

Direct Gradient Analysis (constrained ordination)

Canonical Correlation Analysis,
Redundancy Analysis and Canonical
Correspondence Analysis

Direct Gradient Analysis

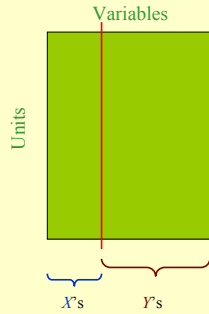
- Direct gradient analysis utilizes external environmental data in addition to the species data.
- In its simplest form, direct gradient analysis is a regression technique.
- Direct analysis tells us if species composition is related to our measured variables.

Direct Gradient Analysis

- Ideally, it will be able to do this even if we did not measure the most important gradients (Palmer 1993).
- Direct analysis allows us to test the null hypothesis that species composition is unrelated to measured variables.
-
- A special case of direct gradient analysis is when our 'measured variables' are experimentally imposed treatments.

Multivariate Statistics with Two Groups of Variables

- Look at relationships between two groups of variables
 - species variables vs environment variables (community ecology)
 - genetic variables vs environmental variables (population genetics)



Canonical Correlation Analysis

- Multivariate extension of correlation analysis
- Looks at relationship between two *sets* of variables

Canonical Correlation Analysis

Given a linear combination of X variables:

$$F = f_1X_1 + f_2X_2 + \dots + f_pX_p$$

and a linear combination of Y variables:

$$G = g_1Y_1 + g_2Y_2 + \dots + g_qY_q$$

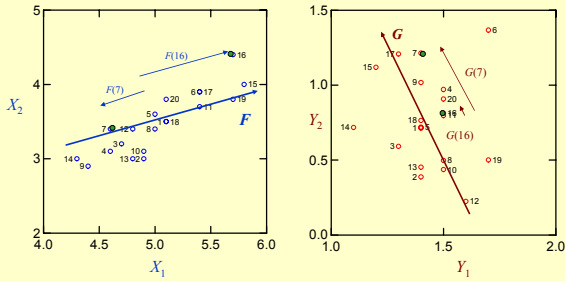
The **first canonical correlation** is:

Maximum correlation coefficient between F and G ,
for all F and G

$F_1 = \{f_{11}, f_{12}, \dots, f_{1p}\}$ and $G_1 = \{g_{11}, g_{12}, \dots, g_{1q}\}$
are corresponding **canonical variates**

Canonical Correlation Analysis

Maximize $r(F, G)$



Canonical Correlation Analysis

The **first canonical correlation** is:

Maximum correlation coefficient between F and G ,
for all F and G

$F_1 = \{f_{11}, f_{12}, \dots, f_{1p}\}$ and $G_1 = \{g_{11}, g_{12}, \dots, g_{1q}\}$
are corresponding **first canonical variates**

The **second canonical correlation** is:

Maximum correlation coefficient between F and G ,
for all F , orthogonal to F_1 , and G , orthogonal to G_1

$F_2 = \{f_{21}, f_{22}, \dots, f_{2p}\}$ and $G_2 = \{g_{21}, g_{22}, \dots, g_{2q}\}$
are corresponding **second canonical variates**

etc.

Canonical Correlation Analysis

- So each canonical correlation is associated with a pair of canonical variates
- Canonical correlations decrease sequentially
- Canonical correlations are *higher* than generally found with simple correlations
 - as coefficients are chosen to *maximize* correlations

Canonical Correlation Analysis

- What are the canonical correlations?
- Are they, all together, significantly different from zero?
- Are some significant, others not? Which ones?
- What are the corresponding canonical variates?
- How does each original variable contribute towards each canonical variate (use **loadings**)?
- How much of the joint covariance of the two sets of variables is explained by each pair of canonical variates?

Redundancy Analysis

- $y_1 \Leftrightarrow y_2$ Correlation Analysis
- $x \Rightarrow y$ Simple Regression Analysis
- $\mathbf{X} \Rightarrow y$ Multiple Regression Analysis
($\mathbf{X}=\{x_1, x_2, \dots\}$)
- $\mathbf{Y}_1 \Leftrightarrow \mathbf{Y}_2$ Canonical Correlation Analysis
- $\mathbf{X} \Rightarrow \mathbf{Y}$ Redundancy Analysis

How one set of variables (\mathbf{X}) may explain another set (\mathbf{Y})

Redundancy Analysis

- “Redundancy” expresses how much of the variance in one set of variables can be explained by the other

Redundancy Analysis

Output:

canonical variates describing how **X** explains **Y**

results may be presented as a biplot:

two types of points representing the units and
X-variables, vectors giving the **Y**-variables

Hourly records of sperm whale behaviour

- Variables:
 - Mean cluster size
 - Max. cluster size
 - Mean speed
 - Heading consistency
 - Fluke-up rate
 - Breach rate
 - Lobtail rate
 - Spyhop rate
 - Sidefluke rate
 - Coda rate
 - Creak rate
 - High click rate
- Data collected:
 - Off Galapagos Islands
 - 1985 and 1987
- Units:
 - hours spent following sperm whales
 - 440 hours

Hourly records of sperm whale behaviour

- Variables:
 - Mean cluster size
 - Max. cluster size
 - Mean speed
 - Heading consistency
 - Fluke-up rate
 - Breach rate
 - Lobtail rate
 - Spyhop rate
 - Sidefluke rate
 - Coda rate
 - Creak rate
 - High click rate
 - Data collected:
 - Off Galapagos Islands
 - 1985 and 1987
 - Units:
 - hours spent following sperm whales
 - 440 hours
- Physical*
- Acoustic*

Canonical Correlation Analysis: *Physical* vs. *Acoustic* Behaviour

	1	2	3
Canonical correlations	0.72	0.49	0.21
P-values	0.00	0.00	0.06
Redundancies:			
$V(\textit{Acoustic}) V(\textit{Physical})$	34%	20%	<1%
$V(\textit{Physical}) V(\textit{Acoustic})$	32%	8%	<1%

Physical vs. *Acoustic* Behaviour

Canonical correlations	1	2
<i>Loadings:</i>		
Mean cluster size	-0.95	0.07
Max. cluster size	-0.85	0.47
Mean speed	0.21	0.06
Heading consistency	0.32	-0.27
Fluke-up rate	0.73	0.23
Breach rate	-0.16	0.02
Lobtail rate	-0.22	0.03
Spyhop rate	-0.18	0.32
Sidefluke rate	-0.21	0.35
Coda rate	-0.64	0.64
Creak rate	-0.50	0.79
High click rate	0.76	0.64

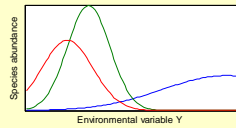
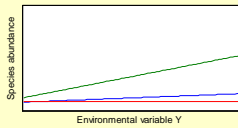
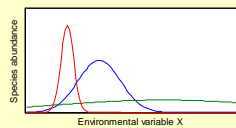
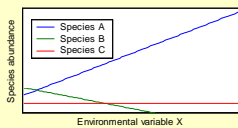
Canonical Correspondence Analysis

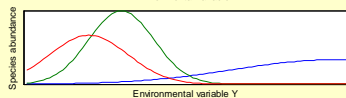
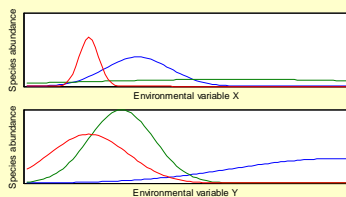
- *Canonical correlation analysis* assumes a linear relationship between two sets of variables
- In some situations this is not reasonable (e.g. community ecology)
- *Canonical correspondence analysis* assumes Gaussian (bell-shaped) relationship between sets of variables
- “Species” variables are Gaussian functions of “Environmental” variables

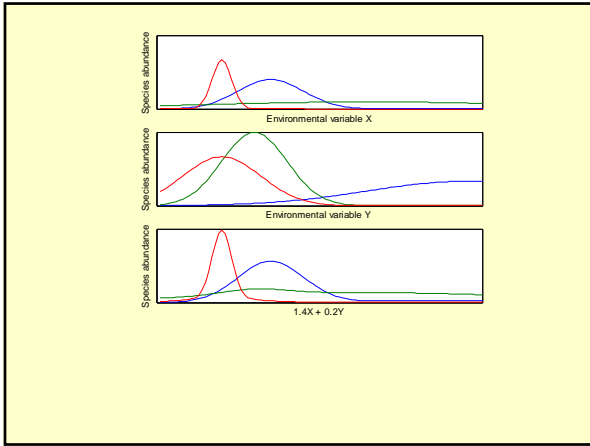
If a combination of environmental variables is strongly related to species composition, CCA will create an axis from these variables that makes the species response curves most distinct

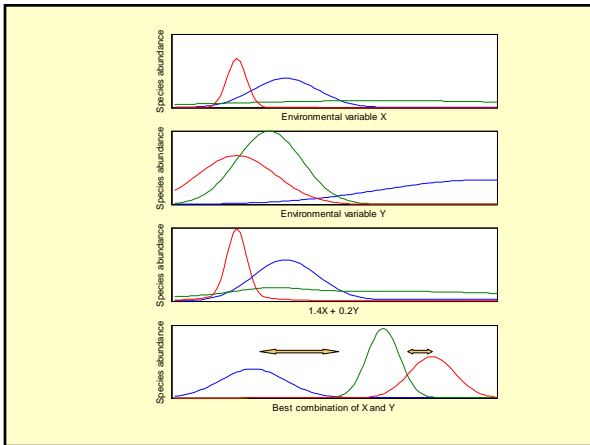
Canonical Correlation Analysis

Canonical Correspondence Analysis





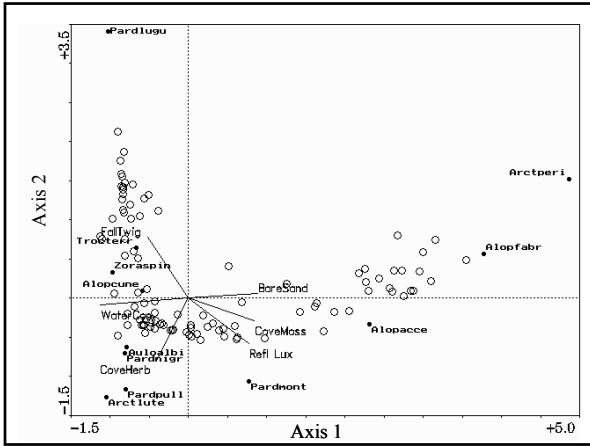




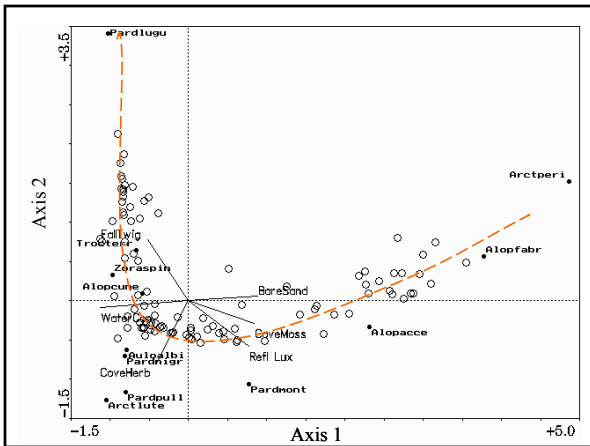
Canonical correspondence analysis: Dutch spiders

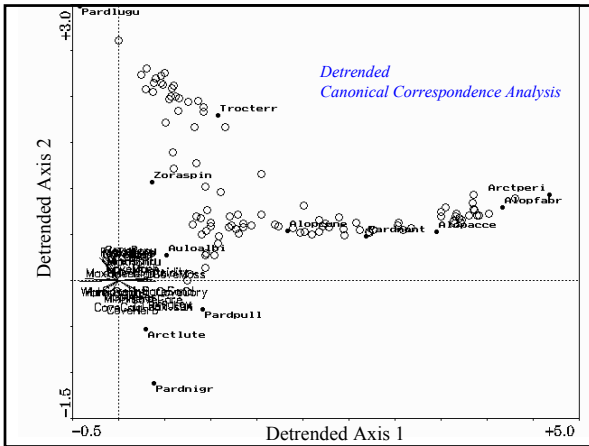
- 26 environmental variables
- 12 spider species
- 100 samples (pit-fall traps)

Axes	1	2	3	4
Eigenvalues	.535	.214	.063	.019
Species-environment correlations	.959	.934	.650	.782
Cumulative percentage variance of species data	46.6	65.2	70.7	72.3
of sp-env relationship	63.2	88.5	95.9	98.2



Canonical correspondence analysis can be *detrended*





Advantages of CCA

- It is possible that patterns result from the combination of several explanatory variables; these patterns would not be observable if explanatory variables are considered separately.
- Many extensions of multiple regression (e.g. stepwise analysis and partial analysis) also apply to CCA.
- It is possible to test hypotheses (though in CCA, hypothesis testing is based on randomization procedures rather than distributional assumptions).
- Explanatory variables can be of many types (e.g. continuous, ratio scale, nominal) and do not need to meet distributional assumptions.

- Variables that contribute little to environmental variance may have a strong impact on species composition
- CCA is not hampered by high correlation between species or environmental variables.
- Can test the significance of environmental variables-Monte Carlo test

Disadvantages of CCA

- In observational studies one cannot necessarily infer direct causation.
- The independent effects of highly correlated variables are difficult to disentangle. However, CCA (and univariate regression) can test the null hypothesis that such variables are completely redundant.

Disadvantages cont.

- The interpretability of the results is directly dependent on the choice and quality of the explanatory variables.
- Although both multiple regression and CCA find the best linear combination of explanatory variables, they are not guaranteed to find the true underlying gradient (which may be related to unmeasured or unmeasurable factors), nor are they guaranteed to explain a large portion of variation in the data. Some ecologists have rejected CCA and other direct gradient analysis techniques because of this, but finding relationships between *measured* variables and species composition is actually a desirable attribute.

- Canonical Correlation Analysis
 - Examines relationship between two sets of variables
- Redundancy Analysis
 - Examines how set of dependent variables relates to set of independent variables
- Canonical Correspondence Analysis
 - Counterpart of Canonical Correlation and Redundancy Analyses when relationship between sets of variables is Gaussian not linear
