PCA & tb-PCA (linear unconstrained ordination)

Principal component analysis (**PCA**) is a linear unconstrained ordination method. It is implicitly based on Euclidean distances among samples, which is suffering from double-zero problem. As such, PCA is not suitable for heterogeneous compositional datasets with many zeros (so common in case of ecological datasets with many species missing in many samples). It can be applied to quantitative variables (these could be also negative), and also presence-absence data, but it cannot handle qualitative variables. **Transformation-based principal component analysis** (**tb-PCA**) is PCA applied on pre-transformed species composition data (using e.g. Hellinger, chord or other transformation) and is implicitly based on distance other than Euclidean (Hellinger, chord or other), which is immune to double-zero problem.

Simplified description of PCA algorithm

(a) Use the matrix of samples × species (or, generally, samples × descriptors, where descriptors could be also environmental variables), and display each sample into the multidimensional space where each dimension is defined by an abundance of one species (or descriptor). In this way, the samples will produce a cloud located in the multidimensional space.
(b) Calculate the centroid of the cloud.
(c) Move the centres of axes to this centroid.
(d) Rotate the axes in such a way that the first axis goes through the cloud in the direction of the highest variance; the positions of samples on this axis become *sample scores*. The second axis is constructed in the way to be perpendicular to the first axis, which means that the correlation of the sample scores on the first axis and sample scores on the second axis is zero. If more axes can be constructed (which is not the case of this example since the original space defined by two species is only two-dimensional), then each higher ordination axis is perpendicular to all previous ones).

Fig. 1 (from Legendre & Legendre 1998) illustrates this algorithm on a very simple case with only two species (descriptors) and five samples. Fig. 2 illustrates the same logic on the data cloud in three-dimensional space (three species/descriptors).

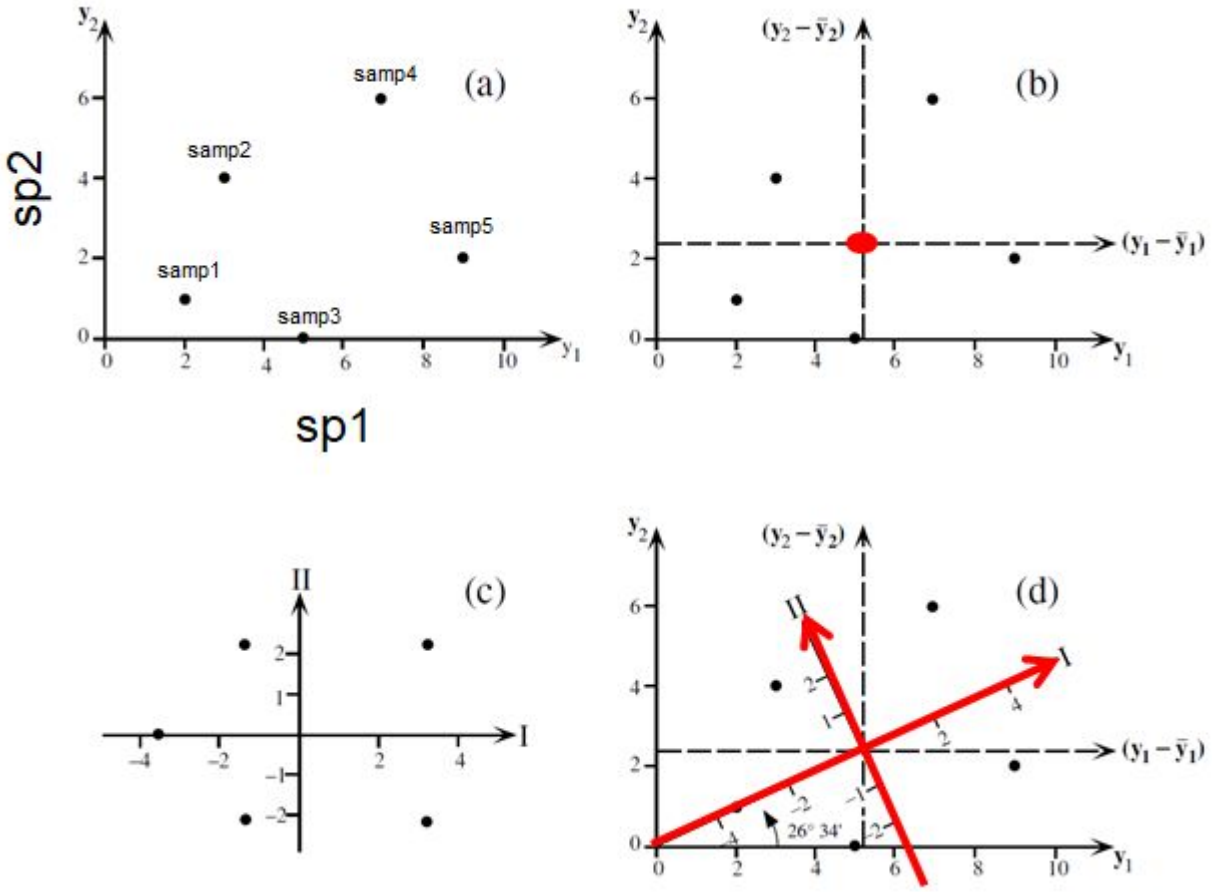


Figure 1: PCA ordination of five samples and two species. (Fig. 9.2 from Legendre & Legendre 1998.)

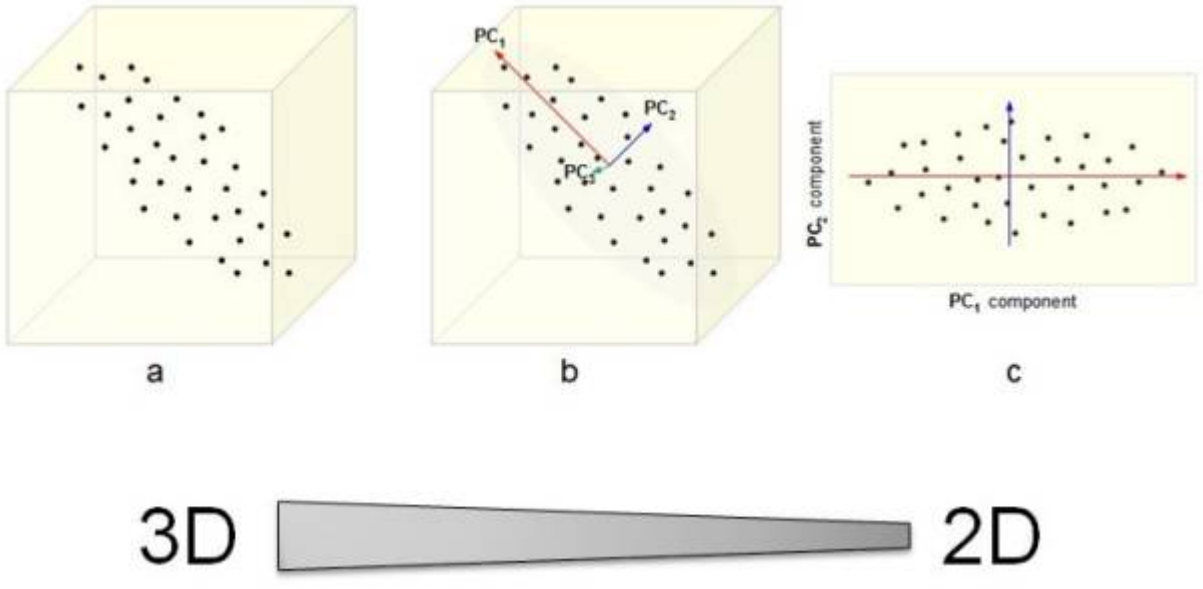


Figure 2: 3D schema of PCA ordination algorithm

Important outputs to consider

- Eigenvalues of individual axes, which represent the amount of variance given axis represents from the total variance (total inertia). One can calculate the proportion of variance explained by given axis as the axis eigenvalue divided by the total variance. If few main axes explain most of the variance, the ordination was successful (multidimensional information was successfully reduced to few main

dimensions). You can plot a barplot with each eigenvalue as a bar to see how steadily/sharply the eigenvalues of higher axes decline.
- Scores of samples and sites along ordination axes (this information is then used to draw the ordination diagram). Each PCA axis is a linear combination of all descriptors.
- Factor loadings, also known as component loadings – correlation of the variable (species, or generally descriptors) with individual PCA axes. If standardized, factor loadings can be compared between variables, and help interpret which descriptors are mostly associated with which PCA axis.
- The correlation among variables is described by angles between variables vectors, not by the distance between the apices of the vectors. This is true only if the scaling of the ordination diagram is set to 2 (correlation biplot; see the note about scaling below).

Main application of PCA on ecological data

When considering ecological data, PCA has three main applications:

1) **Describe correlation structure between different variables**, e.g. environmental variables measured for each sample, or species characteristics (traits) measured for individual species. In this case, the variables need to be standardized to zero mean and unit standard deviation, otherwise, the variable with higher absolute values or variance would be more important in the analysis. Resulting PCA ordination can show the main dimensions of variation in the data. This information can be further processed in several ways:

- Use the sample scores on PCA axes as a "complex" variables representing several real variables highly associated with them, and use the set of few PCA in further analysis in place of many real (and possibly highly correlated variables).
- Use few main PCA axes and from the real variables select one the most correlated with each PCA axis; in this way, we can reduce a large number of (often highly correlated) variables into few with possibly low correlation (PCA axes are from definition not correlated with each other).
- Groups of highly correlated variables can be obtained by clustering applied on the correlation matrix among variables, converted into distances (either as D = 1 - cor (var), or D = 1 - abs (cor (var))).

2) **Analysis of relatively homogeneous species composition data**. "Relatively homogeneous" means that in these data, we assume that species response along the (hypothetical) environmental gradient can be described by a linear relationship. Such data should contain few zeros, thus lowering the issue of the double zero problem, to which Euclidean distance is sensitive (see Ecological resemblance > Distance indices > Euclidean distance). If applied on heterogeneous dataset with many zeros, the result often shows strong horseshoe artefact, when sites with no species in common appear very close to each other in the ordination diagram.

3) Relatively recently was suggested that **PCA applied on pre-transformed species composition data** (e.g. by Hellinger transformation) can solve the problem of Euclidean distances in PCA and double zeros. In the case of Hellinger transformation, Euclidean distance (implicit to PCA) applied on Hellinger-transformed raw species composition data results in PCA representing Hellinger distances between samples, which is not influenced by double zero problem. This method is called transformation-based PCA (tb-PCA) and is described in a separate section. Note, however, that not everybody agrees that this is a good idea (see ESA 2010 presentation of Peter Minchin & Lauren Rennie on this topic).

What means scaling in PCA ordination biplot?

There is no single way how to display sites and variables (species) in the same biplot diagram (i.e. diagram showing two types of results, here sites and variables), that's why there are two ways of scaling results[1)]:

- Scaling 1 - distances among objects (sites) in the biplot are approximations of their Euclidean distances in multidimensional space; the angles among descriptor (species) vectors are meaningless. **Choose this scaling if the main interest is to interpret relationships among objects** (Fig. 3 left).

- Scaling 2 - distances among objects in the biplot are not approximations of their Euclidean distances; the angles between descriptor (species) vectors reflect their correlations. **Choose this scaling if the main interest focuses on the relationships among descriptors (species)** (Fig. 3 right).

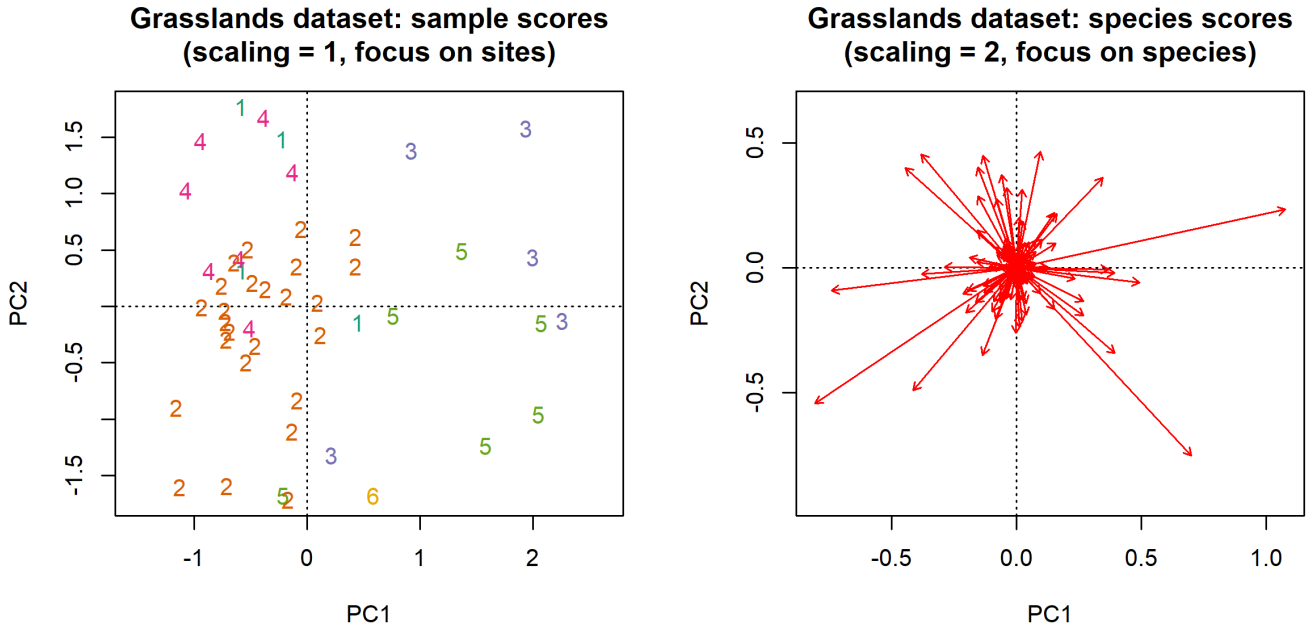


Figure 3: Ordination diagrams of PCA calculated on log-transformed grasslands dataset. Diagram on left is using scaling = 1 with focus on samples, while the right diagram is using scaling = 2 with focus on variables/species..

The circle of equilibrium contribution

The circle sometimes projected onto ordination diagram to estimate the importance of individual species/descriptors/variables. The radius is calculated as $\sqrt{(d/p)}$, where d is the number of displayed PCA axes (usually $d = 2$) and p is the number of variables (columns in the dataset). The descriptor with a vector of the same length as the circle radius contributes equally to all axes in PCA; vectors extending the circle radius make a higher contribution than average to the current display and can be interpreted with confidence (in the context of given number of ordination axes, here two, Fig. 4).

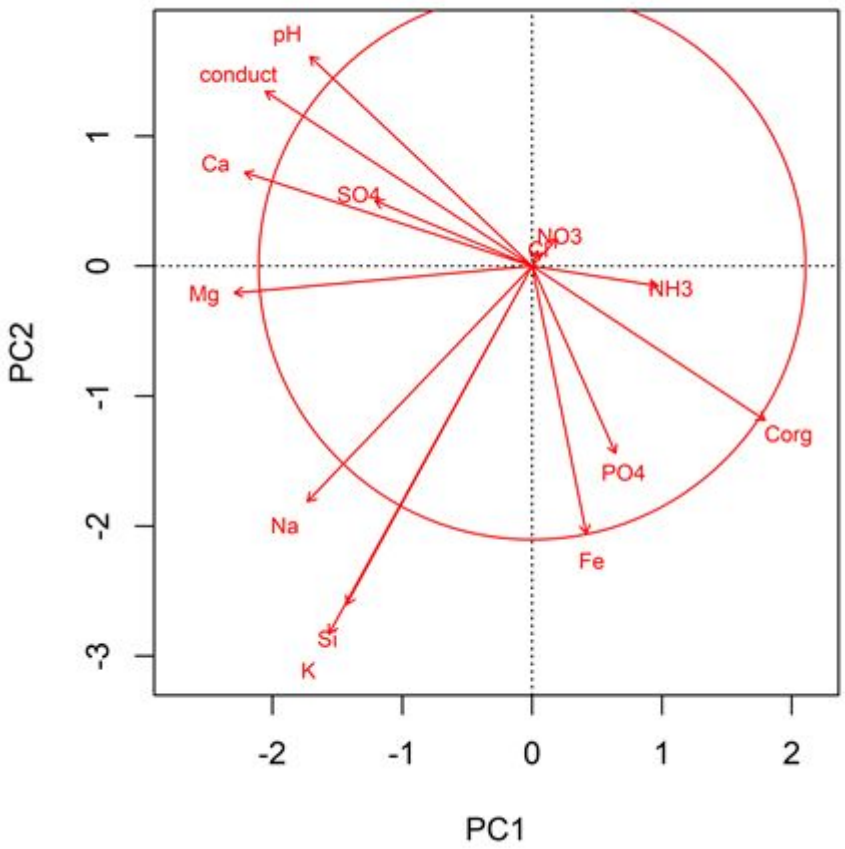


Figure 4: Circle of equilibrium contribution projected onto PCA ordination diagram. PCA based on wetland water chemistry dataset.

1)

In CANOCO, Scaling 1 corresponds to the option *Focus scaling on intersample distances* option, and Scaling 2 corresponds to the option *Focus scaling on inter-species correlations*.

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

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

Last update: **2021/03/03 19:51**