Ecological Community Structure: Ordination in R

Abel Valdivia, Ph. D
http://www.unc.edu/~abelvald/
abelvald@live.unc.edu
https://twitter.com/AbelValdivia

Common goal of multivariate analysis

Discover and summarize the main patterns of variation in a set of variables measured over a number of sampled location

- -R mode focus on variables
  - Association among species
  - Correlations among env variables
- -Q mode focus on samples
  - Relationship among samples

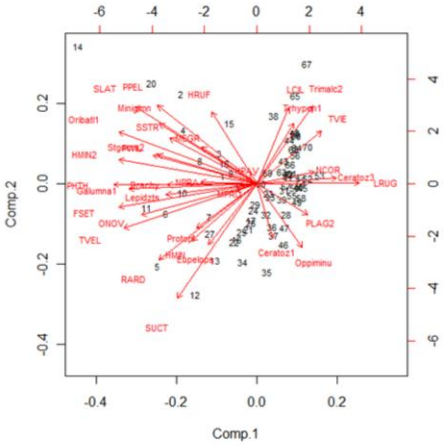
Indirect versus direct methods

- Indirect (axes are hypotetical)
  - Principal Component Analysis (PCA)
  - Correspondence Analysis (CA)
  - Detrended Correspondence Analysis (DCA)
  - Non-metric Multidimensional Scaling (NMDS)

- Direct (axes are linear combination of enviromental variables)
  - Canonical Correspondence Analysis (CCA)
  - Redundant Correspondence Analysis (RDA)

Principal Component Analysis (PCA)

- Oldest multivariate technique
- Based on correlation or covariance among variables
- Assumes multi-normality and linearity (rare in ecological data)
- Better to use with environmental variables
- Eigenvalues express the amount of variance
- Can be used as preliminary Analysis



-Principal components analysis is by far the oldest multivariate technique, dating back to the early 1900's; ecologists have used PCA since the 1950's.
-PCA is based on covariance or correlation among variables
-Thus assumes that the data are multinormal (each variable is normally distributed and so are the joint distributions)

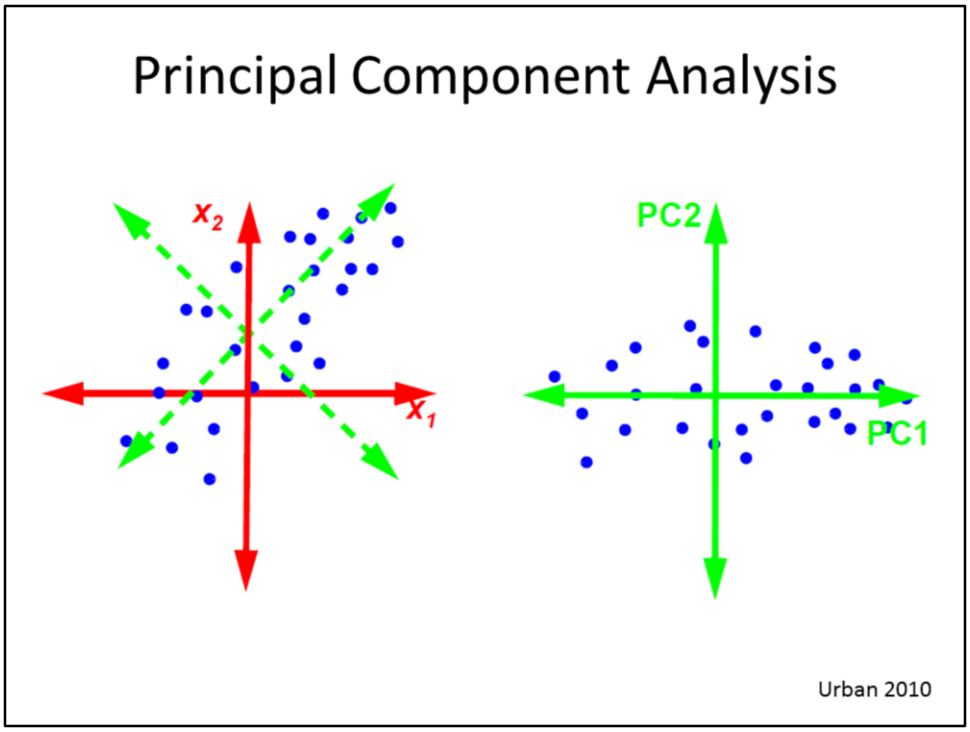
- Also assumes that the relationships among variables are linear.
- Neither assumption is common in ecological data
- Transformation of the raw data can improve normality and interpretation at least is not compromise

- The choice of standardizing the data or not (*i.e*., using the covariance matrix or the correlation matrix as the
starting point) has important implications for the analysis.
- Using covariance means that the variables with the greatest variance tend to dominate the analysis and would be expected to load
on the first PC's.
-Using the correlation matrix treats all the variables equitably, with each variable contributing unit variance.

- The amount of variance on each PC is expressed as its eigenvalue;
- The eigenvalue for the *i*th PC, divided by the sum of all eigenvalues, is the proportion of the variance expressed on that PC.

- The eigenvalues of the principal components indicate the amount of variance attributable to each axis, essentially indicating the relative

lengths of the axes.
- PCA might provide a smaller set of composite variables that captured most of the variation in the data.



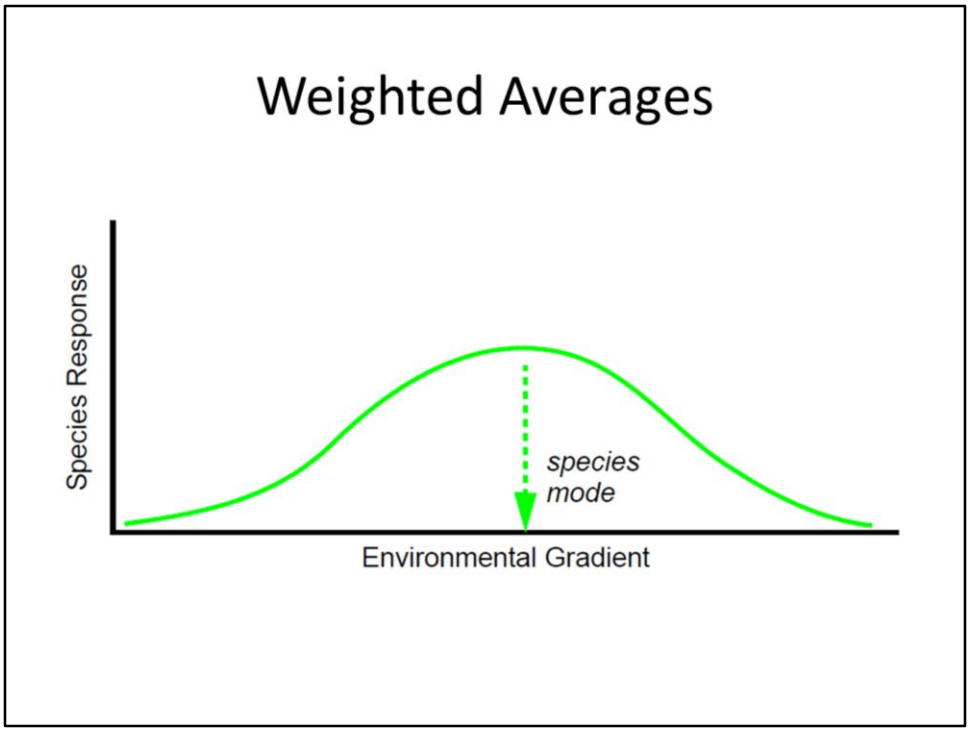
Geometric schematic of PCA as re-projecting a data swarm along new coordinate axes that maximize variance along subsequent orthogonal axes
Arrows in red are the original variables. PC 1 in green is drawn along the longest axis of the data cloud. PC2 is orthogonal to the first and so on

This is as a multiple regression problem, in which the line of PC1 is fitted by least-squares: PC1 minimizes the sum of squared deviations from each point to the line. Subsequent components are fitted to the residuals in each case. In terms of regression, note that this is essentially a regression of the data on *itself* (there are no dependent and independent variables identified).

Note that PCA is a linear model and will summarize patterns of associate among variables as linear relationships.
Initial screening will indicate whether this is a reasonable thing to do. If relationships are profoundly nonlinear, other techniques such as
Ordinations might be more appropriate.

One nice feature of PCA is that it reproduces *all* of the variance in the original dataset.

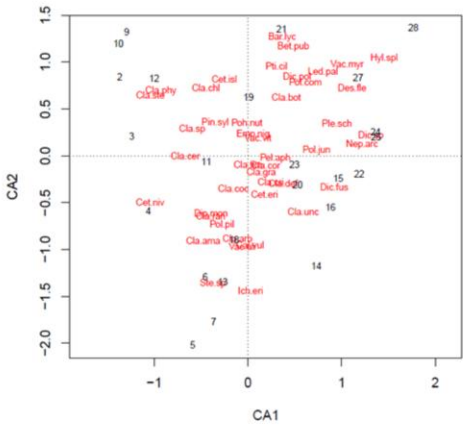
No variance goes unaccounted; it is merely repackaged onto
new components



An example of a unimodal and symmetric species response (*e.g.*, as abundance) to an environmental gradient. In this case it makes sense to summarize the
species response in terms of its abundance-weighted average.
Reciprocally, it might also make sense to summarize a sample location in terms of the species that occur on it.
Given an appropriate weighting variable for species, weighted averaging to produce sample scores yields a one-dimensional
ordering of the samples, a direct ordination. The use of two indicator variables would produce a two-dimensional ordination, and so on.
ecological data are inherently low-dimensional
That is, two or three ordination axes typically suffice to capture the major trends in the data, and these axes are
readily interpretable ecologically.ave site affinities such that their presence indicates something ecologically about that site

Correspondance Analyis (CA) or Reciprocal Analysis

- Better and more robust than PCA
- Disadvantages
  - Arc effects with long ecological gradients
  - Sites are packed closely at gradient extremes
  - Rare species have high influence in the results

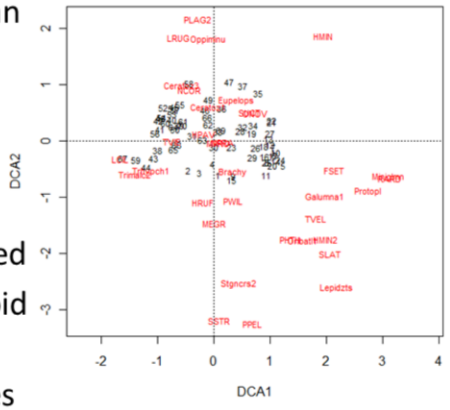


-The use of weighted averages, either as species scores or as samples scores, plays a central role in multivariate analysis
-CA is much better and more robust for community ordination than PCA
-The process begins by assigning arbitrary weights to the samples (*e.g.*, their sample numbers) and using these to compute species scores as
weighted averages. The new species scores are then used to recompute the sample scores, then these are used to recompute the species scores, and so on, until the scores stabilize.
-However single long ecological gradients appear as curves or arcs in ordination (arc effect):
-Sites are packed more closely at gradient extremes than at the centre
-Rare species seem to have an unduly high inuence on the results
In practice, RA has suffered from a numerical problem that subsequent axes are forced to
be linearly independent of prior axes.
In particular, the method of taking residuals from prior axes has the quirky result that the second
axis has a quadratic dependence on the first, the third has a cubic dependence, and so on.
This is responsible for the so-called “arch effect” with RA: a plot of the first two axes reveals an arch
that expresses the quadratic relationship between the axes. When projected onto a

single axis,
this also compresses the spacing between samples near either end of the first axis. This makes it
difficult to interpret RA ordinations, which led to a numerical “fix” in the method, detrended
correspondence analysis.

Detrended Correspondence Analysis (DCA)

- Better and more robust than PCA and CA
- Unimodal species response
- Eigenvalues are defined
- Advantages over CCA
  - The arc effect is detrended
  - Rescaling the axes to avoid extremes
  - Down-weight rare species



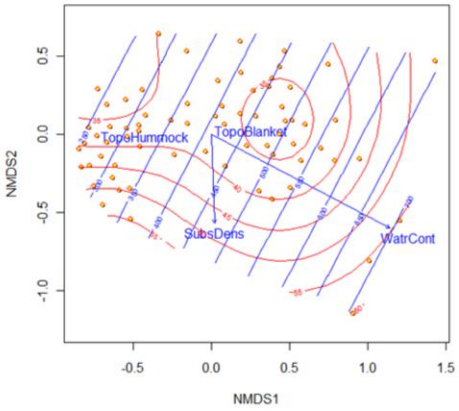
-Correspondence analysis is a much better and more robust method for community ordination than principal components analysis.
-Eigenvalues are dened as shrinkage values in weighted averages, similarly as in cca above.
-The arc effect is detrended. The solution is to detrend the later axes by making their means equal along segments of previous axes
-This basically flatten the arc effect
-In heterogenous data with a clear arch effect the changes are often more dramatic, this data is more homogenoeus
-The solution of the packed extremes is to rescale the axes to equal variances of species scores
-In this, the first axis is divided into short segments. Sample scores within each segment are centered to have a mean of 0.0 on the second axis; this
centering by segments realigns the arch into a straight line with samples scores scattered evenly above and below the line.
- The solution for high influence of rare species is to downweight them
The compression of sample scores near the ends of the axes are fixed by rescaling the segments separately.
- Some authors (especially Minchin 1987 *et seq.*) argue that the detrending seriously degrades the solution, in
part because sample separation in ordination space does not necessarily reflect ecological dissimilarity—it reflects detrending and rescaling instead.

The eigenvalues express the relative lengths of the ordination axes. Because of the way the scores are computed, the largest eigenvalue is always less than 1.0, and eigenvalues larger than 0.5 tend to be interpreted as meaningful.

Note that RA and DCA extract only a few axes (typically no more than four), and so the total variance in the data set is not recovered (*cf* PCA, which recovers all the variance)

Non-metric Multidimensional Scaling (NMDS)

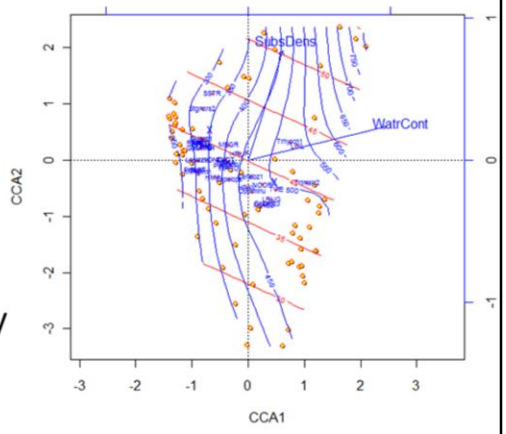
- Based on ecological distance (Bray-Curtis) matrices
- The result will depend of the ecological distance used
- No assumptions about linearity or modality
- Info is converted to ranks
- Iterative process
- Stress value represent the goodness of the test



-Non-metric multidimensional scaling is a good ordination method because it can use ecologically meaningful ways of measuring community dissimilarities.
-Only uses rank information and maps ranks non linearly onto ordination space
-Because all ordination axes are fitted simultaneously there is no real ordering of the axes in terms of relative importance or proportion
of variance accounted; the order in which axes are displayed is arbitrary.
-Thus, a strong underlying gradient which would appear as the first axis in DCA might not appear as the first axis
in a multi-axis NMS solution but rather, might show up on axis 2 or 3.
-It makes no assumptions about linearity (as with PCA),
nor does it presume any underlying model of species response to gradients (as with RA, DCA,
In practice, NMS makes a powerful
companion analysis to other ordinations by providing an unbiased assessment of the assumptions of other methods.
For example, if a NMS solution corroborates a CCA ordination,
that would provide powerful support for the latter.

Constrained Correspondence Analysis (CCA)

- Assumes a non-linear species response model and unimodal distribution along env gradient
- Species dispersions are similar across species
- Sample scores are regressed on the ancillary env variables
- It is like a multiple regression



CCA, like RA and DCA, assumes a nonlinear species response model in which species vary along environmental gradients in a unimodal manner.
As the constrained version of RA,
CCA amounts to the simple insertion of a regression into the RA algorithm. In this, sample scores
are computed by weighted averaging from species scores. Then, these sample scores are
regressed on the ancillary environmental variables. These scores (LC-scores, for linear combinations derived by regression), are then used to recompute new species scores, and so
on.
It might be helpful to consider CCA as a special form of multiple regression, in which the
dependent variables (species abundances) meet a number of assumptions:

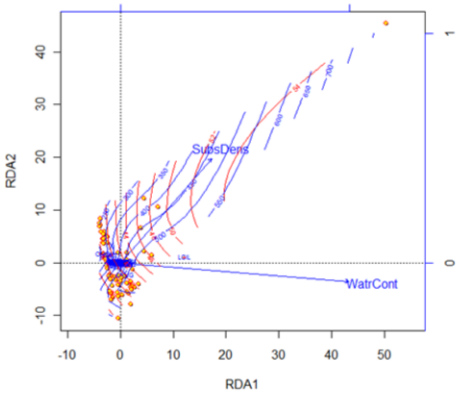
1 that the species environmental responses are unimodal and symmetric (*i.e.,* bell-shaped curves),
2. species dispersions (the widths of the curves) and maxima are similar across species.

In this case, CCA is essentially a multiple regression problem in which the regression curve is fitted as

gaussian.
• the species ordination is summarized by their relative positions along each axis;
• sample locations indicate their compositional similarity to each other;
• samples tend to be dominated by the species that are located near them in ordination
space;
• the *length* of the arrow (vector) for each environmental variable indicates its importance
to the ordination (*i.e.*, how much it contributes to the axes);
• the *direction* of an environmental vector indicates its correlation with each of the illustrated axes (*i.e.*, a vector aligned parallel to an axis is highly correlated with it; a vector at an
angle between the two axes is correlated with both);
• angles between environmental vectors indicate the correlation between the environmental variables themselves;
• the location of samples near environmental vectors suggests the environmental conditions of those samples; and
• the location of species scores near environmental vectors suggests the environmental
affinities of the species.

Redundancy Analysis (RDA) constrained or direct

- Assumes a non-linear species response model and unimodal distribution along env gradient
- Species abundances are regressed on the ancillary env variables



The counterpart technique of interest here is redundancy analysis (RDA). In RDA, species abundances are regressed on the ancillary
environmental variables. These regressions are then used to predict species responses
(abundances), and then PCA is performed on the predicted species responses. This amounts to
a summary of patterns in species response that can be accounted by environment. Importantly,
the underlying species response is linear, in accord with the assumptions of regression and PCA.

Choosing the right ordination

- All techniques introduce bias
- Conservative approach would use a combination of complimentary techniques
- A good ordination make sense and will depend on your questions
- The right technique will depend of the data

- All ordination are biased because they are all try to represent a multidimensional reality into a low-dimensional summary
- As we discussed every technique has advantage and disadvantages, with assumptions and a prior exploratory analysis is required to chose the right technique than can best describe more faithfully the distribution of your data.
- A conservative approach would use a combination of complimentary techniques and search for similarities in the solutions
- In general a good technique will be the one that make most sense of the data and can be useful to answer the question in hand
- Some applications should call for specific ordination techniques
-

Comparison of Ordination Techniques

Technique	Data	Response	Ancillary Data
PCA	R or C	linear	post facto
DCA	primary	unimodal	post facto
CCA	primary	unimodal	constrained
PO	D	n/a	post facto
NMS	D	n/a	post facto
PCoA	D	n/a	post facto
RDA	primary	linear	constrained
dbRDA	D	n/a	constrained

R: correlation matrix
C: covariance matrix on variables
D: distance matrix on samples

- **Data** can be primary (species x samples) or secondary (R: correlation matrix, C: covariance matrix, D: distance matrix)
- **Response** model of how species are assumed to respond to underlying gradients
    - PO, NMS, PCoA and distance based RDA make no assumptions based on species response
- **Ancillary data** are typically a matrix of environmental data associated with species data
    - CCA, RDA and dbRDA are constrained by environmental variables
    - PCA, DCA, PO, NMS and PCoA are added after the fact

Key to Ordination techniques

1a. Ancillary data (ENV) not available, or not used to constrain the ordination 2
1b. Ancillary data (ENV) available and used in ordination 4

2a. Species response assumed linear..... PCA
2b. Species response not linear or unknown 3

3a. Species response assumed nonlinear and unimodal ..... DCA
3b. Species response not linear, nor nonlinear and unimodal, or unknown ..... NMS

4a. Species response assumed linear ..... RDA
4b. Species response not linear, or unknown 5

5a. Species response nonlinear and unimodal 6
5b. Species response not linear, nor nonlinear and unimodal, or unknown .....dbRDA

6a. A single (or few) environmental factors known (or required) as constrains .....DO
6b. Several factors suspected .....CCA

Urban 2010

References

- Legendre, P., and L. Legendre. 1998. *Numerical ecology (2nd English edition).* Elsevier, Amsterdam, The Netherlands
- McCune, B., and J.B. Grace. 2002. *Analysis of ecological communities*. MjM Software Design, Gleneden Beach, Oregon.