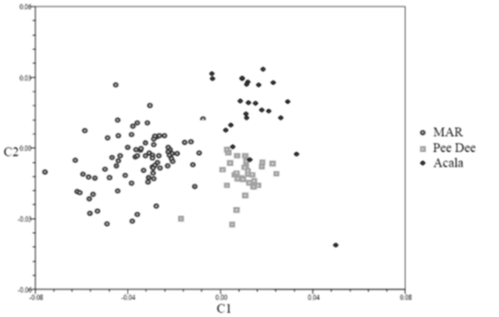
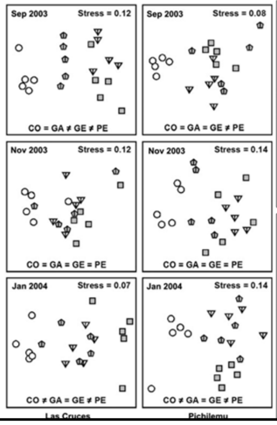
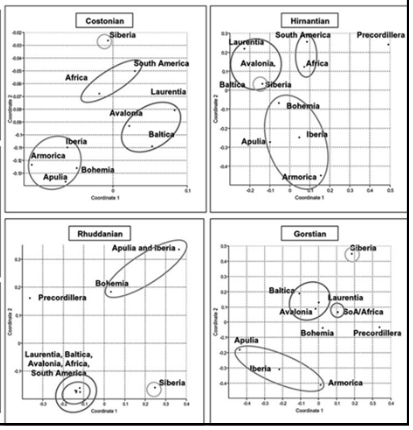


Figure 3. Principal Coordinate Analysis plot of the clustering of the cultivars/germplasm lines of three breeding programs; MAR, Pee Dee, and Acala.

Ordination

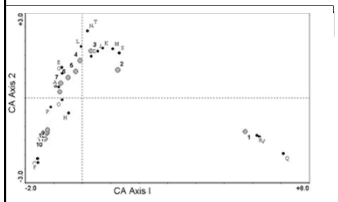
- NMDS – non-metric multidimensional scaling. A very useful and commonly employed method for ordination. Dimensions /axes are called ”coordinates” or ”axes”. Stress: >0.2 poor, <0.1 good, <0.05 excellent

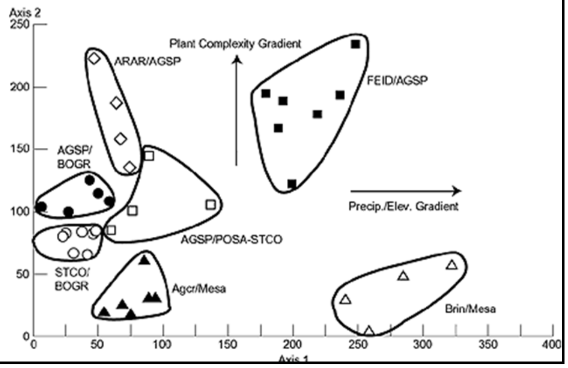




Ordination

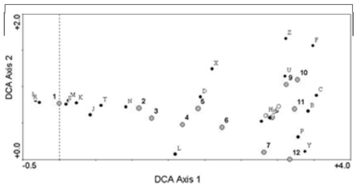
- CA – Correspondence analysis. Useful, but sometimes suffers from distortion – called the ”arch effect”. Dimensions called ”CA axes”.

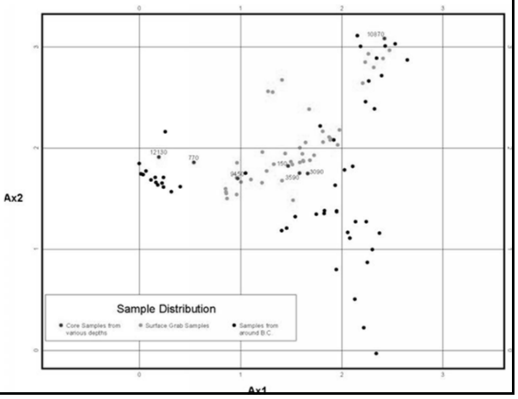




Ordination

- DCA – Detrended correspondence analysis. Very useful, eliminates distortion. Dimenions called ”DCA axes”.





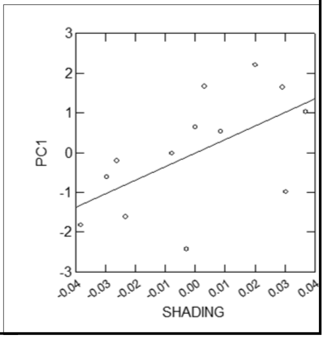
Types of ordination

Very important distinction!

- All of these ordination methods (PCA, NMDS, PCoA, CA, DCA) are ”indirect” – they are naïve to underlying causal (e.g. environmental) variation. Grouping/gradient only based on species composition data.
- Also ”direct” ordination methods, that seek covariation between variable data one one hand and extroneous/environmental/causal variables on the other!

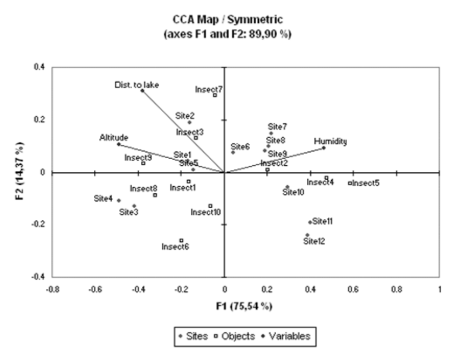
Direct ordination

- ”Simple” linear regression between derived variables (dimensions/axes) and causal variables can be used, but power is typically quite low (not ideal). Better but more complex multivariate methods: Redundancy analysis (RDA), Canonical correlation analysis, Canonical correspondence analysis (CCA), Detrended canonical correspondence analysis, Partial least squares analysis...



Direct ordination

- Canonical correspondence analysis (CCA) – in essence: produces a bivariate plot where both the samples and the “reponse” variables are ordinated and the “explanatory” variables (e.g., environmental) are represented by vectors. Tests for an association between e.g. species-environment by means of a resampling test!



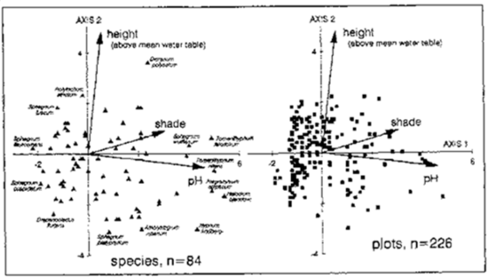
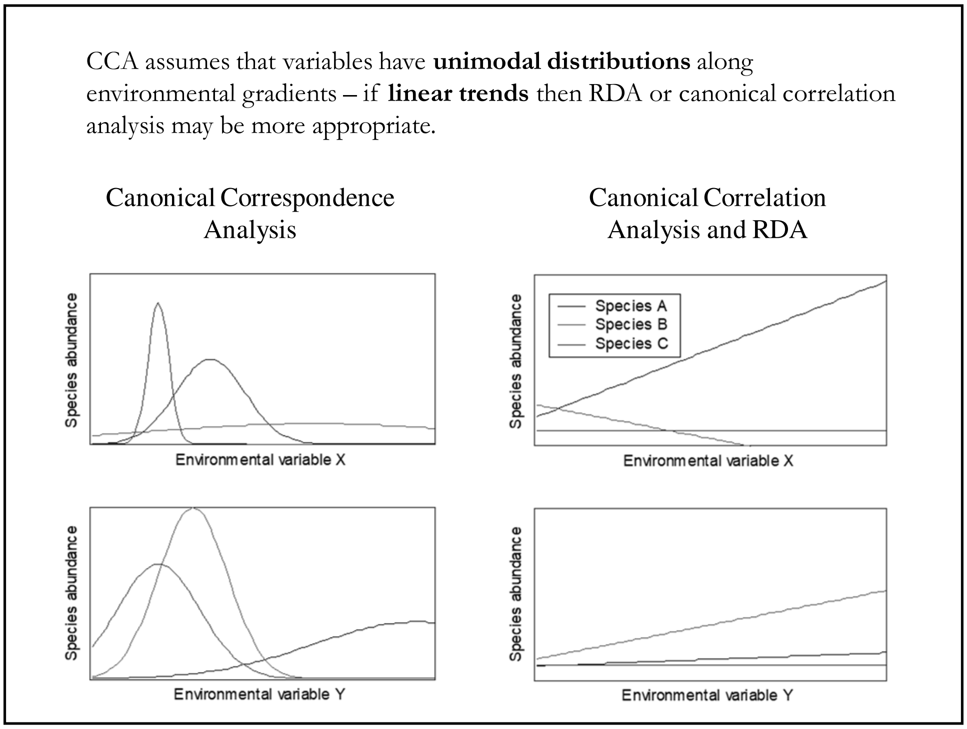


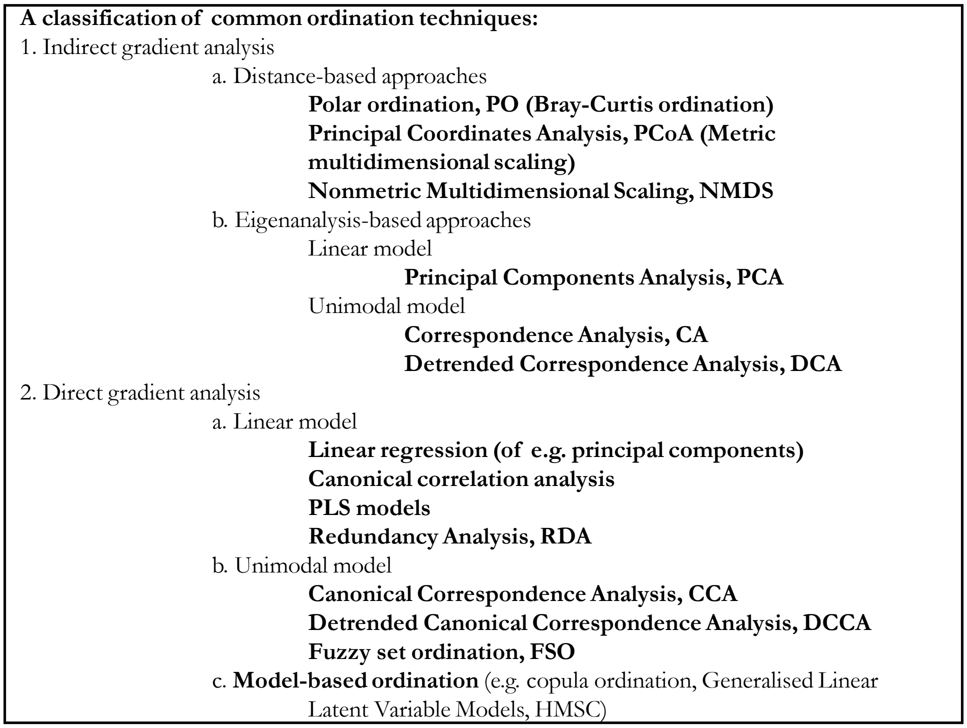
Figure 1. CCA species-environment and sample-environment biplots. The environmental variables are represented by arrows. (Eigenvalues of canonical (constrained) axes: λ_1=0.5474, λ_2=0.3946, and λ_3=0.2333.)



CCA assumes that variables have **unimodal distributions** along environmental gradients – if **linear trends** then RDA or canonical correlation analysis may be more appropriate.

Canonical Correspondence Analysis

Canonical Correlation Analysis and RDA



A classification of common ordination techniques:

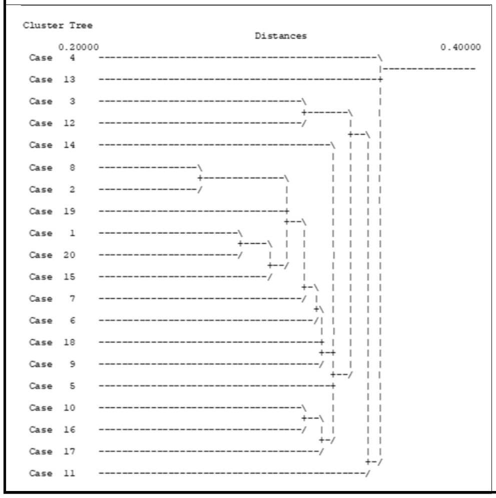
1. Indirect gradient analysis
   - a. Distance-based approaches
     - **Polar ordination, PO (Bray-Curtis ordination)**
     - **Principal Coordinates Analysis, PCoA (Metric multidimensional scaling)**
     - **Nonmetric Multidimensional Scaling, NMDS**
   - b. Eigenanalysis-based approaches
     - Linear model
       - **Principal Components Analysis, PCA**
     - Unimodal model
       - **Correspondence Analysis, CA**
       - **Detrended Correspondence Analysis, DCA**
2. Direct gradient analysis
   - a. Linear model
     - **Linear regression (of e.g. principal components)**
     - **Canonical correlation analysis**
     - **PLS models**
     - **Redundancy Analysis, RDA**
   - b. Unimodal model
     - **Canonical Correspondence Analysis, CCA**
     - **Detrended Canonical Correspondence Analysis, DCCA**
     - **Fuzzy set ordination, FSO**
   - c. **Model-based ordination** (e.g. copula ordination, Generalised Linear Latent Variable Models, HMSC)

Some final notes...

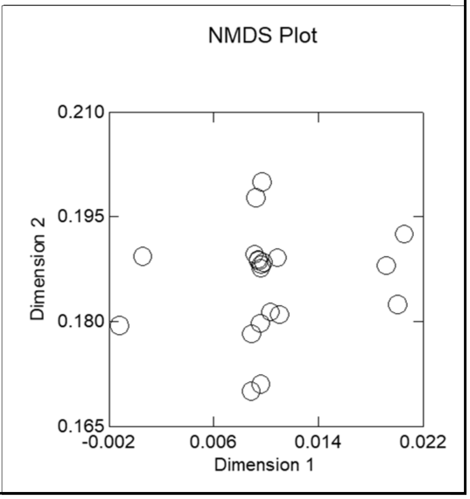
- Cluster analysis (mostly) and ordination typically used for *exploratory data analysis* (**not meant for hypothesis testing !**) - are useful tools for description and visualization... This said, some null hypotheses can be tested with appropriate resampling tests in direct ordinations.

To illustrate this point!

- Data for abundance of 20 species at 20 sites...



```
Cluster Tree
                                        Distances
      0.20000                                                            0.40000
 Case   4    -------------------------------------------------\
                                                              |----------------
 Case  13    -------------------------------------------------+
                                                              |
 Case   3    ------------------------------------\            |
                                                 +-------\    |
 Case  12    ------------------------------------/       |    |
                                                         +--\ |
 Case  14    -----------------------------------------\  |  | |
                                                      |  |  | |
 Case   8    -----------------\                       |  |  | |
                              +---------------\       |  |  | |
 Case   2    -----------------/               |       |  |  | |
                                              |       |  |  | |
 Case  19    ---------------------------------+       |  |  | |
                                              +--\    |  |  | |
 Case   1    -------------------------\       |  |    |  |  | |
                                      +----\  |  |    |  |  | |
 Case  20    -------------------------/    |  |  |    |  |  | |
                                           +--/  |    |  |  | |
 Case  15    ------------------------------/     |    |  |  | |
                                                 +-\  |  |  | |
 Case   7    ------------------------------------/ |  |  |  | |
                                                   +\ |  |  | |
 Case   6    --------------------------------------/| |  |  | |
                                                    | |  |  | |
 Case  18    ---------------------------------------+ |  |  | |
                                                    +-+  |  | |
 Case   9    ---------------------------------------/ |  |  | |
                                                      +--/  | |
 Case   5    -----------------------------------------+     | |
                                                      |     | |
 Case  10    ------------------------------------\    |     | |
                                                 +--\ |     | |
 Case  16    ------------------------------------/  | |     | |
                                                    +-/     | |
 Case  17    ---------------------------------------/       | |
                                                            +-/
 Case  11    -----------------------------------------------/
```



Some final notes…

- Reading: Chapter 18 in course book and a lot on the internet – see “Ordination methods” at http://ordination.okstate.edu/ (*very* useful site for anyone interested in ordination, especially biologists – lots of resources!)
- Cluster analysis and ordination can be made with most general software packages and special software (primarily Canoco, TWINSPAN, CAP, PC-ORD, NTSYS and a few others) – http://ordination.okstate.edu/ for a list. Many packages in R! (vegan, and many, many others)