Correspondence Analyses

- *Objectives:*

Compare Direct and Indirect Ordination Methods

Introduce CA - Correspondence Analysis

Introduce DCA - Detrended Correspondence Analysis

Introduce CCA - Canonical Correspondence Analysis

Briefly Discuss RDA - Redundancy Analysis

Common ordination techniques

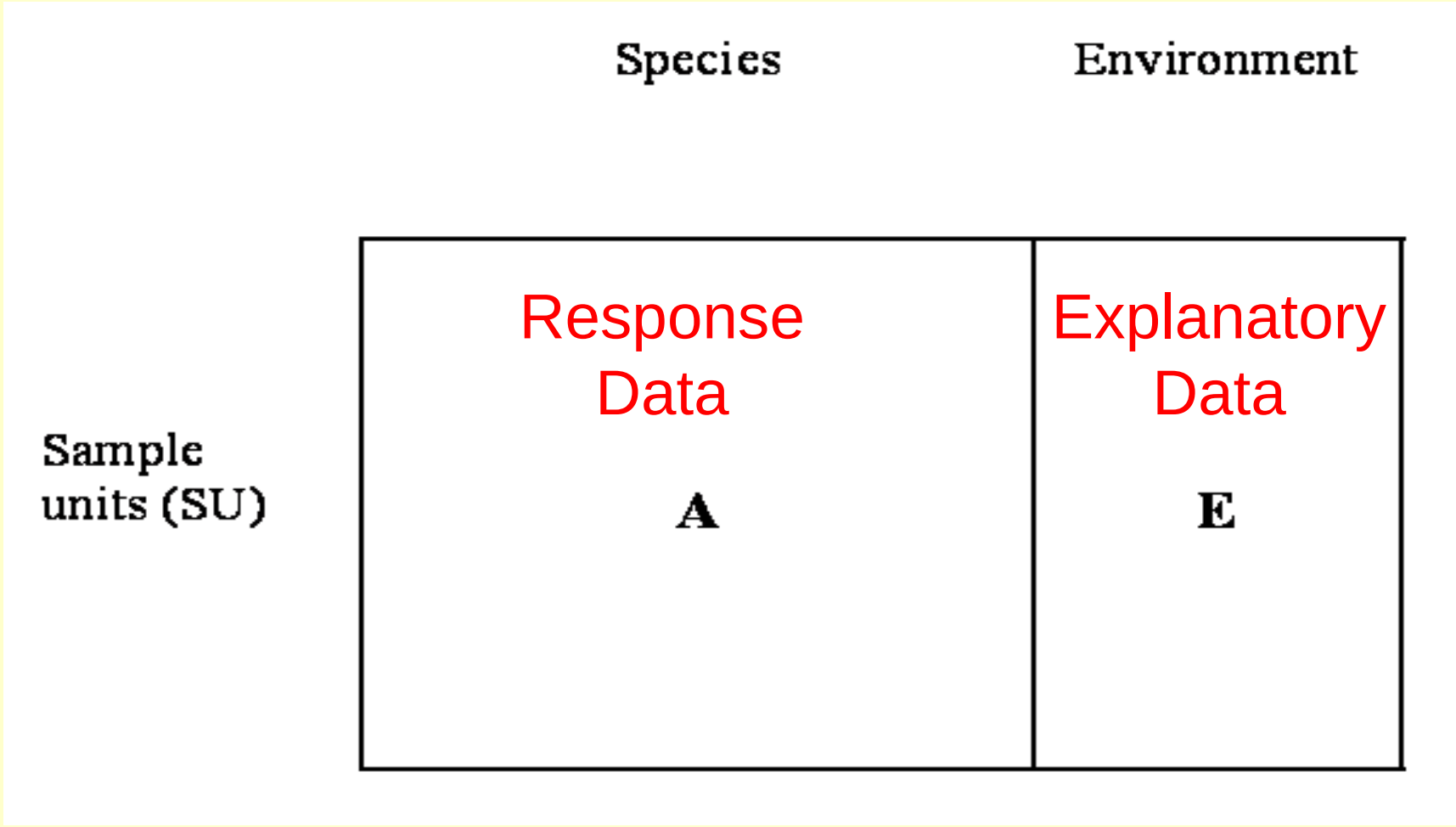
- Dichotomy between indirect / direct gradient analysis (Gauch 1982, ter Braak & Prentice 1988):

- Indirect gradient analyses use only species sample. If there is information about the environment, it is used after indirect gradient analysis, as an interpretative tool.
- *When we perform an indirect analysis, we are asking the species to define the most important gradients.*

- Direct gradient analysis uses external environmental data in addition to the species data.
- *When we perform a direct analysis, we are asking if the species composition is related to the environmental variables measured concurrently.*

Unconstrained (Indirect) Ordination

- Order “Species” Records into Fewer Dimensions
- Correlate Resulting Axes with Environmental Data



Common Indirect ordination techniques
(ter Braak and Prentice 1988)

- Indirect Gradient Analysis

Distance-based approaches

Polar ordination, PO (Bray-Curtis ordination)

Nonmetric Multidimensional Scaling, NMDS

Eigenanalysis-based approaches

Linear model

Principal Components Analysis, PCA

Unimodal model

Correspondence Analysis, CA

Detrended Correspondence Analysis, DCA

Correspondence Analysis (CA) – Concept

Conceptually similar to PCA, but applies to categorical rather than continuous data. It provides a means of displaying or summarizing a set of data in fewer independent dimensions.

Correspondence analysis performed on a contingency table.

Creates orthogonal components and, for each item in the table, a set of scores (called factor scores).

CA decomposes chi-squared statistic associated with table into orthogonal factors. Treats rows / columns equivalently.

All data should be nonnegative (≥ 0) and on the same scale.

Applied to species counts – via CCA and DCCA.

Correspondence Analysis (CA) - Background

- A Brief History:
  - Popular in 1970s – multi-D gradients (species / plots)
  - Formerly known as Reciprocal Averaging (RA)
- Limitations:
  - Distance measure is Chi-square
  - Echoes: Provides multi-D patterns, but often unreal
  - Distortion: Stretches samples at edges of gradients
- When to Use?

NEVER

> Because of the faults described above, there should be no regular application of CA to ecological community data. It is, however, conceivably useful if (a) you need to superimpose species space and sample unit space, (b) your system is essentially one-dimensional, and (c) you consider chi-square distances conceptually appropriate (see below). Usefulness of CA, as compared to distribution-free techniques such as NMS, has not been argued convincingly in the literature.

(McCune & Grace 2002)

Detrended Correspondence Analysis (CA)

- A Brief History:
  - Popular in 1970s – multi-D gradients (species / plots)
  - Formerly known as DECORANA
- Limitations:
  - Distance measure is Chi-square
  - Not robust – compared with NMDS (Minchin 1987)
  - Sample sequence affects results (Tausch et al. 1995)
- When to Use?

NEVER

> DCA unnecessarily imposes assumptions about the distribution of sample units and species in environmental space. Other methods, especially nonmetric multidimensional scaling, perform as well or better without making those assumptions. There is no need to use DCA.

(McCune & Grace 2002)

Constrained (Direct) Ordination

- The final ordination scores influenced by two factors:
  (i) the patterns inherent in the response data and
  (ii) the relationship of response data to explanatory variables

"the dough":
the ecological patterns inherent in response data

"cookie cutter":
relationship of community to explanatory variables

Implication: CCA ignores community patterns not related to environmental variables in Second Matrix

Common Direct ordination techniques
(ter Braak and Prentice 1988)

- Direct Gradient Analysis

Linear model

Redundancy Analysis, RDA

Unimodal model

Canonical Correspondence Analysis, CCA

Detrended Canonical Correspondence Analysis, DCCA

CCA – Why is it special ?

- CCA is a widely used ordination tool in ecology, but it is complex and limited by assumptions that can be difficult to meet with biotic community (species) datasets.

- Thus, CCA is a tool for more advanced users.

- Yet, you are likely to encounter it in literature.

- Why Discuss Today?

  - Illustrates 'constrained' ordination.

  - Gateway into group comparisons.

CCA – Why is it special ?

- CCA allows simultaneous ordering of sample units and species, based on linear relationship between unimodal patterns of redundant co-occurrence in species data and explanatory variables.

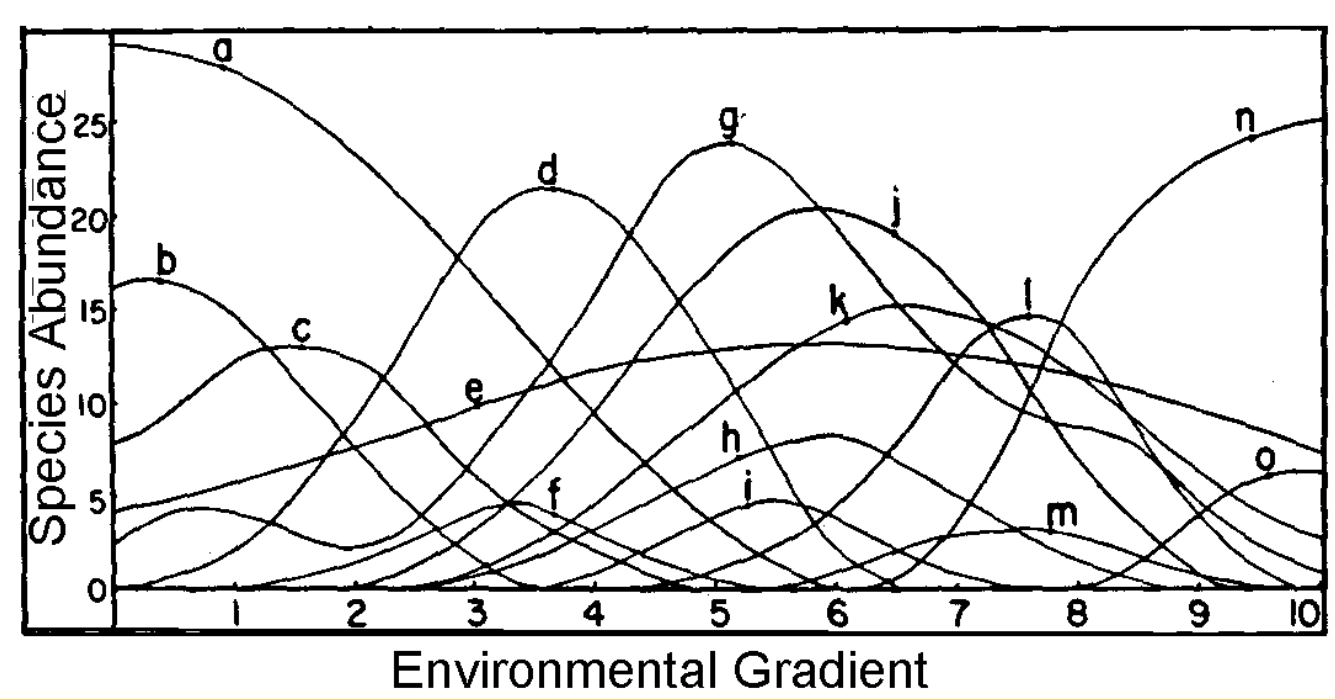
Linear relationships:

Unimodal patterns:

Redundant co-occurrence:

CCA – Assumptions

- Redundant Co-occurrence: Cross-correlated, Multi-species
- Linear Relationships of Species X Environment
- Unimodal Species Distributions



CCA – Assumptions

- Because of these assumptions, CCA is applicable to a limited set of scenarios and ecological questions

Table 21.1. Questions about the community (**A**) and environmental or experimental design (**E**) matrices that are appropriate for using CCA.

Matrix relationships	Questions for which CCA is OK	Questions for which CCA is not OK
A only	Not applicable.	What are the strongest gradients in species composition?
H_o: no linear relationship between **A** ← → **E**	Are any aspects of community structure related to these environmental variables?	Are the strongest community gradients related to these environmental variables?
Describe **A** ← → **E**	How is the community structure related to these environmental variables?	How are the strongest gradients in community structure related to these environmental variables?

CCA – Is it For You ?

- What you need

"Small" Second Matrix consisting of one to several explanatory variables that are not strongly cross-correlated but are associated with the patterns in your species data.

These are called "design variables" (i.e., environmental conditions that vary among samples due to sampling).

NOTE: Coding variables (e.g., water-mass, island) need to be converted into quantitative (Q) variables, since categorical (C) variables are not considered by the analysis.

Remember: Your choice of CCA implies acceptance of *Chi-square* distance measure: this implies influence of common species de-emphasized while rare species emphasized.

CCA – How does it Work ?

Canonical correspondence begins with a set of arbitrary sample unit scores, then calculates species scores as weighted averages of these sample unit scores.

Next, CCA calculates new sample unit scores as weighted averages of species scores in an iterative refinement process.

These sample unit scores are known as the 'WA scores'.

This is followed by a weighted least-squares multiple regression of the sample unit scores on the explanatory variables in the Second Matrix.

The regression coefficients are used to fit new sample unit scores (linear combinations), known as the 'LC scores'.

This process is repeated until LC scores do not vary … (until we have arrived at a robust result - “stable”)

CCA – A Hybrid Approach

- Approach:

WA scores map sample unit coordinates in species space (i.e., gradient obtained through an unconstrained ordination).

LC scores are strongly constrained by explanatory variables and represent a direct gradient.

- Implication:

Therefore, LC scores interpret ordination scores for samples in 'environmentally-shaped species space'.

Because both species and environmental data are used, the multivariate space is not defined solely by either.

Yet, the greater influence of the explanatory variables means it is closer to environmental space than species space.

Weighted Averaging – Calculating Weights

- First, uses arbitrary sample weights to score species

$$w_j = \frac{\sum_{i=1}^{n} a_{ij} v_i}{\sum_{i=1}^{n} a_{ij}}$$

- Then, uses species weights to revise sample weights

$$v_i = \frac{\sum_{j=1}^{p} a_{ij} w_j}{\sum_{j=1}^{p} a_{ij}}$$

CCA – An Example

- Imagine an industry is dumping pollutants in a river, and we want to ascertain whether there is a change in community composition.

- However, we did not collect baseline data (before the impact).

- Because replication is impossible (there is only one river), the key issue is determining a link between the community structure and the spatial gradient in water quality properties downstream from the pollution site.

CCA – Results

- Examine correlations among environmental variables:
  - Pollutant Concentration: LogPoll
  - Two Other Unrelated Environmental Variables: Var2 & Var3

	LogPoll	Var2	Var3
LogPoll	1	0.107	-0.119
Var2	0.107	1	-0.039
Var3	-0.119	-0.039	1

CCA – Results

- Stability of the LC result with iterations.

```
ITERATION REPORT
-----------------------------------------------------------------
Calculating axis 1
Residual =   0.53E+04 at iteration   1
Residual =   0.96E-01 at iteration   2
Residual =   0.47E-01 at iteration   3
Residual =   0.19E-01 at iteration   4
Residual =   0.84E-02 at iteration   5
Residual =   0.43E-02 at iteration   6
Residual =   0.24E-02 at iteration   7
Residual =   0.14E-02 at iteration   8
Residual =   0.88E-03 at iteration   9
Residual =   0.54E-03 at iteration  10
Residual =   0.46E-05 at iteration  20
Residual =   0.40E-07 at iteration  30
Residual =   0.34E-09 at iteration  40
Residual =   0.30E-11 at iteration  50
Residual =   0.69E-13 at iteration  58
Solution reached tolerance of 0.100000E-12 after  58 iterations.
-----------------------------------------------------------------
Calculating axis 2
Residual =   0.20E+01 at iteration   1
Residual =   0.30E-03 at iteration   2
```

etc....

CCA – Results

- Axis summary statistics.

	Axis 1	Axis 2	Axis 3
Eigenvalue	0.636	0.044	0.016
Variance in species data			
% of variance explained	14.4	1.0	0.4
Cumulative % explained	14.4	15.4	15.8
Pearson Correlation, Spp-Envt	0.900	0.307	0.213
Kendall (Rank) Corr., Spp-Envt	0.717	0.167	0.158

CAA – Results

- Sample unit scores from species scores (WA scores)
- Raw data totals (weights) are also given

	WA scores			Raw Data
	Axis 1	Axis 2	Axis 3	Totals
Site1	1.298381	1.555888	-0.98204	1131
Site2	1.17872	1.19812	-1.26412	1000
Site3	0.808255	0.145479	-1.10749	721
Site4	0.335053	-1.16647	-0.34654	635
Site5	0.204182	-1.40531	-0.05847	735
...				
Site99	-1.15441	0.044354	0.100729	580
Site100	-1.31167	-0.45384	-0.23881	748

CAA – Results

- Sample unit scores are linear combinations of environmental variables.
- These are the LC Scores.

	Axis 1	Axis 2	Axis 3
Site1	0.857	0.213	0.012
Site2	0.423	-0.100	-0.103
Site3	0.646	0.103	-0.024
Site4	0.474	-0.238	-0.104
Site5	-0.107	-0.297	0.078
...			
Site99	-0.807	0.159	0.147
Site100	-1.405	0.011	-0.084

CAA – Results

- Correlations of environmental variables with ordination axes.

  "interset correlations" are correlations of environmental variables with the WA scores.

  "intraset correlations" are correlations of environmental variables with the LC scores.

CAA – Results

- Biplot scores and correlations for environmental variables with ordination axes. Biplot scores used to plot vectors in the ordination space.

- Two kinds of correlations shown: interset and intraset.

WA – environmental correlation w samples

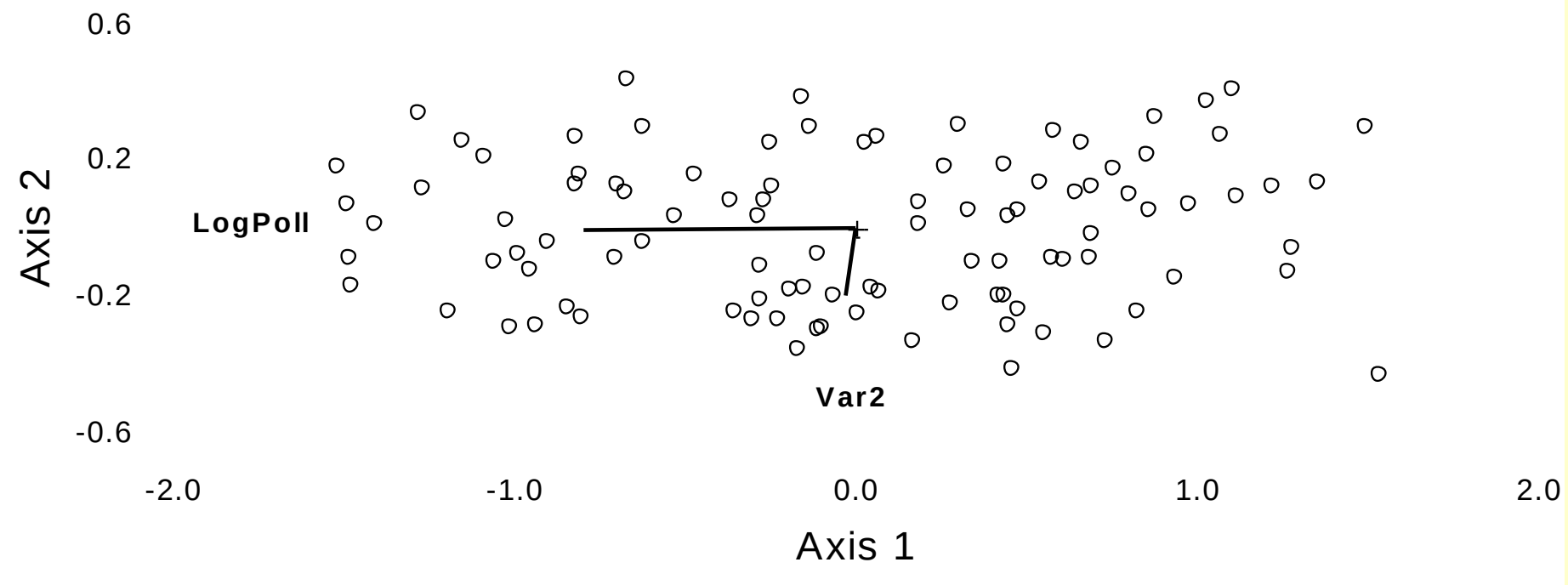
LC – environmental correlation w gradients

Variable	Axis 1	Axis 2	Axis 3
BIPLOT scores			
LogPoll	-0.797	-0.008	0.002
Var2	-0.028	-0.196	0.045
Var3	0.073	0.081	0.115
INTRASET correlations			
LogPoll	-0.999	-0.038	0.018
Var2	-0.035	-0.933	0.357
Var3	0.092	0.386	0.918
INTERSET correlations			
LogPoll	-0.899	-0.012	0.004
Var2	-0.032	-0.286	0.076
Var3	0.083	0.118	0.195

CAA – Results

- Environmental variables represented as lines radiating from the centroid of the ordination (point 0, 0).

- Biplot scores give coordinates of the tips of radiating lines.



CAA – Results

- Montecarlo significance tests (randomizations):

$H_{o\ eigenvector}$: No structure in main matrix.

For this hypothesis, elements in the main matrix are randomly reassigned *within* columns.

$H_{o\ spp\text{-}env}$: No linear relationship between matrices.

For this hypothesis, the rows in the second matrix are randomly reassigned within the second matrix.

CAA – Results

- Montecarlo significance tests (999 runs):

Axis	Real data	Randomized data			*p*
		Mean	Minimum	Max.	
	Eigenvalue				
1	0.636	0.098	0.033	0.217	0.001
2	0.044	0.046	0.009	0.112	
3	0.016	0.020	0.004	0.076	
	Spp-Envt Corr.				
1	0.900	0.378	0.224	0.553	0.001
2	0.307	0.287	0.155	0.432	
3	0.213	0.218	0.107	0.396	

CCA – Pros / Cons

➢ Advantages:

- Focuses on species pattern driven by environmental data
- Allows for test of multiple linear gradients
- Provides predictive capabilities (via the LCs)

➢ Disadvantages:

- Can only include few environmental variables to constrain ordination (maximum: “number of samples” – 1)
- Several assumptions (linear, unimodal), complex methods
- Problematic distance measure (Chi-squared)

CCA - Recommendations

- Tricky ordination method and should be used with caution.

- Ability to focus on environmental – species pattern, while ignoring other environmental variability is attractive.

- Because results depend upon input from a Second Matrix of variables, CCA can only predict species responses when also measuring those explanatory variables.

For example: If you have hypothesized that the community composition will shift in a certain way as a response to a linear change in a particular suite of explanatory variables, you can use CCA to see if the hypothesized pattern is observed.

- NOTE: 2 Tests: main matrix & spp - env relationship.

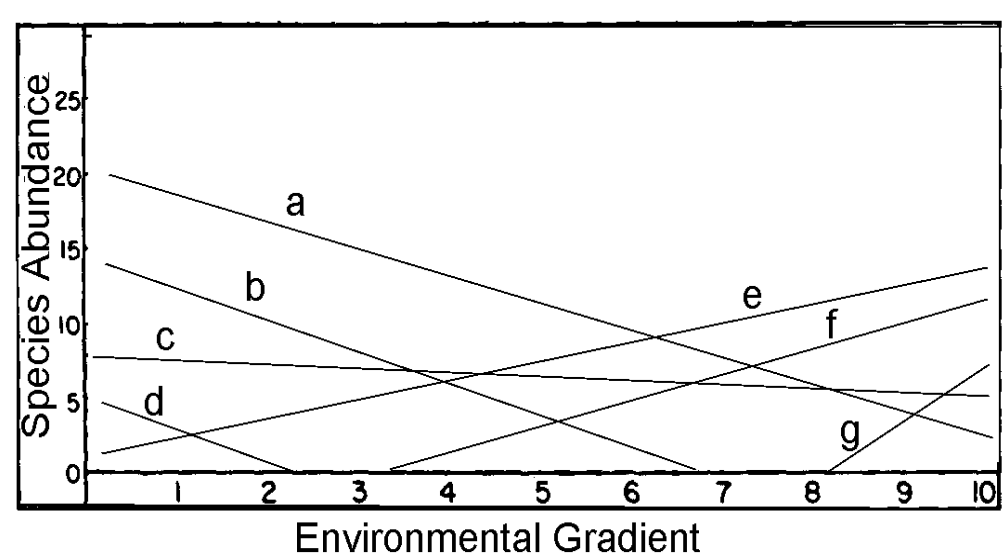
Redundancy Analysis (RDA)

- New(er) MVS approach(PC-ORD 6.19); not much literature.
- RDA applies to the same conceptual problem as canonical correspondence analysis (CCA).
- Both RDA and CCA model a set of response variables (species abundances) as function of a set of predictor (environmental) variables based on linear models.
- But:
    - RDA assumes interactions among response variables and between response / predictor variables.

Redundancy Analysis (RDA)

- Because RDA assumes that response variables are linearly related to predictors, it is not generally suitable for community data (McCune & Grace 2002).

- RDA suitable where linear relationships among response variables and between response and predictor variables expected / demonstrated.



- NOTE: If unsure whether data show linear relationships, use scatterplots to evaluate if linear approximation reasonable.

RDA – How to Implement ?

- Look for linear relationships between response variables and predictors. Note that a linear model is not good if you see strongly hump-shaped, asymptotic or exponential shapes.

  Note: If you see those shapes but wish to use RDA, consider decreasing range of measurements to improve linearity.

- RDA is conceptually, ideal for “short” gradients (did not capture unimodal distribution, but only linear change)

- RDA represents species and environmental variables using arrows. It is best to represent the two sets of arrows in two separate figures for ease of display.

- RDA can use 'species' measured in different units.

  If so, the data must be centered and standardized.

RDA – How to Implement it ?

- PC-ORD implements RDA using sequential multiple regression and PCA. Calculates the full solution – for all axes.

Basic steps of RDA (following Legendre & Legendre 1998) are:

• 1) Center response variables (main matrix) and predictors (second matrix). In other words, subtract the mean of each column from each element in the column. (Note: the response variables can also be standardized, that is expressed as standard deviations away from the column mean).

• 2) Regress each response variable against all predictors using multiple linear regression with simultaneous entry of the predictors.

• 3) Calculate fitted values for each response variable and each sample unit by applying the regression equations to the input data for the predictors.

• 4) Calculate a covariance matrix (or alternatively a correlation matrix, if the variables are standardized) among the fitted values.

• 5) Conduct PCA on covariance matrix.

RDA – How to Implement it ?

• 6) This extracts the strongest linear patterns in fitted values in orthogonal axes. Number of axes extracted is the minimum of the following:

  – number of columns in main (response) matrix (1)
  – number of predictors
  – sample size – 1.

• 7) Express **eigenvalues** as proportion of total variance in original response matrix. The **eigenvectors** contain **scores for the response variables**.

• 8) Calculate scores for sites (sample units) in space of the response variables by multiplying original response matrix by eigenvectors. These are called simply **site scores** or **scores in response variable space**

• 9) Calculate **fitted site scores** for sites (sample units) in space of the predictors by multiplying fitted response matrix by eigenvectors.

• 10) Calculate the correlation between the two sets of scores.

• 11) Calculate **canonical coefficients** that express weight of predictors in calculation of the fitted site scores.

• 12) Calculate biplot scores, depending on choice of axis scaling options.

RDA - Recommendations

- Tricky ordination method and should be used with caution.

- Ability to focus on environmental – species pattern, while ignoring other environmental variability is attractive.

- Because results depend upon input from a Second Matrix of variables, RDA can only predict species responses when varying precisely those explanatory variables.

- 3 randomization tests available:

- Eigenvalues for individual axes (like PCA).
- Overall relationship between predictor / response matrices.
- Correlation between fitted scores and scores in response variable space. Quantifies species-environment correlation.

References

Gauch, H. G., Jr. 1982. Noise reduction by eigenvalue ordinations. Ecology 63:1643-1949.

Legendre, P. & L. Legendre. 1998. Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. Adv. Ecol. Res. 18:271-313.