Appendix A
A Euclidean Viewpoint on Statistics

This appendix gives the main algebraic and geometric principles used in the descriptive statistic methods presented in this book.

A.1 Inner and Dot Products

Let us consider two vectors $\mathbf{x}$ and $\mathbf{y}$ of $\mathbb{R}^n$. The inner product is a function that associates a real number to the pair of vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\langle \,|\, \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

with the following properties:

- **Symmetric:** $\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
- **Bilinear:**
  - $\langle \mathbf{x}|\mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{x}|\mathbf{z}\rangle, \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$
  - $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha \langle \mathbf{x}|\mathbf{y}\rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\forall \alpha \in \mathbb{R}$
- **Positive definite:** $\langle \mathbf{x}|\mathbf{x}\rangle \geq 0, \forall \mathbf{x} \in \mathbb{R}^n$
- **Non-degenerate:** $\langle \mathbf{x}|\mathbf{x}\rangle = 0 \Rightarrow \mathbf{x} = \mathbf{0}, \forall \mathbf{x} \in \mathbb{R}^n$

In $\mathbb{R}^n$, the dot product is the inner product defined in the standard basis by:

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{y}^\top \mathbf{x}$$

If $\mathbf{A}$ is an $n \times n$ symmetric positive definite matrix, then the bilinear form $\mathbf{y}^\top \mathbf{A}\mathbf{x}$ satisfies the four properties and defines thus an inner product denoted:

J. Thioulouse et al., *Multivariate Analysis of Ecological Data with ade4*,
https://doi.org/10.1007/978-1-4939-8850-1

$$\langle \mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_j y_i$$

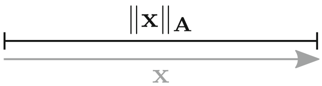
In $\mathbb{R}^n$, the usual dot product defined in the standard basis is obtained by setting $\mathbf{A}$ to the identity matrix $\mathbf{I}_n$. More generally, given a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{R}^n$, the matrix $\mathbf{A}$ defined by $a_{ij} = \langle \mathbf{v}_i|\mathbf{v}_j\rangle_{\mathbf{A}}$ is the unique matrix representing the dot product $\langle \mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}}$. Indeed, we have $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{v}_i$ and $\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{v}_i$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and:

$$\begin{aligned}\langle \mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}} &= \left\langle \sum_{i=1}^{n} x_i \mathbf{v}_i \middle| \sum_{j=1}^{n} y_j \mathbf{v}_j \right\rangle_{\mathbf{A}} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i \langle \mathbf{v}_i|\mathbf{v}_j\rangle_{\mathbf{A}}\, y_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} y_j = \mathbf{y}^\top \mathbf{A}\mathbf{x}\end{aligned}$$

A.2 Length, Projection, Angle and Distance

The norm (or length) of a vector $\mathbf{x}$ is defined by:

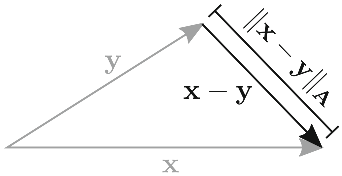
$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle_{\mathbf{A}}}$$



Note that $\|\alpha\mathbf{x}\|_{\mathbf{A}} = |\alpha|\, \|\mathbf{x}\|_{\mathbf{A}}$.

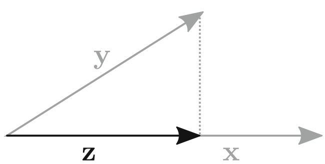
The distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is the norm of their difference:

$$d_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{A}}$$



The projection of $\mathbf{y}$ on the nonzero vector $\mathbf{x}$ is a vector $\mathbf{z}$ parallel to $\mathbf{x}$ so that $\mathbf{y} - \mathbf{z}$ is orthogonal to $\mathbf{x}$. It is given by:

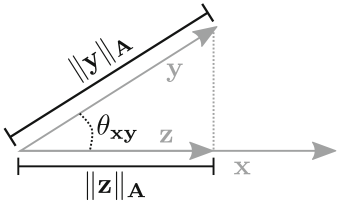
$$\mathbf{z} = \frac{\langle \mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}}}{\langle \mathbf{x}|\mathbf{x}\rangle_{\mathbf{A}}}\mathbf{x}$$



It follows that

$$\cos(\theta_{\mathbf{xy}}) = \frac{\|\mathbf{z}\|_{\mathbf{A}}}{\|\mathbf{y}\|_{\mathbf{A}}} = \frac{\langle\mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}}}{\|\mathbf{x}\|_{\mathbf{A}}\,\|\mathbf{y}\|_{\mathbf{A}}}$$

$$0 \leq \theta_{\mathbf{xy}} \leq \pi$$



and thus

$$\langle\mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}} = \|\mathbf{x}\|_{\mathbf{A}}\,\|\mathbf{y}\|_{\mathbf{A}}\cos(\theta_{\mathbf{xy}})$$

Hence, two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\langle\mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}} = 0$. Moreover, we have

$$\left|\langle\mathbf{x}|\mathbf{y}\rangle_{\mathbf{A}}\right| \leq \|\mathbf{x}\|_{\mathbf{A}}\,\|\mathbf{y}\|_{\mathbf{A}} \text{ (Cauchy-Schwartz inequality)}$$

and

$$\|\mathbf{x}+\mathbf{y}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{A}} + \|\mathbf{y}\|_{\mathbf{A}} \text{ (triangular inequality)}$$

A.3 Mean and Variance

The observed values of a quantitative variable for n individuals are stored in $\mathbf{x} = (x_1, \cdots, x_n)^\top$, a vector of $\mathbb{R}^n$. The mean of $\mathbf{x}$ is equal to

$$\mathrm{m}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and its variance is

$$\mathrm{v}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mathrm{m}(\mathbf{x}))^2$$

Let us consider the uniform inner product of $\mathbb{R}^n$ associated to the diagonal matrix $\frac{1}{n}\mathbf{I}_n$. In a geometric viewpoint, the standard mean is computed by an inner product and corresponds to a Euclidean projection (Fig. A.1):

$$\mathrm{m}(\mathbf{x}) = \langle\mathbf{x}|\mathbf{1}_n\rangle_{\frac{1}{n}\mathbf{I}_n}$$

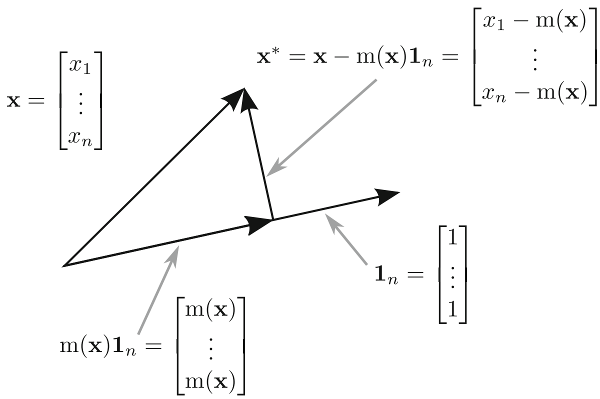


Fig. A.1 Centring a variable seen as an orthogonal projection.

The variance is equal to the squared norm of the centred vector $\mathbf{x}^*$

$$\mathrm{v}(\mathbf{x}) = \|\mathbf{x} - \mathrm{m}(\mathbf{x})\mathbf{1}_n\|^2_{\frac{1}{n}\mathbf{I}_n} = \left\|\mathbf{x}^*\right\|^2_{\frac{1}{n}\mathbf{I}_n}$$

A.4 Weighted Mean and Varianc

A weighting function can be defined to give some individuals more influence on the result than other individuals. Weights for the n individuals are stored in a vector $\mathbf{w}$ of $\mathbb{R}^n$. They are positive and their sum is equal to 1:

$$\mathbf{w} = (w_1 \cdots w_n)^\top \text{ with } \sum_{i=1}^{n} w_i = 1 \text{ and } w_i > 0$$

Using $\mathbf{w}$, the weighted mean of $\mathbf{x}$ is

$$\mathrm{m}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i$$

and the weighted variance equals

$$\mathrm{v}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{n} w_i \left(x_i - \mathrm{m}_{\mathbf{w}}(x)\right)^2$$

Considering the diagonal matrix $\mathbf{D}_{\mathbf{w}} = \mathrm{diag}(\mathbf{w})$ as the inner product of $\mathbb{R}^n$, the weighted mean and variance are given by:

$$\mathrm{m}_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{x} | \mathbf{1}_n \rangle_{\mathbf{D_w}}$$

and

$$\mathrm{v}_{\mathbf{w}}(\mathbf{x}) = \| \mathbf{x} - \mathrm{m}_{\mathbf{w}}(\mathbf{x}) \mathbf{1}_n \|^2_{\mathbf{D_w}}$$

The standard mean and variance (m($\mathbf{x}$), v($\mathbf{x}$)) correspond to the particular cases of weighted statistics ($\mathrm{m}_{\mathbf{w}}(\mathbf{x})$, $\mathrm{v}_{\mathbf{w}}(\mathbf{x})$) when uniform weights $w_i = \frac{1}{n}$ are chosen. From a geometric viewpoint, computing standard or weighted statistics corresponds to the same operation (i.e., a projection) but using different inner products.

A.5 Covariance and Correlation

The values of two quantitative variables are stored in the vectors $\mathbf{x}$ and $\mathbf{y}$. This information can be considered either as n points (individuals) in $\mathbb{R}^2$ or as 2 points (variables) of $\mathbb{R}^n$. In the first case, data centring corresponds to moving the origin of the system of axes (Fig. A.2a). In the second case, it corresponds to two orthogonal projections on the vector $\mathbf{1}_n$ (Fig. A.2b).

The vectors $\mathbf{x}^*$ and $\mathbf{y}^*$ contain centred data. The standard covariance is equal to

$$\begin{aligned} \mathrm{cov}(\mathbf{x}, \mathbf{y}) &= \mathrm{cor}(\mathbf{x}, \mathbf{y}) \sqrt{\mathrm{v}(\mathbf{x})} \sqrt{\mathrm{v}(\mathbf{y})} \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathrm{m}(\mathbf{x}))(y_i - \mathrm{m}(\mathbf{y})) = \frac{1}{n} \sum_{i=1}^{n} x_i^* y_i^* \end{aligned}$$

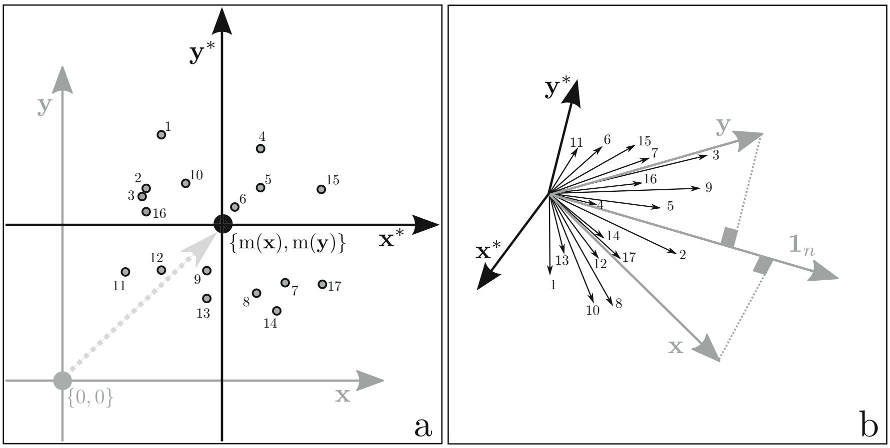


Fig. A.2 Two geometric viewpoints on centring (example with 2 variables and 17 individuals). It corresponds to (**a**) move the origin in $\mathbb{R}^2$ and to (**b**) two orthogonal projections in $\mathbb{R}^{17}$.

It can be rewritten as:

$$\begin{aligned}\mathrm{cov}(\mathbf{x},\mathbf{y}) &= \langle \mathbf{x}^*|\mathbf{y}^*\rangle_{\frac{1}{n}\mathbf{I}_n} \\ &= \|\mathbf{x}\|_{\frac{1}{\mathbf{n}}\mathbf{I_n}}\,\|\mathbf{y}\|_{\frac{1}{\mathbf{n}}\mathbf{I_n}}\cos(\theta_{\mathbf{xy}})\end{aligned}$$

As $\|\mathbf{x}\|_{\frac{1}{\mathbf{n}}\mathbf{I_n}} = \sqrt{\mathrm{v}(\mathbf{x})}$ and $\|\mathbf{y}\|_{\frac{1}{\mathbf{n}}\mathbf{I_n}} = \sqrt{\mathrm{v}(\mathbf{y})}$, it follows that:

$$\mathrm{cor}(\mathbf{x},\mathbf{y}) = \cos(\theta_{\mathbf{xy}})$$

Hence, the covariance is equal to the dot product between the two vectors whereas the correlation is the cosine of the angle formed by the two vectors. Note that weighted covariance and correlation could be obtained by using the appropriate inner product $\mathbf{D_w}$.

A.6 Linear Regression

The linear model that aims to explain the variation of $\mathbf{y}$ by the dependent variable $\mathbf{x}$ can be written as:

$$\mathbf{y} = \beta\mathbf{x} + \alpha\mathbf{1}_n + \boldsymbol{\epsilon}$$

Estimates of α and β are chosen to minimise the sum of squared residuals $\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ (Fig. A.3a). The least squares estimates of parameters are given by:

$$\hat{\alpha} = \mathrm{m}(\mathbf{y}) - \hat{\beta}\mathrm{m}(\mathbf{x}) \text{ and } \hat{\beta} = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\mathrm{v}(\mathbf{x})}$$

Considering the centred variables $\mathbf{x}^*$ and $\mathbf{y}^*$, the estimate of the slope can be rewritten as:

$$\hat{\beta} = \frac{\langle \mathbf{x}^*|\mathbf{y}^*\rangle_{\frac{1}{n}\mathbf{I}_n}}{\|\mathbf{x}^*\|^2_{\frac{1}{n}\mathbf{I}_n}}$$

Thus, the vector of predicted values $\hat{\mathbf{y}}$ can be decomposed as follows:

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{\beta}\mathbf{x} + (\mathrm{m}(\mathbf{y}) - \hat{\beta}\mathrm{m}(\mathbf{x}))\mathbf{1}_n \\ &= \hat{\beta}(\mathbf{x}^* + \mathrm{m}(\mathbf{x})\mathbf{1}_n) + (\mathrm{m}(\mathbf{y}) - \hat{\beta}\mathrm{m}(\mathbf{x}))\mathbf{1}_n \\ &= \hat{\beta}\mathbf{x}^* + \mathrm{m}(\mathbf{y})\mathbf{1}_n\end{aligned}$$

$$= \frac{\langle \mathbf{x}^* | \mathbf{y}^* \rangle_{\frac{1}{n}\mathbf{I}_n}}{\|\mathbf{x}^*\|^2_{\frac{1}{n}\mathbf{I}_n}} \mathbf{x}^* + \langle \mathbf{y} | \mathbf{1}_n \rangle_{\frac{1}{n}\mathbf{I}_n} \mathbf{1}_n$$

The previous equation shows that the vector of predicted values $\mathbf{y}$ can be computed as the sum of two vectors (Fig. A.3b). The first vector corresponds to the projection of the centred variable $\mathbf{y}^*$ on $\mathbf{x}^*$. The second vector corresponds to the projection of $\mathbf{y}$ on $\mathbf{1}_n$.

As $\mathbf{x} = \mathbf{x}^* + \mathrm{m}(\mathbf{x})\mathbf{1}_n$, the three vectors $\mathbf{x}$, $\mathbf{x}^*$ and $\mathbf{1}_n$ are linearly dependent and thus lie in the same plane. It follows that vector of fitted values $\hat{\mathbf{y}}$ corresponds to the orthogonal projection of $\mathbf{y}$ on the plane spanned by the vectors $\mathbf{x}$ and $\mathbf{1}_n$. Applying the Pythagorean theorem to the triangle formed by the vectors $\mathbf{y}$, $\hat{\mathbf{y}}$ and $\boldsymbol{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$, we obtained the well-known decomposition of variance (Fig. A.4):

$$\|\mathbf{y}\|^2_{\frac{1}{n}\mathbf{I}_n} = \underbrace{\|\hat{\mathbf{y}}\|^2_{\frac{1}{n}\mathbf{I}_n}}_{\text{explained variance}} + \underbrace{\|\mathbf{y} - \hat{\mathbf{y}}\|^2_{\frac{1}{n}\mathbf{I}_n}}_{\text{residual variance}}$$

The coefficient of determination, $R^2_{\mathbf{y}|\mathbf{x}}$ measures the proportion of variance of the dependent variable $\mathbf{y}$ explained by the explanatory variable $\mathbf{x}$. Geometrically, it is the cosine of the angle formed by the vectors $\hat{\mathbf{y}}$ and $\mathbf{y}$ (Fig. A.4):

$$R^2_{\mathbf{y}|\mathbf{x}} = \frac{\|\hat{\mathbf{y}}\|^2_{\frac{1}{n}\mathbf{I}_n}}{\|\mathbf{y}\|^2_{\frac{1}{n}\mathbf{I}_n}} = \cos(\theta_{\hat{\mathbf{y}}\mathbf{y}})$$

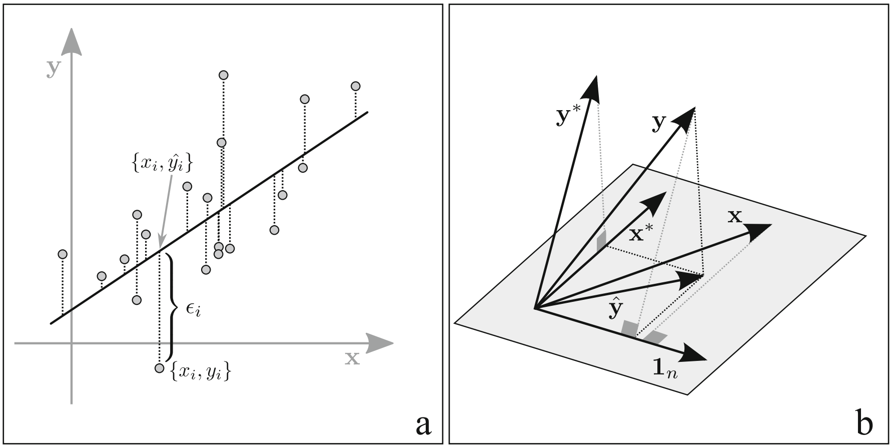


Fig. A.3 Two geometric viewpoints on linear regression with intercept (example with one explanatory variable and 20 individuals). In $\mathbb{R}^2$ (**a**), the usual representation shows that the regression line minimises the residual sum of squares. In $\mathbb{R}^{20}$ (**b**), fitted values are obtained by orthogonal projection of $\mathbf{y}$ on the plane spanned by vectors $\mathbf{x}$ and $\mathbf{1}_n$.

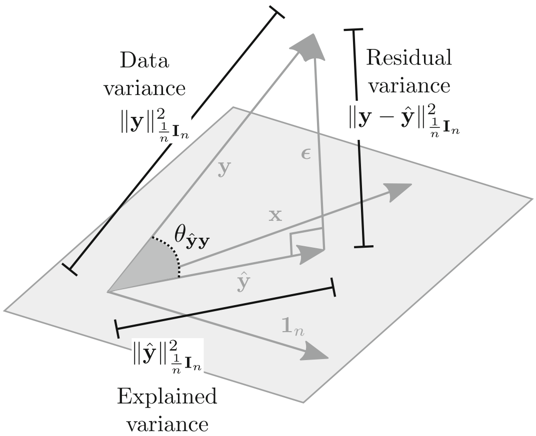


Fig. A.4 Geometric decomposition of the variance using the Pythagorean theorem.

A.7 Categorical Variables

A categorical variable is a variable that can take one of a finite number of possible values, each individual being assigned to a particular group (category, level or class). If we consider a categorical variable with m categories measured for n individuals, the information can be coded as a vector $\mathbf{q}$ of integers. An $n \times m$ table $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_m]$ of dummy variables can be built. For the k-th category, the dummy variable $\mathbf{x}_k$ is equal to 1 if the individual belongs to this category and 0 otherwise:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ n \end{array} \begin{array}{c} \text{Category} \\ \begin{bmatrix} \text{blue} \\ \text{red} \\ \text{blue} \\ \text{green} \\ \vdots \\ \text{black} \end{bmatrix} \end{array} \rightarrow \begin{array}{c} \mathbf{q} \\ \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \\ \vdots \\ m \end{bmatrix} \end{array} \rightarrow \begin{array}{c} \begin{array}{ccccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \cdots & \mathbf{x}_m \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}$$

Whereas a quantitative variable corresponds to a vector, a categorical variable defines a subspace spanned by vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. If we consider a diagonal matrix of weights $\mathbf{D_w}$, the dummy variables are orthogonal by definition (i.e., $\langle \mathbf{x}_i | \mathbf{x}_j \rangle_{\mathbf{D_w}}$ for $i \neq j$). The weight w_k^+ associated to the k-th category is equal to the sum of the weights of the individuals belonging to this category. It is equal to the squared norm of the associated dummy variable, $w_k^+ = \|\mathbf{x}_k\|^2_{\mathbf{D_w}}$.

Let us consider a quantitative variable $\mathbf{y}$. The projection of $\mathbf{y}$ on the k-th dummy variable is equal to:

$$\mathcal{P}_{\mathbf{x}_k}(\mathbf{y}) = \frac{\langle \mathbf{y} | \mathbf{x}_k \rangle_{\mathbf{D_w}}}{\|\mathbf{x}_k\|^2_{\mathbf{D_w}}} \mathbf{x}_k = \frac{\sum\limits_{i/q_i=k} w_i y_i}{w_k^+} \mathbf{x}_k = \mathrm{m}_{\mathbf{w}/k}(\mathbf{y}) \mathbf{x}_k$$

The value $\mathrm{m}_{\mathbf{w}/k}(\mathbf{y})$ is the conditional mean of $\mathbf{y}$ given k (i.e., the weighted mean of the variable $\mathbf{y}$ computed only on the individuals belonging to the k-th category). Hence, the vector $\mathcal{P}_{\mathbf{x}_k}(\mathbf{y})$ takes the value $\mathrm{m}_{\mathbf{w}/k}(\mathbf{y})$ for the individuals of the k-th category and 0 otherwise.

It follows that the projection of the centred variable $\mathbf{y}^* = \mathbf{y} - \mathrm{m}_{\mathbf{w}}(\mathbf{y})\mathbf{1}_n$ on $\mathbf{x}_k$ is simply given by:

$$\mathcal{P}_{\mathbf{x}_k}(\mathbf{y}^*) = (\mathrm{m}_{\mathbf{w}/k}(\mathbf{y}) - \mathrm{m}_{\mathbf{w}}(\mathbf{y}))\mathbf{x}_k$$

As the dummy variables are orthogonal, the projection on the subspace spanned by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ is simply the sum of the individual projections on each vector $\mathbf{x}_k$:

$$\mathcal{P}_{\mathbf{X}}(\mathbf{y}^*) = \sum_{k=1}^{m} \mathcal{P}_{\mathbf{x}_k}(\mathbf{y}^*)$$

After some substitutions, the squared norm of this projection can be rewritten as:

$$\left\|\mathcal{P}_{\mathbf{X}}(\mathbf{y}^*)\right\|^2_{\mathbf{D_w}} = \sum_{k=1}^{m} w_k^+ (\mathrm{m}_{\mathbf{w}/k}(\mathbf{y}) - \mathrm{m}_{\mathbf{w}}(\mathbf{y}))^2 = b(\mathbf{y})$$

The quantity $b(\mathbf{y})$ is the between-group variance that measures the differences among categories. Using the Pythagorean theorem, the within-group variance is defined by

$$w(\mathbf{y}) = \left\|\mathbf{y}^* - \mathcal{P}_{\mathbf{X}}(\mathbf{y}^*)\right\|^2_{\mathbf{D_w}}$$

and we obtain the standard ANOVA decomposition of variance:

$$\left\|\mathbf{y}^*\right\|^2_{\mathbf{D_w}} = \underbrace{\left\|\mathcal{P}_{\mathbf{X}}(\mathbf{y}^*)\right\|^2_{\mathbf{D_w}}}_{\text{between-group variance}} + \underbrace{\left\|\mathbf{y}^* - \mathcal{P}_{\mathbf{X}}(\mathbf{y}^*)\right\|^2_{\mathbf{D_w}}}_{\text{within-group variance}}$$

The correlation ratio $\eta^2(\mathbf{q}, \mathbf{y}) = \frac{b(\mathbf{y})}{\mathrm{v}_{\mathbf{w}}(\mathbf{y})}$ measures the proportion associated to the between-group variance. It varies between 0 and 1. Geometrically, it is the cosine of the angle formed by the vectors $\mathcal{P}_{\mathbf{X}}(\mathbf{y}^*)$ and $\mathbf{y}^*$.

A.8 Weighted Multiple Regression

Multiple regression aims to explain the variation of a response variable $\mathbf{y}$ by several dependent variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ stored in column in an $n \times p$ table $\mathbf{X}$

($\mathbf{X} = [\mathbf{x}_1|\dots|\mathbf{x}_p] = [x_{ij}]$). For a given weighting matrix $\mathbf{D_w}$, the aim of multiple regression is to predict the observation y_i by a linear model:

$$\hat{y}_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \alpha = y_i - \epsilon_i$$

The weighted least-squares estimation leads to minimise the residual sum of squares:

$$RSS = \sum_{i=1}^{n} w_i(\hat{y}_i - y_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2_{\mathbf{D_w}}$$

The minimisation of the RSS is provided by the orthogonal projection of $\mathbf{y}$ on the subspace spanned by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{1}_n$. The vector $\mathbf{1}_n$ is added to consider the intercept in the model so that

$$\hat{\mathbf{y}} = \beta_1\mathbf{x}_1 + \cdots + \beta_p\mathbf{x}_p + \alpha\mathbf{1}_n$$

The vector of predicted values $\hat{\mathbf{y}}$ exists and is unique. The uniqueness of the coefficients $\beta_1, \cdots, \beta_p, \alpha$ is ensured only if the vectors $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{1}_n$ are independent (i.e., no multicollinearity). This independence is obtained if and only if the centred vectors $\mathbf{x}_1^*, \dots, \mathbf{x}_p^*$ are independent, with $\mathbf{x}_i^* = \mathbf{x}_i - \mathrm{m}_{\mathbf{w}}(\mathbf{x}_i)\mathbf{1}_n$. If the centred vectors are independent, the covariance matrix $\mathbf{X}^{*\top}\mathbf{D_w}\mathbf{X}^*$ is invertible (with $\mathbf{X}^* = [\mathbf{x}_1^*|\dots|\mathbf{x}_p^*]$). In this case, we have:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathcal{P}_{\mathbf{X}^*}(\mathbf{y}) + \mathcal{P}_{\mathbf{1}_n}(\mathbf{y}) \\ &= \mathcal{P}_{\mathbf{X}^*}(\mathbf{y}^*) + \mathcal{P}_{\mathbf{1}_n}(\mathbf{y}^*) + \mathcal{P}_{\mathbf{X}^*}(\mathrm{m}_{\mathbf{w}}(\mathbf{y})\mathbf{1}_n) + \mathcal{P}_{\mathbf{1}_n}(\mathrm{m}_{\mathbf{w}}(\mathbf{y})\mathbf{1}_n)\end{aligned}$$

By definition, the centred vectors $\mathbf{x}_1^*, \dots, \mathbf{x}_p^*, \mathbf{y}^*$ are orthogonal to $\mathbf{1}_n$ so that the previous equation simplifies to

$$\hat{\mathbf{y}} = \mathcal{P}_{\mathbf{X}^*}(\mathbf{y}^*) + \mathcal{P}_{\mathbf{1}_n}(\mathrm{m}_{\mathbf{w}}(\mathbf{y})\mathbf{1}_n)$$

In the standard basis, the projection operator $\mathcal{P}_{\mathbf{X}^*}(.)$ is simply equal to $\mathbf{X}^*(\mathbf{X}^{*\top}\mathbf{D_w}\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{D_w}$ and the previous equation can be rewritten as:

$$\hat{\mathbf{y}} = \mathbf{X}^*(\mathbf{X}^{*\top}\mathbf{D_w}\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{D_w}\mathbf{y}^* + \mathrm{m}_{\mathbf{w}}(\mathbf{y})\mathbf{1}_n$$

The estimates of the parameters are then obtained by

$$\begin{bmatrix}\hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p\end{bmatrix} = (\mathbf{X}^{*\top}\mathbf{D_w}\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{D_w}\mathbf{y}^*$$

and

$$\hat{\alpha} = \mathrm{m}_{\mathbf{w}}(\mathbf{y}) - \hat{\beta}_1 \mathrm{m}_{\mathbf{w}}(\mathbf{x}_1^*) - \cdots - \hat{\beta}_p \mathrm{m}_{\mathbf{w}}(\mathbf{x}_p^*)$$

As in simple regression, the part of variance explained by the model is equal to the ratio of two squared norms (and thus the cosine of the angle formed by these two vectors):

$$R^2_{\mathbf{y}|\mathbf{X}} = \frac{\|\hat{\mathbf{y}}\|^2_{\mathbf{D_w}}}{\|\mathbf{y}\|^2_{\mathbf{D_w}}}$$

Appendix B
Graphical User Interface

Abstract This chapter is a short presentation of **ade4TkGUI**, a Tcl/Tk Graphical User Interface (GUI) package for some basic functions of **ade4**. The **ade4TkGUI** package tries to mix the advantages of a GUI (ease of use, no need to learn numerous commands) with the possibility to use **R** expressions in the dialog boxes, to generate understandable **R** commands, and to manage a session `.Rhistory` file.

B.1 Introduction

This chapter is based on the paper by Thioulouse and Dray (2007), but only the most interesting features of **ade4TkGUI** are detailed here. The **ade4** package is a part of a previous software that was written in C. This software was mainly used by ecologists, and it had a rich and very useful GUI, written in HyperTalk and based successively on HyperCard, WinPlus and MetaCard (see Chapter 1, Thioulouse et al. 1997).

Switching to **R** and to the command line interface of **ade4** was a hard task for many users, and we decided to make it easier by providing them with a GUI. The first aim of **ade4TkGUI** was to give the users of "Classical ADE-4" an easy access to the main functions of **ade4**. As most users would also be new to **R**, we wanted it to be easy to install, and using **Tcl/Tk** was a guarantee of easiness and multi-platform compatibility.

Only one-table and two-table methods are currently available in **ade4TkGUI** and graphical functions are limited to the basic classes. *K*-table methods are not included.

We decided to use the **Tcl/Tk** language to implement **ade4TkGUI** because the **tcltk** package is available in **R**, and included by default in the base distribution. Many other GUI development systems are available, but they do not offer the same level of availability and platform independence as **tcltk**.

J. Thioulouse et al., *Multivariate Analysis of Ecological Data with ade4*,
https://doi.org/10.1007/978-1-4939-8850-1

B.2 Overview of the ade4TkGUI Package

It is not possible to give here a detailed description of all the functions of **ade4TkGUI**, and only the main characteristics will be presented. The core of the package is the `ade4TkGUI()` function, which opens the main GUI window (Fig. B.1).

In the main GUI window, buttons are grouped in 6 rows, according to their function: Data sets, One table analyses, One table analyses with groups, Two tables analyses, Graphic functions, and Advanced graphics. To avoid cluttering this window, only a limited subset of functions is displayed. Less frequently used functions are available through the menus of the menu bar, located at the top of the window. Right-clicking the buttons opens the **ade4** help window for the corresponding function. The question-head button opens the help window of **ade4TkGUI**.

The `ade4TkGUI()` function takes two boolean arguments, `show` and `history`. The first one determines whether the **R** commands generated by the GUI

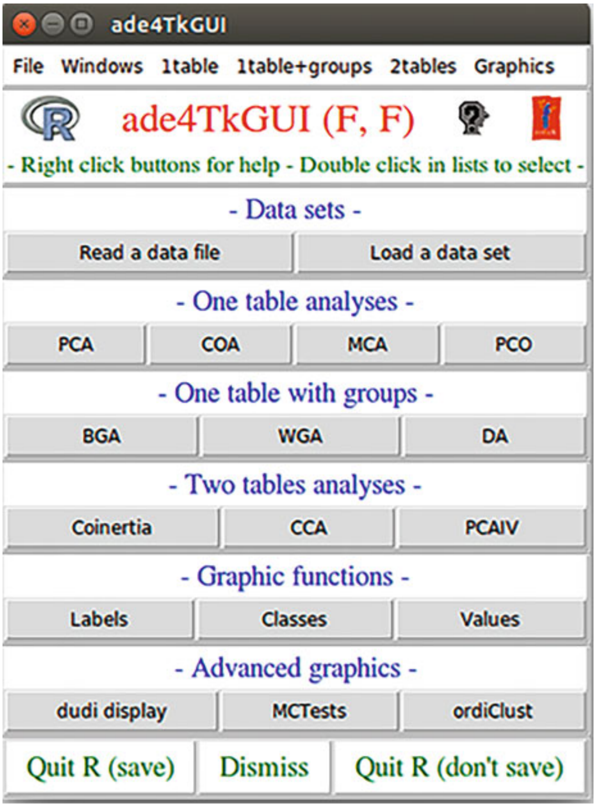


Fig. B.1 The main **ade4TkGUI** window.

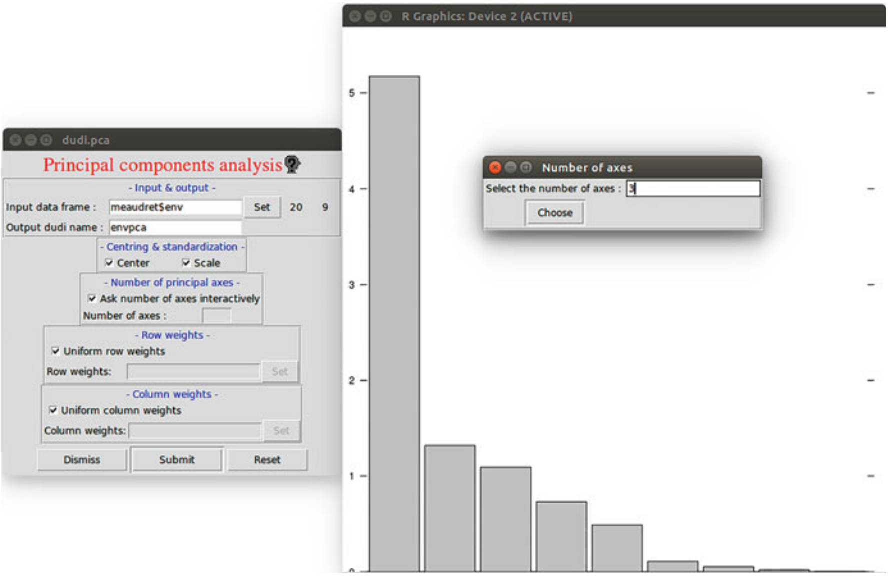


Fig. B.2 The `dudi.pca` function GUI window (left), the eigenvalues barchart (right) and the selection of the number of axes by the user (top-right).

should be printed in the console. When users interact with the GUI, they modify the status of **tcltk** widgets, and when they click on the "`Submit`" button, an **R** command is generated from the status of these widgets. This command is executed and can optionally be displayed in the console. If the `history` argument is set to TRUE, the commands generated by the GUI are also stored in the `.Rhistory` file, where they can easily be retrieved by users. The state of the two parameters is recalled in the main window heading "`ade4TkGUI(T,T)`".

The "`Read a data file`" button opens a dialog window that can be used to set the parameters of the `read.table` command to read a data text file. The "`Load a data set`" just displays the list of **ade4** data sets. This list can be used to choose a particular data set and to load it in memory using the `data` command.

When the "`PCA`" button is clicked, a new window appears (Fig. B.2): this is the GUI window of the `dudi.pca` function.

In this new window, the "`Set`" button can be used to choose the PCA input data frame through a listbox showing the list of data frames in the user global environment. After the "`Input data frame`" text field has been filled by the user, the number of rows and columns (20, 9) are displayed next to it. The output of the `dudi.pca` function is an object of class `dudi` and the user can type the name of this object in the "`Output dudi name`" field. If this field is left empty, the name "`Untitled1`" is used automatically.

The remaining widgets can be used to set particular options for the PCA: centring and standardisation, number of principal axes used to compute row and column coordinates, and row and column weights.

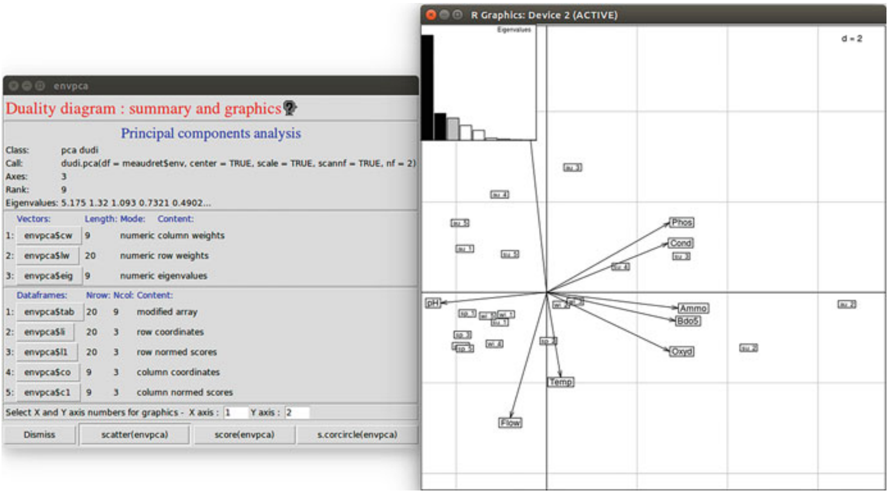


Fig. B.3 The `dudi` object display window (left) and the biplot obtained by clicking on the "`scatter`" button (right).

Most of the windows created by **ade4TkGUI** are non-blocking, which means that the user can do other things in the GUI or in the **R** console before taking the action required by this window. This was designed to make the interface more flexible and easier to use.

Clicking the "`Submit`" button starts the PCA computations. When they are completed, the barplot of eigenvalues is displayed (right of Fig. B.2) and, if this option was chosen in the previous dialog window, the user is asked to select the number of axes on which the row and column scores should be computed.

After scores are computed, the `dudi` window is displayed (Fig. B.3, left). This window shows a summary of the analysis, and displays the elements of the `dudi` object under the form of buttons. All these buttons can be used to draw graphs of the corresponding elements. For example, the row and column coordinates buttons draw the classical factor maps. In the lower part of the window, the user can choose which axes are used to draw these graphs.

The last row of buttons gives access to special graphs, according to the particular properties of the `dudi` that is displayed. For example, in the case of a normed PCA, the "`s.corcircle`" button allows to draw a correlation circle. The "`scatter`" button draws a biplot, with a small barchart for eigenvalues (Fig. B.3, right). These additional buttons are adapted to the type of `dudi` that is displayed, and they allow to draw graphs that illustrate particular properties of this `dudi`.

An example of GUI for one of the graphical functions of **adegraphics** is given in Fig. B.4. This is the `s.class` function, which allows to draw factor maps with groups of individuals. The user can choose the data frame containing the row scores (here they come from the "`envpca`" `dudi`), and the factor that should be used

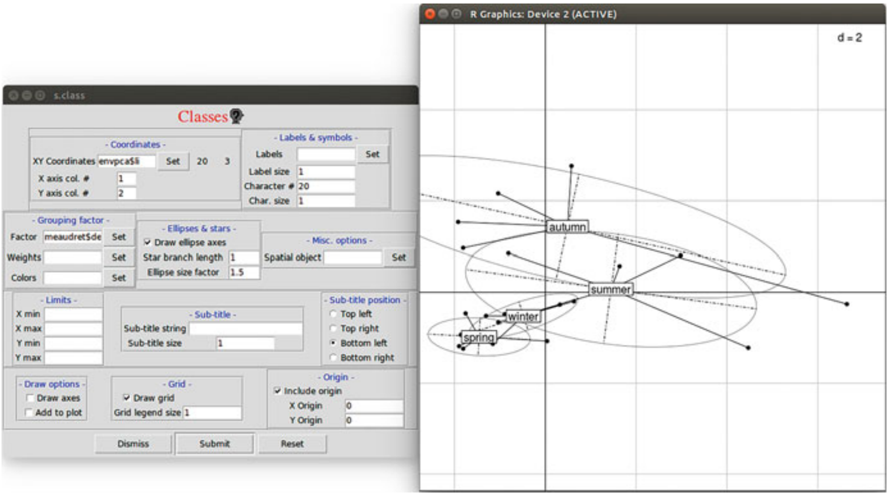


Fig. B.4 The `s.class` GUI window (left) and the corresponding graph (right).

to draw the groups on the factor map. Many other options can be set to enrich the graphs.

B.3 Conclusion

The main advantage of a GUI is the ease of use for beginners, occasional users, or teachers and students. It makes easier learning how to use a software by making the learning curve smoother, or to get back to work after a long period. This is particularly important in the case of **ade4** in ecological data analysis, because ecologists are mostly occasional users of **R**.

An important feature in **ade4** is the `dudi`, a complex **R** object containing all the information relating to a duality diagram. The `dudi` GUI window (Fig. B.3) was designed to display all the components of a `dudi`, and to draw automatically default graphs for each of these components. Therefore, it offers a centralised and synthetic view of an analysis, and it allows to see rapidly and interactively many graphs. In command line mode, the user must know all the components of a `dudi`, and remember which one is needed to draw a particular graph; this is particularly difficult for occasional users.

The **ade4TkGUI** package also facilitates the use of **ade4** by pre-selecting the type of objects that are proposed to users when they must do a selection. For example, in the `dudi.pca` dialog window (Fig. B.2), when the user clicks on the “`Set`” button to select a data frame, the dialog box contains only data frames present in the global environment (or in lists present in the global environment). In the same window, if the user wants to set non-uniform row weights, the “`Set`” button for row weights displays only vectors of length equal to the number of rows of the data frame. More generally, lists are filtered to propose only objects with properties consistent with the aim of the action. In the same way, in the `dudi` window, the buttons and their functions are coherent with the type of `dudi` and with the mathematical properties of its components.

Obviously, a GUI is not well adapted to scripting, and even to simple repetitive tasks. It is also not good for batch, online or remote use, and it is not easy to integrate into Sweave or R Markdown documents and vignettes. This is probably the main drawback of GUIs: they are made for *personal* and *instant* use, while the command line interface (CLI) allows many operations like scripting, re-doing the same analysis later, sharing pieces of code among colleagues, and batch use for time-consuming computations.

GUIs and CLIs should not be opposed, but considered as complementary. GUIs make the learning curve smoother for beginners, and can be used in education to introduce students to CLI mode. CLI mode is more powerful, it allows to build more complex analyses, particularly when using several packages jointly. When possible, the joint use of both CLI and GUI is attractive, as the user gets the benefits of the two approaches. Joint use can be very intimate: for example, it is possible to use **R** expressions in the GUI dialogs, and the GUI can return **R** expressions that can be copied and pasted in the console. In the case of **ade4TkGUI**, the strings typed by the user in the text fields of the GUI are parsed, and it is therefore possible to use **R** expressions, for example, to specify a subset of a data frame in a PCA.

When `ade4TkGUI` is called with argument "`show = TRUE`", **R** commands built by the GUI are echoed to the console. It is then possible to copy/paste these commands and execute them when needed in the console. This is also an effective way for beginners to learn how to use elaborate **R** function calls. Occasional users can thus analyse these command lines and possibly adapt them to their needs, with the additional benefit of gradually learning the **R** language.

In addition to this **Tcl/Tk** GUI, we are developing a new Shiny application to use the main `ade4` functionalities through a Web application. Shiny is an **R** package that makes very easy building interactive Web Apps in **R**. The main advantage is that users do not have to install **R** and multiple packages: they only need a web browser. As an example, a first piece of this "in progress" work is deployed on the *shinyapps.io* web site. It can be used to perform a PCA at this URL: https://ade4.shinyapps.io/ShinyPCA. We plan to develop this approach in the near future and hope to be able to propose a complete Shiny GUI to the `ade4` package.

Appendix C
Index of Boxes

C.1 Chapter 3: The `dudi` Class

C.2 Chapter 4: Multivariate Analysis Graphs

C.3 Chapter 5: Description of Environmental Variables Structures

C.4 Chapter 6: Description of Species Structures

J. Thioulouse et al., *Multivariate Analysis of Ecological Data with ade4*,
https://doi.org/10.1007/978-1-4939-8850-1

C.5 Chapter 7: Taking into Account Groups of Sites

C.6 Chapter 8: Description of Species-Environment Relationships

C.7 Chapter 9: Analysing Changes in Structures

C.8 Chapter 10: Analysing Changes in Co-structures

C.9 Chapter 11: Relating Species Traits to Environment

C.10 Chapter 12: Analysing Spatial Structures

C.11 Chapter 13: Analysing Phylogenetic Structures

C.12 Chapter 14: Analysing Patterns of Biodiversity

References

Abdi H, Williams LJ, Valentin D, Bennani-Dosse M (2012) STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. Wiley Interdiscip Rev Comput Stat 4:124–167

Abouheif E (1999) A method for testing the assumption of phylogenetic independence in comparative data. Evol Ecol Res 1:895–909

Ackerly DD, Donoghue MJ (1998) Leaf size, sapling allometry, and corner's rules: phylogeny and correlated evolution in maples (acer). Am Nat 152(6):767–791

Anselin L (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer MM, Scholten HJ, Unwin D (eds) Spatial analytical perspectives on GIS. Taylor and Francis, London, pp 111–125

Auda Y (1983) Rôle des méthodes graphiques en analyse des données: application au dépouillement des enquêtes écologiques. Thèse de 3ème cycle, Université Lyon 1

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol Model 157:101–118

Austin M, Noy-Meir I (1971) The problem of non-linearity in ordination: experiments with two-gradient models. J Ecol 59:763–773

Bady P, Dolédec S, Dumont B, Fruget JF (2004) Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. C R Biol 327:29–36

Bauman D, Drouet T, Dray S, Vleminckx J (2018) Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography. https://doi.org/10.1111/ecog.03380

Beals EW (1973) Ordination: mathematical elegance and ecological naivete. J Ecol 61:23–35

de Bélair G, Bencheikh-Lehocine M (1987) Composition et déterminisme de la végétation d'une plaine côtière marécageuse: La Mafragh (Annaba, Algérie). Bull d'Ecologie 18(4):393–407

Benzécri JP (1969) Statistical analysis as a tool to make patterns emerge from data. In: Watanabe S (ed) Methodologies of pattern recognition. Academic, New York, pp 35–60

Bertin J (1967) Les diagrammes, les réseaux, les cartes. Mouton & Gautier-Villars, Paris

Bertrand F, Maumy M (2010) Using partial triadic analysis for depicting the temporal evolution of spatial structures: assessing phytoplankton structure and succession in a water reservoir. Case Stud Bus Ind Gov Stat 4:23–43

Bivand R, Lewin-Koh N (2017) maptools: tools for reading and handling spatial objects. https://CRAN.R-project.org/package=maptools, R package version 0.9-2

Bivand R, Piras G (2015) Comparing implementations of estimation methods for spatial econometrics. J Stat Softw 63(18):1–36

J. Thioulouse et al., *Multivariate Analysis of Ecological Data with ade4*,
https://doi.org/10.1007/978-1-4939-8850-1

Bivand RS, Pebesma E, Gomez-Rubio V (2013) Applied spatial data analysis with R, 2nd edn. Springer, New York

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717–745

Blondel J, Farré H (1988) The convergent trajectories of bird communities along ecological successions in European forests. Oecologia 75:83–93

Borcard D, Gillet F, Legendre P (2011) Numerical Ecology with R. Springer, New York

Bougeard S, Dray S (2018) Supervised multiblock analysis in R with the ade4 package. J Stat Soft 86(1):1–17. http://doi.org/10.18637/jss.v086.i01

Bouroche JM (1975) Analyse des données ternaires: la double analyse en composantes principales. PhD thesis, Thèse de troisième cycle, Université Paris VI

ter Braak CJF (1983) Principal components biplots and alpha and beta diversity. Ecology 64(3):454–462

ter Braak C (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179

ter Braak CJF (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal reponse model. Biometrics 41:859–873

ter Braak CJF (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69:69–77

ter Braak CJF, Looman CWN (1986) Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65:3–11

ter Braak CJF, Verdonschot PFM (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat Sci 57:255–289

ter Braak C, Cormont A, Dray S (2012) Improved testing of species traits-environment relationships in the fourth corner problem. Ecology 93:1525–1526

Cardoso GC, Price TD (2010) Community convergence in bird song. Evol Ecol 24:447–461

Chessel, D. (1992) Echanges interdisciplinaires en analyse des données écologiques. Mémoire d'Habilitation à Diriger des Recherches. 107 pages, Université Lyon 1, 69622 Villeurbanne Cedex, France.

Chessel D, Hanafi M (1996) Analyse de la co-inertie de K nuages de points. Rev Stat Appl 44:35–60

Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package - I: one-table methods. R News 4:5–10

Cheverud JM, Dow MM (1985) An autocorrelation analysis of genetic variation due to lineal fission in social groups of Rhesus macaques. Am J Phys Anthropol 67:113–121

Cheverud JM, Dow MM, Leutenegger W (1985) The quantitative assessment of phylogenetic constaints in comparative analyses: sexual dimorphism in body weights among primates. Evolution 39:1335–1351

Choler P (2005) Consistent shifts in Alpine plant traits along a mesotopographical gradient. Arct Antarct Alp Res 37(4):444–453

Culhane A, Perriere G, Considine E, Cotter T, Higgins D (2002) Between-group analysis of microarray data. Bioinformatics 18:1600–1608

Curtis JT, McIntosh RP (1951) An upland forest continuum in the prairie-forest border region of Wisconsin. Ecology 32(3):476–496

Dale M, Fortin MJ (2010) From graphs to spatial graphs. Ann Rev Ecol Evol Syst 41(1):21–38

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 Oecol Oecol Generalis 8:403–426

Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw Biol 31:277–294

Dolédec S, Chessel D, ter Braak CJF, Champely S (1996) Matching species traits to environmental variables: a new three-table ordination method. Environ Ecol Stat 3:143–166

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46
Dragulescu AA, Arendt C (2018) xlsx: read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. https://CRAN.R-project.org/package=xlsx, R package version 0.6.1
Dray S (2008) On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. Comput Stat Data Anal 52:2228–2237
Dray S, Dufour A (2007) The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 22(4):1–20
Dray S, Jombart T (2011) Revisiting Guerry's data: introducing spatial constraints in multivariate analysis. Ann Appl Stat 5:2278–2299
Dray S, Josse J (2015) Principal component analysis with missing values: a comparative survey of methods. Plant Ecol 216:657–667
Dray S, Legendre P (2008) Testing the species traits-environment relationships: the fourth-corner problem revisited. Ecology 89:3400–3412
Dray S, Chessel D, Thioulouse J (2003) Co-inertia analysis and the linking of ecological data tables. Ecology 84:3078–3089
Dray S, Legendre P, Peres-Neto PR (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecol Model 196:483–493
Dray S, Dufour A, Chessel D (2007) The ade4 package–II: two-table and K-table methods. R News 7:47–52
Dray S, Saïd S, Débias F (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J Veg Sci 19:45–56
Dray S, Pélissier R, Couteron P, Fortin MJ, Legendre P, Peres-Neto PR, Bellier E, Bivand R, Blanchet FG, De Caceres M, Dufour AB, Heegaard E, Jombart T, Munoz F, Oksanen J, Thioulouse J, Wagner HH (2012) Community ecology in the age of multivariate multiscale spatial analysis. Ecol Monogr 82(3):257–275
Dray S, Choler P, Dolédec S, Peres-Neto PR, Thuiller W, Pavoine S, ter Braak CJ (2014) Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. Ecology 95(1):14–21
Dray S, Pavoine S, Aguirre de Carcer D (2015) Considering external information to improve the phylogenetic comparison of microbial communities: a new approach based on constrained Double Principal Coordinates Analysis (cDPCoA). Mol Ecol Resour 15:242–249
Dray S, Blanchet G, Borcard D, Clappe S, Guenard G, Jombart T, Larocque G, Legendre P, Madi N, Wagner HH (2018) adespatial: multivariate multiscale spatial analysis. https://CRAN.R-project.org/package=adespatial, R package version 0.2-0
Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308:1635–1638
Escofier B, Pagès J (1994) Multiple Factor Analysis (AFMULT package). Comput Stat Data Anal 18:121–140
Escoufier Y (1973) Le traitement des variables vectorielles. Biometrics 29:750–760
Escoufier Y (1987) The duality diagram: a means of better practical applications. In: Legendre P, Legendre L (eds) Development in numerical ecology. NATO advanced Institute, Serie G. Springer, Berlin, pp 139–156
Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15
Foucart T (1978) Sur les suites de tableaux de contingence indexées par le temps. Statistique et Analyse des données 3:67–85
Franquet E, Chessel D (1994) Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. Comptes Rendus de l'Académie des sciences Série 3 317:202–206
Franquet E, Dolédec S, Chessel D (1995) Using multivariate analyses for separating spatial and temporal effects within species-environment relationships. Hydrobiologia 300:425–431

Friday L (1987) The diversity of macroinvertebrate and macrophyte communities in ponds. Freshw Biol 18:87–104

Friendly M, Dray S (2014) Guerry: maps, data and methods related to Guerry (1833) "Moral statistics of France". http://CRAN.R-project.org/package=Guerry, R package version 1.6-1

Gabriel K (1971) The biplot graphical display of matrices with application to principal component analysis. Biometrika 58:453–467

Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5(3):115–145

Gimaret-Carpentier C, Dray S, Pascal JP (2003) Broad-scale biodiversity pattern of the endemic tree flora of the Western Ghats (India) using canonical correlation analysis of herbarium records. Ecography 26:429–444

Gittleman JL, Kot M (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool 39:227–241

Gleason HA (1926) The individualistic concept of the plant association. Bull Torrey Bot Club 53(1):7–26

Goodall DW (1954) Objective methods for the classification of vegetation. III. An essay on the use of factor analysis. Aust J Bot 2:304–324

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338

Gower JC (1984) Distance matrices and their euclidean approximation. In: Diday E, Jambu M, Lebart L, Pagès J, Tomassone R (eds) Data analysis and informatics III. Elsevier, Amsterdam, pp 3–21

Gower J, Legendre P (1986) Metric and euclidean properties of dissimilarity coefficients. J Classif 3:5–48

Green RH (1971) A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of Central Canada. Ecology 52:543–556

Green RH (1974) Multivariate niche analysis with temporally varying environmental factors. Ecology 55:73–83

Greenacre M (1984) Theory and applications of correspondence analysis. Academic, London

Griffith DA (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Can Geogr 40(4):351–367

Harvey PH, Pagel M (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Hill M (1973) Reciprocal averaging: an eigenvector method of ordination. J Ecol 61:237–249

Hill MO (1974) Correspondence Analysis: a neglected multivariate method. J R Stat Soc Ser C (Appl Stat) 23:340–354

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: Nolan D, Speed T (eds) Festschrift for David Freedman. IMS, Beachwood, OH, pp 1–14

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441

Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377

Jaffrenou P (1978) Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. PhD thesis, Thèse de troisième cycle, Université Lyon 1

Jaromczyk JW, Toussaint GT (1992) Relative neighborhood graphs and their relatives. Proc IEEE 80(9):1502–1517

Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F, Nesme X, Etienne J, Vandenesch F (2002) Relationships between staphylococcus aureus genetic background, virulence factors, agr type (alleles), and human disease type. Infect Immun 70:631–641

Jolliffe I (2002) Principal component analysis, 2nd edn. Springer, New York

Jombart T, Dray S, Dufour AB (2009) Finding essential scales of spatial variation in ecological data: a multivariate approach. Ecography 32:161–168

Jombart T, Balloux F, Dray S (2010a) adephylo: new tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26:1907–1909

Jombart T, Pavoine S, Dufour AB, Pontier D (2010b) Putting phylogeny into the analysis of biological traits: a methodological approach. J Theor Biol 264:693–701
de Jong P, Sprenger C, van Veen F (1984) On extreme values of Moran's I and Geary's c. Geogr Anal 16(1):17–24
Krasnov BR, Mouillot D, Khokhlova IS, Shenbrot G, Poulin R (2012) Compositional and phylogenetic dissimilarity of host communities drives dissimilarity of ectoparasite assemblages: geographical variation and scale-dependence. Parasitology 139:338–347
Kroonenberg PM (1983) Three-mode principal component analysis: theory and applications. DWO Press, Leiden
Kroonenberg PM (1989) The analysis of multiple tables in factorial ecology. III. Three-mode Principal Component Analysis: "Analyse triadique complète". Acta Oecol Oecol Generalis 10:245–256
Kroonenberg PM, Lombardo R (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivar Behav Res 34(3):367–396
Lafosse R, Hanafi M (1997) Concordance d'un tableau avec K tableaux: définition de K + 1 uples synthétiques. Revue de Statistique Appliquée 45:111–126
Lavit C (1988) Analyse conjointe de tableaux quantitatifs. Masson, Paris,
Lavit C, Escoufier Y, Sabatier R, Traissac P (1994) The ACT (STATIS method). Comput Stat Data Anal 18:97–119
Lebart L, Morineau A, Warwick K (1984) Multivariate descriptive statistical analysis. Wiley, New York
Lebreton JD, Chessel D, Prodon R, Yoccoz NG (1988a) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. Acta Oecol Oecol Generalis 9(1):53–67
Lebreton JD, Chessel D, Richardot-Coulet M, Yoccoz NG (1988b) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. Acta Oecol Oecol Generalis 9(2):137–151
Lebreton P, Lebrun P, Martinot J, Miquet A, Tournier H (1999) Approche écologique de l'avifaune de la vanoise. Travaux scientifiques du Parc national de la Vanoise 21:7–304
Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74(6):1659–1673
Legendre P, Anderson M (1999) Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. Ecol Monogr 69:1–24
Legendre P, Legendre L (1998) Numerical ecology. Elsevier, Amsterdam
Legendre P, Galzin R, Harmelin-Vivien ML (1997) Relating behavior to habitat: solutions to the fourth-corner problem. Ecology 78(2):547–562
LeRoux B, Rouanet H (2004) Geometric data analysis. Academic, Dordrecht
Martins EP, Diniz-Filho JAF, Housworth EA (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56(1):1–13
McIntosh RP (1978) Matrix and plexus techniques. In: Whittaker RH (ed) Ordination of plant communities. Springer, Netherlands, pp 151–184
Mendes S, Gómez JF, Pereira MJ, Azeiteiro UM, Galindo-Villardón MP (2010) The efficiency of the Partial Triadic Analysis method: an ecological application. Biometr Lett 47:83–106
Miller JK (1975) The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: a Monte Carlo study. Multivar Behav Res 10(2):233–244
Miller JK, Farr SD (1971) Bimultivariate redundancy: a comprehensive measure of interbattery relationship. Multivar Behav Res 6(3):313–324
Moran PAP (1948) The interpretation of statistical maps. J R Stat Soc Ser B-Methodol 10:243–251
Munoz F (2009) Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecol Model 220(20):2683–2689
Neuwirth E (2014) RColorBrewer: ColorBrewer palettes. https://CRAN.R-project.org/package=RColorBrewer, R package version 1.1-2
Nishisato S (1980) Analysis of categorical data: dual scaling and its applications. University of Toronto Press, Toronto

Obadia J (1978) L'analyse en composantes explicatives. Revue de Statistique Appliquée 26(4): 5–28

Oksanen J (1987) Problems of joint display of species and site scores in correspondence analysis. Vegetatio 72:51–57

Oksanen J, Minchin PR (2002) Continuum theory revisited: what shape are species responses along ecological gradients? Ecol Model 157:119–129

Ollier S, Couteron P, Chessel D (2005) Orthonormal transform to detect and characterize phylogenetic signal. Biometrics 62:471–477

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pavoine S, Bailly X (2007) New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. BMC Evol Biol 7:156

Pavoine S, Dolédec S (2005) The apportionment of quadratic entropy: a useful alternative for partitioning diversity in ecological data. Environ Ecol Stat 12:125–138

Pavoine S, Dufour AB, Chessel D (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. J Theor Biol 228:523–537

Pavoine S, Blondel J, Baguette M, Chessel D (2007) A new technique for ordering asymmetrical three-dimensional data sets in ecology. Ecology 88:512–523

Pavoine S, Ollier S, Pontier D, Chessel D (2008) Testing for phylogenetic signal in life history variable: Abouheif's test revisited. Theor Popul Biol 73:79–91

Pavoine S, Vallet J, Dufour AB, Gachet S, Daniel H (2009) On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos 118(3):391–402

Pavoine S, Vela E, Gachet S, de Bélair G, Bonsall MB (2011) Linking patterns in phylogeny, traits, abiotic variables and space: a novel approach to linking environmental filtering and plant community assembly. J Ecol 99(1):165–175

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572

Pebesma EJ, Bivand RS (2005) Classes and methods for spatial data in R. R News 5(2):9–13

Pegaz-Maucet D (1980) Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés benthiques d'un cours d'eau. Comparaison avec le benthos. PhD thesis, Thèse de troisième cycle, Université Lyon 1

Pélabon C, Gaillard JM, Loison A, Portier C (1995) Is sex-biased maternal care limited by total maternal expenditure in polygynous ungulates? Behav Ecol Sociobiol 37(5):311–319

Pélissier R, Couteron P, Dray S, Sabatier D (2003a) Consistency between ordination techniques and diversity measurements: two alternative strategies for species occurrence data. Ecology 84:242–251

Pélissier R, Couteron P, Dray S, Sabatier D (2003b) Consistency between ordination techniques and diversity measurements: two strategies for species occurrence data. Ecology 84(1):242–251

Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625

R Hackathon et al (2017) phylobase: base package for phylogenetic structures and comparative data. https://CRAN.R-project.org/package=phylobase, R package version 0.8.4

Rao CR (1964) The use and interpretation of Principal Component Analysis in applied research. Sankhya: Indian J Stat Ser A 26:329–358

Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21:24–43

Revell LJ, Collar DC (2009) Phylogenetic analysis of the evolutionary correlation using likelihood. Evolution 63(4):1090–1100

Rolland A, Bertrand F, Maumy M, Jacquet S (2009) Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. Water Res 43:3155–3168

Sarkar D (2008) Lattice: multivariate data visualization with R. Springer, New York

Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592–593

Schliep, Klaus, Potts, J A, Morrison, A D, Grimm, W G (2017) Intertwining phylogenetic trees and networks. Meth Ecol Evol 8(10):1212–1220

Schutten GJ, hong Chan C, Leeper TJ (2016) readODS: read and write ODS files. https://CRAN.R-project.org/package=readODS, R package version 1.6.4

Siberchicot A, Julien-Laferrière A, Dufour AB, Thioulouse J, Dray S (2017) adegraphics: an S4 lattice-based package for the representation of multivariate data. R J 9(2):198–212

Simier M, Blanc L, Pellegrin F, Nandris D (1999) Approche simultanée de k couples de tableaux: applications à l'étude des relations pathologie végétale-environnement. Revue de Statistique Appliquée 47:31–46

Slimani N, Guilbert E, El Ayni F, Jrad A, Boumaiza M, Thioulouse J (2017) The use of STATICO and COSTATIS, two exploratory three-ways analysis methods. An application to the ecology of aquatic Heteroptera in the Medjerda watershed (Tunisia). Environ Ecol Stat 24:269–295

Swan J (1970) An examination of some ordination problems by use of simulated vegetational data. Ecology 51:89–102

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

Tennekes M (2017) treemap: treemap visualization. https://CRAN.R-project.org/package=treemap, R package version 2.4-2

Thioulouse J (1989) Statistical analysis and graphical display of multivariate data on the Macintosh. Comput Appl Biosci 5(4):287–292

Thioulouse J (1990) Macmul and Graphmu: two Macintosh programs for the display and analysis of multivariate data. Comput Geosci 16(8):1235–1240

Thioulouse J (1996) Outils logiciels, méthodes statistiques et implications biologiques: une approche de la biométrie. Mémoire d'habilitation à diriger des recherches, Université Lyon 1

Thioulouse J (2011) Simultaneous analysis of a sequence of paired ecological tables: a comparison of several methods. Ann Appl Stat 5:2300–2325

Thioulouse J, Chessel D (1987) Les analyses multitableaux en écologie factorielle. I: de la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Oecol Oecol Generalis 8:463–480

Thioulouse J, Chessel D (1992) A method for reciprocal scaling of species tolerance and sample diversity. Ecology 73(2):670–680

Thioulouse J, Dray S (2007) Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. J Stat Softw 22(5):1–14

Thioulouse J, Chessel D, Dolédec S, Olivier J (1997) ADE-4: a multivariate analysis and graphical display software. Stat Comput 7(1):75–83

Thioulouse J, Simier M, Chessel D (2004) Simultaneous analysis of a sequence of paired ecological tables. Ecology 85:272–283

Tucker LR (1966) Some mathematical notes on three-mode factor analysis. Psychometrika 31:279–311

Turner R (2018) deldir: Delaunay triangulation and Dirichlet (Voronoi) tessellation. https://CRAN.R-project.org/package=deldir, R package version 0.1-15

Turroni F, Foroni E, Pizzetti P, Giubellini V, Ribbera A, Merusi P, Cagnasso P, Bizzarri B, de'Angelis GL, Shanahan F, Sinderen Dv, Ventura M (2009) Exploring the diversity of the bifidobacterial population in the human intestinal tract. Appl Environ Microbiol 75:1534–1545

Valet J, Daniel H, Beaujouan V, Rozé F, Pavoine S (2010) Using biological traits to assess how urbanization filters plant species of small woodlands. Appl Veg Sci 13:412–424

Verneaux J (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse de Doctorat d'Etat, Besançon

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. Geogr Anal 17(4):263–283

Wesuls D, Oldeland J, Dray S (2012) Disentangling plant trait responses to livestock grazing from spatio-temporal variation: the partial RLQ approach. J Veg Sci 23:98–113

Whittaker R (1956) Vegetation of the great smoky mountains. Ecol Monogr 26(1):1–80

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21:213–251

Wickham H, Hester J, Chang W (2018) devtools: tools to make developing R packages easier. https://CRAN.R-project.org/package=devtools, R package version 1.13.5

van den Wollenberg AL (1977) Redundancy analysis, an alternative for canonical analysis. Psychometrika 42(2):207–219

Index

J. Thioulouse et al., *Multivariate Analysis of Ecological Data with ade4*,
https://doi.org/10.1007/978-1-4939-8850-1