Section: Ordination analysis

CA & DCA (unimodal unconstrained ordination)

Theory R functions Examples

Correspondence analysis (**CA**, previously know also as reciprocal averaging, RA), is a unimodal unconstrained ordination method. An interesting property, which attracted ecologists to this method, is the fact that it can calculate and display correspondence between samples and species in the same ordination space. In the space of all ordination axes, the distances among samples (and also among species) are represented by chi-square distance metric, which does not suffer from the double-zero problem (but is blamed by some for being too much influenced by rare species, see below). The data must be non-negative integers or presences-absences. Correspondence analysis suffers from creating often strong *arch artefact* in ordination diagrams, which is caused by a non-linear correlation between first and higher axes. Arch can be removed by detrending, which is the base of the **detrended correspondence analysis** (**DCA**). Distribution of samples along the first (D)CA axis is used as a base of TWINSPAN classification algorithm.

Simplified description of CA algorithm

Although nowaday's software is using matrix algebra to calculate CA (either using singular value decomposition or eigenvalue decomposition of the $\bar{Q}$ matrix), the original algorithm is based on reciprocal averaging of column and row scores, which starts from random values, and by iterative row- and column-averaging converge into a unique solution, which represents the sample and species scores.

It has the following five calculation steps:

1. Start with arbitrary (random) sample scores (x_i) along ordination axes.
2. Calculate species scores (u_i) as **a mean of sample scores** (x_i) **weighted by species abundances in samples**.
3. Calculate new sample scores (x_i) as **a mean of species scores** (u_i) **weighted by species abundances in samples**.
4. Standardize the sample scores (stretch the axis, which shrinked due to weighted-averaging).
5. If the newly calculated sample scores are the same as the old sample scores (or almost identical), stop, if it differs, continue by step 2.

After calculating the sample and species scores for the first axis, one can continue to the second and higher axes, while maintaining linear independence from all previously calculated axes.

The following table (modified Table 4-5 from Šmilauer & Lepš 2014) shows a simple example of how to calculate sample and species scores:

	Cirsium	***Glechoma***	***Rubus***	***Urtica***	initial score	x.WA1	x.WA1resc	x.WA2
Sample 1	**0**	**5**	**6**	**8**	0	1.095	0	0.41
Sample 2	**0**	**2**	**2**	**1**	4	1.389	0.422	0.594
Sample 3	**3**	**1**	**0**	**0**	10	8.063	10	7.839
u.WA1	10	2.25	1	0.444				
u.WA2	10	1.355	0.105	0.047				

u.WA3	10	1.312	0.062	0.028				
u.WA4	10	1.31	0.06	0.027				
Calculation steps: 1. Initial scores (0, 4, and 10) 2. Species scores: $u.WA1_{Cirsium}$ = (0*0 + 0*4 + 3*10)/(0 + 0 + 3) = 30 $u.WA1_{Glechoma}$ = (5*0 + 2*4 + 1*10)/(5 + 2 + 1) = 2.25 $u.WA1_{Rubus}$ = (6*0 + 2*4 + 0*10)/(6 + 2 + 0) = 1 $u.WA1_{Urtica}$ = (8*0 + 1*4 + 0*10)/(8 + 1 + 0) = 0.444 3. Sample scores: $x.WA1_{Sample\ 1}$ = (0*10 + 5*2.25 + 6*1 + 8*0.444)/(0 + 5 + 6 + 8) = 1.095 $x.WA1_{Sample\ 2}$ = (0*10 + 2*2.25 + 2*1 + 1*0.444)/(0 + 2 + 2 + 1) = 1.389 $x.WA1_{Sample\ 3}$ = (3*10 + 1*2.25 + 0*1 + 0*0.444)/(3 + 1 + 0 + 0) = 8.063 4. Rescale to the original range (0-10 here) 5. Continue by step 2 until the values converge.								

Important property of this algorithm is that it actually does not depends on the arbitrary choice of initial scores, as can be seen on Fig. 1 (in the example table above, the initial scores were preselected in the way that the convergence is faster; if they are random values, the convergence will still occur but will happen later).

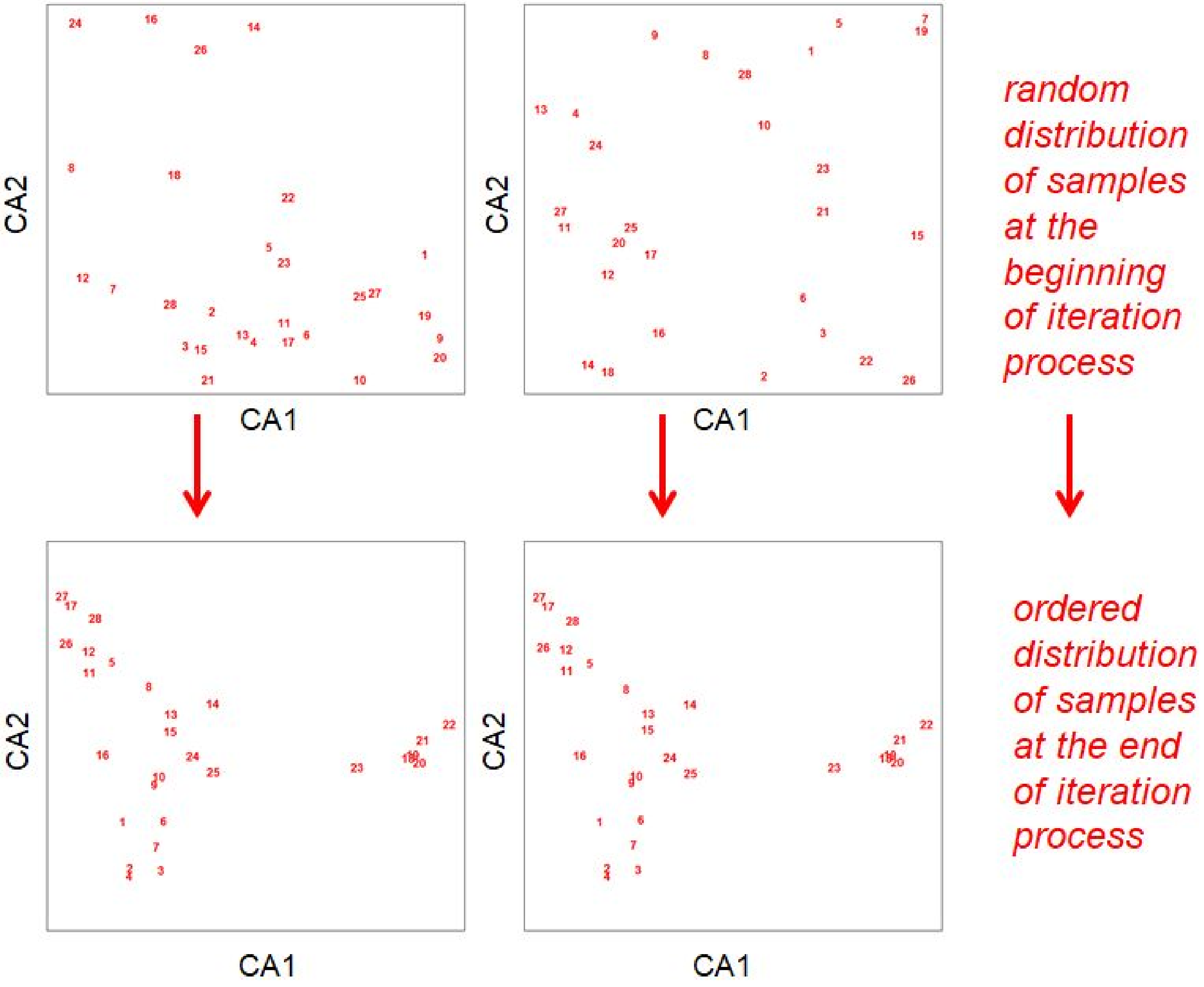

Figure 1: CA algorithm as iteration starting from two different random configuration (upper left and upper right diagram), both converging to the same configuration (lower left and lower right diagram).

Arch artefact and removal by detrending

CA algorithm has, however, two unpleasant properties: it produces a more or less pronounced arch artefact, and it compresses the samples at the 1st-axis ends relative to the middle (see an example on Fig. 2).

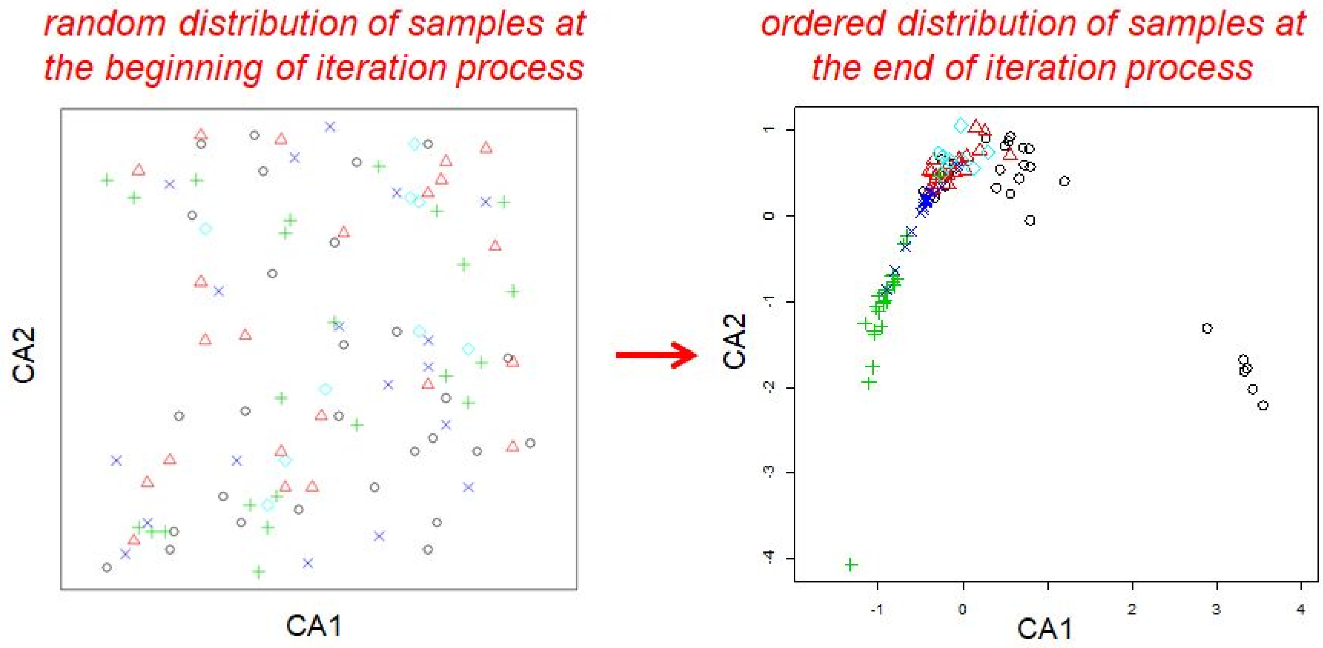

Figure 2: Arch artefact as a result of CA ordination. Samples of the same colour and symbol belong to the same community type (vegetation type in the case of this dataset).

A **detrended correspondence analysis (DCA)** attempts to remove the arch effect from ordination (Fig. 3). The method was (and still is) very popular, especially among vegetation ecologists, because it often returns meaningful distribution of samples in ordination diagrams. Additionally, it has one interesting property: the length of the first DCA axis (in SD units) refers to the heterogeneity or homogeneity of the dataset, and can be used to decide whether data should be analysed by linear (axis shorter than 3 SD) or unimodal (axis longer than 4 SD) ordination methods (details here). However, detrending (by segments) is a brute-force approach which resembles using a hammer on data - arch is hammered by cutting the first axis into segments and moving the sample points up and down along the second axis (you may see rescaling from CA to DCA here). For this and other reasons, the method is criticized and not recommended for use by some of the researchers (see e.g. Legendre & Legendre 1998, Borcard et al. 2011, or Jari Oksanen), while defended by others (e.g. ter Braak & Šmilauer 2015). A technical note: since the DCA implemented in both *vegan* or *CANOCO* is calculated by (the updated version of) the original DECORANA algorithm written by Mark O. Hill (and published in Hill & Gauch 1980), it still has the same limitations; one of them is that it calculates species and sample scores only for the first four ordination axes.

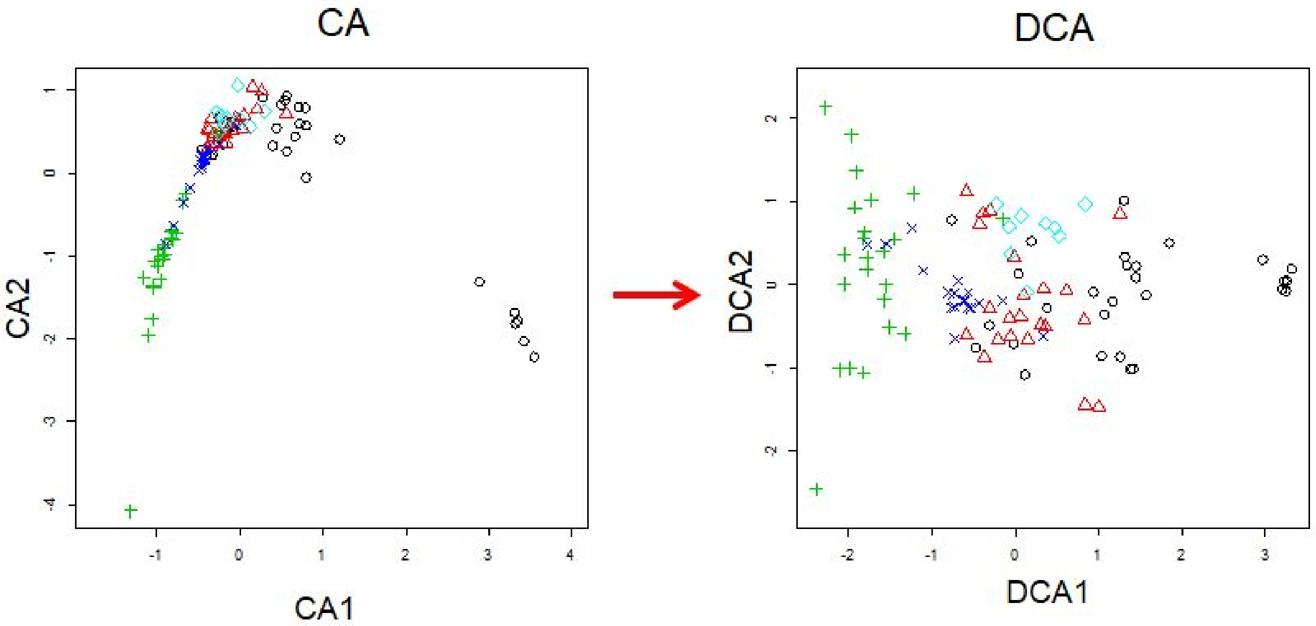

Figure 3: Removing of the arch artefact by detrending by segments.

An alternative solution, proposed by ter Braak (1986), is to remove arch artefact by applying linear constraints (explanatory variables) by calculating canonical correspondence analysis (CCA).

Influence of rare species

Traditionally, CA (and CCA) method was criticized for being too sensitive to descriptors (e.g. species) with very low total abundance, i.e. species that occur with very low frequency or in very few samples. As a result, rare objects are often located as outliers in CA ordination diagrams, and as such, they appear as highly influential. However, since they also have low weight (due to low total abundance), their effect on the result is reduced. In fact, deleting rare species before conducting (C)CA analysis (as often done previously to reduce the computational time) has minimal effect on the results of the calculation. More details about this are in Greenacre (2013), who also suggests using alternative scaling method when plotting results of (C)CA, so-called contribution biplot, where species coordinates are directly proportional to the species contribution to the solution.

Scaling in CA/DCA

In CA, both objects and species are represented by points in the ordination diagram (compare to PCA, where species/descriptors are vectors and sites are points). Similarly to PCA, two types of scaling are available (Fig. 4, Borcard et al. 2011):

- Scaling 1 - the distances **among samples (sites)** in the reduced ordination space approximate chi-square distances among samples in the full-dimensional space; any object found near the point representing a species is likely to contain a high contribution of that species. Sample scores are calculated as the means of species scores occurring in the sample, weighted by species abundances. This is why species scores are often displayed outside of the range of sample scores in the ordination diagram.
- Scaling 2 - the distances **among species** in the reduced ordination space approximate chi-square distances among species in the full-dimensional space; any species that lies close to the point representing an object is more likely to be found in that object or to have higher

frequency there. Species scores are calculated as the means of sample scores in which species occur, weighted by species abundance in each plot. This is why samples are often displayed outside the range of species scores in the ordination diagram.

In the case of DCA, only one scaling type, equivalent to scaling 1 in CA, is available.

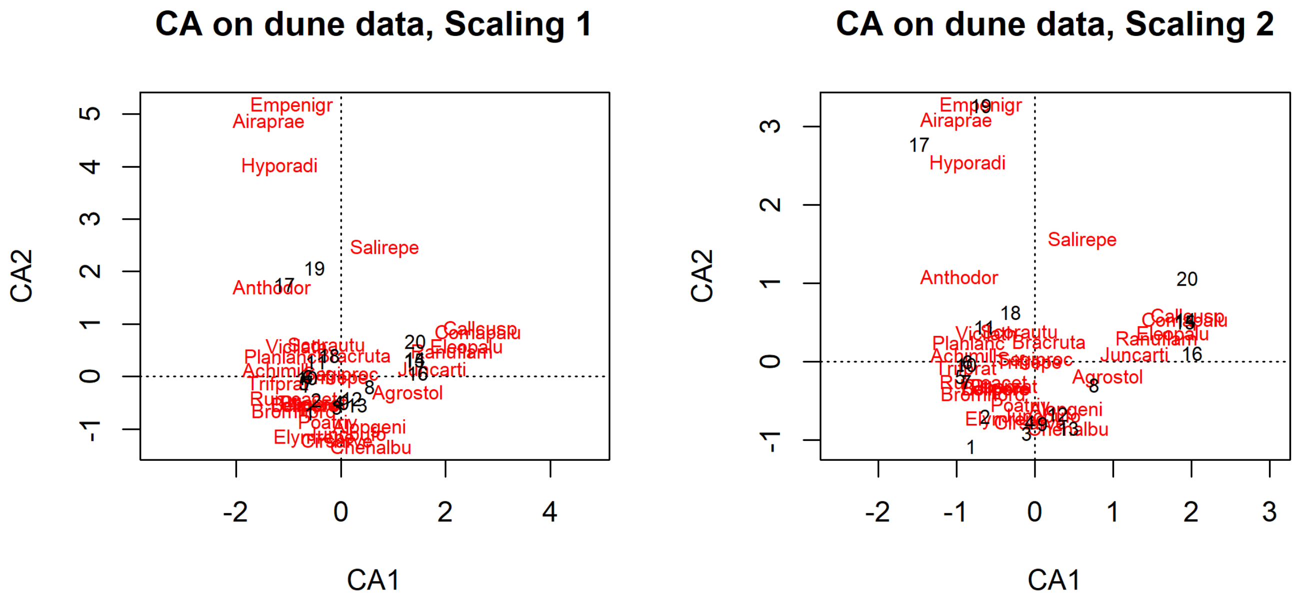

Figure 4: CA calculated on dune dataset, displayed with Scaling 1 (left) and Scaling 2 (right).

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:ca_dca

Last update: **2021/03/03 11:52**