Analysis of Multivariate Ecological Data

School on Recent Advances in Analysis of Multivariate Ecological Data

24-28 October 2016

Prof. Pierre Legendre
Dr. Daniel Borcard
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre Ville
Montréal QC
H3C 3J7 Canada
pierre.legendre@umontreal.ca
daniel.borcard@umontreal.ca
Brain storming in Pierre Legendre's lab

Daniel Borcard  
Pierre Legendre  
Pedro Peres-Neto  
Stéphane Dray
Day 1
Introduction to data analysis
Main references


• http://www.numericalecology.com/
Introduction to data analysis

# Import data into R
spe <- read.csv("DoubsSpe.csv", row.names=1)
env <- read.csv("DoubsEnv.csv", row.names=1)

# Hellinger-transform the species dataset
spe.hel <- decostand(spe, "hellinger")

# Redundancy analysis (RDA)
# *************************
# RDA of the Hellinger-transformed fish species data, constrained
# by all the environmental variables contained in env2
(spe.rda <- rda(spe.hel ~ ., env)) # Observe the shortcut
# formula
summary(spe.rda) # Scaling 2 (default)
# Canonical coefficients from the rda object
coef(spe.rda)
# R^2 retrieved from the rda object
Rsquare Adj(spe.rda)
Software

R 3.3.1 for Windows, Mac OS or Linux
Numerical ecology

“The field of quantitative ecology devoted to the numerical analysis of ecological data sets. (...) The purpose of numerical ecology is to describe and interpret the structure of data sets by combining a variety of numerical approaches. Numerical ecology differs from descriptive or inferential ecological statistics in that it combines relevant multidimensional statistical methods with heuristic techniques (e.g. cluster analysis) (...) ” (Legendre & Legendre, 2012).
Definitions

Numerical ecology

Let us add that a great number of the methods used in numerical ecology, especially the new approaches developed since the 1980’s, have been developed by ecologists (and not pure statisticians!), in response to specific ecological problems.
Definitions

**Multivariate, multidimensional analysis:** methods of numerical analysis addressing *whole data tables* where every observation, every sampling unit is characterised by several variables:

- species abundances
- climatic measures
- and so on...
The data

Table I - Structure of an ecological data table

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Objects</th>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Variable j</th>
<th>Variable p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td>$y_{11}$</td>
<td>$y_{12}$</td>
<td>...</td>
<td>$y_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>Object 2</td>
<td>$y_{21}$</td>
<td>$y_{22}$</td>
<td>...</td>
<td>$y_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object $i$</td>
<td>$y_{i1}$</td>
<td>$y_{i2}$</td>
<td>...</td>
<td>$y_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object $n$</td>
<td>$y_{n1}$</td>
<td>$y_{n2}$</td>
<td>...</td>
<td>$y_{nj}$</td>
<td>...</td>
</tr>
</tbody>
</table>

The **objects** are the observations (sites, relevés...).
The data

Species

Descriptors

Environnementsal variables

Spatial variables

Objects

$p$

$n$

$m$

$q$

$n$

$n$

$n$
Aims of multivariate analysis

- Measurement of resemblance among objects or variables of a data table;
- Clustering of the objects or variables according to these resemblances;
- Ordination in a reduced space allowing to emphasise their main structures (especially gradients);
- Modelling of the relationships between response data tables and explanatory variables;
- Tests of these relationships for statistical significance.
Ordination in reduced space
Ordination in reduced space

1. Generalities

<table>
<thead>
<tr>
<th>Obj.</th>
<th>Var.1</th>
<th>Var.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj.1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Obj.2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Obj.3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Obj.4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Obj.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Obj.6</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
2. Principal component analysis (PCA)
Ordination in reduced space

2. Principal component analysis (PCA)

![Diagram of principal component analysis (PCA)]
2. Principal component analysis (PCA)

One can conduct a PCA and display its results in different ways:

**Covariance or correlation?**

PCA on the raw(centred) variables (= PCA on a covariance matrix) is only appropriate when the variables are dimensionally homogeneous.

PCA on a correlation matrix: when the variables are dimensionally heterogeneous or when one wants an equal weight for all variables.
Ordination in reduced space

2. Principal component analysis (PCA)

One can conduct a PCA and display its results in different ways:

Scaling:

*PCA Scaling 1 = Distance biplot:* the eigenvectors are scaled to unit length; the main properties of the biplot are the following:

1. Distances among objects in the biplot are approximations of their Euclidean distances in multidimensional space.
2. Projecting an object at right angle on a descriptor approximates the position of the object along that descriptor.
2. Principal component analysis (PCA)

One can conduct a PCA and display its results in different ways:

Scaling:

*PCA Scaling 1 = Distance biplot:* the eigenvectors are scaled to unit length; the main properties of the biplot are the following:

(3) Since descriptors have length 1 in the full-dimensional space, the length of the projection of a descriptor in the reduced space indicates how much it contributes to that space.

(4) The angles among descriptor vectors do not reflect their correlations.
**PCA Scaling 1 = Distance biplot:** the eigenvectors are scaled to unit length; the main properties of the biplot are the following:

1. Distances among objects in the biplot are approximations of their Euclidean distances in multidimensional space.
2. Projecting an object at right angle on a descriptor approximates the position of the object along that descriptor.
3. Since descriptors have length 1 in the full-dimensional space, the length of the projection of a descriptor in the reduced space indicates how much it contributes to that space.
4. The angles among descriptor vectors do not reflect their correlations.
Ordination in reduced space

2. Principal component analysis (PCA)

**PCA Scaling 2 = correlation biplot:** the eigenvectors are scaled to the square root of their eigenvalue.

1) *Distances among objects in the biplot are not approximations of their Euclidean distances in multidimensional space.*

2) *Projecting an object at right angle on a descriptor approximates the position of the object along that descriptor.*
Ordination in reduced space

2. Principal component analysis (PCA)

(3) (3a) PCA on a covariance matrix: descriptors have length \( s_j \) (= their standard deviation) in the full-dimensional space; therefore, the length of the projection of a descriptor in the reduced space is an approximation of its standard deviation.

(4) (3b) PCA on a correlation matrix: all the descriptors have unit variance \( (s = 1) \); the length of the projection of a descriptor in the reduced space reflects its contribution to that space.

(5) The angles between descriptors in the biplot reflect their correlations.
PCA: scaling 2

Circle of equilibrium contribution

Circle of radius 1

PCA on a correlation matrix; Hellinger-transformed species density data; scaling type 2.
Ordination in reduced space

2. Principal component analysis (PCA) - in summary

One can conduct a PCA and display its results in different ways:

Scaling:

If the main interest of the analysis is to interpret the relationships among objects, choose scaling 1.

If the main interest focuses on the relationships among descriptors, choose scaling 2.
Ordination in reduced space

2. Principal component analysis (PCA)

Equilibrium contribution circle

• In all the options except covariance PCA + scaling 2.

• Equilibrium contribution: length that a descriptor-vector would have if it contributed equally to all the dimensions (principal axes) of the PCA.
Ordination in reduced space

2. Principal component analysis (PCA)

Equilibrium contribution circle

- The circle has a radius equal to $\sqrt{\frac{d}{p}}$, where $d$ equals the number of dimensions of the reduced space considered (usually $d=2$) and $p$ equals the total number of descriptors (and hence of principal components) in the analysis.

- In a covariance PCA+scaling, the equilibrium contribution must be computed separately for each descriptor, and is equal to $s_j\sqrt{\frac{d}{p}}$, where $s_j$ is the standard deviation of the descriptor considered.
PCA on a correlation matrix; Hellinger-transformed species density data; scaling type 2.
Ordination in reduced space

2. Principal component analysis (PCA)

Interpretation of meaningful components

Not all ordination axes are meaningful. Which ones are?

1. **Arbitrary** decision (ex.: interpretation of axes representing 75% variance)

2. Interpretation of axes whose eigenvalues are larger than the **average** of all eigenvalues (Kaiser-Guttman criterion)
Ordination in reduced space

2. Principal component analysis (PCA)

Interpretation of meaningful components

Not all ordination axes are meaningful. Which ones are?

3. **Broken stick model**: comparison of eigenvalues to a model predicting the (decreasing) sizes of the pieces of a stick broken in as many pieces as there are eigenvalues.
   
   1. Comparison of individual eigenvalues;
   2. Comparison of cumulative eigenvalues.
Ordination in reduced space

2. Principal component analysis (PCA)

Uses and misuses of PCA

Main application in ecology: ordination of sites on the basis of quantitative environmental variables (physical and chemical descriptors), not untransformed species abundances in most cases (see Chapter 2, double zero problem).
2. Principal component analysis (PCA)

Uses and misuses of PCA

PCA has originally been defined for data with multinormal distributions.

However, PCA is not very sensitive to departure from multinormality, as long as the distributions are not exaggerately skewed. If some are, then the first few PCA axes will display several objects with extreme values on these variables instead of main trends.
Ordination in reduced space

2. Principal component analysis (PCA)

Uses and misuses of PCA

PCA is computed from a matrix of dispersion (linear covariances or correlations).

This means that the method must be applied on a dispersion matrix among descriptors that are quantitative and for which valid estimations of covariances or correlations may be obtained.
Ordination in reduced space

2. Principal component analysis (PCA)

Uses and misuses of PCA

These conditions are violated in the following cases:

(1) The eigenvalues and eigenvectors of a dispersion matrix cannot be estimated using a number of objects \( n \) smaller or equal to the number of descriptors \( p \).

(2) The data matrix must not be transposed, since covariances or correlations among objects are meaningless.

(3) Covariances and correlations are defined for quantitative variables. However, PCA is very robust to variations in the precision of data.
Ordination in reduced space

2. Principal component analysis (PCA)

Uses and misuses of PCA

These conditions are violated in the following cases:

(4) With data sets with many zeros, PCA produce inadequate estimates of the site positions.

(5) Avoid the (all too common) mistake of interpreting the relationships among variables with help of the proximities of the apices (points) of the vector-arrows instead of the angles (when appropriate) on the biplots!
Transformation of species abundance data tables prior to linear analyses
Data transformation

0. Preliminary remarks: general types of transformation for any kind of variables
Data transformation

1. Make comparable descriptors that have been measured in different units:

Ranging: $$y_i' = \frac{y_i - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}$$

$$y_i = \frac{y_i}{y_{\text{max}}}$$

Example: model II regression, ranged major axis

Standardization: $$y_i' = z_i = \frac{y_i - \bar{y}}{s_y}$$

Example: explanatory variables in RDA
Data transformation

2. Normalize the data and stabilize their variance

**Square root transformation:** $y'_i = \sqrt{(y_i+c)}$:

Example: species data with moderate asymmetry
Data transformation

2. Normalize the data and stabilize their variance

Log transformation: \( y'_i = \ln(y_i + c) \):

Example: species data with more severe asymmetry
3. **Recode semi-quantitative variables as quantitative**

**Example:** transformation of Braun-Blanquet's phytosociological scale into quantitative values:

<table>
<thead>
<tr>
<th>Skala Braun-Bl.</th>
<th>Deckung, %</th>
<th>Ordinal-skala x</th>
<th>Transformation ( y = x^w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1eer</td>
<td>0.0</td>
<td>0</td>
<td>0.0 0.00 0.00 0.00 0 0 0 0</td>
</tr>
<tr>
<td>r</td>
<td>()</td>
<td>1</td>
<td>1.0 1.00 1.00 1 1 1</td>
</tr>
<tr>
<td>+</td>
<td>0.1</td>
<td>2</td>
<td>1.0 1.19 1.41 2 4 16</td>
</tr>
<tr>
<td>1</td>
<td>5.0</td>
<td>3</td>
<td>1.0 1.32 1.73 3 9 81</td>
</tr>
<tr>
<td>2m</td>
<td>17.5</td>
<td>5</td>
<td>1.0 1.50 2.24 5 25 625</td>
</tr>
<tr>
<td>2a</td>
<td>37.5</td>
<td>6</td>
<td>1.0 1.57 2.45 6 36 1296</td>
</tr>
<tr>
<td>2b</td>
<td>62.5</td>
<td>7</td>
<td>1.0 1.63 2.65 7 49 2401</td>
</tr>
<tr>
<td>3</td>
<td>87.5</td>
<td>8</td>
<td>1.0 1.68 2.83 8 64 4096</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Skala Braun-Bl.</th>
<th>Deckung, %</th>
<th>Ordinal-skala x</th>
<th>Transformation ( y = x^w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.0</td>
<td>9</td>
<td>1.0 1.73 3.00 9 81 6561</td>
</tr>
</tbody>
</table>
Data transformation

4. Binary coding of nominal variables

<table>
<thead>
<tr>
<th>1 qualitative descriptor</th>
<th>4 binary descriptors (dummy variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modalities</strong></td>
<td><strong>Classes (factors)</strong></td>
</tr>
<tr>
<td>Calcosol</td>
<td>1</td>
</tr>
<tr>
<td>Brunisol</td>
<td>2</td>
</tr>
<tr>
<td>Neoluvisol</td>
<td>3</td>
</tr>
<tr>
<td>Calcisol</td>
<td>4</td>
</tr>
</tbody>
</table>

Translated from F. Gillet (Neuchâtel, Switzerland) course notes.

For $k$ levels, $(k - 1)$ dummy variables needed $<->$ $(k - 1)$ d.f.

$\rightarrow$ later: see orthogonal Helmert contrasts
Data transformation

5. Transformations to obtain ecologically meaningful relationships among sites while using linear techniques
1. The double-zero problem

Community composition data sampled over variable environmental conditions, e.g. along long environmental gradients, typically contain many zero values because species are known to generally have unimodal distributions along environmental gradients and to be absent from sites far from their optimal living conditions.

The proportion of zeros is greater when the environmental conditions are more variable across the sampling sites.
### 1. The double-zero problem

<table>
<thead>
<tr>
<th>Brachysp</th>
<th>Phthirsp</th>
<th>Hoplcfpa</th>
<th>Rhysardu</th>
<th>Atrostri</th>
<th>Protopsp</th>
<th>Malacfeg</th>
<th>Malacfpr</th>
<th>Malacosp</th>
<th>Eniominu</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>7</td>
<td>16</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>7</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
1. The double-zero problem

The zero value in a matrix of species abundances (or presence-absence) is tricky to interpret:

The \textit{presence} of a species at a given site (i.e., non-zero value) generally implies that this site provides a set of minimal conditions allowing the species to survive (the dimensions of its ecological niche).
1. The double-zero problem

The zero value in a matrix of species abundances (or presence-absence) is tricky to interpret:

The absence of a species from a site may be due to a variety of causes: the species' niche may be occupied by a replacement species, or the absence of the species is due to adverse conditions on any of the important dimensions of its ecological niche, or the species has been missed because of a purely stochastic component of its spatial distribution.
1. The double-zero problem

Double zero: absence of a given species from two sites.

A double zero cannot readily be interpreted as a resemblance between the two sites.

The methods of analysis of species data must take this problem into account.

Most methods of analysis of species data work by comparing the sites, explicitly or implicitly, on the basis of an association measure.
1. The double-zero problem

- The association coefficients that consider the double zero as a resemblance (as any other value) are said to be **symmetrical**.

- The association coefficients that do not consider the double zero as a resemblance (as any other value) are said to be **asymmetrical**.

It is preferable to use **asymmetrical** coefficients when analysing species data.
2. Pre-transformation of species data

Since it is a linear method working in a Euclidean space, PCA is not adapted to raw species abundance data, since the zero is treated as any other value. The Euclidean distance imbedded in PCA is a symmetrical measure.

Therefore…. 
Transformation of species data

2. Pre-transformation of species data


*Pre-transform the species data* in such a way that, after PCA, the distance preserved among objects is no more the Euclidean distance, but an ecologically meaningful one.
Ordination in reduced space

2. Pre-transformation of species data

Chord distance ($D_3$) → $y_{ij}' = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}$

$\chi^2$ metric ($D_{15}$) → $y_{ij}' = \frac{y_{ij}}{y_i+\sqrt{y_{+j}}}$

$\chi^2$ distance ($D_{16}$) → $y_{ij}' = \sqrt{y_{++}} \frac{y_{ij}}{y_i+\sqrt{y_{+j}}}$

Distance between species profiles → $y_{ij}' = \frac{y_{ij}}{y_i+}$

Hellinger distance ($D_{17}$) → $y_{ij}' = \sqrt{\frac{y_{ij}}{y_{i+}}}$
Ordination in reduced space

2. Pre-transformation of species data: illustration

The species abundance paradox (Orlóci, 1978)

In these data, the Euclidean distance between sites 2 and 3 (no species in common) is smaller than the distance between sites 1 and 2.

<table>
<thead>
<tr>
<th></th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>0</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Site 2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Site 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ D_1 = \begin{bmatrix} 0.0000 & 7.6158 & 9.0000 \\ 7.6158 & 0.0000 & 1.7321 \\ 9.0000 & 1.7321 & 0.0000 \end{bmatrix} \]
### Euclidean Distance

\[
D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}
\]

### Chord Distance

\[
D_3(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left( \frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2}
\]

### Species Profiles Distance

\[
D_{18}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}
\]

### Hellinger Distance

\[
D_{17}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left[ \sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}
\]

### \(\chi^2\) Distance

\[
D_{16}(x_1, x_2) = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left( \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}
\]

### Transforms

\[
y_{ij}' = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}
\]

### Example

**Figure 7.8** Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 2 and 3, which have no species in common, is smaller than that between sites 1 and 2 which share species 2 and 3. This results in an incorrect assessment of the ecological relationships among sites. With the other coefficients in this figure, which are asymmetrical, the distance between sites 2 and 3 is larger than that between sites 1 and 2, and the distance between sites 1 and 3 is the same as between sites 2 and 3, or very nearly so.

### Matrix Transformations

<table>
<thead>
<tr>
<th>(D_1)</th>
<th>(D_3)</th>
<th>(D_{18})</th>
</tr>
</thead>
<tbody>
<tr>
<td>([0.0000 , 7.6158 , 9.0000])</td>
<td>([0.0000 , 0.3204 , 1.4142])</td>
<td>([0.0000 , 0.2357 , 1.2472])</td>
</tr>
<tr>
<td>(7.6158 , 0.0000 , 1.7321)</td>
<td>([0.3204 , 0.0000 , 1.4142])</td>
<td>([0.2357 , 0.0000 , 1.2247])</td>
</tr>
<tr>
<td>(9.0000 , 1.7321 , 0.0000)</td>
<td>(1.4142 , 1.4142 , 0.0000)</td>
<td>(1.2472 , 1.2247 , 0.0000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(D_1)</th>
<th>(D_3)</th>
<th>(D_{18})</th>
</tr>
</thead>
<tbody>
<tr>
<td>([0.0000 , 0.1697 , 1.4142])</td>
<td>([0.1697 , 0.0000 , 1.4142])</td>
<td>([0.1697 , 0.0000 , 1.4142])</td>
</tr>
<tr>
<td>(1.4142 , 1.4142 , 0.0000)</td>
<td>(1.4142 , 1.4142 , 0.0000)</td>
<td>(1.4142 , 1.4142 , 0.0000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(D_1)</th>
<th>(D_3)</th>
<th>(D_{18})</th>
</tr>
</thead>
<tbody>
<tr>
<td>([0.0000 , 0.3600 , 4.0092])</td>
<td>([0.3600 , 0.0000 , 4.0208])</td>
<td>([0.3600 , 0.0000 , 4.0208])</td>
</tr>
<tr>
<td>(4.0092 , 4.0208 , 0.0000)</td>
<td>(4.0208 , 4.0208 , 0.0000)</td>
<td>(4.0208 , 4.0208 , 0.0000)</td>
</tr>
</tbody>
</table>
2. Pre-transformation of species data

The **chord transformation** consists in transforming the species abundances from each object into a vector of length 1.

```r
y.chord <- decostand(y, method="normalize")
```

If one transforms a data table with this formula, any method that preserves the Euclidean distance among the objects (e.g., PCA, RDA, ANOVA) is now accessible for species abundance data.
2. Pre-transformation of species data

The Hellinger transformation consists in taking the squareroot of the relative species abundances.

```r
y.hel <- decostand(y, method="hellinger")
```

If one transforms a data table with this formula, any method that preserves the Euclidean distance among the objects (e.g., PCA, RDA, ANOVA) is now accessible for species abundance data.
2. Pre-transformation of species data

The chord and Hellinger transformations appear to be the best for general use.

They can both be applied to presence-absence data as well, with identical results.
Correspondence analysis (CA)
Ordination in reduced space

4. Correspondence analysis (CA)

CA is actually a PCA on a species data table that has been transformed into a table of Pearson's $\chi^2$ statistic (the one used in contingency tables).

The $\chi^2$ distance ($D_{16}$) is preserved among sites or species.

CA is a method adapted to the analysis of species abundance data.
4. Correspondence analysis (CA)

The data must be dimensionally homogeneous and 0 or positive. CA produces one axis less than \( \min[n,p] \).

The variation of the data is measured by a quantity named inertia. Inertia is partitioned among ordination axes as in PCA. The weight of each axis is also measured by an eigenvalue.

The objects and the species are represented as points on the same joint plot.
Ordination in reduced space

4. Correspondence analysis (CA)

Scalings:

CA scaling type 1: rows (objects) are at the centroids of columns.

The most appropriate if one is primarily interested in the ordination of objects (sites).

In the multidimensional space, $\chi^2$ distance is preserved among objects.
Ordination in reduced space

4. Correspondence analysis (CA)

CA scaling type 1: interpretation:

(1) The distances among objects in the reduced space approximate their $\chi^2$ distance. Thus, object points that are close to one another are likely to be relatively similar in their species relative frequencies.

(2) Any object found near the point representing a species is likely to have a high contribution of that species. For presence-absence data, the object is more likely to possess the state "1" for that species.
Ordination in reduced space

4. Correspondence analysis (CA)

Scalings:

**CA scaling type 1**

### Table VII - Artificial data for CA

<table>
<thead>
<tr>
<th>Spec.1</th>
<th>Spec.2</th>
<th>Spec.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj.1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Obj.2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Obj.3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Obj.4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Obj.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Obj.6</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Any object found near the point representing a species is likely to have a high contribution of that species. For presence-absence data, the object is more likely to possess the state "1" for that species.

CA scaling type 2:

- Columns are at the centroids of rows. This scaling is the most appropriate if one is primarily interested in the ordination of species. In the multidimensional space, $\chi^2$ distance is preserved among species.

Interpretation:

1. The distances among species in the reduced space approximate their $\chi^2$ distance. Thus, species points that are close to one another are likely to have relatively similar relative frequencies in the objects.
2. Any species that lies close to the point representing an object is more likely to be found in that object, or to have a higher frequency there than in objects that are further away in the joint plot.
library(vegan)
data(mite)
mite.ca <- cca(mite)
plot(mite.ca, scaling=1, main="Oribatid mites, Lac Geai")
Ordination in reduced space

4. Correspondence analysis (CA)

Scalings:

*CA scaling type 2*: columns (species) are at the centroids of rows.

The most appropriate if one is *primarily interested in the ordination of species*.

In the multidimensional space, $\chi^2$ distance is preserved among variables (species).
Ordination in reduced space

4. Correspondence analysis (CA)

CA scaling type 2: interpretation:

(1) The distances among species in the reduced space approximate their $\chi^2$ distance. Thus, species points that are close to one another are likely to have relatively similar relative frequencies in the objects.

(2) Any species that lies close to the point representing an object is more likely to be found in that object, or to have a higher frequency there than in objects that are further away in the joint plot.
Ordination in reduced space

4. Correspondence analysis (CA)

Scalings:

*CA scaling type 2*

<table>
<thead>
<tr>
<th></th>
<th>Spec.1</th>
<th>Spec.2</th>
<th>Spec.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj.1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Obj.2</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Obj.3</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Obj.4</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Obj.5</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Obj.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Any object found near the point representing a species is likely to have a high contribution of that species. For presence-absence data, the object is more likely to possess the state “1” for that species.
- **CA scaling type 2:** columns are at the centroids of rows. This scaling is the most appropriate if one is primarily interested in the ordination of species. In the multidimensional space, $\chi^2$ distance is preserved among species.

Interpretation:

1. The distances among species in the reduced space approximate their $\chi^2$ distance. Thus, species points that are close to one another are likely to have relatively similar relative frequencies in the objects.
2. Any species that lies close to the point representing an object is more likely to be found in that object, or to have a higher frequency there than in objects that are further away in the joint plot.

The following example (Table VII) will be submitted to a correspondence analysis:

<table>
<thead>
<tr>
<th></th>
<th>Spec.1</th>
<th>Spec.2</th>
<th>Spec.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec.1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Spec.2</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Spec.3</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Spec.4</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Spec.5</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Spec.6</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

• $\lambda_1 = 0.2295$
• $\lambda_2 = 0.0857$
library(vegan)
data(mite)
mite.ca <- cca(mite)
plot(mite.ca, scaling=2, main="Oribatid mites, Lac Geai")
Ordination in reduced space

4. Correspondence analysis (CA)

Words of caution

• **Graphically**, correspondence analysis seems **very sensitive to rare species**, which take extreme values **in plots**.

• Rare species have a small influence on the extraction of the principal axes, however.
Ordination in reduced space

4. Correspondence analysis (CA)

Words of caution

- **Strong gradients** generate **arch effects:**

  Environmental gradients often support a **succession** of species that tend to have **unimodal distributions.** At the ends of the gradients, series are **distorted.** Distortion occurs also in PCA (horseshoe effects).
Ordination in reduced space

4. Correspondence analysis (CA)

Words of caution

Species packing model:

CA approximates a Gaussian ordination.

Position of species along axes approximate their ecological optimum.
Ordination in reduced space

Arch and horseshoe effects

CA: arch effect
PCA: horseshoe effect
Ordination in reduced space

4. Correspondence analysis (CA)

*Detrended correspondence analysis (DCA):*

*Detrending by segments:*

Axis I is divided into an arbitrary number of segments. Within each segment, the mean of the object scores along axis 2 is made equal to zero.

This method has been strongly rejected by many authors. The scores on the second axis are essentially meaningless.
Day 1

Principal coordinate analysis (PCoA)
Ordination in reduced space

5. Principal coordinate analysis (PCoA)

Allows to obtain a Euclidean representation of a set of objects whose relationships are measured by any similarity or dissimilarity coefficient chosen by the user.

PCoA produces a set of orthogonal ordination axes whose importance is measured by eigenvalues (as in PCA and CA).

PCoA can only represent the relationships among objects (Q mode) or variables (R mode), but not both at the same time, unless scores for variables are computed after a Q-mode PCoA (using correlations or weighted averages).
5. Principal coordinate analysis (PCoA)

In the case of *Euclidean association measures*, PCoA will behave in a *Euclidean manner*.

If the association coefficient used is nonmetric, semimetric or has other problems of "non-Euclideanarity", then PCoA will react by producing several *negative eigenvalues* in addition to the positive ones.

If one wants to retain all the variance of the original association matrix, there are technical solutions to this problem (e.g. Lingoes or Caillez correction).
5. Principal coordinate analysis (PCoA)

1 × 2 axis of a PCoA on a $D_{14}$ (percentage difference, aka Bray-Curtis) matrix of the 6 × 3 data used in the first CA example. The two first axes have eigenvalues of 0.359 and 0.095; they represent 59.5% and 20.2% variance respectively. This PCoA gives 4 positive, one zero and one negative eigenvalue.
mite.bray <- vegdist(mite, "bray")
mite.PCoA <- cmdscale(mite.bray, k=ncol(mite)-1, eig=TRUE)
ordiplot(scores(mite.PCoA)[,c(1,2)], type="t", main="Oribatid mites, D14, PCoA")
abline(h=0, lty=3)
abline(v=0, lty=3)
sp.wa <- wascores(mite.PCoA$points[,1:2], mite)
text(sp.wa, rownames(sp.wa), cex=0.7, col="red")

Species scores have been computed after the PCoA analysis by weighted averages of site scores
PCoA biplot
Response variables projected as in PCA with scaling 1

PCoA on the Doubs fish data. Scaling 1.
Species scores have been computed after the PCoA analysis by correlation with the ordination axes, as in PCA.
5. Principal coordinate analysis (PCoA)

Sometimes used as an intermediate, technical step in more complex analyses.

Examples:
• Raw variables neither quantitative nor orthogonal $\Rightarrow$ PCoA
  $\Rightarrow$ one obtains a set of quantitative and orthogonal variables (the PCoA axes)!
• Spatial analysis using dbMEM (distance-based Moran's eigenvector maps $\Rightarrow$ Day 5
Unconstrained ordination of species data

(a) Classical approach

\[ Y = \text{Raw data} \]
\[ \text{(sites x species)} \]

Short gradients: CA or PCA
Long gradients: CA

(b) Transformation-based approach (tb-PCA)

Raw data
\[ \text{(sites x species)} \]

\[ Y = \text{Transformed data} \]
\[ \text{(sites x species)} \]

PCA

(c) Distance-based approach (PCoA)

Raw data
\[ \text{(sites x species)} \]

Distance matrix

PCoA

Representation of elements:
Species = arrows
Sites = symbols

Ordination biplot

Ordination of sites

Representation of elements:
Sites = symbols
Constrained ordination of species data

(d) Classical approach

\[ Y = \text{Raw data} \quad (\text{sites x species}) \quad \text{X} = \text{Explanatory variables} \]

Short gradients: CCA or RDA

Long gradients: CCA

---

(e) Transformation-based approach (tb-RDA)

Raw data

\[ Y = \text{Transformed data} \quad (\text{sites x species}) \quad \text{X} = \text{Explanatory variables} \]

RDA

---

(f) Distance-based approach (db-RDA)

Raw data

\[ \text{Distance matrix} \]

\[ Y = \text{(sites x principal coord.)} \quad \text{X} = \text{Explanatory variables} \]

PCoA

RDA

---

Canonical ordination triplot

Representation of elements:
Species = arrows
Sites = symbols
Explanatory variables = lines