Ordination Analysis I - Indirect Gradient Analysis

John Birks

Quantitative Methods in Palaeoecology and Palaeoclimatology
PAGES Valdivia October 2010

INDIRECT GRADIENT ANALYSIS

Introduction
Principal components analysis (PCA)
Correspondence analysis (CA)
Detrended correspondence analysis (DCA)
PCA or CA/DCA?
Metric scaling or principal co-ordinates analysis (PCOORD)
Non-metric scaling (NMDS)
Is there an ideal method?
Conclusions
Principal curves
Comparing different ordinations
Software

INTRODUCTION

Basic ideas of ordination analysis

GRADIENT MODEL OF REAL WORLD

- Environment varies continuously in space and time
- Gradients are real or abstract environmental factors
- Transects are concrete entities in the landscape
  You can have continuous gradients but boundaries on transects (e.g. geological change)
- Resource gradients are factors that are used by organisms (e.g. nitrogen)
- Direct gradients influence organisms (e.g. pH) but do not use them
- Indirect gradients (e.g. altitude) influence organisms but many environmental variables may be important (e.g. temperature, wind, snow-fall)
- Complex gradients are correlated direct and/or resource gradients
  In field data sets you can have mainly complex gradients
- Complex gradients are dependent on the data set (landscape concept)
- Stratigraphical data are complex temporal gradients
- Species have curvilinear relations to gradients
Advantages of INDIRECT GRADIENT ANALYSIS
1. Species compositions easier to determine than full range of environmental conditions. Many possible environmental variables. Which are important?
2. Overall composition is often a good reflection of overall environment.
3. Overall composition often of greater concern that individual species.
   Global, holistic picture, in contrast to regression which gives a local, individualist reductionist view.

Constrained canonical ordinations stand between INDIRECT GRADIENT ANALYSIS and REGRESSION. Many species, many environmental variables.

Aims of INDIRECT GRADIENT ANALYSIS
2. Uncover the fundamental underlying structure of data. Assume that there is underlying LATENT structure. Occurrences of all species are determined by a few unknown environmental variables, LATENT VARIABLES, according to a simple response model. In ordination trying to recover and identify that underlying structure.

UNDERLYING RESPONSE MODELS
A straight line displays the linear relation between the abundance value (y) of a species and an environmental variable (x), fitted to artificial data (•). (a = intercept; b = slope or regression coefficient).

A Gaussian curve displays a unimodal relation between the abundance value (y) of a species and an environmental variable (x). (u = optimum or mode; t = tolerance; c = maximum = exp(a)).

Indirect gradient analysis can thus be viewed as being like regression analysis but with the MAJOR difference that in ordination the explanatory variables are not known environmental variables but theoretical ‘latent’ variables.

Constructed so that they ‘best’ explain the species data.
As in regression, each species is a response variable but in contrast to regression, consider all response variables simultaneously.

PRINCIPAL COMPONENTS ANALYSIS PCA
CORRESPONDENCE ANALYSIS CA
& relative DCA

PCA - linear response model
CA - unimodal response model

PRINCIPAL CO-ORDINATES ANALYSIS PCOORD
NON-METRIC MULTIDIMENSIONAL SCALING NMDSCAL
TOWARDS A PHILOSOPHY OF USING ORDINATION METHODS IN DATA ANALYSIS

Ockham's (or Occam's) razor: invaluable philosophical concept because of its strong appeal to common sense

William of Ockham, 14th century English philosopher, stated the principle as:

PLURALITAS NON EST PONENDA SINE NECESSITATE
Plurality must not be posited without necessity
Entities should not be multiplied without necessity
It is vain to do with more what can be done with less
An explanation of the facts should be no more complicated than necessary
Among competing hypotheses, favour the simplest one that is consistent with the data

PRINCIPAL COMPONENTS ANALYSIS

Extension of fitting straight lines and planes by least-squares regression. (Pearson 1901)

Fit moisture to all the species in data by a series of least-squares regressions of $y = b_0 + b_x + i$, we obtain for each regression the RESIDUAL SUM OF SQUARES, the sum of squared vertical distances between observed and fitted line.

Total of the separate residual sum of squares for all species, total residual SS, is a measure of how badly moisture explains the data of all species.

What is the best possible fit that is theoretically possible with straight-line regression?

$y = b_0 + b_x + i$

or, if we have centred the data (subtracted the mean)?

$y = b_x + i$

Defines an ORDINATION problem: construct the single hypothetical variable (latent variable) that gives the best fit to the species data according to our linear equation.

PCA is the ordination technique that constructs the theoretical variable that minimises the total residual SS after fitting straight lines or planes to the species data.

OCKHAM'S RAZOR

What has this to do with ordination?
Data consist of few variables and few objects, perhaps easiest to present and interpret as tables or diagrams.
Typical data of many variables and many objects, purpose of ordination is to assist in the application of Ockham's razor.
A few dimensions are easier to understand than many dimensions
A good ordination technique will be able to determine the most important dimensions or gradients of variation in a data set and help us to ignore, 'noise' or chance variation.
i.e., ordination helps us avoid being vain by doing with more what can be done with less.
Want to show as much as possible with only a few axes. A few important axes are likely to be more easily explicable than many less important axes.

Straight lines for several Dune Meadow species, fitted by PCA to species abundances. Also shown are the abundances of Lolium perenne and their deviations from the fitted straight line. The horizontal axis is the first principal component. Fitting straight lines by least-squares regression of the abundances of the species on the site scores of the first PCA axis gives the same results. The slope equals the species score (b) of the first axis. The site scores are shown by small vertical lines below the horizontal axis.

Fitted $y = b_x$, where $y = abundance$, $b_x$ = slope for species with respect to the latent variable, and $x$ is site score on latent variable, where latent variable is first PCA axis.
How good a fit is the latent variable to the data?
Total SS of regressions is so-called EIGENVALUE of first PCA axis = goodness of fit. Express as % of total variance in species data.
Three dimensional view of a plane fitted by least-squares regression of responses (+) on two explanatory variables PCA axis 1 and PCA axis 2. The residuals, i.e. the vertical distances between the responses and the fitted plane are shown. Least squares regression determines the plane by minimizing the sum of these squared distances.

PCA ordination diagram of the Dune Meadow Data in covariance biplot scaling with species represented by arrows. The b scale applies to species, the x scale to sites. Species not represented in the diagram lie close to the origin (0,0).

Total sum-of-squares (variance) = 1598 = SUM OF EIGENVALUES
Axis 1 471 = 29 %
Axis 2 344 = 22 %
Each axis = vector of species slopes or scores (b) EIGENVECTORS
[Regression $E_i = b_1 b_1 x_1 + b_2 x_2$ (if centred data)]

PCA BIPLOTS

- Axes must have identical scales
- Species loadings and site scores in the same plot: graphical order 2 is an approximation of the data
- Origin: species averages. Points near the origin are average or are poorly represented
- Species increase in the direction of the arrow, and decrease in the opposite direction
- The longer the arrow, the stronger the increase
- Angles between vector arrows approximate their correlations
- Approximation: project site point onto species vector
- Distance from origin reflects magnitude of change
Biplot interpretation for Agrostis stolonifera.
In correlation (covariance) biplot, site scores scaled to unit sum of squares and sum of squared species scores equals eigenvalue \( r = \text{Cost} = \text{Correlation} \)

**AN ALTERNATIVE GEOMETRICAL REPRESENTATION OF PCA**

1. Move the axis origin to the centroid of the species space (species averages).
2. Rotate the axes so that the first axis becomes
   (a) as close to all observations as possible, which means that it
   (b) explains as much of the variance as possible.
   • First rotated axes show the configuration as well as possible.
   (PCA = reduced major axis regression in Model II regression)

**EXPLAINING THE VARIATION IN PCA**

- **Total variation**: sum of squared distances from the origin
- **Explained variation**: sum of squared projections onto principal components = \( \lambda \)
- **Residual variation**: sum of squared orthogonal distances from the principal components
- **Only a rotation**: all axes (min \( M \), \( N \)) explain everything

**DATA TRANSFORMATIONS**

1. Centred species data = PCA variance-covariance matrix. Species implicitly weighted by the variance of their values.
2. Standardised PCA = PCA correlation matrix. Centre species and divide by standard deviation (zero mean, unit variance). All species receive equal weight, including rare species. Use when data are in different units, e.g. pH, LOI, Ca.
4. Log \((y + 1)\) transformation for abundance data.
5. Log transformation and centre by species and samples = log-linear contrast PCA for closed % data (few variables, e.g. blood groups).
STANDARDISED OR NON-STANDARDISED PCA

Standardising species to unit variance (using the correlation coefficient) makes all species equally important, instead of concentrating on the abundant species with largest variances (using covariance matrix).

HOW MANY ORIENTATION AXES TO RETAIN FOR INTERPRETATION?

PCA - applicable to CA, PCOORD, TRDA, CCA, TIDCA

Real data - 3 data-sets
1) Correlations between variable 0.6 - 0.9
2) 0.0 - 0.7
3) 0.3 - 0.4
Simulated data
1000 observations; 4, 12, 32
r = 0.0 or 0.3 or 0.8 for off-diagonal

i.e. no dimensions, weak one-dimension, strong one-dimension

Adding 'Unknown' Samples

Passive samples

A PCA biplot showing the scores of the first and second components of the modern pollen spectra, the vectors of the pollen taxa, and the means and standard deviations of the five pollen zones from the Lateral Pond fossil site (zone 1 is the oldest); o represents the projection of the origin.

Useful in palaeoecology

Assessment of eigenvalues:
\[ \lambda > 1.0 \]
Bootstrap \( \lambda \) to derive confidence intervals encompassing the \( \lambda = 1.0 \) criterion.

Scree plot.

Broken-stick.

Total variance \( (\sum \lambda) \) divided randomly amongst the axes, eigenvalues follow a broken stick distribution.

\[ b_k = \sum_{i=1}^{p} \frac{1}{r} \]
\( p = \) number of variables \( (= np^2) \)
\( b_k = \) size of eigenvalue

e.g. 6 eigenvalues

\% variance - 40.8, 24.2, 15.8, 10.7, 6.1, 2.8
Proportion of variance  75%  95%

Statistical tests
- Bartlett’s test of sphericity: is each eigenvalue different from remaining ones?
- Bartlett’s equality test of \( \lambda_1 \)
- Lawley’s test of \( \lambda_1 \)
- Bootstrap eigenvalue 100 times
  Mean, minima, maxima, 95% confidence intervals
- If overlap between pairs of successive eigenvalues, eigenvalues indistinguishable

Applications:
1. \( \lambda > 1.0 \)  Gave too many components
2. Bootstrap \( \lambda \)  Gave too many components
3. Scree plot  Gave too many components
4. Broken stick  Correct
5. Proportion of variance  Unreliable
6. Bartlett & Lawley tests  Overestimated
7. Bootstrap eigenvalue  Correct except for poorly structured data
   (Eigenvalue > 0)

BROKEN STICK - simple to calculate, easy to follow

---

SCALING OF ORDINATION SCORES IN PRINCIPAL COMPONENTS ANALYSIS, as implemented by CANOCO 3.1 and later

CHOICES
1. Euclidean distance biplot
2. Correlation biplot
3. Symmetric scaling

Negative (-1, -2, -3) for covariance-based scores

Negative scalings use raw data in which variables have different variances.
Resulting species scores are covariances between species and eigenvalues in -2.
Positive values adjust species scores for their variances by standardising data by species.
Resulting species scores are correlations between species and
eigenvalues in scaling 2.

1. Euclidean distance biplot - focus on samples, distance between samples
   approximates Euclidean distance in species-space.
2. Correlation biplot - focus on species. Reflects correlations between species.
3. Intermediate between 1 and 2.

No extra mathematical properties. Compromise.

But see Gabriel (2002) Blometrika - may be optimal!

---

GENERAL INTERPRETATION OF PCA PLOTS

1. Samples close together are inferred to resemble one
   another in species composition.
2. There is a tacit assumption that samples with similar
   species composition have similar environments.
**PRINCIPAL COMPONENTS BIPILOTS & DIVERSITY**


In non-centred PCA of proportional data and distance scaling ('distance biplot') the squared length of each site vector \( p_i \) is

\[
\| p_i \|^2 = \sum_{j=1}^{n} p_{ij}^2 = SI
\]

and is exactly the formula for Simpson's Index. It can be transformed into a measure of alpha-diversity (within-site diversity) as \( 1/2SI \) (Hill's N2).

Squared Euclidean distances between sites

\[
d_{ij}^2 = \sum_{k=1}^{n} (p_{ik} - p_{jk})^2
\]

is a measure of dissimilarity between sites and can be regarded as a measure of beta-diversity (between-site diversity). 'Beta or between-habitat diversity refers to the degree of contrast in species composition among samples of a set taken from a landscape' (Whittaker, 1973).

Beta-diversity is simply the dissimilarity between two samples. If there are more than two samples, their mean dissimilarity is a measure of their beta-diversity.

**ALPHA-DIVERSITY (WITHIN-SITE DIVERSITY)**

<table>
<thead>
<tr>
<th></th>
<th>1920</th>
<th>1978</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simpson Index</td>
<td>A B C D E F</td>
<td>a b c d e f</td>
</tr>
<tr>
<td></td>
<td>0.46 0.41 0.23 0.16 0.15 0.12</td>
<td>0.74 0.93 0.88 0.78 0.32 0.60</td>
</tr>
</tbody>
</table>

In PCA biplot samples, the Simpson Index values are approximated by the origin position (0.2777) plus the squared distance from the projection of the origin.

D, E, and F are close to asterisk, hence have high alpha diversity (Simpson = 0.12 - 0.16), whereas A is far away and has low diversity (Simpson 0.46), a, b, c, d, and f are far from asterisk, and hence have high Simpson values and low alpha diversity.

More efficient to use a species-centred PCA than a non-centred PCA to represent beta-diversities. Can calculate where the true origin of the axes is if the PCA was non-centred so as to represent alpha-diversity.
BETA-DIVERSITY
(BETWEEN-SITE DIVERSITY)

CORRESPONDENCE ANALYSIS (CA)

Invented independently numerous times:

1. Correspondence Analysis: Weighted Principal Components with Chi-squared metric.

2. Optimal or Dual Scaling: Find site and species scores so that (i) all species occurring in one site are as similar as possible, but (ii) species at different sites are as different as possible, and (iii) sites are dispersed as widely as possible relative to species scores.

3. Reciprocal Averaging: species scores are weighted averages of site scores, and simultaneously, site scores are weighted averages of species scores.

CORRESPONDENCE ANALYSIS

Weighted averaging regression

\[ \hat{u}_i = \frac{\sum_{k=1}^{n} y_{ik} x_i}{\sum_{k=1}^{n} y_{ik}} \]

Species optimum (species score)

Can also have weighted averaging calibration

\[ \hat{x}_i = \frac{\sum_{k=1}^{n} y_{ik} \hat{u}_i}{\sum_{k=1}^{n} y_{ik}} \]

Environmental variable (site score)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>dist (x100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>119</td>
<td>45</td>
<td>21</td>
<td>89</td>
<td>85</td>
<td>78</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>138</td>
<td>56</td>
<td>30</td>
<td>108</td>
<td>104</td>
<td>97</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>132</td>
<td>52</td>
<td>27</td>
<td>103</td>
<td>99</td>
<td>92</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>75</td>
<td>3</td>
<td>14</td>
<td>47</td>
<td>34</td>
<td>34</td>
<td>31</td>
<td>41</td>
<td>38</td>
<td>32</td>
<td>30</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>104</td>
<td>38</td>
<td>13</td>
<td>74</td>
<td>70</td>
<td>63</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Samples close together have low pairwise squared Euclidean distances, hence low beta-diversity.

1920 sample (A-F) have mean squared distance = 0.41

1978 samples (a-f) have mean squared distance = 0.12

Indicates decrease in alpha-diversity of pools and the beta-diversity between pools has decreased between 1920 and 1978. Interpreted as result of acidification.
Two-way weighted averaging

Species optimum estimation by WA regression

\[ \hat{u}_k = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i} \]

Environmental variable estimation by WA calibration

\[ \bar{x}_k = \frac{\sum_{i=1}^{m} y_i u_{ik}}{\sum_{i=1}^{m} y_i} \]

Important in environmental reconstructions

CORRESPONDENCE ANALYSIS

Simultaneously ordinates both species and samples.
Usual to standardise site scores to weighted mean of 0 and variance 1.

Dispersion of species scores

\[ \sigma^2 = \frac{\sum_{i=1}^{m} y_i u_{ik}^2}{\sum_{i=1}^{m} y_i} - \frac{1}{\sum_{i=1}^{m} y_i} \]

Dispersion steadily increases and stabilises at maximum value. Maximum dispersion = CA axis 1. Maximises dispersion of species scores or 'optima'.

Can extract axis 2 in similar way but with the complication that the trial scores of axis 2 are made uncorrelated with scores of axis 1. Eigenvalue is maximised dispersion of species scores on ordination axis. \( \lambda_1 > \lambda_2 \), etc. Lies between 0 and 1.

Eigenvectors are species scores or 'optima'.
UNIMODAL RESPONSE MODEL

- Optimal scaling tries to pack species occurrences into tight packets: unimodal species-packing response model
- Eigenvalue tells the success of packing - but too high a value (-1) indicates disjunct subsets of sites

J. Oksanen (2002)

HILL’S SCALING IN CA

We standardised site scores to unit variance. Want them to be spread out most along the most important axes.

Hill proposed that the average width of the species curves should be the same for each axis. Calculate, for each species, the variance of the site scores containing the species and take weighted average of these variances.

Equalise average curve width among different axes by dividing all scores of an axis by its average curve width.

i.e. divide site scores by \( \frac{1}{\lambda} \) and species scores by \( \frac{1}{\sqrt{\mu}} \)

Scores so obtained are in multiples of one standard deviation or “TURNOVER”. Sites differing by 4SD in scores tend to have no species in common.

TRANSFORMATIONS
Log, square root
GOOD APPROXIMATION to
ML Gaussian ordination

CA Joint Plot

CA ordination diagram of the Dune Meadow Data in Hill’s scaling. In this and the following ordination diagrams, the first axis is horizontal and the second axis vertical; the sites are represented by crosses.
\( \lambda_1 = 0.26, \lambda_2 = 0.17 \)

CA: Joint Plot Interpretation

Joint plot with weighted Chi-squared metric: species and sites in the same plot with Hill’s scaling.

- Distance from the origin: Chi-squared difference from the profile
- Points at the origin either average or poorly explained
- Distant species often rare, close species usually common
- Unimodal centroid interpretation: species optima and gradient values - at least for well-explained species
- Can also construct CA biplots
- Samples close together are inferred to resemble one another in species composition
- Samples with similar species composition are assumed to be from similar environments

J. Oksanen (2002)
WHAT IS THE CHI-SQUARED METRIC?

Implicit in correspondence analysis, just as Euclidean distance or standardised Euclidean distance is implicit in PCA.

CA used ‘expected’ abundances as metric. Expected abundances from the marginal totals. Exactly like in χ² analysis of contingency tables.

Species profile is the average proportion of species in the data

Site profile is the average proportion of sites

Euclidean: \( d_{ij} = f_{ij} \cdot e_{ij} \)

Chi-squared: \( \chi_{ij} = (f_{ij} \cdot e_{ij}) / e_{ij} \)

Weights in CA are proportional to the square root of site totals

Euclidean distance implicit in PCA involves **absolute** differences of species between sites.

Chi-squared distance involves **proportional** differences in abundances of species between sites.

Differences in site and species totals are therefore less influential in CA than in PCA unless some transformation is used in PCA to correct for this effect.
DETRENDED CORRESPONDENCE ANALYSIS (DCA)

Aim to correct three ‘artefacts’ or ‘faults’ in CA:

1. **Detrending** to remove ‘spurious’ curvature in the ordination of strong single gradients
2. **Rescaling** to correct shrinking at the ends of ordination axes resulting in packing of sites at gradient ends
3. **Downweighting** to reduce the influence of rare species

Implemented originally in DECORANA and now in CANOCO

---

DETRENDING CA: THE ARGUMENT

PCA and CA both produce a curve from a single, ideal gradient, but the shapes have one important difference:

- Horseshoe in PCA curved inwards at ends: wrong order along the first axis
- Arch in CA preserves the correct ordering along the first axis: worth detrending

---

THE CAUSE OF THE ARCH IN CA

- There is a curve in the species space, and PCA shows it correctly.
- CA may be able to deal with unimodal responses, but if there is one dominant gradient, the second axis is the first axis folded. Occurs when the first axis is at least twice as long as the second ‘real’ axis.
- Problems clearly arise when there is one strong dominant gradient.
Defining terms and concepts.

**Detrending by segments - Basic Idea**

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).

In the two-way CA algorithm, replace the orthogonalisation step by detrending to remove any systematic dependences between axes.

**Alternative detrending procedures**

Detrend by polynomials - in the two-way CA algorithm, trial site scores for a particular axis are made to be uncorrelated to the ordination axes already extracted (orthogonalisation).

In detrending-by-polynomials they are also made uncorrelated to the $k$th order polynomials of the axes already extracted ($k = 2, 3, \text{ or } 4$) and to first-order cross-products of these axes.

When there is an arch, CA axis 2 is approximately a quadratic function (= second-order polynomial) of CA axis 1. Detrending by second-order polynomials should therefore remove the arch effect.

But this may not be enough because when there is a very dominant first gradient, CA axis 3 can be a function of CA axis 1, namely a cubic function, and CA axis 4 can be a quartic function, thereby obscuring the true second underlying gradient.

Promise of detrending-by-polynomials not supported by experiments with simulated or real data.

Detrending-by-segments consistently performed best.

**Detrending artefacts**

DCA configuration often a triangle, diamond, or ‘trumpet’.

DCA detrends by twisting the space so that empty space is left in the middle of the points. Not really seen in 2-dimensional plots.

In 2-dimensional plots, one end of DCA axis 1 (usually the upper end) has a wider range along DCA axis 2. This does not mean that there is greater variation at that end of the gradient.

Twisting often exposes this variation.

Any variation at the opposite end can get twisted into DCA axis 3.

Carefully inspect DCA plots (2-D and 3-D) and their eigenvalues.
NON-LINEAR RESCALING IN DCA

Assume a species-packing model, variance of optima of species at a site (‘within-site variance’) is an estimate of average response curve breadth (‘tolerance’) of those species. Because of edge effect, species curves are narrower at edges of axes than in centre and within-site variance is correspondingly smaller in sites near the ends.

Rescale by equalising within-site variance at all points along axis by dividing into small segments, expand those with small within-site variance and contract those with large within-site variance.

Site scores then calculated as WA of species scores and standardised so that within-site variance is 1.

Length of axis is range of site scores in ‘standard deviation’ units.

DCA - close approximation to ML Gaussian ordination with simulated data based on two-dimensional species-packing model.

DOWNWEIGHTING OF RARE TAXA IN CA AND DCA

Downweighting - to reduce the abundances of rare species so that they have even less weight on the ordination than if they were not downweighted.

CA and DCA sensitive to species that occur only in a few species-poor sites.

If AMAX is the frequency of the commonest species, effect of downweighting is to reduce the abundance of species rarer than (AMAX/5) in proportion to their frequency.

Rare species can only distort the analysis if they appear in samples with few other more common species. These are, by definition, ‘deviant samples’.

The same effect can often be achieved, more elegantly, by deleting these deviant samples (unpopular!) or by making them passive or supplementary.

Do not take part in the analysis but positioned in the final results as if they had taken part.

DCA PLOTS

• Axes scaled so that the smallest sample score is 0. Species scores have a wider dispersion than samples scores. Some will be negative.
• Axes taken as gradients, and scaled in ecologically meaningful ‘sd’ units.
• Species scores taken as species optima.
• Species scaled so that site scores are their direct weighted averages ($w = 1$).

![DCA ordination of the Dune Meadow Data. The scale marks are in multiples of standard deviations (s.d.). ($w_1 = 0.53, w_2 = 0.29$; axis 1 = 3.7 sd, axis 2 = 3.1 sd)](image)

J. Oksanen (2002)
**WHEN DOES DCA APPROXIMATE THE GAUSSIAN RESPONSE MODEL?**

**CAJO TER BRAAK** (1985 Biometrics 41, 870)

“Four conditions (equal tolerances, equal or independent maxima, and equally-spaced or uniformly distributed optima and sample points) are needed to show that (detrended) correspondence analysis provides an approximate solution to the unimodal models.”

- (D)CA can approximate the Gaussian model if we have infinite species packing gradients. Inevitable that some distortions occur, especially at gradient ends. Still a good approximation.

Despite its heuristic nature, DCA usually very useful.
**ORDERED VEGETATION TABLE**

<table>
<thead>
<tr>
<th>Plant</th>
<th>PCA</th>
<th>CCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 2</td>
<td>3 1</td>
<td>4 1</td>
</tr>
<tr>
<td>3 1 4 2</td>
<td>3 1</td>
<td>4 1</td>
</tr>
<tr>
<td>3 1 4 2</td>
<td>3 1</td>
<td>4 1</td>
</tr>
<tr>
<td>3 1 4 2</td>
<td>3 1</td>
<td>4 1</td>
</tr>
</tbody>
</table>

J. Oksanen (2002)

**SCALING OF CA ORDINATION SCORES IN CORRESPONDENCE ANALYSIS, as implemented by CANOCO 3.1 and later**

**Choices:**
1. Sample scores are weighted mean species scores.
2. Species scores are weighted mean sample scores.
3. Symmetric scaling. May be good compromise
   - Negative (-1, 2, 3) for Hill scaling
   - Positive scalings standardise scores to \( \lambda^\alpha \)
   - Negatives standardise scores to \( \lambda^\alpha / (1 - \lambda) \)
   - Where \( \alpha = 0, 0.5 \) or 1.

**PCA or CA/DCA?**

**PCA** - linear response model

**CA/DCA** - unimodal response model

**How to know which to use?**

**Gradient lengths important**
- If short, good statistical reasons to use LINEAR methods.
- If long, linear methods become less effective, UNIMODAL methods become more effective.

**Range 1.5-3.0 standard deviations both are effective.**

**In practice:**
- Do a DCA first and establish gradient length.
- If less than 2SD, responses are monotonic. Use PCA.
- If more than 2SD, use CA or DCA.

When to use CA or DCA more difficult.
- Ideally use CA (fewer assumptions) but if arch is present, use DCA.

**DCA results can be unstable when eigenvalues 1 and 2 are close to each other (e.g. 0.55, 0.54) (Oksanen 1988) Vegetatio 74, 29-32). Always do a CA to assess the effect of downtrending on the data-set.**

**RARE SPECIES IN CA AND DCA**

Rare species cannot have an average profile and are extreme in CA and DCA, but still have small weights.

Solutions to the problem:
1. Remove rare species: unpopular
2. Downweight rare species: CANOCO option
3. Do not show rare species: popular
4. Make rare species "passive", rarely done
5. Make deviant samples "passive", rarely done

J. Oksanen (2002)
The response curves of 3 species along a gradient; 12 quadrats are located at the numbered points marked with arrowheads (artificial data).

Ordinations of the 12 data points by PCA (hollow dots, dashed line) and by CA (solid dots, solid line). Both ordinations exhibit the arch-effect. The CA ordination also shows scale contractions at both extremities.

(a) The response curves of 3 species along a gradient; 12 quadrats are located at the numbered points marked with arrowheads (artificial data).

(b) Ordinations of the 12 data points by PCA (hollow dots, dashed line) and by CA (solid dots, solid line). Both ordinations exhibit the arch-effect. The CA ordination also shows scale contractions at both extremities.

Outline of ordination techniques. DCA (detrended correspondence analysis) was applied for the determination of the length of the gradient (LG). LG is important for choosing between ordination based on a linear or on a unimodal response model. In cases where LG < 3, ordination based on linear response models is considered to be the most appropriate. PCA (principal component analysis) visualizes variation in species data in relation to best fitting theoretical variables. Environmental variables explaining this variation are deduced afterwards, hence, indirectly. RDA (redundancy analysis) visualizes variation in species data directly in relation to quantified environmental variables. Before analysis, covariables may be introduced in RDA to compensate for systematic differences in experimental units. After RDA, a permutation test can be used to examine the significance of effects.
Curving of Gradients

- Eigenvector methods project species space into ordination space
- If there is a curvature in the ordination, there may not be a curvature in the data
- The model behind ordination may be different from the vegetation model
- PCA assumes linear relation between axes and latent variables
- CA assumes unimodal model
- In both cases a curvature ("horseshoe", "arch") can be produced from a "straight" gradient
- All ordination methods are based on dissimilarity measures (hidden in CA)
- If there are no species in common between two sites, it is difficult to estimate their similarity (cf. extended dissimilarity approach)

HOW TO HANDLE CURVATURE

1. Accept and acknowledge probably only one dominant gradient (try Principal Curves to check).
2. Degree of absence - if the curve is caused by the 'naughty noughts', estimate how many and how much the species are absent.
3. Extended dissimilarity - if all distant points have nothing in common, try to estimate dissimilarity using stepping-stone points.
4. Detrend as in DCA - remove curvature in ordination. May also remove other structure and introduce a 'trumpet' or 'triangle' concentration of sample points with a wider range of sample scores on DCA axis 2 at one end (usually the upper end) of DCA axis 1. Does not mean that there is greater variation in the data at that end!
5. Monotonic regression as in non-metric scaling - does not require linear relation between dissimilarities and ordination distances.
6. Constrain - use of external variables to constrain ordination axes seems to remove curvature. Direct gradient analysis (CCA, RDA).

TRANSFORMATION OF SPECIES ABUNDANCES IN ORDITION

Not a well-explored topic.

<table>
<thead>
<tr>
<th>Method</th>
<th>Transformation</th>
<th>Reason 1</th>
<th>Reason 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCA/CA/CCA</td>
<td>Work well with raw data, +/- data, log transformed, square root transformed</td>
<td>Stabilise variance</td>
<td>Dampen effects of very abundant species</td>
</tr>
<tr>
<td>PCA/RDA</td>
<td>Square root with % data</td>
<td>Log double centring for % data with few variables</td>
<td>Log for abundance data</td>
</tr>
</tbody>
</table>

Suggest CA/DCA with untransformed, square root, and log (y + 1).

Look at total variance (= inertia) and relative sizes of axes 1, 2, and 3. See which reduces variance without destroying axis 1, 2, 3 structure.

MULTIDIMENSIONAL SCALING (MDS) RATIONALE

- Map points into low-dimensional ordination space so that the distances in the ordination space are as close to the original dissimilarities as possible
  - regression from measured dissimilarities to ordination distances
  - principal co-ordinates analysis or metric scaling (PCORD)
- Non-metric MDS: Real distances have the same rank order as the ordination distances
  - Monotonic regression
  - No specified shape of the regression
  - Goodness of fit measured by 'stress'
  - NMSRCA

Crucial: How to measure distances in a meaningful way?

Question of selection of dissimilarity or proximity coefficient. Critical problem, as in classification procedures

**DISSIMILARITIES FOR ECOLOGICAL DATA**

- Presence/absence indices based on the number of common species \( J \) on two sites compared to species richness \( A, B \) of sites.
- Similarity index \( s \) can be transformed to a dissimilarity index \( d = 1 - s \).
- Quantitative generalisations:
  1. **Manhattan style:** use common part of abundance and sums.
  2. **Euclidean style:** use cross products and sums of squares.

\[
\text{Jaccard: } J = \frac{a}{a + b + c} \\
\text{Sørensen: } S = \frac{2a}{2a + b + c}
\]

\( J = a, A = a + b, B = a + c \)

**J. Oksanen (2002)**

**POSSIBLE PROXIMITY MEASURES**

**Manhattan style**

\[
J = \sum_1^t \min(x_{ij}, x_{ik})
\]

**Euclidean style**

\[
J = \sum_1^t \sqrt{x_{ij}^2 + x_{ik}^2}
\]

- **0/1**
- **Manhattan style**
- **Euclidean style**

**J. Oksanen (2002)**

**PRINCIPAL CO-ORDINATES ANALYSIS - PCORD**

J.C. Gower 1966 Biometrika 53: 325-338

Start with a matrix of similarities \( A \) or dissimilarities \( D \)

\[
a_{ij} = -{1/2}d_{ij}^2
\]

Transform \( A \) by subtracting column and row means and adding matrix mean.

Extract eigenvalues and eigenvectors, scale each eigenvector so that sum of squares = eigenvalue (multiply by \( \sqrt{\lambda} \)).

Scaled eigenvectors are co-ordinates of objects whose distances apart are given by best-fitting possible approximations to original distances in chosen number of dimensions.

Measure of fit

\[
\sum_{i=1}^2 a_i / \sum_{i=1}^n a_i
\]

Species can be added as their weighted sums (cf. PCA) or weighted averages (cf. CA) of the site scores.

Note that \( d_{ij}^2 \) is Euclidean distance, PCORD = PCA

If \( d_{ij}^2 \) is squared chi-squared distance, PCORD = CA

**R**
NON-METRIC MULTIDIMENSIONAL SCALING - NMDSCAL

Metric scaling
\[ d_i \sim d'_i \]
or \[ d_i^2 \sim d'_i^2 \]

Non-metric scaling
\[ d_i > d_k > d_j \Rightarrow d_i' = d_k' = d_j' \]
i.e. monotonic relationship between original and new distances.

Use ranks, try to find a low-dimensional representation that minimises 'stress', a measure of how monotonic are \( d' \) with \( d \). No unique solution. Iterative procedures.

NON-METRIC MULTIDIMENSIONAL SCALING (NMDSCAL)

- Rank-order relation with community dissimilarity and ordination distance: no specified form of regression, but the best shape is found from the data. Monotonic regression.
- Non-linear regression can cope with non-linear species responses of various shapes: not dependent on Gaussian model.
- Iterative solution: no guarantee of convergence
- Must be solved separately for each number of dimensions: a lower dimensional solution is not a subset of a higher, but each case is solved individually.

MONOTONIC REGRESSION IN KRUSKAL’S NMDSCAL ALGORITHM

- Measured community dissimilarities and ordination distances have similar rank ordering.
- No specified shape, but can cope with different response shapes.
- Sum of squared residuals from the regression: Stress.
- The basic model: finds gradients if dissimilarities meaningful.
- Iterative solution: no guaranteed convergence.
- Shepard plot comparing dissimilarities with ordination distance.

J. Oksanen (2002)

NMDSCAL: Algorithm

- Take arbitrary initial site scores
- Move points so that stress decreases
- Stop when stress is stable: Minimum found
- Must be solved separately for each dimensionality
- The minimum may be local instead of global
- Use several starting configurations and compare the solutions to find the stable one with minimum stress
- Species scores can be estimated as weighted averages or weighted sums and can be “stretched” to make them comparable with CA or PCA species scores

NMDSCAL: Scores, stress and plots

- Stress cannot be meaningfully compared between data sets
- Site scores are arbitrary
- Site scores can be scaled to “half change”
- Species scores are supplementary

NMDSCAL plot is a map:
- All directions equal: No axes
- No origin
- However, often rotated to principal components
**RECOMMENDED PROCEDURE IN NMDSCAL**

1. Use adequate dissimilarity indices: an adequate index gives a good rank-order relation between community dissimilarity and gradient distance
2. No convergence guaranteed: start with several random starts and inspect those with lowest stress
3. Satisfied only if minimum stress configurations are similar

**NMDSCAL PLOT**

- All that counts is the configuration: axes have no direct meaning
- No origin, just a map
- Species can be added as their weighted averages or weighted sums

**HOW MANY DIMENSIONS TO USE IN NMDSCAL?**

- In eigenvector methods axes are orthogonal and previous axes remain unchanged when new axes are evaluated
- NMDSCAL solutions are separate for each number of dimensions
- Adequate number of dimensions difficult to know; after sudden drop of stress, reached adequate number of dimensions

**METRIC AND NON-METRIC SCALINGS ONLY PROVIDE MAPS**

- Scaling tries to find an underlying configuration from dissimilarities
- Scaling tries to draw a map using distance data
- Metric scaling (PCORD) assumes linear relation, but NMDSCAL finds a ranked relation
- Only the configuration counts: - no origin, but only the constellations - no axes or natural directions, but only a framework for points

J. Oksanen (2002)
Scaling of NMDSCAL Axes

NMDSCAL axes have no unique scaling or direction, and all rotations and scaling are equally good solutions.

- Customary to rotate to principal components: first dimension most important
- Half-change scaling gives ecologically meaningful units for the axes

Comparing Ordination Methods

- Ordination methods cannot be compared with real data sets: the truth is unknown
  - The correct structure is inferred from the data, and the comparison may be biased towards the favoured result
- Comparison needs external criteria (e.g. environmental variables)
- Simulated community pattern:
  - Assume an interpretable gradient pattern, and see if the method can find this pattern; reliable only if it finds the known pattern
  - Robustness is the ability to work even when the assumptions are violated
  
  COMPAS, COENOFLEX, R (Vegan)

Methods of Unconstrained Ordination of a Multivariate Set of Data, Y.

<table>
<thead>
<tr>
<th>Name of method (acronyms, synonyms)</th>
<th>Distance measure preserved</th>
<th>Relationship of ordination axes with original variables</th>
<th>Criterion for drawing ordination axes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Components Analysis (PCA)</td>
<td>Euclidean distance</td>
<td>Linear</td>
<td>Finds axes that maximise the total variance (or, equivalently, that minimises the total residual variation)</td>
</tr>
<tr>
<td>Correspondence Analysis (CA, reciprocal averaging, dual scaling)</td>
<td>Chi-square distance</td>
<td>Unimodal (approximately Gaussian)</td>
<td>Finds axes that maximise dispersion of species scores (which are themselves weighted averages of site scores)</td>
</tr>
<tr>
<td>Principal Coordinates Analysis (PCO, PCAs, metric multidimensional scaling, classical scaling, Torgerson scaling)</td>
<td>Any chosen distance or dissimilarity measure</td>
<td>Unknown; depends on distance measure chosen</td>
<td>Euclidean distances in new full-dimensional space are equal to original distances (or dissimilarities)</td>
</tr>
<tr>
<td>Nonmetric Multidimensional Scaling (NMD, NMDSCAL)</td>
<td>Any chosen distance or dissimilarity measure</td>
<td>Unknown; depends on distance measure chosen</td>
<td>The number of dimensions for the new space is chosen a priori (reduced); Euclidean distances in new space are monotonically related to original distances</td>
</tr>
</tbody>
</table>
**IS THERE AN ‘IDEAL’ METHOD FOR INDIRECT GRADIENT ANALYSIS?**

The ‘ideal’ ordination method should possess the following qualities:

1. It recovers gradients without distortion.
2. If clusters exist in nature, these should be obvious in the ordination.
3. It gives the same result every time for a given data set.
4. There is a unique solution.
5. Ecological similarity is related to proximity in ordination space.
6. Scaling of axes is related to beta-diversity.
7. The method is not sensitive to noise.
8. ‘Signal’ and ‘noise’ are easily separated.
9. You do not need to pre-specify the number of axes.
10. The solution is the same, no matter how many dimensions one selects to examine.
11. Unless by choice, all objects are treated equally.
12. The solution does not take too much computer time.
13. The method is robust - works well for short or long gradients; for low or high noise; for sparse or full data matrices; for big or small data sets; for species-poor and species-rich data.
14. To the statistician it is elegant.
15. To the ecologist it is available, easy to use, and inexpensive.

**CONCLUSIONS**

Non-metric scaling  Usefull with ‘poor’ data, missing values. Q-mode scaling (n x n matrix). 400-500 objects.

Metric scaling + principal coordinates analysis Any distance measure, sample scores only, species scores can be roughly estimated, interpretation difficult. Linear or unimodal - depends on DC.

Correspondence analysis & DCA Specific $\chi^2$ distance, unimodal response model, species and site scores, interpretation moderate, relative abundances, ‘shape’.

Principal components analysis Euclidean distance, species and site scores, linear response model, interpretation via biplots easy, absolute abundances, ‘size’.

**CURRENT STATUS OF INDIRECT GRADIENT ANALYSIS METHODS**

<table>
<thead>
<tr>
<th></th>
<th>PCA</th>
<th>CA</th>
<th>DCA</th>
<th>PCORD</th>
<th>NMSICAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good gradient recovery and no distortion</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Find clusters</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Same result every time</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Unique solution</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ecological similarity</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scaling of axes, diversity</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Not sensitive to noise</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Signal/noise separation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Do not have to specify axes number</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Solution same, irrespective of axis number</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Objects treated equally</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Computer time</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Can be slow</td>
<td>Slow</td>
</tr>
<tr>
<td>Robust</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Elegant</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ecologist - available, easy to use, inexpensive</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Current status (15)</td>
<td>11.5</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>6.0</td>
</tr>
</tbody>
</table>

What to do with data containing many zero values and long gradients, and to avoid the inherent problems of the Chi-squared distance metric implicit in CA and CCA?

These are:

1. Implicit use of relative abundances
2. A difference between abundance values for a common species contributes less to the distance than the same difference for a rare species, so rare species may have an unduly large influence on the analysis

Solutions:

1. Delete rare species
2. Empirical downweighting of rare species as in CANOCO
3. Data transformations that preserve the Euclidean distances and balance rare and common species


Software: www.fas.umontreal.ca/biol/legendre
(1) Chord distance

$$D_{ij} = \sqrt{\frac{1}{n} \sum_{j} \left( y_{ij} \right)^2}$$

(2) Chi-square metric

$$D_{ij} = \sqrt{\frac{1}{n} \sum_{j} \frac{(y_{ij} - \bar{y}_j)^2}{\bar{y}_j}}$$

where \( y_{ij} \) is the column (species) sum for species \( j \), and \( \bar{y}_i \)

is the row (sample) sum for sample \( i \).

(3) Species profiles

$$D_{ij} = \sqrt{\frac{1}{n} \sum_{j} \frac{(y_{ij} - \bar{y}_j)^2}{\bar{y}_j}}$$

(4) Hellinger distance

$$D_{ij} = \sqrt{\frac{1}{n} \sum_{j} \frac{(y_{ij} - \bar{y}_j)^2}{\bar{y}_j}}$$

Note that with percentage or proportional data, Hellinger distance is the same as
square-root transformation and Euclidean distance (also called chord distance!)

PRINCIPAL CURVES

Principal curves are smooth one-dimensional curves in a high-dimension space.
Form of non-linear PCA, analogous to LOESS smoothers as a non-linear regression tool.
Principal curves minimise sum of squares distances from data (as does PCA) but to a curve, not to a line or plane as in PCA.

Two species along single gradient
Principal curve showing gradient location

\( R \), \( S-Plus \)

Relation of distances along principal curve and original gradient. Rates high when species abundances change rapidly (centre of gradient), low when species constant (ends of gradient).
FITTING PRINCIPAL CURVES

25 sites, 4 species, Gaussian responses, one gradient. Plotted on first two PCA axes. Iterative fitting of principal curves.

(a) Data using PCA axis 2 (39.4%) as initial curve
(b) First iteration, smooth spline 3 d.f.
(c) Convergence with 3 d.f.
(d) Improved and final fit with 7 d.f. 98.3% variance
(e) Result of using PCA axis 1 (50.4%) as start and 3 d.f.
(f) As (e) but 7 d.f.

PRINCIPAL CURVES AND REAL DATA

12 hunting spiders at 28 sites and 6 environmental variables

Principal curve superimposed on PCA biplot. Numbers are locations along the gradient. Principal curve captures 90% of species variance. Modelled environmental variable values for 6 locations show PC is mainly moisture, sand, moss and twig gradient.

SPECIES RESPONSES ALONG PRINCIPAL CURVES

All are unimodal and optima well approximated by intersection of species vectors with the principal curve. Curves have approximately equal tolerances.

Ideal for finding one-dimensional gradients that explain species composition as well, or better than, higher dimensional ordination methods. Have been extended to two-dimensional gradients as principal surfaces.

Less restrictive in assumptions that PCA, CA, or DCA. Only assume that responses are smooth. Very neutral method (cf. LOESS in regression).

Computationally difficult, hardly used yet...
CURRENT USES OF INDIRECT ORDINATION METHODS


“Ordination is a rather artificial technique. The idea that the world consists of a series of environmental gradients, along which we should place our vegetation samples, is attractive. But this remains an artificial view of vegetation. In the end the behaviour of vegetation should be interpreted in terms of its structure, the autecology of its species and, above all, the time factor. At this level, trends become unimportant and multivariate analysis is perhaps irrelevant. Ordination is useful to provide a first description but it cannot provide deeper biological insights.”

COMPARING DIFFERENT ORDINATIONS

PROCURSTES ROTATION

Two configurations of points in ordinations representing the same n samples.

Take one configuration as fixed, move the other to match as closely as possible and to minimise the sum of squared distances of the transformed points from the respective points of the fixed configuration

1. Translation of origin - shift the origins of the co-ordinate axes
2. Rotation and/or reflection of axes
3. Uniform scaling (deflate or inflate the axis scale)

Single points can move a lot although the sum of squared distances can stay fairly constant, especially in large data sets.

R (Vegan)

MERITS AND DRAWBACKS OF INDIRECT ORDINATION METHODS

1. Can distract attention from individual species responses by focussing on the overall multivariate response only.
2. As it is a correlative method, it can help with hypothesis generation.
3. It can rarely, if ever, demonstrate causality.
4. Ordinations can provide useful low-dimensional representations of complex data. Valuable for summarisation and for hypothesis generation.
5. Ordination is a tool and a means to an end. It is not an end in itself.
Yield and pH explain about 65% of the sum of squares in the PCA ordination. Some plots do not fit well e.g. 14a, 14d, 11/1b.

Results of a Procrustes rotation of the yield/pH scatter diagram on to the two-dimensional PCA for Park Grass plots

<table>
<thead>
<tr>
<th>Rotation matrix H</th>
<th>-0.411 -0.912</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaling parameter p</td>
<td>3.97</td>
</tr>
</tbody>
</table>

Sum of squares
- Fitted yield/pH configuration: 260
- Residual: 128
- Total for PCA ordination: 428

Configuration of Park Grass plots for standardised yield and pH: rotated, reflected and scaled to fit PCA ordination.

Lines connect labelled points for the rotated yield/pH ordination to their respective fixed points from the PCA.
GENERALISED PROCRUSTES ROTATION

Any number of configurations. Basic idea is to find a consensus or centroid configuration so that the fit of ordinary Procrustes rotation to this centroid over all configurations, is optimal. Minimise \( m^2 \) where \( m^2 = \sum m_i^2 \) where \( m_i^2 \) is Procrustes statistic for each pair-wise comparison.

1. Translation
2. Rotation and/or reflection
3. Scaling

SOFTWARE

<table>
<thead>
<tr>
<th>Technique</th>
<th>Program</th>
<th>Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>CANOCO, NT-SYS, R, R (Vegan)</td>
<td>CANOCO, DECORANA, NT-SYS, R (Vegan)</td>
</tr>
<tr>
<td>CA</td>
<td>CANOCO, DECORANA, NT-SYS, R (Vegan)</td>
<td>CANOCO, DECORANA, R (Vegan)</td>
</tr>
<tr>
<td>DCA</td>
<td>CANOCO, DECORANA, R (Vegan)</td>
<td>CANOCO, DECORANA, NT-SYS, R (Vegan)</td>
</tr>
<tr>
<td>PCORD</td>
<td>PCORD (+DCMAT), PROCOORD, NT-SYS, Dist-PCoA, R, R (Vegan)</td>
<td>PCORD (Marti Andersen)</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.stat.auckland.ac.nz/~mja/">www.stat.auckland.ac.nz/~mja/</a></td>
<td></td>
</tr>
<tr>
<td>Non-metric scaling</td>
<td>EUSCAL, NT-SYS, WINKYST, R, R (Vegan)</td>
<td>EUSCAL, NT-SYS, WINKYST, R, R (Vegan)</td>
</tr>
<tr>
<td>Simulations</td>
<td>COMPAS, COENOFLEX, R (Vegan)</td>
<td>COMPAS, COENOFLEX, R (Vegan)</td>
</tr>
<tr>
<td>Principal curves</td>
<td>5-PLUS, R</td>
<td>5-PLUS, R</td>
</tr>
<tr>
<td>Procrustes rotation</td>
<td>GENSTAT, SOLCOMP, R (Vegan)</td>
<td>GENSTAT, SOLCOMP, R (Vegan)</td>
</tr>
</tbody>
</table>

Example - to compare results of 12 different ordination methods to the same data

Location of ordination methods on the two principal co-ordinates axes; these two axes represent 75% of the variation in statistics.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Non-metric scaling</td>
</tr>
<tr>
<td>1,2</td>
<td>Presence/absence data only</td>
</tr>
<tr>
<td>C</td>
<td>Correspondence analysis</td>
</tr>
<tr>
<td>P</td>
<td>Principal co-ordinates analysis</td>
</tr>
<tr>
<td>3</td>
<td>Log abundance data</td>
</tr>
<tr>
<td>4</td>
<td>All joint absences ignored</td>
</tr>
<tr>
<td>5</td>
<td>Abundance data</td>
</tr>
</tbody>
</table>

MAJOR PLAYERS IN INDIRECT GRADIENT ANALYSIS

Karl Pearson 1901
Invented PCA

David W. Goodall 1954
First use of PCA in ecology

John C. Gower 1966
Popularised PCOORD, Invented Procrustes rotation

Mark O. Hill 1973
Popularised CA in ecology, 1980 DCA

Cajo J.F. ter Braak 1985
Unified PCA and CA in terms of response models

Jari Oksanen 1990s
Continuous questioning about ordination methods, championing NMDSCAL