Lecture 9: Indirect gradient analysis IV: weighted averaging, Correspondence Analysis (CA) and Detrended Correspondence Analysis (DCA)

1. Why go further than PCA?

2. Weighted averaging

3. Correspondence Analysis (CA)

4. Detrended Correspondence Analysis (DCA)

5. Biplots

6. Review of all ordination techniques

PCA- Why go further?

- The linear model just doesn't fit… species come and go across a gradient. This unimodal response is true with almost any type of normally-distributed data if you sample it over a large enough gradient.

- Correspondence analysis is another eigenvector-based technique that assume a unimodal, rather than linear, relationship among the variables.

- As far as what you can do with it, what the graphing looks like, etc. it is very similar to PCA, the big change is how the axes are derived

Weighted Averaging

- Developed by Curtis and McIntosh in their studies of upland forests in southern Wisconsin
- Important precursor to the application of more advanced multivariate techniques
- Knowledge of species response along a known environmental gradient is used to order stands of vegetation along the same gradient
- Species are weighted according to the position of their peak abundance along the known gradient
- These weightings were then used to provide relative positions of the stands on the axis.

Weighted averaging: an example

- The *weighted average* is calculated for each stand by multiplying the abundance of each tree species times the weighting factor for that species and summing the scores for all species and dividing by the sum of abundances of all species
- For example: suppose that we had a set of vegetation samples, and we knew a lot about the wetland status of all the species within. We could then calculate a *wetland rating* for each sample by taking the abundance of each species and using that to weight it according to its wetland ranking

Weighted averaging: an example (cont.)

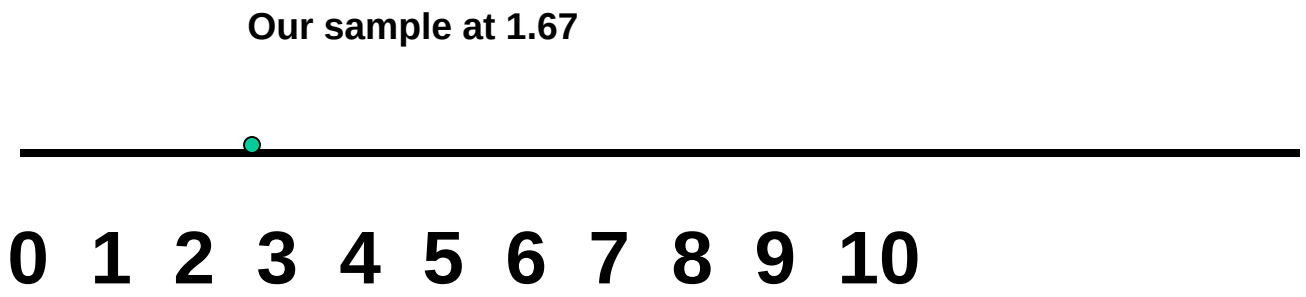
SPECIES	WETLAND RANKING	COVER	WRXC
Juntri	1	2	2
Bigglu	1	4	4
Lupplu	3	3	9
	sum:	9	15

In this example the wetland ranking for this sample is 15/9 or 1.67, reflecting that it is towards the dry each of the gradient

Weighted averaging: an example

- The *weighted average* is calculated for each stand by multiplying the abundance of each tree species times the weighting factor for that species and summing the scores for all species and dividing by the sum of abundances of all species
- The result is a one-dimensional ordination of the plots along the environmental axis (in our example, it would be only one point on a line because we only have one sample or plot).

Weighted averaging: an example (cont.)



Weighted averaging: an example (cont.)

- **INTERPRETATION**
  - Species cover values or environmental factor values can be substituded for plot number along the axis to show distribution of a species' cover or environmental factors along the gradient.
  - The method shows only one axis and is useful for situations where there is only one primary environmental gradient.
  - The major criticism of the method is that the original weights assigned to each species is based on a subjective assessment of the position of the species with respect to the gradient, and this score could vary from one worker to another.

Reciprocal averaging (Correspondence Analysis)

- Reciprical averaging works on much of the same principle as weighted averaging, but rather than forcing an extrenal structure into the results, it finds inherent structure within a data set.
- Provides the basis for more advanced methods of ordinations developed after 1970.
- Also at the heart of two-way indicator species analysis (TWINSPAN).
- Papers by Hill (1973, 1974) first made CA well known to ecologists.

Basic idea of CA:

- Based on the *original* matrix, so you get a simultaneous ordination of species and samples (or variables and plots…). If you arrange a set of species and samples according to their first axis CA order, they will look similar to an arranged table, with dominant species in the middle, and rare ones on the ends.
- The method of weighted averages is applied to a data matrix such that quadrat scores are derived from species scores and weightings. These are carried out successively using an interative procedure. The scores eventually stablize to get a set of scores for quadrats which give axes for a quadrat ordination and a set of scores for species that provide axes for a species ordination
- It is this two-way weighted averaging and is only slightly more complex than one-way weighted averaging

Calculation of the first axis

- Calculate the row and column totals
- Allocate weights to the species
- Reciprical averaging then commences.
- The averaging pricess is then applied in reverse to give a new set of scores for the species using the quadrat scores.
- To avoid calculation with very small numbers these new speceis scores are rescaled from 1 to 100
- The species scores of the final interation are the positions of the species along the firest axis (0 to 100) of the ordination, and the quadrat scores are the positions of the quadrats along the first axis.
- The contraction in the range of species scores in one interation is the eigenvalue.

Reciprocal averaging (an example):

SPECIES	Sample 1 cover	Sample 2 cover	Sample 3 cover
Juntri	2	8	1
Bigglu	4	6	3
Lupplu	8	2	1
Average:	4.7	5.3	1.7
Rescale:	83.5	100	1

Rescaling: Example:(speciesvalue-lowest species value)/range of species values x 100

Reciprocal averaging (an example):

SPECIES	Sample 1 cover	Sample 2 cover	Sample 3 cover	Weighted average
Juntri	2	8	1	88
Bigglu	4	6	3	72
Lupplu	8	2	1	79
Average:	4.7	5.3	1.7	
Rescale:	83.5	100	1	

Th next step is to calculate an average value for each species, weighted by the first axis value.

So: ((83.5*2)+(100*8)+(1*1))/11=88 etc.

Reciprocal averaging (an example):

SPECIES	Sample 1 cover	Sample 2 cover	Sample 3 cover	Weighted average
Juntri	2	8	1	88
Bigglu	4	6	3	72
Lupplu	8	2	1	79
Wt. Av.	78.3	80.1	76.6	
Rescale	49	100	1	

This new vector is used to calculate a new weighted average which is then rescaled.

So: ((88*2)+(72*4)+(79*8))/14=78.3 etc.

Reciprocal averaging (an example):

Keep going...

SPECIES	Sample 1 cover	Sample 2 cover	Sample 3 cover	Weighted average
Juntri	2	8	1	81.7
Bigglu	4	6	3	61.5
Lupplu	8	2	1	53.9
Wt. Av.	78.3	80.1	76.6	
Rescale	49	100	1	

So: ((49*2)+(100*8)+(1*1))/11=81.7 etc.

Reciprocal averaging (an example):

Keep going...

SPECIES	Sample 1 cover	Sample 2 cover	Sample 3 cover	Weighted average
Juntri	2	8	1	81.7
Bigglu	4	6	3	61.5
Lupplu	8	2	1	53.9
Wt. Av.	60	70.6	64.0	
Rescale	1	100	57	

So: ((81.7*2)+(61.5*4)+(53.9*8))/14=60 etc. Remember to rescale!

Reciprocal averaging (an example):

- Eventually what happens is that **IF there is indeed a diagonal structure to the matrrix (where the species and samples correspond to a diagonal along the middle) then the scores will stabilize.**

- **Once they stabilize, you have the first axis!**

- **The species axis is usually used as a “test” and the sample axis is the weighted average of the species axis.**

- **If the variables are standardized scores then the shrinkage of the axis… although because the cumulative sum of the eigenvalues is not the same as the sum of the variances the eigenvalues are less important than in PCA.**

Calculation of the second axis

- It is possible to extract a second axis. This will likely be necessary if there are plots that lie close together in the first axis but which also have a great deal of differences in species composition. The second axis is extracted by the same interation process, with one extra step in which the trial scores for the second axis are made uncorrelated with the first axis.

- In other words, The linear correlation with the first axis is removed. This is done by taking the trial scores for the second axis and regressing it against the site scores for the first axis. The residuals from this regression are the new trial axis. This is done once for each step. (ie. Only for site scores)

Forming the ordination

- The position of the quadrats and species in the ordination space is determined by species scores and quadrat scores for the first two axes.

Review of reciprocal averaging (correspondence analysis)

- Axis 1
  - Assign random scores to each species.
  - Use these to calculate a weighted average for each sample.
  - Rescale these sample scores.
  - Use the sample scores to calculate a weighted average for each species.
  - Continue until scores converge to a unique solution.

- Axis 2
  - Assign a new set of random scores to each species.
  - Calculate a trial axis as above.
  - Perform a multiple regression between the trial axis and the final axis obtained (above).
  - Take the residual values as the new trial axis.

- Although there's no definitive cut-off, generally you want an eigenvalue of at least .25.

Problems with CA

- THE “ARCH EFFECT”
  - A mathematical artifact corresponding to no real structure in the data.
  - The second axis is a quadratic distortion of the first axis.
  - In data sets where there is no strong controlling gradient for the second axis, the arch effect is likely to occur
- COMPRESSION NEAR THE ENDS OF THE AXES
  - Related to the arch effect and does not show the actual comings and goings of species along the first axis.

Arch effect and compression of the axes: an example

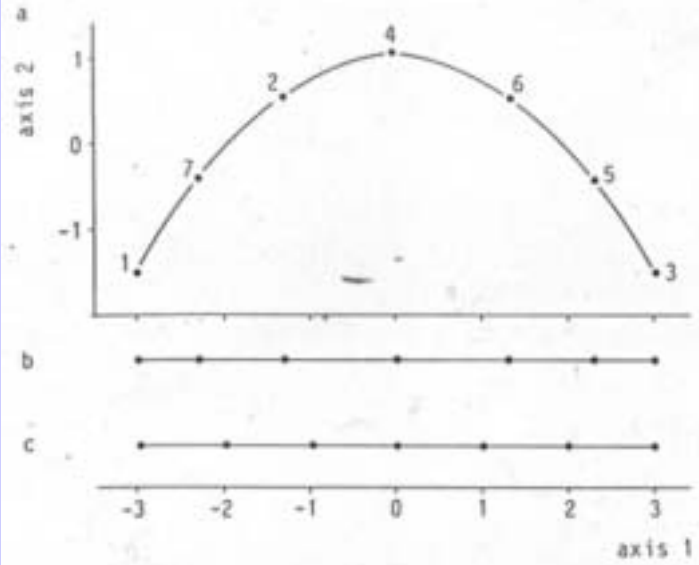


Figure 5.5 Ordination by CA of the two-way Petrie matrix of Table 5.3. a: Arch effect in the ordination diagram (Hill's scaling; sites labelled as in Table 5.3; species not shown). b: One-dimensional CA ordination (the first axis scores of Figure a, showing that sites at the ends of the axis are closer together than sites near the middle of the axis. c: One-dimensional DCA ordination, obtained by nonlinearly rescaling the first CA axis. The sites would not show variation on the second axis of DCA.

Detrended Correspondence Analysis (DCA)

- ***Correcting for the arch effect (detrending)***
  - The first axis is divided into a number of segments and within each segment, the second axis scores are recalculated so that they have an average of zero.
  - In DECORANA, the first axis os divided into many segments and the averaging is achieved through a running averages procedure.
- ***Correcting for the compression effect***
  - Also overcome by segmenting the first axis and rescaling the species ordination (not the lot ordination), such that the coming and going of species is about equal along the gradient.

Detrended Correspondence Analysis (DCA) (cont.)

- ***Scaling of the axes in SD***
  - The axes in DCA are scaled into units that are the average standard deviation of species turnover (SD units).
  - A 50% change in species composition occurs in a distance of about 2 SD unit. Species appear, rise to their modes, and disapear over a distance of about 4 SD units.
  - The more SD units that occur along the axis the more change in species composition is shown. Thus, the axes of DCA are a useful measure of beta diversity.
  - The position of samples along the 1st axis are thus shifted to equalize beta-diversity. Species have, on average, a habitat breadth (as measure by standard deviations) of 1.

Detrended Correspondence Analysis (DCA): an example

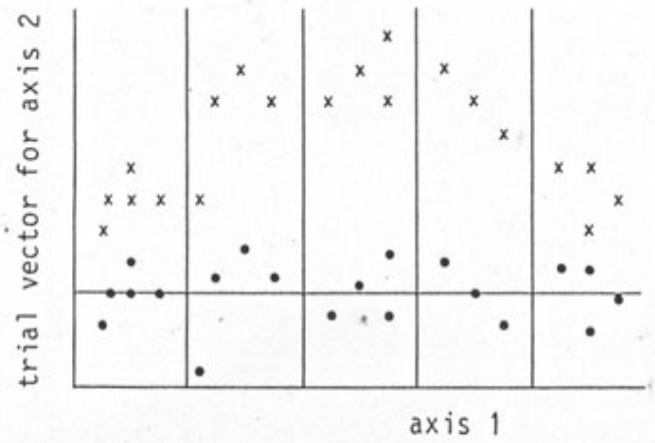


Figure 5.6 Method of detrending by segments (simplified). The crosses indicate site scores before detrending; the dots are site scores after detrending. The dots are obtained by subtracting, within each of the five segments, the mean of the trial scores of the second axis (after Hill & Gauch 1980).

Detrended Correspondence Analysis (DCA): an example

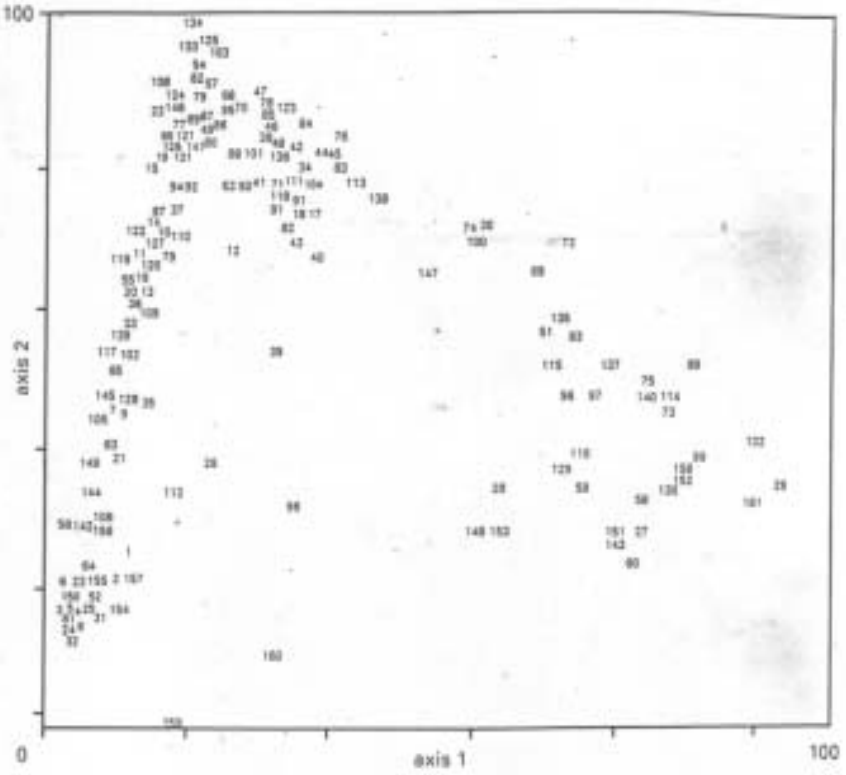


Figure 6.11 Quadrat ordination of vegetation data from the Narrator Catchment, Dartmoor using correspondence analysis/reciprocal averaging (Kent and Wathern, 1980; reproduced with kind permission of *Vegetatio* and Kluwer Academic Publishers)

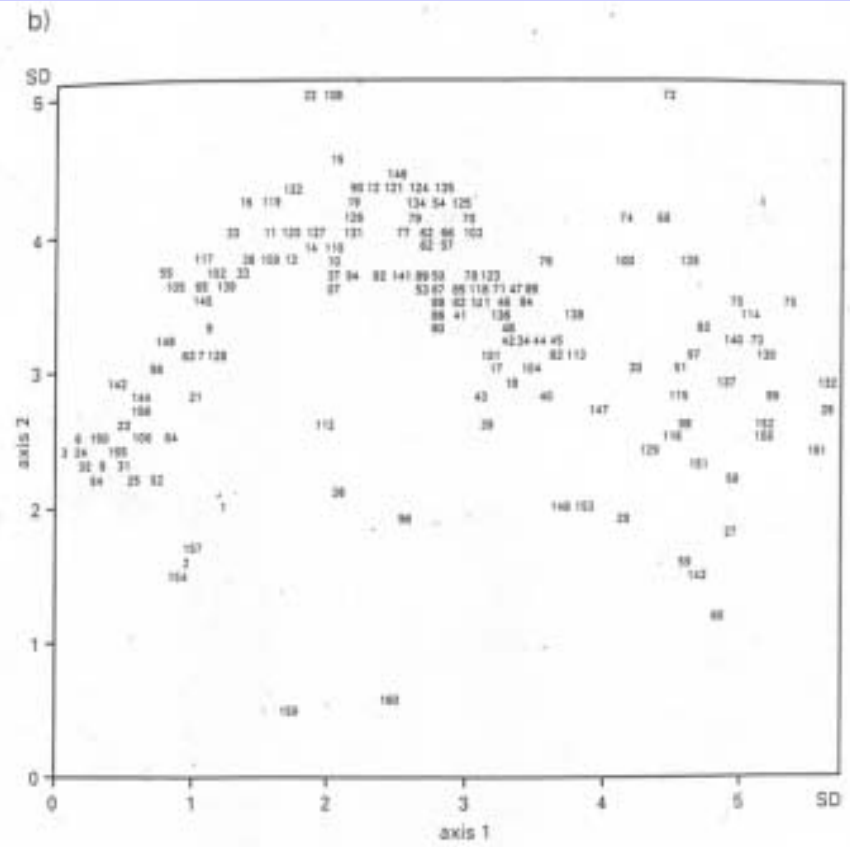


Figure 6.12 The same data as in Figure 6.11 analysed by detrended correspondence analysis

How do you tell if DCA is for you?

- A general rule of thumb is that if the axis is less than 2 units long, you should consider PCA, if it is more than 4 units long then DCA will likely be appropriate, and in between that… you will have to look harder at the data and make some “educated” choices...

Criticism of DCA

- The method used for correcting the arch effect and compression have no emprical or theoritical basis (Wartenberg *et. al* 1987).
- The assumption that species turnover is constant or even along gradients is likely not true.
- It removes, with brute force, if needed any arch effect… even there is really is an arch effect in your data set.
- Because it is done by arbitrarily dividing the axis into pieces, and then shifting those pieces up or down (to get an constant mean), the relationships within the segment are maintained, but other relationships can be “ripped” apart.

DCA OVERALL

- Despite these critisms, tehre does not appear to be a good alternative method at this time, and DCA remains one of the most powerful methods of indirect gradient analysis and is computationally very efficient.

- It is the best method to to use when there are no environmental data

- The interpretation of results from DCA is best carried out with some knowledge of its limitations and comparison with other techniques. By doing this, one can get a better feel for patterns created by actual structure within the data set.

BIPLOTS

- A useful diagram of the species ordination together with the environmental factors (Gabriel, 1971).
- Both species and environmental factors are plotted on the same graph but using different scales.
- Arrows are drawn from the joint centred ordination axes to the points representing species.
  - The direction of the arrow indicates the direction in which the abundance of a variable increases most rapidly.
  - The length of the arrow indicates the rate of change in abundance in that direction.

BIPLOT: an example

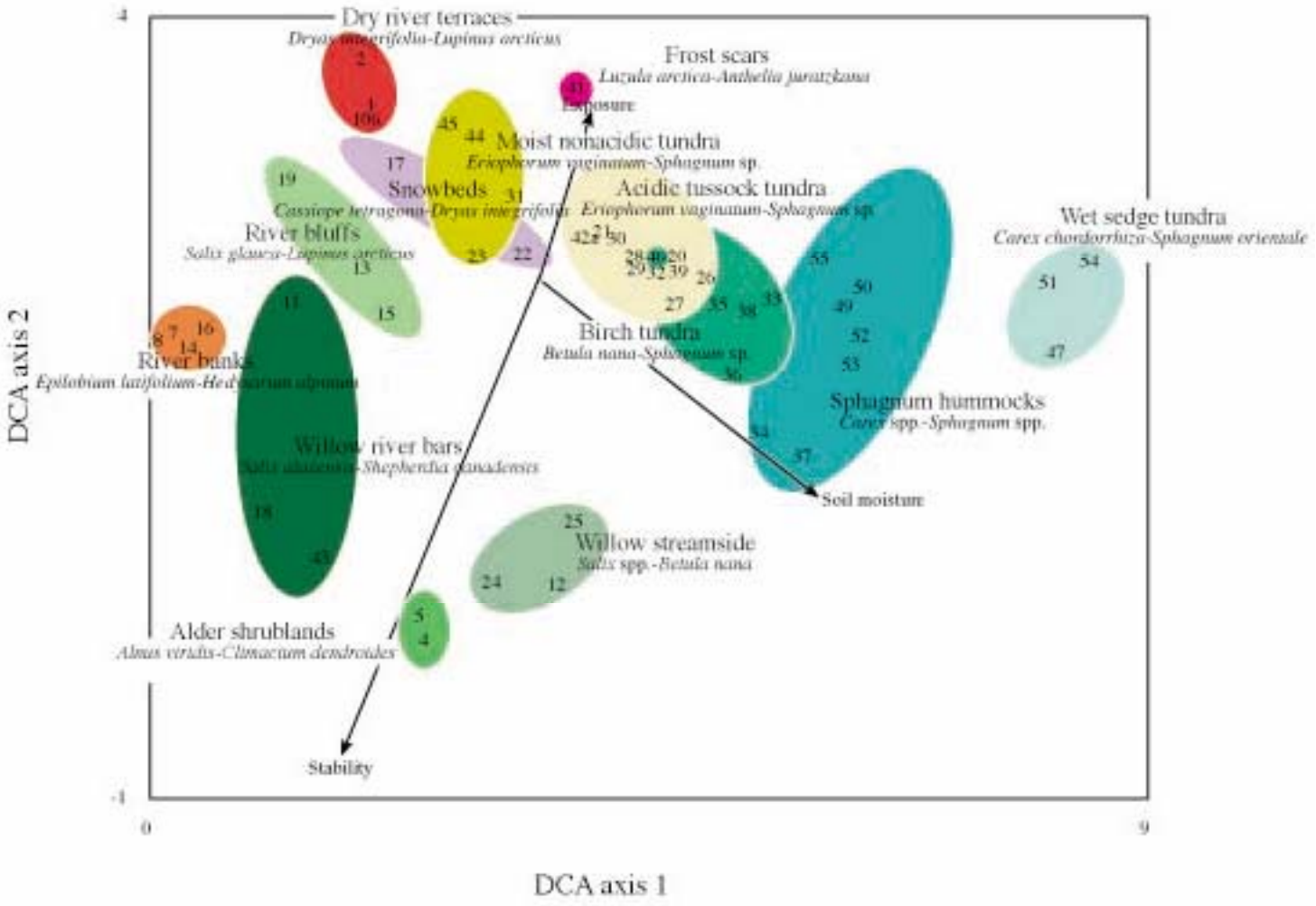


Figure 2. Detrended Correspondance Analysis (DCA) ordination of Happy Valley vegetation relevés. Extreme relevés (HV-3, 46, 48, 9, 42b) were excluded because species composition was too disimilar with other plots. Colors indicate preliminary community types.

Review of all ordination techiques

TECHNIQUE	SUMMARY	WHEN TO USE
Principal Components Analysis	PCA is based on the assumption that there are linear correlations among the data being reduced, i.e., that the species or samples are linearly correlated. The eigenvalues are correlation coefficients, and therefore can be used directly to measure the variance explained by the ordination. Higher order axes are uncorrelated with the first axis, except that in an oblique solution this is not enforced.	Any data set with high linear correlations is appropriate for PCA. This is often a series of related data, such as climate, soils, or biogeochemical data. It may be appropriate for community data that have mostly linear correlations.
Correspondence Analysis	CA is based on the assumption that the data have a unimodal response to an underlying gradient. The eigenvalues vary from 0 to 1 and reflect how well the data fit the ordination model.	Most community data sets are appropriate for CA.
Detrended Correspondence Analysis	DCA is similar to CA except that it corrects two major "faults": (1)the presence of a quadratic relationship between the first two axes and similar relationships with higher order axes, and (2) a compression of scores toward the ends of the axes.	Most community data sets are appropriate for CA.
Canonical Correlation Analysis	Cancor is based on developing linear combinations of 2 sets variables that maximizes the linear combination between those sets.	Cancor is appropriate when there are two distinct sets of data with moderate correlations within them. Lots of high correlations cause problems and unstable solutions