Correspondence analysis

Jens C. Frisvad
Department of Systems Biology
DTU

Correspondence analysis

- = Dual scaling = Reciprocal averaging

Chi-square value

Chi-square distance (can be used in cluster analysis)I

Frequency weighted Chi-square distance:
(the latter used for correspondence analysis)

Chi-square value

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Chi-square distance

(can be used for for clustering)

$$d_{ij} = \sqrt{\sum^{k} \frac{(\frac{x_{ki}}{x_{\cdot i}} - \frac{x_{kj}}{x_{\cdot j}})^2}{x_k}}$$

Frequency weighted Chi-square distance
(between two columns i and j)

$$d_{ij} = \sqrt{\sum^{k} \left(\frac{f_{ki}}{f_{\cdot i}} - \frac{f_{kj}}{f_{\cdot j}}\right)^2 \Big/ f_{k\cdot}}$$

Correspondence analysis

$$f_{ij} = \sqrt{f_{i\cdot}}\, f_{\cdot j}\left(1 + \sum^{k} \sqrt{\lambda_k}\, \Psi_{ik} \Phi_{ik}\right)$$

- One fits the model above to the observed data, where f_{ij} are the observed relative frequences (x_{ij} /x..)
- $f_{i\cdot}$ and $f_{\cdot j}$ are the relative frequences of the rows and columns
- λ_k is the k*th* eigenvalue
- The latter two greek letters are elements of the row and column factor matrices

CA factors

- The CA factors describe the patterns of deviations from independence
- If the rows and columns were perfectly independent then the eigenvalues would all be zero
- Output matrices of factor coordinates rather than the factors themselves:

$$\hat{\Psi}_{ik} = \sqrt{\lambda_k}\,\Psi_{ik} \qquad \hat{\Phi}_{jk} = \sqrt{\lambda_k}\,\Phi_{jk}$$

Results given in CA

- Trace of matrix
- Chi-square distances
- Eigenvalues (principal inertias)
- Relative frequencies (for row and column variables)
- Scores
- Loadings
- Biplots are often used to see the correspondence between objects and variables (and object to object, variable to variable)

Primary uses of correspondence analysis

- Contingency tables
- Presence/absence or abundance data
- Unimodal data (bell shaped dependencies): minimum, optimum, maximum

Unimodal dependencies

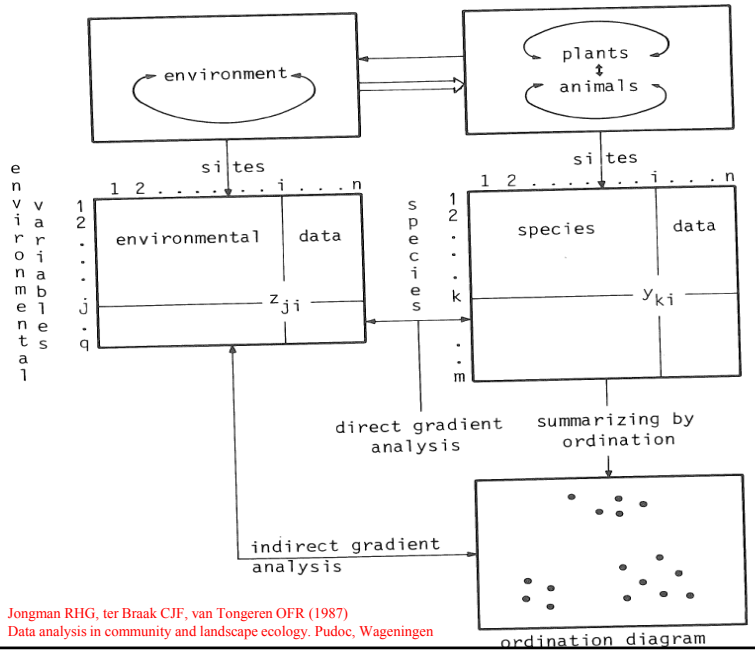
- Examples:
- Enzyme activity and pH
- Bacterial growth and temperature
- Fungal growth and water activity (skewed, however)
- Number of *Bellis* along a pH gradient

Gaussian logit model:

$$\log \frac{p}{1-p} = b_0 + b_1 x + b_2 x^2$$

p = probability of occurrence for binary data

Ordination in community ecology



Jongman RHG, ter Braak CJF, van Tongeren OFR (1987)
Data analysis in community and landscape ecology. Pudoc, Wageningen

Jongman RHG, ter Braak CJF, van Tongeren OFR (1987) Data analysis in community and landscape ecology. Pudoc, Wageningen

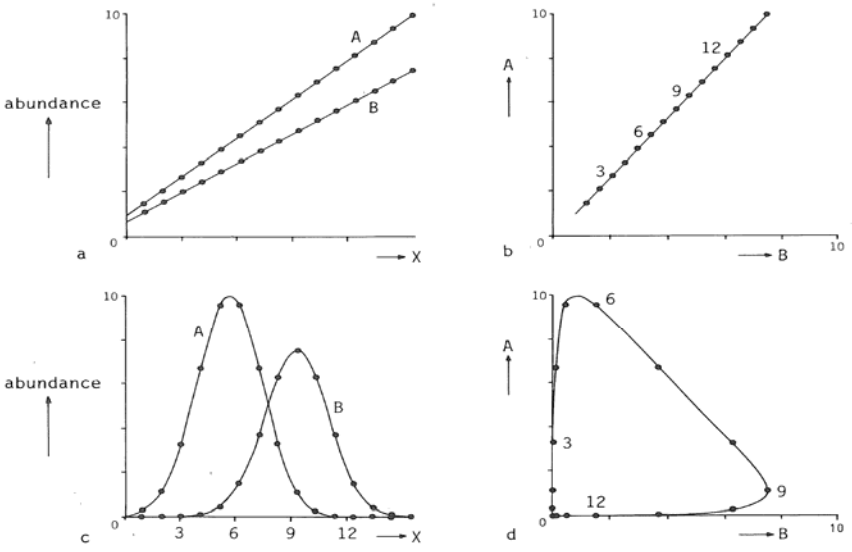
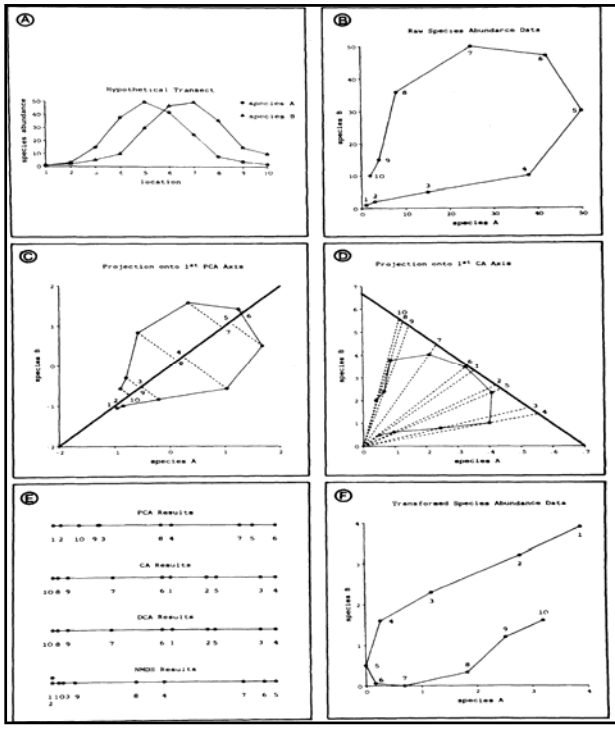
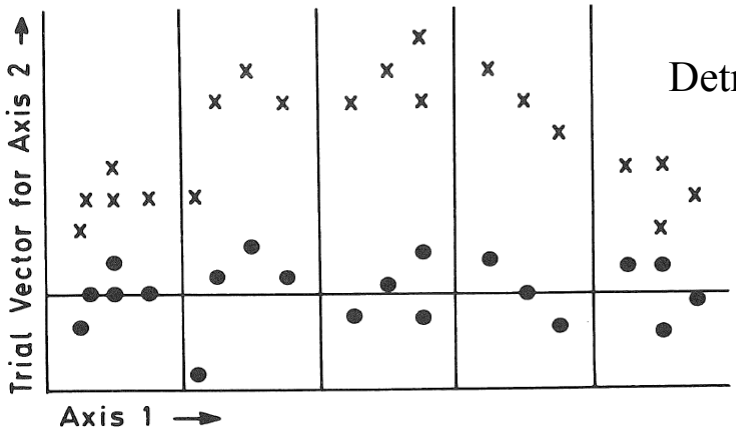


Figure 5.2 Response curves for two species A and B against a latent variable x (a, c) and the expected abundances of the species plotted against each other (b, d), for the straight line model (a, b) and a unimodal model (c, d). The numbers refer to sites with a particular value for x. The ordination problem is to make inferences about the relations in Figures a and c from species data plotted in Figures b and d.



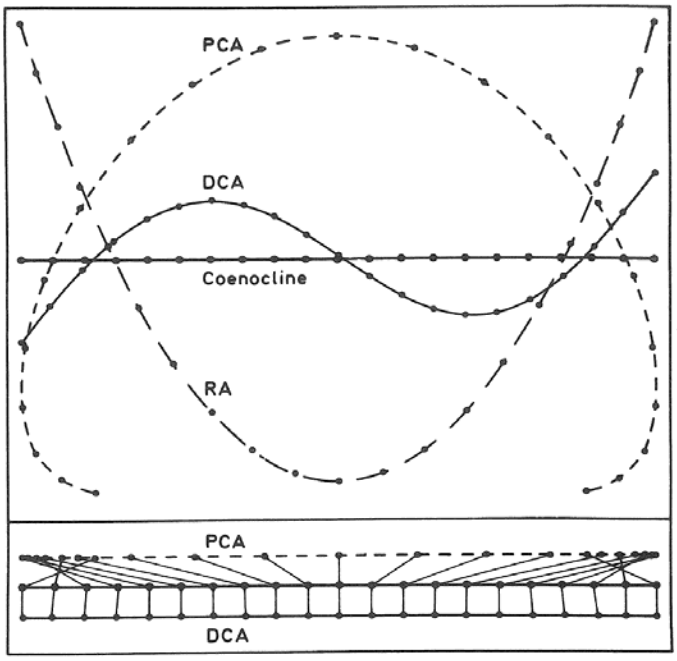
Wartenberg D, Ferson S, Rohlf FJ (1987) Putting things in order: a critique of detrended correspondence analysis. American Naturalist **129**: 434-448.



Gauch HG Jr. (1982) Multivariate analysis in Community ecology, Cambridge University Press, Cambridge.

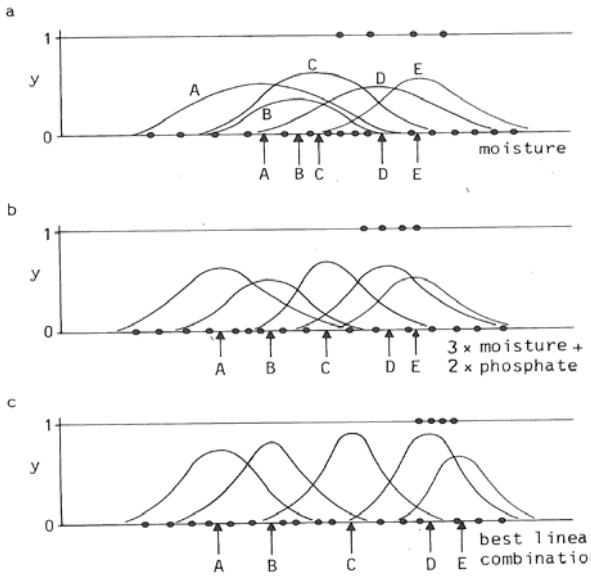
Figure 4.12. Method of detrending used in detrended correspondence analysis. The (×) indicate sample scores before detrending; (●), sample scores after detrending. The first axis is divided into five segments and a local mean is subtracted in each. Note that after detrending there is virtually no systematic relation between the trial vector for axis 2 scores and the axis 1 scores. The computer program for detrended correspondence analysis uses a larger number of segments and a sophisticated running averages procedure, but the essential concept of detrending is displayed by this simplified diagram. (From Hill & Gauch 1980:48)

A simulated coenocline: PCA, CA (RA = reciprocal averaging) and detrended CA



Gauch HG Jr. (1982) Multivariate analysis in community ecology, Cambridge University Press, Cambridge.

Canonical correspondence analysis (CCA)



Jongman RHG, ter Braak CJF, van Tongeren OFR (1987) Data analysis in community and landscape ecology. Pudoc, Wageningen

Figure 5.18 Artificial example of unimodal response curves of five species (A-E) with respect to standardized environmental variables showing different degrees of separation of the species curves. a: Moisture. b: Linear combination of moisture and phosphate, chosen a priori. c: Best linear combination of environmental variables, chosen by CCA. Sites are shown as dots, at $y = 1$ if Species D is present and at $y = 0$ if Species D is absent.

CA for classification

Christensen M, Frisvad JC, Tuthill D (1999)
Taxonomy of the *Penicillium miczynskii* group
Based on morphology and secondary metabolites.
Mycol Res **103**: 527-541.

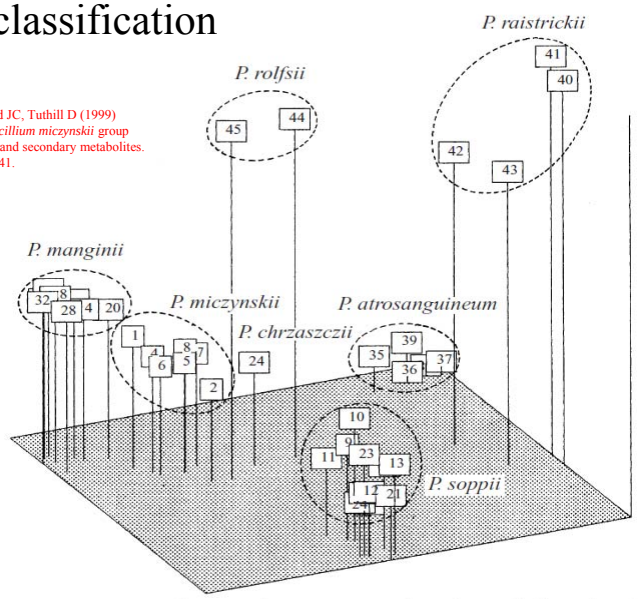


Fig. 4. Correspondence analysis using combined morphological and secondary metabolite characters. The two *ex*-types of *P. syriacum* have been omitted. Isolates are identified by OTU number (Table 1).

A little extra for those that might want to use CA for regression

- Not for examn

CA-PLS

Ter Braak, CJF, Juggins S, Birks HJB, van der Voet H (1993) in *Multivariate Environmental Statistics*, eds. Patil GP and Rao CR, pp. 519-553, Elsevier, Amsterdam

- CA-PLS1 apparently outperforms PLS1 or a mixiumum-likelihood method when used on data on species compositions as related to an environmental gradient
- Especially if there are many variables in **X,** a large number of zeroes in **X**, or an unimodal relationship between **X** and **Y** variables
- Rare occurences is a problem in CA-PLS1 (and PLS1), the standard error of the estimate is inversely proportional to the square root of the number of occurrences

Correspondence analysis (compared to PCA)

- n x m matrix
- i = 1 to n objects and k = 1 to m variables
- For the matrix **X** we calculate:
- **R** = diag (x_{1+}, ..., x_{n+})
- **K** = diag (x_{+1}, ..., x_{+m})
- x_{i+} is the total sum of object i
- x_{+k} is the total sum of variable k

CA and PCA

- In PCA we seek the **w** vector that maximizes:
- **t**‘**t** / **w**‘**w**, where **t** = **Xw**
- In CA we seek the centered weight vector **u** (**1**‘$_m$ **K u** = 0) that maximizes:
- **t**‘**Rt** / **u**‘**Ku**, where **t** = $\mathbf{R}^{-1}\mathbf{Xu}$

CA

- In CA we seek a centered and normalized weight vector (**u**) for the variables in such a way that the vector of the weighted averages, **t** = $\mathbf{R}^{-1}\mathbf{Xu}$, has maximum variance.

PLS & CA-PLS

- In PLS we seek a vector w that maximizes:
- **t'y** / **w'w**, where **t** = **Xw**
- In CA-PLS we seek a centered weight vector **u** (**1'**$_m$ **Ku** = 0) that maximizes:
- **t' Ry** / **u' Ku**, where **t** = **R**$^{-1}$**Xu**
- Later components can be found the same way, but there should be orthogonality (**t**$_i$'**Rt**$_j$ = 0)

Use a PLS program, but:

- For CA-PLS, pre-process:
- **X*** = **R**$^{-1/2}$**X K**$^{-1/2}$and **y*** = **R**$^{1/2}$**y**
- After PLS analysis post-process the results of **X*** and **y*** by:
- **U** = **K**$^{-1/2}$**w**
- **y** is therefor **R** centered at input and input variables should no be centered or standardized in the normal PLS calculation