1.2. Correspondence analysis

Pierre Legendre
Département de sciences biologiques
Université de Montréal

© Pierre Legendre 2018
Outline of the presentation

1. Definition of correspondence analysis
2. Computation steps
3. Data transformations before CA?
4. Scalings in CA biplots
5. CA examples
6. Effects of rare species on CA
7. Arch effect, detrending and DCA
8. Algorithms for CA
9. Choose an ordination method
10. References
Definition of correspondence analysis

Correspondence analysis (CA)

An ordination method preserving the chi-square distance\(^1\) among objects ...

… applicable to multivariate frequency or presence-absence data.

In ecology:
Community composition data (species abundance or presence-absence, biomass data, …) can be analysed by CA.

Mathematical properties of data
Data must be non-negative (i.e. \(\geq 0\)), frequency-like\(^2\), and dimensionally homogeneous.

---

\(^1\) The chi-square distance is described in detail in the course on Dissimilarities.

\(^2\) Examples: biomass or energy data; monetary units (e.g. $, £, €, ¥).
Computation steps

Example: community composition data

\[ \mathbf{Y} = [f_{ij}] = \begin{bmatrix} \begin{array}{c} \text{Site1} \\ \text{Site2} \\ \text{Site3} \end{array} \end{bmatrix} = \begin{bmatrix} [f_{++}] = \begin{bmatrix} 45 & 10 & 15 & 0 & 10 \\ 25 & 8 & 10 & 0 & 3 \\ 7 & 15 & 20 & 14 & 12 \end{bmatrix} \end{bmatrix} \]

\[
\begin{bmatrix} \begin{array}{c} \text{Sp1} \\ \text{Sp2} \\ \text{Sp3} \\ \text{Sp4} \\ \text{Sp5} \end{array} \end{bmatrix} \begin{bmatrix} f_{i+} \end{bmatrix} \]

\[
\begin{bmatrix} \begin{array}{c} 80 \\ 46 \\ 68 \end{array} \end{bmatrix} \]

\[ f_{++} = 194 \]
Compute the matrix of contributions to chi-square

Matrix \( \bar{Q} = [\bar{q}_{ij}] = \left[ \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} / \sqrt{f_{++}} \)

where

- \( p_{ij} = f_{ij} / f_{++} \)
- \( p_{i+} = f_{i+} / f_{++} \)
- \( p_{+j} = f_{+j} / f_{++} \)

The statistics \( (O_{ij} - E_{ij})/\sqrt{E_{ij}} \) in the numerator are called “components of chi-square” in contingency table analysis.

They are the square roots of the statistics that are summed to produce the Pearson chi-square statistic. Reference: Legendre & Legendre (2012, eq. 6.26).
Compute the matrix of contributions to chi-square

\[
\bar{Q} = \begin{bmatrix} \bar{q}_{ij} \end{bmatrix} = \begin{bmatrix} \frac{p_{ij} - p_{i+p+} p_{++}}{\sqrt{p_{i+p} p_{++}}} \end{bmatrix} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} \sqrt{f_{++}}
\]

where

\[
\begin{align*}
p_{ij} &= f_{ij} / f_{++} \\
p_{i+} &= f_{i+} / f_{++} \\
p_{+j} &= f_{+j} / f_{++}
\end{align*}
\]

\[
\bar{Q} = \begin{bmatrix} \bar{q}_{ij} \end{bmatrix} = \begin{bmatrix} \begin{array}{ccccc} Site1 & Site2 & Site3 & Site4 & Site5 \\ 0.169 & -0.070 & -0.059 & -0.173 & -0.007 \\ 0.113 & 0.004 & -0.015 & -0.131 & -0.086 \\ -0.276 & 0.072 & 0.076 & 0.295 & 0.079 \end{array} \end{bmatrix}
\]

Total inertia: \( \sum(\bar{q}_{ij}^2) = 0.285 = \text{sum of the CA eigenvalues} \)
Cross-product matrix:

$$\tilde{Q}'\tilde{Q}_{(c \times c)} = \begin{bmatrix}
0.118 & -0.031 & -0.033 & -0.125 & -0.033 \\
-0.031 & 0.010 & 0.010 & 0.033 & 0.006 \\
-0.033 & 0.010 & 0.010 & 0.035 & 0.008 \\
-0.125 & 0.033 & 0.035 & 0.134 & 0.036 \\
-0.033 & 0.006 & 0.008 & 0.036 & 0.014 \\
\end{bmatrix}$$

Compute eigen-decomposition of $$\tilde{Q}'\tilde{Q}_{(c \times c)}$$: 

$$(\tilde{Q}'\tilde{Q} - \lambda_k I)u_k = 0$$

**Eigenvalues**: $$\lambda_1 = 0.278, \lambda_2 = 0.007$$

**Matrix of eigenvalues**: 

$$\Lambda = \begin{bmatrix}
0.278 & 0.000 \\
0.000 & 0.007 \\
\end{bmatrix}$$

Maximum number of eigenvalues > 0 in CA: 

$$k = \min(r - 1, c - 1)$$

Correspondence analysis
Matrix of eigenvectors of $\mathbf{Q}'\mathbf{Q}_{(c \times c)}$:

$$
\mathbf{U}_{(c \times k)} = \begin{bmatrix}
Sp1 & 0.651 & 0.078 \\
Sp2 & -0.172 & -0.536 \\
Sp3 & -0.181 & -0.272 \\
Sp4 & -0.694 & 0.067 \\
Sp5 & -0.183 & 0.793 \\
\end{bmatrix}
$$

Matrix of eigenvectors of $\mathbf{Q}\mathbf{Q}'_{(r \times r)}$:

$$
\hat{\mathbf{U}}_{(r \times k)} = \mathbf{Q}\mathbf{U}\Lambda^{-0.5} = \begin{bmatrix}
Site1 & -0.481 & 0.597 \\
Site2 & -0.345 & -0.802 \\
Site3 & 0.806 & 0.012 \\
\end{bmatrix}
$$

Correspondence analysis
Do not use the chord, Hellinger or chi-square transformation on the data before subjecting them to CA.

CA must be computed on raw abundance data.

Sometimes, users apply the log(y+1) transformation to frequency data before CA in order to reduce the importance of extremely high abundance values, e.g. in microbial ecology.

The results may still be meaningful, but the mathematics suffer because the row and column sums of the raw abundance data table, which are used in the calculation of the chi-square transformation leading to the $\tilde{Q}$ matrix, do not have their usual mathematical meaning of sums of frequencies.
Scalings in CA biplots

As in PCA, biplots are graphs in which objects and variables are represented together.

Scaling type 1:
• Preserves the chi-square distance among the sites.
  ➢ Plot matrices $F$ for sites and $V$ for species.

Scaling type 2:
• Preserves the chi-square distance among the species.
  ➢ Plot matrices $\hat{V}$ for sites and $\hat{F}$ for species.

Compute the following matrices used in these plots:

\[
\begin{align*}
V_{(c \times k)} &= D(p_{+j})^{-0.5} U \\
\hat{V}_{(r \times k)} &= D(p_{i+})^{-0.5} \hat{U} \\
F_{(r \times k)} &= \hat{V} \Lambda^{0.5} \\
\hat{F}_{(c \times k)} &= V \Lambda^{0.5}
\end{align*}
\]
CA biplot
scaling type 1

In scaling 1 biplots, the sites are at chi-square distances of one another (e.g. \( D(\text{Site.1, Site.3}) \), blue arrow).

Matrices \( F \) and \( V \) form a pair such that the sites (coordinates given by matrix \( F \)) are at the centroid (also called *centre of mass*, or *barycentre*) of the species in matrix \( V \).

Scaling 1 is most appropriate if one is primarily interested in the distance relationships among the sites.

Example computed and drawn using function \texttt{CA.newr()} of the book *Numerical ecology with R, 2*\textsuperscript{nd} \textit{edition} (Borcard et al., 2018).
• The contributions of the species to the sites are reflected by the site-to-species distances in the biplot (e.g. Site.2, red arrows).

• The species are around the sites, in positions reflecting their abundances at each site.

In the example, species 1 is closer to sites 1 and 2 than it is to site 3 because it is more abundant at sites 1 and 2.
CA scaling 1 biplot preserves the chi-square distances among the sites

Can we verify that property?

Compute the chi-square distance among the sites in the raw data matrix \( \mathbf{Y} \)

<table>
<thead>
<tr>
<th></th>
<th>Site.1</th>
<th>Site.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site.2</td>
<td>0.21578</td>
<td></td>
</tr>
<tr>
<td>Site.3</td>
<td>1.11479</td>
<td>1.10026</td>
</tr>
</tbody>
</table>

Compute the Euclidean distance among the rows of \( \mathbf{F} \) (giving the positions of the sites in the CA scaling 1 biplot)
In scaling 2 biplots, the species are at chi-square distances of one another (e.g. $D(\text{Sp.1, Sp.5})$, blue arrow).

- Matrices $\hat{V}$ and $\hat{F}$ form a pair such that the species (in matrix $\hat{F}$) are at the centroids of the sites in matrix $\hat{V}$.
  - Sp. 2 and 3 are the most similar.
  - Sp. 1 and 4 are the most distant.

- The sites are around the species, in positions reflecting sp. abundances at the sites (e.g. Sp.5, red arrows).
  - Species 5 is more abundant at sites 1 and 3 than at site 2.
  - Species 1 is more abundant at sites 1 and 2 than at site 3.
  - Species 4 is only found at site 3 (symbols superposed).
Scaling 2 is most appropriate if one is primarily interested in the distance relationships among the species.
CA scaling 2 biplot preserves the chi-square distances among the species
Can we verify that property?

Compute the chi-square distance among the species in the raw data matrix $Y$

Compute the Euclidean distance among the rows of $\hat{F}$ (giving the positions of the species in CA scaling 2 biplot)

<table>
<thead>
<tr>
<th></th>
<th>Sp.1</th>
<th>Sp.2</th>
<th>Sp.3</th>
<th>Sp.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp.2</td>
<td>0.773</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp.3</td>
<td>0.744</td>
<td>0.065</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp.4</td>
<td>1.905</td>
<td>1.149</td>
<td>1.165</td>
<td></td>
</tr>
<tr>
<td>Sp.5</td>
<td>0.831</td>
<td>0.296</td>
<td>0.242</td>
<td>1.105</td>
</tr>
</tbody>
</table>
Two other types of CA scalings may also be available in software:

**Scaling type 3 –**

- This “symmetric scaling” is a compromise between scalings 1 and 2. Draw together matrices $\mathbf{V}\mathbf{\Lambda}^{0.25}$ (or $\mathbf{F}\mathbf{\Lambda}^{-0.25}$) for sites and $\mathbf{V}\mathbf{\Lambda}^{0.25}$ (or $\mathbf{F}\mathbf{\Lambda}^{-0.25}$) for species.
- This scaling does not preserve the chi-square distances among the species or among the sites.

**Scaling type 4 –**

- Useful in the analysis of a contingency table crossing two qualitative descriptors or two factors. Draw a joint plot using $\mathbf{F}$ which preserves the chi-square distances among the rows, and $\hat{\mathbf{F}}$ which preserves the chi-square distances among the columns of the contingency table.
- This hybrid scaling correctly represents the chi-square distance relationships among the states of each of the two qualitative descriptors.
Compute CA for the spider data using function CA.newr() –

```r
# Load function CA.newr()
# Read the file "Spiders_28x12_spe.txt"
spiders <- read.table(file.choose())

spiders.CA <- CA.newr(spiders)

par(mfrow=c(1,2))
biplot(spiders.CA, scaling=1)
biplot(spiders.CA, scaling=2)
```

CA biplots for the spider data
Compute CA for the Doubs fish data using vegan’s cca() –

```r
library(vegan)
load(Doubs.RData)  # File Doubs.RData
Doubs.ca <- cca(spe[-8,])
par(mfrow=c(1,2))
plot(Doubs.ca, scaling=1, main="Doubs fish, scaling 1")
plot(Doubs.ca, scaling=2, main="Doubs fish, scaling 2")
```
CA biplots for the Doubs fish data

Doubs fish, scaling 1

Doubs fish, scaling 2
R code for the Oribatid mite data using vegan’s cca() –

```r
library(vegan)
data(mite)
mite.ca <- cca(mite)

par(mfrow=c(1,2))
plot(mite.ca, scaling=1, main="Oribatid mites, scaling 1")
plot(mite.ca, scaling=2, main="Oribatid mites, scaling 2")
```

*Exercise: run these lines of code in R and examine the biplots.*
Another application of CA

In vegan, function `vegemite()` prints compact community data tables. The rows and/or columns can be reordered by explicit indexing using values provided by users, by an environmental variable, or using the result of a cluster analysis or an ordination.

Correspondence analysis is the ordination method of choice for reordering the sites and/or the species when producing such a reordered table. The theoretical basis for the production of reordered tables by CA was explained by ter Braak (1987, section 5.2.3).

```
# Example – Reorder the Doubs data using CA results
# Run the following code in R
library(vegan)
Doubs.ca = cca(spe)  # Compute CA results
vegemite(spe, use=Doubs.ca)
```
Effects of rare species on CA

We will now look at the influence on CA biplots of species that are rare in the data set.

1. For a rare species, a given difference in values between two sites increases the chi-square distance between the sites more than the same difference found for a common species.
In that sense, the chi-square distance gives higher weights to the rare than to the common species in the calculation of the distance.

[Property demonstrated in the course on dissimilarities.]

1 Rare species: species with small total abundance or with few occurrences in the data file.
2. In scaling 1 biplots, rare species with few occurrences may take extreme values, meaning that they may be located far from the origin. First example (artificial data):

<table>
<thead>
<tr>
<th></th>
<th>Sp1</th>
<th>Sp2</th>
<th>Sp3</th>
<th>Sp4</th>
<th>Sp5</th>
<th>Sp6</th>
<th>Sp7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site.1</td>
<td>45</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Site.2</td>
<td>25</td>
<td>8</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Site.3</td>
<td>7</td>
<td>15</td>
<td>20</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Site.4</td>
<td>25</td>
<td>10</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Site.5</td>
<td>7</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Site.6</td>
<td>45</td>
<td>8</td>
<td>15</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Sp.sums</td>
<td>154</td>
<td>66</td>
<td>90</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Occur.</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Species 6 and 7 have the smallest numbers of occurrences. Species 5 occurs everywhere but has a small total abundance.
Rare species #6 and 7 (low occurrences) are located far from the center of the plot.

• **Species 6** is found in sites {1, 2, 4, 5}; it pulls these sites to the lower-left corner.

• **Species 7** is found in sites {3, 6}; it pulls these sites towards the upper-right corner of the plot.
• Rare species #5 (small abundances) is near the center on all axes because it is present at all sites.
Second example, the spider data –

Two species are rare in occurrences and total abundance:

<table>
<thead>
<tr>
<th></th>
<th>Occurrences</th>
<th>Total abund.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arct.peri</strong></td>
<td>6</td>
<td>39</td>
</tr>
<tr>
<td><strong>Arct.lute</strong></td>
<td>7</td>
<td>26</td>
</tr>
</tbody>
</table>

- Rare species **Arct.peri** is located far from the center of the plot.
- It only has high abundance at site 26, where it is the only dominant species.
- **Arct.peri** pulls site 26 towards the upper-left corner of the plot.
Correspondence analysis

Rare species *Arct.lute* is near the center of the plot.

- It is found at 7 sites that have a diversified and average species composition.
- 7 species (including *Arct.lute*), found in several sites, are near the centre of the plot.

Second example, the spider data.

Two species are rare in occurrence and total abundance:

<table>
<thead>
<tr>
<th></th>
<th>Occurrences</th>
<th>Total abund.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arct.peri</em></td>
<td>6</td>
<td>39</td>
</tr>
<tr>
<td><em>Arct.lute</em></td>
<td>7</td>
<td>26</td>
</tr>
</tbody>
</table>

Occurrence table:

- *Arct.peri*: Sites 4, 7, 13, 14, 8, 19-21, 15-18
- *Arct.lute*: Sites 2, 12, 10, 11, 9, 26

Correspondence analysis
3. Compared to common species, rare species with small occurrences have a small influence on the first few eigenvalues and axes in CA.

For that reason, rare species can be removed from the data table without major change to the biplot.

An efficient method to select and remove rare species from a data table was proposed by Daniel Borcard. This method is described in Legendre & Legendre (2012, Box 9.2, with permission of D. Borcard).

Example:
Elimination of rare species: an example\textsuperscript{1}

The example concerns fish biomass data:

47 underwater transects \(\times\) 156 fish species collected during underwater surveys by researchers Pierre Labrosse and Eric Clua (\textit{Secretariat of the Pacific Community}) near the village of Manuka in the Tonga Islands, South Pacific.

- The species with the smallest number of occurrences were removed step by step. CA was recomputed at each step.
- For each step, the total inertia was noted, as well as the first few eigenvalues.

\textsuperscript{1}Legendre & Legendre (2012), Box 9.2, pp. 480-481.
61 species with 1 to 4 occurrences (out of 156 species) were removed step by step.

(a) These 61 species generated 24% of the inertia in the data matrix subjected to eigenvalue decomposition by CA.

(b) Decrease of individual eigenvalues: removing these 61 species had little effects on the first four eigenvalues.

This example supports the statement above, that “rare species can be removed from the data table without major change to the biplot”. 
4. An alternative method to reduce the influence of rare species is **down-weighting of rare species**, a method proposed by Hill (1979) and implemented in the DECORANA ordination program. It is also available in CANOCO.

In R: downweighting is an option in function `decorana()` of \{vegan\}. Also available in \{vegan\} in the stand-alone function `downweight()`.
Arch effect, detrending and DCA

Species have unimodal distributions along environmental gradients. Species populations succeeding one another along an environmental gradient form a continuum called coenocline.

A simulated coenocline along an environmental gradient (abscissa). From Whittaker (1972).
Simulate a coenocline with function `coenocline()` of `coenocliner`:
19 species that only differ by the positions of their optimum values.

```r
library(coenocliner)

x <- seq(1, 100, 1)               # Ecological var. values
opt <- seq(from=5, to=95, by=5)  # Species optima along x
tol <- rep(2.5, 19)               # Species tolerance
h <- rep(20, 19)                  # Maximum abundance

Y <- coenocline(x, responseModel = "gaussian",
                 params = cbind(opt = opt, tol = tol, h = h),
                 countModel = "poisson", expectation = TRUE)
plot(Y, type = "l", lty = "solid", xlab="Ecological variable", ylab="Abundance", main="Coenocline")
```

Correspondence analysis
There are 69% zeros in the data file (100 × 19).

CA is an appropriate method for ordination of this type of data because the unimodal distribution of a species resembles a Gaussian function.

**Exercise:** The species symbols along CA axis 1 are in the same order as their ecological optima along the gradient. Generate the coenocline data (R code of the previous slide), analyse them with CA (function cca() of vegan), then represent the species in a biplot.

*Note: the species are not shown in the following two slides.*
Arch effect

Ordination of the sites displaying a quadratic shape with the ends not folded inwards.

Reason: the chi-square distance preserved in CA has an upper bound.

An arch is typical of coenoclines represented in CA ordinations.
Chi-square distances between *Site 1* and selected sites along the coenocline. The distance reaches a maximum after 13 steps and does not increase thereafter. The plot reacts by forming an arch.
Horseshoe effect

An ordination plot displaying a quadratic shape with the ends folded inwards.

- A horseshoe is typical of coenoclines represented in PCA ordinations.

(a) Raw data: a wobbly horseshoe. The small number of species (19) in this simulation, compared to the number of sites (100), produces site vectors with irregular norms. This is reflected by a wobbly line in the PCA plot.

(b) Hellinger-transformed data: regular horseshoe. After Hellinger transformation, site vectors have uniform lengths of 1, producing a regular horseshoe.
Detrending

A post-treatment of CA results to stretch the arch and make it linear.

- **Detrending by segments** (Hill & Gauch, 1980): axis 1 is divided into an arbitrary number of segments. The mean of the group of sites in each segment is moved to the abscissa; that mean becomes 0 on axis 2.

© Legendre & Legendre (2012, Figure 9.12).
Detrending is carried out in detrended correspondence analysis (DCA).

- Comparative and simulation studies have strongly criticized detrending because it distorts the ordination structure with no gain for interpretation. It is seldom used nowadays.


**Alternative method to eliminate the arch effect**

Compute extended dissimilarities (De’ath, 1999). The modified $D_{\text{ext}}$ matrix produces an ordination where the sites are ordered more linearly along the gradient than in the arch or in detrending by segments.

See the documentation files of functions `xdiss()` in `{mvpart}` and `stepacross()` in `{vegan}` and the references therein. Use principal coordinate analysis (PCoA) for ordination of the transformed $D_{\text{ext}}$ matrix.
Algorithms for CA

CA is a statistical method of data analysis (not a statistical test). As in PCA, three different algorithms (or methods of calculation) can be used to implement it:

- Eigenvalue decomposition (EVD); eigen(\(\tilde{Q}'\tilde{Q}\)) in R.
- Singular value decomposition (SVD); svd(\(\tilde{Q}\)) in R.

These two algorithms are interchangeable, although statisticians often prefer svd(), which offers greater numerical accuracy.

- An iterative algorithm developed by Clint & Jennings (1970) was adapted to correspondence analysis by Hill (1973). It was then used by ter Braak in the Canoco ordination package.

Details are found in Legendre & Legendre (2012, Section 9.2.7).
Choose an ordination method

• For community composition data, use correspondence analysis (CA) if you want to emphasize the role of rare species in the ordination plot.

CA is the correct choice if you expect rare species to be indicative of particular environmental conditions.

The validity of this kind of conclusion will depend on the following assumption: the rare species have been estimated without bias, like the more common species. Is that the case?

• To obtain an ordination that gives the same importance to the common and rare species, apply the Hellinger or chord transformation to the community data and use PCA. See the courses on species transformations and on beta diversity.
References


End of section