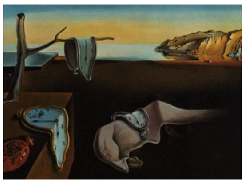


Journal of Statistical Software

September 2007, Volume 22, Issue 4. http://www.jstatsoft.org/

The ade4 Package: Implementing the Duality Diagram for Ecologists

Stéphane Dray
Université de Lyon

Anne-Béatrice Dufour
Université de Lyon

Abstract

Multivariate analyses are well known and widely used to identify and understand structures of ecological communities. The **ade4** package for the R statistical environment proposes a great number of multivariate methods. Its implementation follows the tradition of the French school of "Analyse des Données" and is based on the use of the duality diagram. We present the theory of the duality diagram and discuss its implementation in **ade4**. Classes and main functions are presented. An example is given to illustrate the **ade4** philosophy.

Keywords: **ade4**, duality diagram, ecological data, multivariate analysis, ordination.

1. Introduction

Since the early work of Goodall (1954) who applied principal component analysis (PCA) to vegetation data, multivariate analyses have been and remain intensively used by community ecologists. Multivariate analysis provides methods to identify and summarize joint relationships of variables in large data sets. Community ecologists usually sample a number of sites and aim to analyze the effects of several environmental factors on several species simultaneously. In this context, *the application of multivariate analysis to community ecology is natural, routine and fruitful* (Gauch 1982, p. 1). The diversity of ecological questions, models and types of data has induced the development of a great number of multivariate methods. There are at least three R packages, devoted to ecologists and available on the Comprehensive R Archive Network `http://CRAN.R-project.org`, which implement some of these methods (**ade4**, **labdsv**, **vegan**).

The **ade4** package (**D**ata **A**nalysis functions to analyze **E**cological and **E**nvironmental data in the framework of **E**uclidean **E**xploratory methods) is a complete rewrite for the R environment (R Development Core Team 2007) of the **ADE-4** (in uppercase) software (Thioulouse, Chessel, Dolédec, and Olivier 1997). The '4' in the name of the package is not a version number but means that there are four E in the acronym. The implementation of the **ade4** package follows the tradition of the French school of "Analyse des Données" and is based on the use of a unifying mathematical tool: the **du**ality **di**agram (Cailliez and Pagès 1976; Escoufier 1987; Holmes 2006). Each method is considered as a particular case of the duality diagram: it is called with a function '`dudi.*`' which returns an object of the class `dudi`. The duality diagram theory includes standard methods such as principal component analysis or correspondence analysis. It includes also more recent methods which have been developed in an ecological context like RLQ analysis (Dolédec, Chessel, ter Braak, and Champely 1996) to evaluate the link between species traits and environmental variables. This method has been named 'RLQ' because it finds linear combination of the variables of table **R** (environmental variables) and linear combinations of the variables of table **Q** (species traits) of maximal covariance weighted by species abundance data contained in table **L** (link table). RLQ analysis has been extended into a spatial context (Dray, Pettorelli, and Chessel 2002) for studying the relationships between two data sets that have been sampled at different locations. Other recent approaches include double principal coordinate analysis (Pavoine, Dufour, and Chessel 2004) to compare several communities containing species that differ according to their taxonomic, morphological or biological features or Outlying Mean Index (OMI) analysis (Dolédec, Chessel, and Gimaret-Carpentier 2000) to address the question of niche breadth and niche separation.

In this paper, we explain the **ade4** philosophy. We firstly give a description of the duality diagram theory. The implementation of this theory in the **ade4** package is then discussed. Lastly, we present a worked example based on environmental data which consists in the analysis of a table containing a mix of quantitative and qualitative variables.

2. The duality diagram theory

The duality diagram theory was developed by Pagès and Cazes in a series of lectures in 1969-1970. Its first good description can be found in Cazes's thesis (Cazes 1970). After improvements by Cailliez, Mailles, Nakache and Pagès, a complete synthesis has been published in a French book called 'Introduction à l'analyse de données' (Cailliez and Pagès 1976) (Cazes, personal communication). To our knowledge, only two papers (Escoufier 1987; Holmes 2006) are available for non-French readers.

2.1. Definitions

X is a data table with n rows (individuals) and p columns (variables). This table can be viewed as p points in $\mathbb{R}^n$. In this case, each point corresponds to a variable (column vectors) and each dimension corresponds to an individual. The coordinates of the point are then given by the values taken by the n individuals for the variable considered. Symmetrically, table **X** can be viewed by n points (individuals) in $\mathbb{R}^p$ (Figure 1). In ecology, table **X** could be a floro-faunistic table containing the abundances of p species for n sites or an environmental table with the measurements of p environmental variables for n sites.

Ecologists often want to obtain a summary of these two representations in order to understand:

- what are the relationships between the variables,
- what are the resemblances/differences between the individuals.

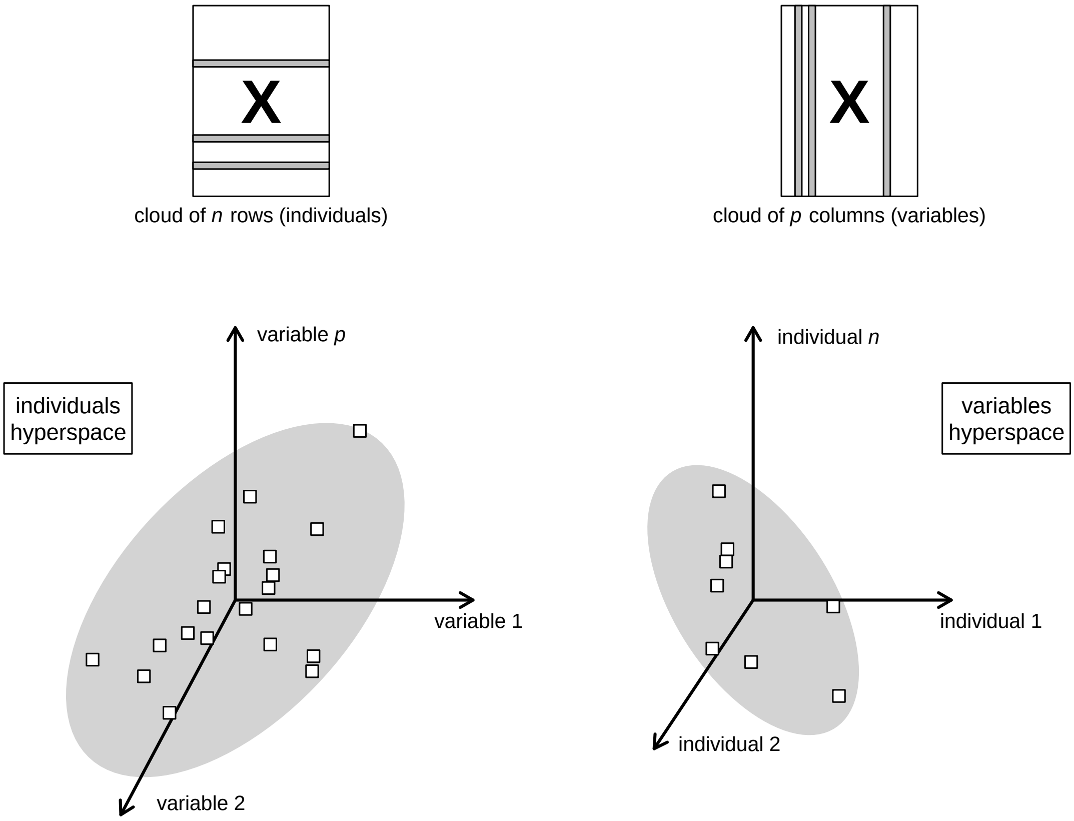


Figure 1: Representation of table $\mathbf{X}$ as a cloud of n points (individuals) in an hyperspace with p dimensions (variables) or as a cloud of p points (variables) in an hyperspace with n dimensions (individuals)

Multivariate methods aim to answer these two questions and seek for small dimension hyperspaces (few axes) where the representations of individuals and variables are as close as possible to the original ones. To answer the two previous questions, we need to define how to compute *relationships* between variables and *differences* between individuals. Thus, we define $\mathbf{Q}$, a $p \times p$ positive symmetric matrix and $\mathbf{D}$, a $n \times n$ positive symmetric matrix. $\mathbf{Q}$ is a metric used as an inner product in $\mathbb{R}^p$ and thus allows to measure distances between the n individuals. $\mathbf{D}$ is a metric used as an inner product in $\mathbb{R}^n$ and thus allows to measure relationships between the p variables.

In practice, the choice for matrices $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{D}$ is closely related to the objectives of the study. For an environmental table with only quantitative variables, considering Euclidean distances between individuals leads to $\mathbf{Q} = \mathbf{I}_p$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix. If we assume that relationships between variables are measured by covariances, then $\mathbf{X} = \left[x_{ij} - \mathsf{m}(\mathbf{x}^j)\right]$ (where $\mathsf{m}(\mathbf{x}^j)$ is the mean for the j-th column of $\mathbf{X}$) and $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$. If we prefer to use correlations, then $\mathbf{X} = \left[\frac{x_{ij} - \mathsf{m}(\mathbf{x}^j)}{\mathsf{sd}(\mathbf{x}^j)}\right]$ (where $\mathsf{sd}(\mathbf{x}^j)$ is the standard deviation for the j-th column of $\mathbf{X}$) and $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$. These two alternatives correspond to PCA on covariance matrix (centered PCA) and PCA on correlation matrix (normed PCA) respectively.

For a floro-faunistic table, several alternatives can be considered. For instance, PCA of species profiles ($\mathbf{X} = \left[\frac{x_{ij}}{x_{\cdot j}}\right]$ with $x_{\cdot j} = \sum_{i=1}^{n} x_{ij}$) removes the effect of the differences in global abundances between species when computing distances between sites. This analysis is focused on the relative distribution of species over the sites and aims to compare the ecological preferences of species.

Various centering of table $\mathbf{X}$ can also be used. Mathematically and geometrically, and also ecologically, centering involves a point of reference for the study. Information is given by a site when it deviates from this hypothetical reference site and a species is taken into account if it departs from the reference distribution over all sites. For noncentered data, the point of reference is the all-zero record: an empty site or a species that is always absent. Centering by species ($\mathbf{X} = [x_{ij} - x_{\cdot j}]$) implies that the reference point is a hypothetical site where the species composition is simply the mean species composition computed for all sites. Centering by site ($\mathbf{X} = [x_{ij} - x_{i\cdot}]$) implies that the reference point is an average species for which the abundance in a site is a constant proportion of the species total abundance in this site. To get a more detailed description of these various transformation, the reader could consult Noy-Meir (1973); Noy-Meir, Walker, and Williams (1975); Legendre and Gallagher (2001); Dray, Chessel, and Thioulouse (2003).

Lastly, various definitions of matrices $\mathbf{Q}$ and $\mathbf{D}$ allow to give more or less weights to species and sites. For instance, setting $\mathbf{Q} = \mathrm{diag}(x_{\cdot 1}, \cdots, x_{\cdot p})$ allows to weight each species with its global abundance when computing distances between sites. This could be useful if we consider that the samples are not representative of the community. Sampling selectivity can be a reason for this nonrepresentativeness because many species are rare in the sample not because they are rare in the studied area, but because the collecting method is not efficient for capturing them (Bayley and Peterson 2001). In this case, information given by an abundant species is more reliable than that given by a rare species and must have more weight when comparing two sites.

Different definitions of matrices $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{D}$ correspond to different multivariate methods including PCA (`dudi.pca`), correspondence analysis (`dudi.coa`), non-symmetric correspondence analysis (`dudi.nsc`), multiple correspondence analysis (`dudi.acm`)... All these methods correspond to the analysis of a particular triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ of three matrices. This information is summarized in the following mnemonic diagram:

$$\begin{tikzcd}
\boxed{p} \arrow[r, "\mathbf{Q}"] & \boxed{p} \arrow[d, "\mathbf{X}"] \\
\boxed{n} \arrow[u, "\mathbf{X}^\top"] & \boxed{n} \arrow[l, "\mathbf{D}"]
\end{tikzcd}$$

It should be noticed that $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ is strictly equivalent to $\left(\mathbf{X}^\top, \mathbf{D}, \mathbf{Q}\right)$.

2.2. Eigendecompositions

As said before, multivariate methods seek for a small dimension hyperspace where the representation of individuals is as close as possible to the original one. This objective is achieved by the diagonalization of $\mathbf{X}^\top\mathbf{DXQ}$. Symmetrically, small dimension hyperspace where the representation of variables is as close as possible to the original one is obtained by the diagonalization of $\mathbf{XQX}^\top\mathbf{D}$. Note that these two operators could be non-symmetric. These two different diagonalizations are linked and will be considered in the diagonalization of the duality diagram. In this section, we show the various relationships between these two diagonalizations and others which consider symmetric operators.

We consider the Cholesky decompositions $\mathbf{Q} = \mathbf{E}^\top\mathbf{E}$ ($\mathbf{E}$ is $q \times p$) and $\mathbf{D} = \mathbf{B}^\top\mathbf{B}$ ($\mathbf{B}$ is $g \times n$). In order to obtain a symmetric operator to diagonalize, we can 'break' the duality diagram:

$$\begin{tikzcd}
\text{Diagonalization 1} \arrow[r, squiggly, no head] & \boxed{q} \arrow[r, "\mathbf{E}^\top"] & \boxed{p} \arrow[r, "\mathbf{X}"] & \boxed{n} \arrow[r, "\mathbf{B}"] \arrow[d, dotted, "\mathbf{D}"] & \boxed{g} \arrow[d, "\mathbf{I}_g"] & \\
& \boxed{q} \arrow[u, "\mathbf{I}_q"] & \boxed{p} \arrow[u, dotted, "\mathbf{Q}"] \arrow[l, "\mathbf{E}"] & \boxed{n} \arrow[l, "\mathbf{X}^\top"] & \boxed{g} \arrow[l, "\mathbf{B}^\top"] \arrow[r, squiggly, no head] & \text{Diagonalization 2}
\end{tikzcd}$$

Let $\mathbf{\Omega} = \mathbf{BXE}^\top$. The $q \times q$ operator $\mathbf{\Omega}^\top\mathbf{\Omega} = \mathbf{EX}^\top\mathbf{B}^\top\mathbf{BXE}^\top$ is symmetric and its eigendecomposition (diagonalization 1 in the previous diagram) leads to:

$$\mathbf{\Omega}^\top\mathbf{\Omega} = \mathbf{V\Lambda V}^\top \text{ with } \mathbf{V}^\top\mathbf{V} = \mathbf{I}_q$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues and $\mathbf{V}$ contains the associated eigenvectors as columns.

Let $\mathbf{F} = \mathbf{E}^\top\mathbf{V}$. After some matrix manipulations, we obtain:

$$\mathbf{QX}^\top\mathbf{DXF} = \mathbf{F\Lambda} \text{ and } \mathbf{F}^\top\mathbf{Q}^{-1}\mathbf{F} = \mathbf{I}_q$$

Let $\mathbf{A} = \mathbf{E}^{-1}\mathbf{V}$. After some matrix manipulations, we obtain:

$$\mathbf{X}^\top\mathbf{DXQA} = \mathbf{A\Lambda} \text{ and } \mathbf{A}^\top\mathbf{QA} = \mathbf{I}_q$$

Symmetrically, the operator $\mathbf{\Omega\Omega}^\top = \mathbf{BXE}^\top\mathbf{EX}^\top\mathbf{B}^\top$ ($g \times g$) is symmetric and its eigendecomposition (diagonalization 2 in the previous diagram) leads to:

$$\mathbf{\Omega\Omega}^\top = \mathbf{U\Lambda U}^\top \text{ with } \mathbf{U}^\top\mathbf{U} = \mathbf{I}_g$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues and $\mathbf{U}$ contains the associated eigenvectors as columns.

Let $\mathbf{G} = \mathbf{B}^\top\mathbf{U}$. After some matrix manipulations, we obtain:

$$\mathbf{DXQX}^\top\mathbf{G} = \mathbf{G\Lambda} \text{ and } \mathbf{G}^\top\mathbf{D}^{-1}\mathbf{G} = \mathbf{I}_g$$

Let $\mathbf{K} = \mathbf{B}^{-1}\mathbf{U}$. After some matrix manipulations, we obtain:

$$\mathbf{XQX}^\top\mathbf{DK} = \mathbf{K\Lambda} \text{ and } \mathbf{K}^\top\mathbf{DK} = \mathbf{I}_g$$

In summary, we have the following general theoretical properties:

- Matrices $\mathbf{\Omega}^\top\mathbf{\Omega}$, $\mathbf{\Omega\Omega}^\top$, $\mathbf{X}^\top\mathbf{DXQ}$, $\mathbf{DXQX}^\top$, $\mathbf{QX}^\top\mathbf{DX}$ and $\mathbf{XQX}^\top\mathbf{D}$ have the same r nonzero eigenvalues and $r \leq \min(n, p, q, g)$.
- r is called the rank of the diagram and the nonzero eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ are stored in the diagonal matrix $\mathbf{\Lambda}_{[r]}$.

2.3. Axes and components

From the previous part, we can see that the diagonalizations of $\mathbf{X}^\top\mathbf{DXQ}$ (representation of the individuals) and $\mathbf{XQX}^\top\mathbf{D}$ (representation of the variables) produce the same eigenvalues. There are also some relationships between the different subspaces that are defined by these diagonalizations. We define the following elements related to the eigendecompositions described above:

- $\mathbf{F} = \left[\mathbf{f}^1, \cdots, \mathbf{f}^r\right]$ is a $p \times r$ matrix containing the r nonzero eigenvectors (in column) of $\mathbf{QX}^\top\mathbf{DX}$ associated to the r eigenvalues of the diagram. $\mathbf{F}$ is $\mathbf{Q}^{-1}$-orthonormalized i.e. $\mathbf{F}^\top\mathbf{Q}^{-1}\mathbf{F} = \mathbf{I}_r$ where $\mathbf{Q}^{-1}$ is the inverse of $\mathbf{Q}$. These r columns define the ***principal factors***.

- $\mathbf{A} = \left[\mathbf{a}^1, \cdots, \mathbf{a}^r\right]$ is a $p \times r$ matrix containing the r nonzero eigenvectors (in column) of $\mathbf{X}^\top\mathbf{DXQ}$ associated to the r eigenvalues of the diagram. $\mathbf{A}$ is $\mathbf{Q}$-orthonormalized i.e. $\mathbf{A}^\top\mathbf{QA} = \mathbf{I}_r$. The r columns define the ***principal axes***.

- $\mathbf{K} = \left[\mathbf{k}^1, \cdots, \mathbf{k}^r\right]$ is a $n \times r$ matrix containing the r nonzero eigenvectors (in column) of $\mathbf{XQX}^\top\mathbf{D}$ associated to the r eigenvalues of the diagram. $\mathbf{K}$ is $\mathbf{D}$-orthonormalized i.e. $\mathbf{K}^\top\mathbf{DK} = \mathbf{I}_r$. The r columns define the ***principal components***.

- $\mathbf{G} = \left[\mathbf{g}^1, \cdots, \mathbf{g}^r\right]$ is a $n \times r$ matrix containing the r nonzero eigenvectors (in column) of $\mathbf{DXQX}^\top$ associated to the r eigenvalues of the diagram. $\mathbf{G}$ is $\mathbf{D}^{-1}$-orthonormalized: $\mathbf{G}^\top\mathbf{D}^{-1}\mathbf{G} = \mathbf{I}_r$ where $\mathbf{D}^{-1}$ is the inverse of $\mathbf{D}$. The r columns contain the ***principal cofactors***.

The term *duality* is justified by the close connections between the four eigendecompositions. Relationships between the four eigendecompositions allow to compute only one system of axes to obtain the three others. For instance, we have the following transition formulas:

$$\mathbf{F} = \mathbf{QA},\ \mathbf{K} = \mathbf{XF}\boldsymbol{\Lambda}_{[r]}^{-(1/2)},\ \mathbf{G} = \mathbf{DK} \text{ and } \mathbf{A} = \mathbf{X}^\top\mathbf{G}\boldsymbol{\Lambda}_{[r]}^{-(1/2)}$$

We summarize this description in the following diagram:

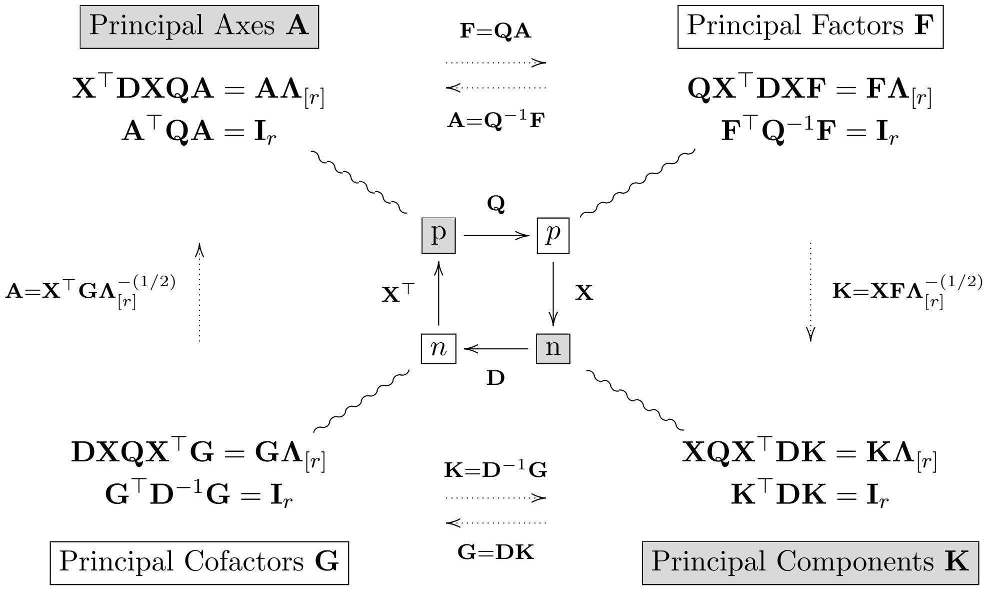
2.4. Properties

There are some fundamental properties linked to the diagonalization of a duality diagram.

- If we search for a $\mathbf{Q}$-normalized vector $\mathbf{a}$ of $\mathbb{R}^p$ maximizing $\parallel \mathbf{XQa} \parallel_{\mathbf{D}}^2$, the solution is unique and is obtained for $\mathbf{a} = \mathbf{a}^1$. The maximum is equal to λ_1. If we search for another vector $\mathbf{Q}$-normalized vector $\mathbf{a}$ of $\mathbb{R}^p$ maximizing the same quantity under the orthogonality constraint $\mathbf{a}^\top\mathbf{Qa}^1 = 0$, the solution is still unique and is obtained with $\mathbf{a^2}$ and the maximum is equal to λ_2 and so on until the last one $\mathbf{a^r}$.

  In summary, the vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^r$ successively maximize, under the $\mathbf{Q}$-orthogonality constraint, the quadratic form $\parallel \mathbf{XQa} \parallel_{\mathbf{D}}^2$.

- The vectors $\mathbf{k}^1, \mathbf{k}^2, \ldots, \mathbf{k}^r$ successively maximize, under the $\mathbf{D}$-orthogonality constraint, the quadratic form $\parallel \mathbf{X}^\top\mathbf{Dk} \parallel_{\mathbf{Q}}^2$.


Principal Axes $\mathbf{A}$
$\mathbf{X}^\top\mathbf{DXQA} = \mathbf{A}\mathbf{\Lambda}_{[r]}$
$\mathbf{A}^\top\mathbf{QA} = \mathbf{I}_r$
$\mathbf{F}$=$\mathbf{QA}$
$\mathbf{A}$=$\mathbf{Q}^{-1}\mathbf{F}$
Principal Factors $\mathbf{F}$
$\mathbf{QX}^\top\mathbf{DXF} = \mathbf{F}\mathbf{\Lambda}_{[r]}$
$\mathbf{F}^\top\mathbf{Q}^{-1}\mathbf{F} = \mathbf{I}_r$
$\mathbf{Q}$
p
p
$\mathbf{X}^\top$
$\mathbf{X}$
n
n
$\mathbf{D}$
$\mathbf{A}$=$\mathbf{X}^\top\mathbf{G}\mathbf{\Lambda}_{[r]}^{-(1/2)}$
$\mathbf{K}$=$\mathbf{XF}\mathbf{\Lambda}_{[r]}^{-(1/2)}$
$\mathbf{DXQX}^\top\mathbf{G} = \mathbf{G}\mathbf{\Lambda}_{[r]}$
$\mathbf{G}^\top\mathbf{D}^{-1}\mathbf{G} = \mathbf{I}_r$
Principal Cofactors $\mathbf{G}$
$\mathbf{K}$=$\mathbf{D}^{-1}\mathbf{G}$
$\mathbf{G}$=$\mathbf{DK}$
$\mathbf{XQX}^\top\mathbf{DK} = \mathbf{K}\mathbf{\Lambda}_{[r]}$
$\mathbf{K}^\top\mathbf{DK} = \mathbf{I}_r$
Principal Components $\mathbf{K}$

- The vectors $\mathbf{g}^1, \mathbf{g}^2, \ldots, \mathbf{g}^r$ successively maximize, under the $\mathbf{D}^{-1}$-orthogonality constraint, the quadratic form $\parallel \mathbf{X}^\top\mathbf{g} \parallel_{\mathbf{Q}}^2$.

- The vectors $\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^r$ successively maximize, under the $\mathbf{Q}^{-1}$-orthogonality constraint, the quadratic form $\parallel \mathbf{Xf} \parallel_{\mathbf{D}}^2$.

- If we search for a pair of vectors $\mathbf{a}$ (a $\mathbf{Q}$-normalized vector $\mathbf{a}$ of $\mathbb{R}^p$) and $\mathbf{k}$ (a $\mathbf{D}$-normalized vector of $\mathbb{R}^n$) which maximize the inner product $\langle \mathbf{XQa}|\mathbf{k}\rangle_{\mathbf{D}} = \left\langle \mathbf{X}^t\mathbf{Dk}|\mathbf{a}\right\rangle_{\mathbf{Q}}$, the solution is unique. It is obtained for $\mathbf{a} = \mathbf{a}^1$ and $\mathbf{k} = \mathbf{k}^1$ and the maximum is equal to $\sqrt{\lambda_1}$. Under the orthogonality constraint, the results can be extended for the other pairs.

These general properties correspond to different statistical criterions in practice. In the case of PCA on covariance matrix (i.e., $\mathbf{X} = \left[x_{ij} - \mathsf{m}(\mathbf{x}^j)\right]$, $\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$), we obtain $\parallel \mathbf{XQa} \parallel_{\mathbf{D}}^2 = \mathsf{VAR}(\mathbf{XQa})$ and $\parallel \mathbf{X}^\top\mathbf{Dk} \parallel_{\mathbf{Q}}^2 = \sum_{j=1}^{p} \mathsf{COV}^2(\mathbf{k}, \mathbf{x}^j)$. Hence, the analysis maximizes simultaneously the variance of the projection of $\mathbf{X}$ onto the principal axes and the sum of the squared covariances between the principal component and the variables of $\mathbf{X}$. Using the Pythagorean theorem, we can show that these two maximizations induce also the minimizations of the distances between the original configurations and those obtained in the subspaces defined by the principal axes and components (Figure 2).

Note that principal cofactors and factors are not useful in the examples presented in this paper but they could have some interest for some particular diagrams. For instance, in the case of canonical correlation analysis (`cancor` in package **stats**) which corresponds to a particular duality diagram, principal factors and cofactors contain the coefficients used to construct linear combination of variables with maximal correlation.

2.5. Matrix approximation

We have demonstrated that the analysis of the individuals and the analysis of the variables are closely related. In this section, we show how it is possible to represent on the same plot

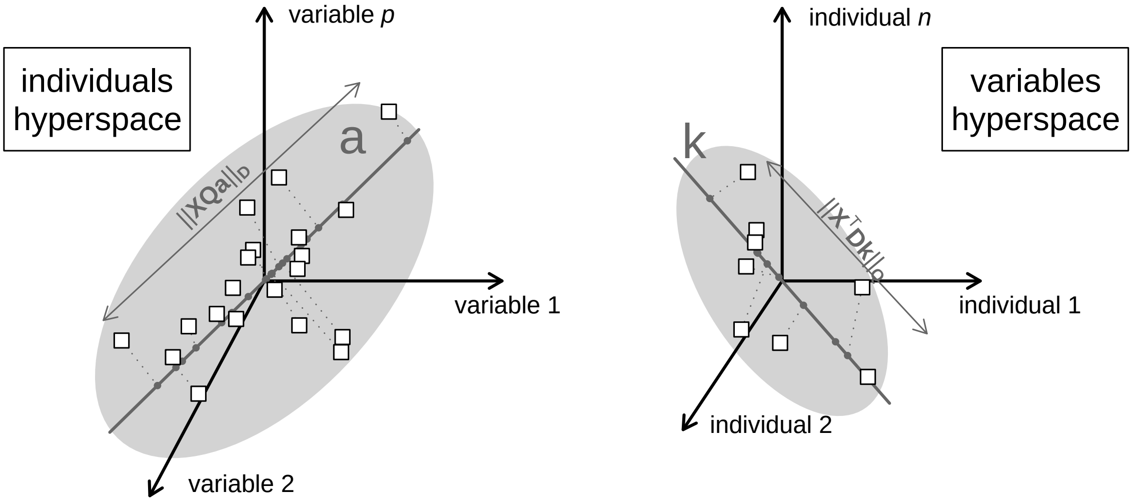


Figure 2: Representation of table $\mathbf{X}$ as a cloud of n points (individuals) in an hyperspace with p dimensions (variables) or as a cloud of p points (variables) in an hyperspace with n dimensions (individuals). In the first hyperspace, the principal axis $\mathbf{a}$ maximizes$\| \mathbf{XQa} \|_{\mathbf{D}}^2$. In the second hyperspace, the principal component $\mathbf{k}$ maximizes $\| \mathbf{X}^\top\mathbf{Dk} \|_{\mathbf{Q}}^2$.

the individuals, the variables and their relationships. This joint representation is linked to the theory of matrix approximation and is considered as the main output of multivariate methods.

We consider the product $\mathbf{K}\mathbf{\Lambda}_{[\mathbf{r}]}{}^{1/2}\mathbf{A}^\top$. Using the transition formulas defined above, we obtain:

$$\mathbf{K}\mathbf{\Lambda}_{[\mathbf{r}]}{}^{1/2}\mathbf{A}^\top = \mathbf{K}\mathbf{K}^\top\mathbf{D}\mathbf{X}$$

Left-multiplication by $\mathbf{K}^\top\mathbf{D}$ leads to:

$$\begin{aligned} \mathbf{K}^\top\mathbf{D}\mathbf{K}\mathbf{\Lambda}_{[\mathbf{r}]}{}^{1/2}\mathbf{A}^\top &= \mathbf{K}^\top\mathbf{D}\mathbf{X} \\ \mathbf{K}\mathbf{\Lambda}_{[\mathbf{r}]}{}^{1/2}\mathbf{A}^\top &= \mathbf{X} \end{aligned}$$

The diagonalization of a duality diagram is thus closely linked to the singular value decomposition of $\mathbf{X}$ (SVD, Eckart and Young 1936). We denote the row scores $\mathbf{L} = \mathbf{XQA}$ (i.e. projection of the rows of $\mathbf{X}$ onto the principal axes) and the column scores $\mathbf{C} = \mathbf{X}^\top\mathbf{DK}$ (i.e. projection of the columns of $\mathbf{X}$ onto the principal components). Using the transition formulas, we demonstrate that $\mathbf{L} = \mathbf{K}\Lambda_{[r]}^{1/2}$, $\mathbf{C} = \mathbf{A}\Lambda_{[r]}^{1/2}$ and $\mathbf{X} = \mathbf{K}\mathbf{\Lambda}_{[r]}^{1/2}\mathbf{A}^\top$. Therefore, $\mathbf{X}$ equals $\mathbf{LA}^\top$ and $\mathbf{KC}^\top$.

For $m < r$, $\mathbf{K}_{[m]}\mathbf{\Lambda}_{[m]}^{1/2}\mathbf{A}_{[m]}{}^\top$ is the best least-squares approximation of $\mathbf{X}$ of rank m (Gabriel 1978). The graphical representation of this approximation is given by a simultaneous plot of $\mathbf{K}_{[m]}$ and $\mathbf{C}_{[m]}$ (or $\mathbf{A}_{[m]}$ and $\mathbf{L}_{[m]}$) (i.e. biplot, Gabriel 1971).

2.6. Decomposition of inertia

Multivariate methods provide graphical representations which summarize the data table $\mathbf{X}$. In order to estimate the quality of this representation, additional tools can be used to (1) measure the part played by a variable or an individual in the construction of the representation and (2) evaluate the quality of the representation of each variable and individual in this new subspace. These tools are derived from the notion of inertia.

If a duality diagram contains at least one matrix of weights (i.e. $\mathbf{Q}$ and/or $\mathbf{D}$ diagonal), it is possible to compute inertia statistics. The inertia of a cloud of points is the weighted sum of the square distances between all the points and the origin. If $\mathbf{D}$ is diagonal, we can compute the inertia for the cloud of rows vectors (in $\mathbb{R}^p$). The total inertia of the diagram is equal to:

$$I(\mathbf{X},\mathbf{Q},\mathbf{D}) = \sum_{i=1}^{n} d_{ii} \parallel \mathbf{x}_i \parallel_{\mathbf{Q}}^2 = Trace(\mathbf{XQX}^\top\mathbf{D}) = \sum_{i=1}^{r} \lambda_i$$

where $\mathbf{d}_{ij}$ is the element at the i-th row and j-th column of $\mathbf{D}$ and $\mathbf{x}_i$ is the i-th row of the matrix $\mathbf{X}$. The rows of $\mathbf{X}$ can be projected onto a $\mathbf{Q}$-normalized vector $\mathbf{a}$ and the projected inertia is then equal to:

$$I(\mathbf{a}) = \mathbf{a}^\top\mathbf{QX}^\top\mathbf{DXQa} = \parallel \mathbf{XQa} \parallel_{\mathbf{D}}^2$$

From the properties of the diagram defined above, it appears that the diagonalization of the diagram consists in finding a set of $\mathbf{Q}$-normalized vector (the principal axes) which maximize the projected inertia. The inertia projected onto the principal axis $\mathbf{a}_k$ is equal to λ_k.

Two inertia statistics are usually used to facilitate the interpretation of results. The absolute contribution measures the contribution by one point to the inertia projected onto one axis. The absolute contribution of the row $\mathbf{x}_i$ to the axis $\mathbf{a}^k$ is equal to:

$$AC_{\mathbf{a}^k}(\mathbf{x}_i) = \frac{\parallel \mathbf{x}_i\mathbf{Q}\mathbf{a}^k \parallel_{\mathbf{D}}^2}{\lambda_k} \text{ and } \sum_{i=1}^{n} AC_{\mathbf{a}^k}(\mathbf{x}_i) = 1$$

The relative contribution (or Cos^2) quantifies the contribution of one axis to the inertia of a point. It measures the quality of representation of one point by its projection onto one axis. The relative contribution of the axis $\mathbf{a}^k$ to the row $\mathbf{x}_i$ is equal to:

$$RC_{\mathbf{x}_i}(\mathbf{a}^k) = \frac{\parallel \mathbf{x}_i\mathbf{Q}\mathbf{a}^k \parallel_{\mathbf{D}}^2}{d_{ii} \parallel \mathbf{x}_i \parallel_{\mathbf{Q}}^2} \text{ and } \sum_{k=1}^{r} RC_{\mathbf{x}_i}(\mathbf{a}^k) = 1$$

3. Implementation in R

The theoretical presentation considers that matrices $\mathbf{Q}$ and $\mathbf{D}$ are positive and symmetric. The duality diagram theory is more general and can consider also one non-positive matrix. The implementation in the function `as.dudi` is more restrictive and considers only diagonal matrices for $\mathbf{Q}$ and $\mathbf{D}$. However, as shown above, Cholesky decompositions of $\mathbf{Q}$ and $\mathbf{D}$ allow to easily obtain a diagram with two diagonal matrices. In this section, we show how the duality diagram theory is implemented in the **ade4**. Usually, the user performs a particular analysis by a call to a '`dudi.*`' function. This function contains a call to the `as.dudi` function and returns an object of the class `dudi`.

3.1. Principle of a '`dudi.*`' function

There is 10 '`dudi.*`' functions in **ade4**:

```
R> library("ade4")
R> apropos("dudi.")
```

```
 [1] "dudi.acm"       "dudi.coa"       "dudi.dec"       "dudi.fca"
 [5] "dudi.fpca"      "dudi.hillsmith" "dudi.mix"       "dudi.nsc"
 [9] "dudi.pca"       "dudi.pco"
```

The reader could consult Chessel, Dufour, and Thioulouse (2004) for a description of these different functions. The `dudi.pco` function performs a principal coordinates analysis (Gower 1966) and takes a distance matrix as argument. It does not use the `as.dudi` function but returns a `dudi` object. It is quite different to other functions and is not considered in the following description. The principles of the other '`dudi.*`' functions consist in:

1. A call is performed by the user. The argument `df` must be filled with a `data.frame` containing the data set of interest. Optional arguments can also be entered.

2. Arguments consistency is checked. For instance, `df` must contains only positive or null values for the `dudi.coa` function.

3. The three basic elements **X**, **Q** and **D** of the duality diagram are created. **X** is obtained by an eventual modification of the argument `df` (e.g., centering and/or scaling of variables for the function `dudi.pca`). Column (**Q**) and row weights (**D**) are computed and stored in two vectors. Note that for some methods, the user can choose its own vectors of weights using the optional arguments `col.w` and `row.w` when available.

4. The function `as.dudi` is called with the three elements defined above as arguments. The duality diagram is diagonalized and a `dudi` object is returned into the environment of the '`dudi.*`' function.

5. Depending on the method, some elements could be added to the `dudi` object.

6. The `dudi` object is returned to the environment where the function '`dudi.*`' has been called.

3.2. The `as.dudi` function

The function `as.dudi` is the core of the implementation of the duality diagram in **ade4**. It has 9 arguments which are described in its help page:

```
R> args(as.dudi)
```

```
function (df, col.w, row.w, scannf, nf, call, type, tol = 1e-07,
    full = FALSE)
NULL
```

The first part of this function consists in checking arguments consistency:

```
if (!is.data.frame(df))
  stop("data.frame expected")
lig <- nrow(df)
col <- ncol(df)
if (length(col.w) != col)
  stop("Non convenient col weights")
if (length(row.w) != lig)
  stop("Non convenient row weights")
if (any(col.w) < 0)
  stop("col weight < 0")
if (any(row.w) < 0)
  stop("row weight < 0")
if (full)
  scannf <- FALSE
```

Then, the diagonalization of the duality diagram is performed. In order to speed up this step, the function diagonalizes in the smaller dimension. If $n > p$ (i.e. `transpose=FALSE`), the matrix $\mathbf{\Omega}^\top\mathbf{\Omega}$ is diagonalized. If $p > n$ (i.e. `transpose=TRUE`), $\mathbf{\Omega}\mathbf{\Omega}^\top$ is diagonalized:

```
transpose <- FALSE
if(lig<col)
  transpose <- TRUE
res <- list(tab = df, cw = col.w, lw = row.w)
df <- as.matrix(df)
df.ori <- df
df <- df * sqrt(row.w)
df <- sweep(df, 2, sqrt(col.w), "*")
if(!transpose){
  df <- crossprod(df,df)
  }
else{
  df <- tcrossprod(df,df)
  }
eig1 <- eigen(df,symmetric=TRUE)
eig <- eig1$values
rank <- sum((eig/eig[1]) > tol)
```

When the diagonalization is performed, the user has to choose the number of axes to keep (`nf`). If the argument `scannf` is set to its default value `TRUE`, the screeplot of eigenvalues is displayed in order to facilitate this choice and to reduce the risk of over- or underestimation of the number of axes to interpret.

```
if (scannf) {
  if (exists("ade4TkGUIFlag") && ade4TkGUIFlag) {
    nf <- chooseaxes(eig, rank)
  }
  else {
    barplot(eig[1:rank])
    cat("Select the number of axes: ")
    nf <- as.integer(readLines(n = 1))
  }
}
if (nf <= 0)
  nf <- 2
if (nf > rank)
  nf <- rank
if (full)
  nf <- rank
res$eig <- eig[1:rank]
res$rank <- rank
res$nf <- nf
```

Lastly, principal axes, principal components, row and column scores are computed. If $n > p$, the function `eigen` returns eigenvectors $\mathbf{V}$.

Principal axes $\mathbf{A}$ are then obtained by $\mathbf{A} = \mathbf{E}^{-1}\mathbf{V}$ and verify $\mathbf{A}^\top\mathbf{Q}\mathbf{A} = \mathbf{I}_r$. Row scores are then computed by $\mathbf{L} = \mathbf{XQA}$. Column scores ($\mathbf{C}$) and principal components ($\mathbf{K}$) are then obtained by rescaling: $\mathbf{C} = \mathbf{A}\boldsymbol{\Lambda}^{(1/2)}$ and $\mathbf{K} = \mathbf{L}\boldsymbol{\Lambda}^{-(1/2)}$. An object of the class `dudi`, which is described in the next section, is returned.

```
col.w[hhich(col.w == 0)] <- 1
row.w[which(row.w == 0)] <- 1
dval <- sqrt(res$eig)[1:nf]
if(!transpose){
  col.w <- 1/sqrt(col.w)
  auxi <- eig1$vectors[, 1:nf] * col.w
  auxi2 <- sweep(df.ori, 2, res$cw, "*")
  auxi2 <- data.frame(auxi2%*%auxi)
  auxi <- data.frame(auxi)

  names(auxi) <- paste("CS", (1:nf), sep = "")
  row.names(auxi) <- names(res$tab)
  res$c1 <- auxi

  names(auxi2) <- paste("Axis", (1:nf), sep = "")
  row.names(auxi2) <- row.names(res$tab)
  res$li <- auxi2

  res$co <- sweep(res$c1,2,dval,"*")
  names(res$co) <- paste("Comp", (1:nf), sep = "")
  res$l1 <- sweep(res$li,2,dval,"/")
  names(res$l1) <- paste("RS", (1:nf), sep = "")


} else {
...
}

res$call <- call
class(res) <- c(type, "dudi")
return(res)
```

As seen above, the `as.dudi` function is called by a number of '`dudi.*`' functions. Note that experienced users can also directly call `as.dudi` to implement a particular analysis that is not available in the package.

3.3. The `dudi` class

The object returned by the `as.dudi` function is a `list` of class `dudi`. This object stores all the elements related to the diagonalization of a duality diagram. It contains at least these different components:

- `tab`: a `data.frame` (n rows, p columns) with the modified table $\mathbf{X}$
- `rw`: a `vector` (length n) of row weights ($\mathbf{D}$)
- `cw`: a `vector` (length p) of column weights ($\mathbf{Q}$)
- `eig`: a `vector` (length r) of eigenvalues ($\mathbf{\Lambda}$)
- `nf`: the number of axes kept
- `rank`: the rank of the duality diagram (r)
- `l1`: a `data.frame` (n rows, `nf` columns) with the principal components ($\mathbf{K}$)
- `c1`: a `data.frame` (p rows, `nf` columns) with the principal axes ($\mathbf{A}$)
- `li`: a `data.frame` (n rows, `nf` columns) with the row scores ($\mathbf{L}$)
- `co`: a `data.frame` (p rows, `nf` columns) with the column scores ($\mathbf{C}$)
- `call`: the matched `call`

There are three methods for the `dudi` class:

```
R> methods(class = "dudi")

[1] print.dudi   scatter.dudi t.dudi
```

The `print.dudi` function prints a `dudi` in a nice way. `t.dudi` transforms the `dudi` corresponding to the triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ into a new one corresponding to $(\mathbf{X}^\top, \mathbf{D}, \mathbf{Q})$. The function `scatter.dudi` provides a graphical representation of a `dudi` by a simultaneous representation of variables and individuals (i.e. biplot, Gabriel 1971).

There are other functions related to the `dudi` class: `is.dudi` tests if an object is of class `dudi`, `inertia.dudi` returns inertia statistics, `reconst` computes the table approximation and `redo.dudi` recomputes an analysis with a new number of axes.

4. An example: `dudi.hillsmith`

In this section, we analyze environmental information of the dune meadow data (Jongman, ter Braak, and Van Tongeren 1987). This data set is available in the **ade4** package. Data on the environment and land-use have been sampled in 20 sites:

```
R> data("dunedata")
R> sapply(dunedata$envir, class)
```

```
$A1
[1] "numeric"
$moisture
[1] "integer"

$manure
[1] "integer"

$use
[1] "ordered" "factor"

$management
[1] "factor"
```

The variables are:

- `A1`: thickness of the A1 horizon.
- `moisture`: moisture content of the soil.
- `manure`: quantity of manure applied.
- `use`: agricultural grassland use. This variable is coded as an ordered factor with levels hayfields < both < grazing.
- `management`: grassland management type. This variable is coded as a factor with levels `"SF"` (standard farming), `"BF"` (biological farming), `"HF"` (hobby farming) and `"NM"` (nature conservation management).

The ordered variable `use` is transformed into a factor:

```
R> dunedata$envir$use <- factor(dunedata$envir$use, ordered = FALSE)
R> summary(dunedata$envir)

       A1             moisture        manure           use       management
 Min.   : 2.800   Min.   :1.0   Min.   :0.00   hayfield:7   BF:3
 1st Qu.: 3.500   1st Qu.:1.0   1st Qu.:0.00   both    :8   HF:5
 Median : 4.200   Median :2.0   Median :2.00   grazing :5   NM:6
 Mean   : 4.850   Mean   :2.9   Mean   :1.75                SF:6
 3rd Qu.: 5.725   3rd Qu.:5.0   3rd Qu.:3.00
 Max.   :11.500   Max.   :5.0   Max.   :4.00
```

PCA on correlation matrix (`dudi.pca` with the arguments `scale` and `center` equal to `TRUE`) is a natural choice to analyze a table of quantitative variables measured in different units. Multiple correspondence analysis (MCA, `dudi.acm`, Tenenhaus and Young 1985) is devoted to the analysis of a table of qualitative variables. In this example, we use Hill-Smith analysis (`dudi.hillsmith`, Hill and Smith 1976) to summarize the environmental table which contains a mix of quantitative and qualitative variables. This method is a compromise between PCA and MCA and is equivalent to PCA when there are only quantitative variables and to MCA if there are only qualitative variables. The function `dudi.hillsmith` creates firstly the basic elements of the triplet. Suppose that the original table contains p_q quantitative variables and p_f factors ($p = p_f + p_q$). To construct the table $\mathbf{X}$, a quantitative variable is not modified while a qualitative variable with m levels is coded by m dummy variables. If the number of levels for each factor is $m_1, \cdots, m_{p_f}$, the resulting table $\mathbf{X}$ has $p_q + m_1 + \cdots + m_{p_f}$ columns.

By default, $\mathbf{D} = (1/n)\mathbf{I}_n$. Columns weights are computed and stored in $\mathbf{Q}$. If the j-th column of $\mathbf{X}$ (denoted $\mathbf{x}^j$) corresponds to a quantitative variable then $q_{jj} = 1$. If $\mathbf{x}^j$ corresponds to a dummy variable coding the l-th level of the f-th factor, then $q_{jj} = (\mathbf{x}^j)^\top \mathbf{D}\mathbf{x}^j = n_{f(l)}/n$ where $n_{f(l)}$ is the number of individuals of the l-th level of the f-th factor.

Lastly, table $\mathbf{X}$ is modified. If $\mathbf{x}^j$ corresponds to a quantitative variable, it is scaled to mean $(\mathbf{x}^j)^\top \mathbf{D}\mathbf{1}_n = 0$ and variance $(\mathbf{x}^j)^\top \mathbf{D}\mathbf{x}^j = 1$ where $\mathbf{1}_n$ is the unit vector of with n rows. If $\mathbf{x}^j$ corresponds to a dummy variable, it is transformed into $(\mathbf{x}^j - q_{jj}\mathbf{1}_n)/q_{jj} = (n/n_{f(l)})\mathbf{x}^j - \mathbf{1}_n$.

```
R> dd1 <- dudi.hillsmith(dunedata$envir, scannf = FALSE, nf = 2)
R> dd1

Duality diagramm
class: mix dudi
$call: dudi.hillsmith(df = dunedata$envir, scannf = FALSE, nf = 2)
$nf: 2 axis-components saved
$rank: 8
eigen values: 2.542 1.858 1.231 0.9899 0.6927 ...
  vector length mode    content
1 $cw    10     numeric column weights
2 $lw    20     numeric row weights
3 $eig   8      numeric eigen values

  data.frame nrow ncol content
```

```
1 $tab       20   10   modified array
2 $li        20   2    row coordinates
3 $l1        20   2    row normed scores
4 $co        10   2    column coordinates
5 $c1        10   2    column normed scores
other elements: assign index cr
```

This analysis seeks for principal axes ($\mathbf{a}^j$) which maximize the quadratic form $\| \mathbf{l}^j \|_{\mathbf{D}}^2 = \| \mathbf{XQa}^j \|_{\mathbf{D}}^2$ with orthogonality constraints $((\mathbf{a}^j)^\top \mathbf{Qa}^j = 1$ and $(\mathbf{a}^i)^\top \mathbf{Qa}^j = 0$ for $i \neq j$. In other words, the analysis finds coefficients ($\mathbf{a}^j$) to obtain a linear combination of variables ($\mathbf{l}^j = \mathbf{XQa}^j$) which maximizes $\| \mathbf{l}^j \|_{\mathbf{D}}^2 = \mathsf{VAR}(\mathbf{l}^j) = \lambda_j$.

Simultaneously, the analysis seeks for principal components ($\mathbf{k}^j$) which maximize the quadratic form $\| \mathbf{c}^j \|_{\mathbf{Q}}^2 = \| \mathbf{X}^\top \mathbf{Dk}^j \|_{\mathbf{Q}}^2$ with orthogonality constraints $((\mathbf{k}^j)^\top \mathbf{Dk}^j = 1$ and $(\mathbf{k}^i)^\top \mathbf{Dk}^j = 0$ for $i \neq j)$.

If the i-th column of $\mathbf{X}$ corresponds to a quantitative variable, the quantity $\| (\mathbf{x}^i)^\top \mathbf{Dk}^j \|_{\mathbf{Q}}^2$ is equal to $\mathsf{COR}^2(\mathbf{x}^i, \mathbf{k}^j)$.

If the i-th column of $\mathbf{X}$ corresponds to a dummy variable, the quantity $\| (\mathbf{x}^i)^\top \mathbf{Dk}^j \|_{\mathbf{Q}}^2$ is equal to $(n_{f(l)}/n) \cdot \mathsf{m}(\mathbf{k}^j, f(l))$ where $\mathsf{m}(\mathbf{k}^j, f(l))$ is the mean of $\mathbf{k}^j$ computed for the individuals of l-th level of the factor f. Summing these quantities for the m_f dummy variables of the factor f leads to $\sum_{l=1}^{m_f} (n_{f(l)}/n) \cdot \mathsf{m}(\mathbf{k}^j, f(l)) = \eta^2(\mathbf{k}^j, f)$ (i.e. a correlation ratio). In other words, the analysis finds principal components which maximizes the sum of squared correlations (for quantitative variables) and correlation ratios (for qualitative variables).

Results of the analysis are summarized on the biplot using the function `scatter`. By default, the first two principal axes and row scores are represented (Figure 3). Setting the argument `permute` to `TRUE` allows to represent principal components and column scores.

```
R> scatter.dudi(dd1)
```

The first axis of the analysis discriminates sites with high level of manure which is related to standard farming (sites 1, 3, 4, 12, 13 and 16) from conserved sites (14, 15, 17, 18, 19, 20). The second axis separates sites with high moisture and A1 horizon (15,14) from dry sites managed as hobby or biological farming (2, 7, 10, 11). The analysis highlights the main environmental variations and provides a synthetic typology of the sites which could be useful for conservation purposes. One could expect to relate this structure to variations in species richness or species composition (gradient analysis). Results obtained by this analysis can also be useful to improve the sampling protocol. For instance, results show that the information given by the thickness of the A1 horizon and the moisture content of the soil is quite redundant. If one wants to reduce the cost of future sampling sessions without losing important information, he could then choose to measure only one of these two variables.

5. Conclusions

This presentation focuses on the analysis of one table using a `dudi.*` function. The duality diagram theory is more general and several other methods can also be considered. These

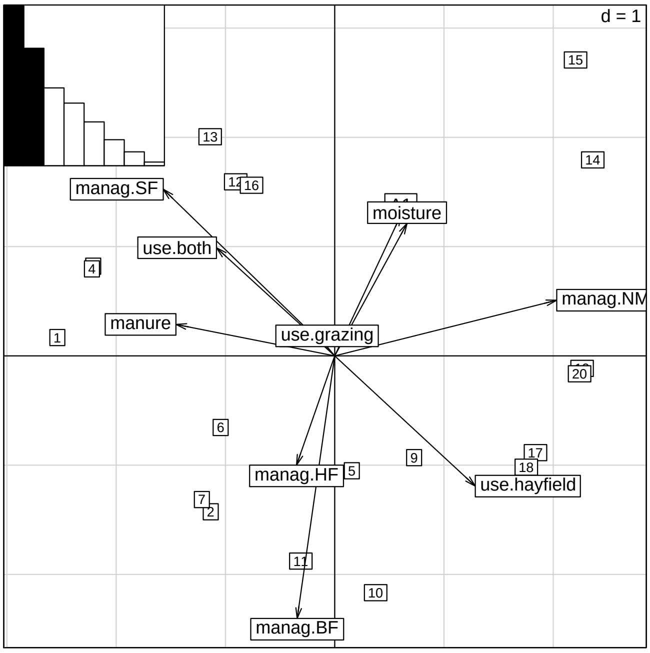


Figure 3: Hill-Smith analysis of the dune meadow data. Representation of the first two principal axes and row scores on a biplot.

include, for instance, methods to take into account a partition of individuals such as discriminant analysis (`discrimin`) or within-between classes analysis (`within` and `between`, Dolédec and Chessel 1987), methods to analyze a pair of tables such as co-inertia analysis (`coinertia`, Dolédec and Chessel 1994; Dray *et al.* 2003) or principal component analysis on instrumental variables (`pcaiv`, Rao 1964; Lebreton, Sabatier, Banco, and Bacou 1991) including redundancy analysis (van den Wollenberg 1977) and canonical correspondence analysis (`cca`, ter Braak 1986). All these methods are particular duality diagrams and their implementations in **ade4** contain a call to the `as.dudi` function.

The duality diagram theory allows to easily define and compare methods using a well established mathematical framework. It also simplifies the implementation of methods as functions are always based on the same skeleton. The reader could consult Chessel *et al.* (2004) and Dray, Dufour, and Chessel (2007) for a more detailled description of the contents of the package **ade4**. It should be noticed that **ade4** contains also more than 100 data sets to illustrate the different methods and that theoretical elements as well as ecological illustrations are presented in the pedagogical ressources available to French readers at `http://pbil.univ-lyon1.fr/R/enseignement.html`.

Acknowledgments

The authors are indebted to Daniel Chessel for working on the **ade4** package and for teaching so much about multivariate analysis and to Simon Penel for maintaining the package. Comments of Thomas Kneib and two anonymous reviewers on a first draft of this manuscript greatly improve the quality of the paper.

References

Bayley PB, Peterson JT (2001). "An Approach to Estimate Probability of Presence and Richness of Fish Species." *Transactions of the American Fisheries Society*, **130**, 620–633.

Cailliez F, Pagès JP (1976). *Introduction à l'Analyse des Données.* SMASH, Paris.

Cazes P (1970). *Application de l'Analyse des Données au Traitement de Problèmes Géologiques.* Thèse de 3ème cycle, Faculté des Sciences de Paris.

Chessel D, Dufour AB, Thioulouse J (2004). "The **ade4** Package – I: One-Table Methods." *R News*, **4**(1), 5–10.

Dolédec S, Chessel D (1987). "Rythmes Saisonniers et Composantes Stationnelles en Milieu Aquatique I- Description d'un Plan d'Observations Complet par Projection de Variables." *Acta Oecologica – Oecologia Generalis*, **8**(3), 403–426.

Dolédec S, Chessel D (1994). "Co-Inertia Analysis: An Alternative Method for Studying Species-Environment Relationships." *Freshwater Biology*, **31**, 277–294.

Dolédec S, Chessel D, Gimaret-Carpentier C (2000). "Niche Separation in Community Analysis: a New Method." *Ecology*, **81**(10), 2914–2927.

Dolédec S, Chessel D, ter Braak CJF, Champely S (1996). "Matching Species Traits to Environmental Variables: a New Three-Table Ordination Method." *Environmental and Ecological Statistics*, **3**, 143–166.

Dray S, Chessel D, Thioulouse J (2003). "Co-Inertia Analysis and the Linking of Ecological Data Tables." *Ecology*, **84**, 3078–3089.

Dray S, Dufour AB, Chessel D (2007). "The **ade4** Package – II: Two-Table and K-Table Methods." *R News*, **7**(2). Forthcoming.

Dray S, Pettorelli N, Chessel D (2002). "Matching Data Sets From Two Different Spatial Samples." *Journal of Vegetation Science*, **13**, 867–874.

Eckart C, Young G (1936). "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika*, **1**(3), 211–218.

Escoufier Y (1987). "The Duality Diagram : A Means of Better Practical Applications." In P Legendre, L Legendre (eds.), "Developments in Numerical Ecology," volume 14, pp. 139–156. Springer Verlag, Berlin.

Gabriel KR (1971). "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis." *Biometrika*, **58**(3), 453–467.

Gabriel KR (1978). "Least Squares Approximation of Matrices by Additive and Multiplicative Models." *Journal of the Royal Statistical Society B*, **40**(2), 186–196.

Gauch HG (1982). *Multivariate Analysis in Community Ecology.* Cambridge University Press, Cambridge.

Goodall DW (1954). "Objective Methods for the Classification of Vegetation III. An Essay on the Use of Factor Analysis." *Australian Journal of Botany*, **2**, 304–324.

Gower JC (1966). "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika*, **53**(3–4), 325–338.

Hill M, Smith A (1976). "Principal Component Analysis of Taxonomic Data with Multi-State Discrete Characters." *Taxon*, **25**, 249–255.

Holmes S (2006). "Multivariate Analysis: The French Way." In D Nolan, T Speed (eds.), "Festschrift for David Freedman," IMS, Beachwood, OH.

Jongman R, ter Braak C, Van Tongeren O (1987). *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen.

Lebreton J, Sabatier R, Banco G, Bacou A (1991). "Principal Component and Correspondence Analyses with Respect to Instrumental Variables : An Overview of Their Role in Studies of Structure-Activity and Species-Environment Relationships." In J Devillers, W Karcher (eds.), "Applied Multivariate Analysis in SAR and Environmental Studies," pp. 85–114. Kluwer Academic Publishers.

Legendre P, Gallagher E (2001). "Ecologically Meaningful Transformations for Ordination of Species Data." *Oecologia*, **129**, 271–280.

Noy-Meir I (1973). "Data Transformation in Ecological Ordination. I. Some Advantages of Non-Centring." *Journal of Ecology*, **61**, 329–341.

Noy-Meir I, Walker D, Williams WT (1975). "Data Transformation in Ecological Ordination. II. On the Meaning of Data Standardization." *Journal of Ecology*, **63**, 779–800.

Pavoine S, Dufour AB, Chessel D (2004). "From Dissimilarities Among Species to Dissimilarities Among Communities: a Double Principal Coordinate Analysis." *Journal of Theoretical Biology*, **228**, 523–537.

Rao CR (1964). "The Use and Interpretation of Principal Component Analysis in Applied Research." *Sankhya A*, **26**, 329–359.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Tenenhaus M, Young FW (1985). "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data." *Psychometrika*, **50**(1), 91–119.

ter Braak CJF (1986). "Canonical Correspondence Analysis: a New Eigenvector Technique for Multivariate Direct Gradient Analysis." *Ecology*, **67**, 1167–1179.

Thioulouse J, Chessel D, Dolédec S, Olivier J (1997). "**ADE-4**: A Multivariate Analysis and Graphical Display Software." *Statistics and Computing*, **7**, 75–83.

van den Wollenberg A (1977). "Redundancy Analysis, an Alternative for Canonical Analysis." *Psychometrika*, **42**(2), 207–219.

Affiliation:

Stéphane Dray, Anne-Béatrice Dufour
Laborartoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS
Université de Lyon
université Lyon 1
43, Boulevard du 11 Novembre 1918
F-69622 Villeurbanne Cedex, France
E-mail: `dray@biomserv.univ-lyon1.fr`, `dufour@biomserv.univ-lyon1.fr`
URL: `http://biomserv.univ-lyon1.fr/~dray/`

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 22, Issue 4 *Submitted:* 2007-02-13
September 2007 *Accepted:* 2007-06-04