Supervised Multiblock Analysis in R with the ade4 Package

Stéphanie Bougeard, Stéphane Dray

Journal of Statistical Software

August 2018, Volume 86, Issue 1. doi: 10.18637/jss.v086.i01

Supervised Multiblock Analysis in R with the ade4 Package

Stéphanie Bougeard
French Agency for
Food, Occupational and Health Safety

Stéphane Dray
Université Lyon 1

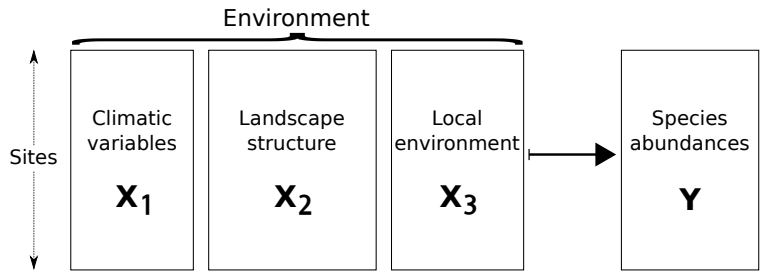
Abstract

This paper presents two novel statistical analyses of multiblock data using the R language. It is designed for data organized in $(K+1)$ blocks (i.e., tables) consisting of a block of response variables to be explained by a large number of explanatory variables which are divided into K meaningful blocks. All the variables – explanatory and dependent – are measured on the same individuals. Two multiblock methods both useful in practice are included, namely multiblock partial least squares regression and multiblock principal component analysis with instrumental variables. The proposed new methods are included within the **ade4** package widely used thanks to its great variety of multivariate methods. These methods are available on the one hand for statisticians and on the other hand for users from various fields in the sense that all the values derived from the multiblock processing are available. Some relevant interpretation tools are also developed. Finally the main results are summarized using overall graphical displays. This paper is organized following the different steps of a standard multiblock process, each corresponding to specific R functions. All these steps are illustrated by the analysis of real epidemiological datasets.

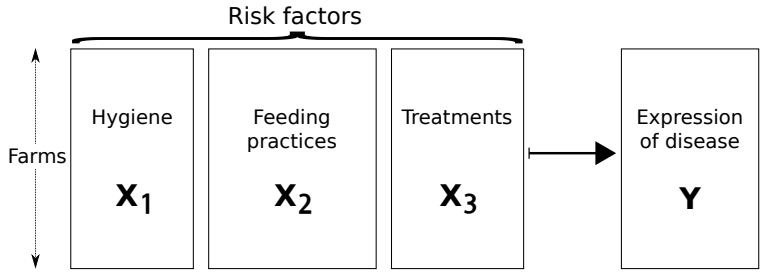
Keywords: multivariate analysis, multiblock partial least squares regression, multiblock principal component analysis with instrumental variables, R, **ade4**.

1. Introduction

Multivariate methods are widely used to summarize data tables where the same experimental units (individuals) are described by several variables. In some circumstances, variables are divided in different sets leading to multiblock (or multitable) data stored in K different tables $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$. In this case, each data table provides a typology of individuals based on different descriptors and several statistical techniques have been developed and implemented in



(a) Ecological multiblock data.



(b) Veterinary epidemiological multiblock data.

Figure 1: Examples of $(K+1)$ multiblock data.

R (R Core Team 2018) to identify the similarities and discrepancies between these typologies (e.g., Dray, Dufour, and Chessel 2007; Lê, Josse, and Husson 2008). A more complex situation is encountered when $\mathbf{Y}$ is a block of several variables to be explained by a large number of explanatory variables organized in K blocks $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$. This $(K+1)$ multiblock data are found in various fields including process monitoring (e.g., Kourti 2003), chemometrics (Kohonen, Reinikainen, Aaljoki, Perkio, Vaananen, and Hoskuldsson 2008), sensometrics (Måge, Menichelli, and Naes 2012), social sciences, market studies, ecology (Hanafi and Lafosse 2001) or epidemiology (Bougeard, Lupo, le Bouquin, Chauvin, and Qannari 2012). Some examples of multiblock data are given in Figure 1. In community ecology, the environmental filtering hypothesis suggests that species abundances would be filtered hierarchically, first by large-scale environmental factors (e.g., climate), and subsequently by landscape structure and fine-scale environmental factors. In veterinary epidemiology, the expression of an animal disease could be related to different factors including feeding practices, hygiene, husbandry practices or treatments.

Multiblock methods preserve the original structure of the data and thus allows one to analyze the $(K+1)$ tables simultaneously. They can be used to select explanatory variables in the datasets $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$, generally numerous and quasi-collinear, that are strongly related with the dependent variables in $\mathbf{Y}$. After this selection step, multiblock methods can also be used to identify the complex links between explanatory and dependent tables both at the variable and block levels. Although methods designed for the analysis of $(K+1)$ tables have been available for a few years, with a straightforward and single eigensolution, there are few published applications. The main reason for this lack of interest is probably the poor availability of free statistical software implementing these methods. Multiblock partial least squares regression is only available in the free **Multi-Block Toolbox** (Van den Berg 2004) of the commercial

software package MATLAB (The MathWorks, Inc. 2015). However, this method is designed for more complex multiblock data (K explanatory datasets to explain K' dependent ones) and there is no proof of the convergence of the associated iterative algorithm. In R (R Core Team 2018), no methods are implemented to deal with $(K+1)$ tables and two unsatisfactory solutions can be envisaged. A first alternative is to ignore the multiblock structure of the explanatory variables so that a two-table technique can be applied. For instance, partial least squares regression (**pls** package; Mevik and Wehrens 2007) or redundancy analysis (`pcaiv` function in **ade4**; Dray and Dufour 2007) can be used to study the link between the merged explanatory dataset $\mathbf{X}$ and the dependent table $\mathbf{Y}$. On the other hand, the user may also use methods developed for more complex data structures, such as partial least squares path modeling (**plspm** package; Sanchez, Trinchera, and Russolillo 2017). However, this method is not specifically designed for a $(K+1)$ structure and the iterative algorithm convergence is only practically encountered but no formal proofs have been provided (Henseler 2010).

We implemented new statistical and graphical functionalities to analyze multiblock $(K+1)$ data in the **ade4** package for R. This package provided classes, methods and functions to handle and analyze multivariate datasets organized in one (Dray and Dufour 2007), two or K tables (Dray *et al.* 2007). We propose additional tools for the statistical analysis of $(K+1)$ datasets with explanatory and modeling methods in **ade4**. We implemented two multiblock methods that are based on the optimization of a criterion with a direct eigensolution. The first method is multiblock partial least squares regression (MBPLS) applied to the particular case of a single response dataset $\mathbf{Y}$ (Wold 1984). The second one is multiblock principal component analysis with instrumental variables (MBPCAIV) also called multiblock redundancy analysis (Bougeard, Qannari, and Rose 2011). We detail preliminary data manipulation, present the two selected multiblock methods and give advice to select the most relevant. We illustrate the main advantages provided by the methods and the overall descriptive graphical displays. Multiblock methods are also devoted to modeling purpose and we thus propose a cross-validation procedure to select the optimal model dimension and diagnostic plots to describe its quality. Lastly, we detail the optimal model interpretation at the variable and at the block levels. The whole procedure is illustrated by the analysis of a real epidemiological dataset.

2. Data manipulation

We consider a response dataset $\mathbf{Y}$ with M variables and K explanatory datasets $\mathbf{X}_k$ with P_k variables $(k = 1, \ldots, K)$. The merged dataset $\mathbf{X}$ is defined as $\mathbf{X} = [\mathbf{X}_1 | \ldots | \mathbf{X}_K]$ and contains $P = \sum_{k=1}^{K} P_k$ explanatory variables. All these variables are measured on the same N individuals and are centered.

Note that we added some restrictions concerning the number of individuals ($N \geq 6$; defined from our experience in multiblock analyses), the explanatory variables ($P_k \geq 2$) for $k = (1, \ldots, K)$. The dependent block $\mathbf{Y}$ may contain a single variable ($M \geq 1$). No missing values are allowed.

The **ade4** package provides the class '`ktab`' that should be used to store the multiblock explanatory datasets ($\mathbf{X}_k$). Variables from the same block must be contiguous. Different procedures can be used to create a '`ktab`' object. In the following example, we illustrate the use of the `ktab.data.frame` function. The data $[\mathbf{Y}|\mathbf{X}_1|\mathbf{X}_2]$ are stored in the `douds` object available in **ade4**. The dataset $\mathbf{Y}$ contains 27 variables, the first explanatory table called `"River"` contains 4 variables and the second one called `"Chemicals"` has 7 variables.

```
R> library("ade4")
R> data("doubs", package = "ade4")
R> Y <- doubs$fish
R> X <- doubs$env
R> blo <- c(4, 7)
R> tab.names <- c("River", "Chemicals")
R> ktabX <- ktab.data.frame(df = X, blocks = blo, tabnames = tab.names)
```

We implemented new functions to manipulate easily multiblock data. For instance, `ktabX[1, 1:5, 1:3]` can be used to select data for the first five individuals and the first three variables of the first table. The dependent dataset $\mathbf{Y}$ should be analyzed by a one-table method providing an object of class '`dudi`'. For instance, `dudi.pca` can be used to apply principal component analysis.

```
R> dudiY <- dudi.pca(Y, center = TRUE, scale = TRUE, scannf = FALSE)
```

Note that the transformation selected for the dependent variables in the call of the `dudi.pca` function (centering and scaling in this example) will also be applied in the subsequent multiblock analysis.

In the rest of the paper, we consider a real dataset (`chickenk` in **ade4**) to illustrate the use of multiblock methods. This example concerns the overall risk factors for losses in broiler chickens ($\mathbf{Y}$) described by ($M = 4$) variables (the first-week mortality rate, the mortality rate during the rest of the rearing, the mortality rate during the transport to the slaughterhouse and the condemnation rate at slaughterhouse). The ($P = 20$) explanatory variables are organized in ($K = 4$) thematic blocks related to the successive production stages of broiler chickens: the farm structure ($\mathbf{X}_1$, 5 variables, `"FarmStructure"`), the flock characteristics at placement ($\mathbf{X}_2$, 4 variables, `"OnFarmHistory"`), the flock characteristics during the rearing period ($\mathbf{X}_3$, 6 variables, `"FlockCharacteristics"`) and the transport-lairage conditions, slaughterhouse and inspection features ($\mathbf{X}_4$, 5 variables, `"CatchingTranspSlaught"`). All these variables are measured on ($N = 351$) broiler chicken flocks. See `?chickenk` for further details and Lupo *et al.* (2009) for a complete description.

```
R> data("chickenk", package = "ade4")
R> losses <- chickenk[[1]]
R> dudiY <- dudi.pca(losses, center = TRUE, scale = TRUE, scannf = FALSE)
R> ktabX <- ktab.list.df(chickenk[2:5])
```

Several questions can be associated to this epidemiological dataset:

1. Are there some relationships between losses in broiler chickens $\mathbf{Y} = (y_1, \ldots, y_Q)$ and variables measured at the different production stages $\mathbf{X} = (x_1, \ldots, x_P)$?

2. Do the chicken flocks (N) have the same features in terms of their production stage description ($\mathbf{X}$) in relation with losses ($\mathbf{Y}$)?

3. Are there significant links between all the variables describing the production stages $\mathbf{X} = (x_1, \ldots, x_P)$ and each type of losses $\mathbf{Y} = (y_1, \ldots, y_Q)$?

4. Is it possible to sort the effects of all the variables describing the production stages $\mathbf{X} = (x_1, \ldots, x_P)$ in relation with the overall losses ($\mathbf{Y}$)?

5. Is it possible to sort the effects of the various production stages $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ in relation with the overall losses ($\mathbf{Y}$)?

Multiblock methods help to answer these different questions. The first two questions relate to a descriptive aim and will be handled in Section 4. The three last questions pertain to the modeling framework and will be treated in Section 6.

3. Multiblock methods

We implemented multiblock partial least squares regression (MBPLS) applied to the particular case of a single response dataset $\mathbf{Y}$ (Wold 1984) and multiblock principal component analysis with instrumental variables (MBPCAIV; Bougeard *et al.* 2011). In comparison with MBPCAIV, the method MBPLS is less sensitive to multicollinearity within explanatory blocks. For the case where only one block of variables is explained, Westerhuis, Kourti, and MacGregor (1998) among others, showed that the solution obtained from the iterative algorithm, originally devoted to K' datasets to be explained with K other ones, is equivalent to the solution obtained from a standard PLS regression of $\mathbf{Y}$ and the merged dataset $\mathbf{X}$. The method MBPCAIV considers the multiblock structure of data and leads to a model with a better fitting ability than MBPLS but it is unstable when explanatory blocks contain quasi-collinear variables. The user has to select the most relevant method by making a trade-off between stable and predictive model according to the data structure and the aims of the study.

Both MBPCAIV and MBPLS methods can be considered as the analysis of $(K+1)$ triplets, a dependent one $(\mathbf{Y}, \mathbf{Q}_\mathbf{Y}, \mathbf{D})$ and K explanatory ones $(\mathbf{X}_k, \mathbf{Q}_{\mathbf{X}_k}, \mathbf{D})$ for $k = (1, \ldots, K)$. One can notice that the following algebraic presentation of multiblock methods as the analysis of $(K+1)$ triplets is consistent with the one of all the methods developed in **ade4** for one, two and K tables (Dray and Dufour 2007; Dray *et al.* 2007). The simultaneous analysis of these triplets is provided by the crossing products of $\mathbf{X}_k$ and $\mathbf{Y}$, i.e., $\mathbf{Y}^\top\mathbf{D}\mathbf{X}_k$, with the metric $\mathbf{D} = \frac{1}{N}\mathbf{I}_N$ where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Multiblock methods seek a smaller dimension space to represent the main relationships between variables and individuals. They are based on the analysis of the K triplets $(\mathbf{Y}^\top\mathbf{D}\mathbf{X}_k, \mathbf{Q}_\mathbf{Y}, \mathbf{Q}_{\mathbf{X}_k})$ where $\mathbf{Q}_\mathbf{Y}$ is usually equal to $\mathbf{I}_M$ and $\mathbf{Q}_{\mathbf{X}_k} = \mathbf{I}_{P_k}$ for MBPLS or $\mathbf{Q}_{\mathbf{X}_k} = (\mathbf{X}_k^\top\mathbf{D}\mathbf{X}_k)^-$ for MBPCAIV (where $^-$ stands for the generalized inverse). Since the generalized inverse leads to the non-uniqueness of the eigenvalue decomposition, results for MBPCAIV may slightly differ when replicated in case of high multicollinearity. The solution is given by the diagonalization of $\sum_k(\mathbf{Y}^\top\mathbf{D}\mathbf{X}_k)\mathbf{Q}_{\mathbf{X}_k}(\mathbf{Y}^\top\mathbf{D}\mathbf{X}_k)^\top\mathbf{Q}_\mathbf{Y}$. For the first dimension, the eigenvector $\mathbf{v}^{(1)}$ associated to first the eigenvalue $\lambda^{(1)}$ maximizes the quantity $\mathbf{v}^{(1)^\top}\left(\sum_k \mathbf{Q}_\mathbf{Y}^\top\mathbf{Y}^\top\mathbf{D}\mathbf{X}_k\mathbf{Q}_{\mathbf{X}_k}\mathbf{X}_k^\top\mathbf{D}\mathbf{Y}\mathbf{Q}_\mathbf{Y}\right)\mathbf{v}^{(1)} = \lambda^{(1)}$ with $\|\mathbf{v}^{(1)}\|_{\mathbf{Q}_\mathbf{Y}} = 1$.

For MBPCAIV, this quantity can be rewritten as $\sum_k \mathsf{VAR}(\mathbf{P}_{\mathbf{X}_k}\mathbf{D}\mathbf{Y}\mathbf{v}^{(1)})$ where $\mathbf{P}_{\mathbf{X}_k}$ is the projector onto $\mathbf{X}_k$. The constraint $\|\mathbf{t}_k^{(1)}\|_\mathbf{D} = 1$ is added and the latent variables, derived from the eigensolution, are given by $\mathbf{t}_k^{(1)} = \mathbf{X}_k\mathbf{w}_k^{(1)} = \mathbf{P}_{\mathbf{X}_k}\mathbf{u}^{(1)}/\|\mathbf{P}_{\mathbf{X}_k}\mathbf{u}^{(1)}\|_\mathbf{D}$ with $\mathbf{u}^{(1)} = \mathbf{Y}\mathbf{v}^{(1)}$ and $\mathbf{t}^{(1)} = \mathbf{X}\mathbf{w}^{(1)} = \sum_k \mathbf{P}_{\mathbf{X}_k}\mathbf{u}^{(1)}/\sqrt{\sum_k \|\mathbf{P}_{\mathbf{X}_k}\mathbf{u}^{(1)}\|^2_\mathbf{D}}$. For MBPLS, the quantity maximized is equal to $\sum_k \mathsf{VAR}(\mathbf{X}_k^\top\mathbf{D}\mathbf{Y}\mathbf{v}^{(1)}) = \mathsf{VAR}(\mathbf{X}^\top\mathbf{D}\mathbf{Y}\mathbf{v}^{(1)})$. In this case, the constraints $\|\mathbf{w}_k^{(1)}\|_{\mathbf{I}_{P_k}} = 1$ and $\|\mathbf{w}^{(1)}\|_{\mathbf{I}_P} = 1$ are added so that the first order solution is also given by $\lambda^{(1)} = \mathsf{VAR}(\mathbf{Y}^\top\mathbf{D}\mathbf{X}\mathbf{w}^{(1)}) = \mathsf{VAR}(\mathbf{Y}^\top\mathbf{D}\mathbf{t}^{(1)})$.

To obtain the second order solution, for MBPCAIV and MBPLS, the variables in the datasets $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ are deflated by a regression onto the first global component $\mathbf{t}^{(1)}$. The same maximization is then performed but the original datasets are replaced by the residuals obtained in the deflation step. Subsequent components are obtained by reiterating this process. The reader can consult Bougeard *et al.* (2011) for more details.

With **ade4**, MBPCAIV is obtained by:

```
R> res <- mbpcaiv(dudiY, ktabX, scale = TRUE, option = "uniform",
+    scannf = FALSE, nf = 5)
```

The MBPLS is performed by replacing the call to `mbpcaiv` by a call to the `mbpls` function. The first two arguments refer to the dependent and the explanatory datasets. Variable scaling of the explanatory dataset can be set by the `scale` argument (default is `TRUE`). The scaling of the dependent dataset has been previously defined in the first call to `dudi.pca`. The argument `option` defines the block weighting. The `"none"` option corresponds to no block weighting, `"uniform"` corresponds to the case where the merged explanatory dataset $\mathbf{X}$ (resp. $\mathbf{Y}$) has a total variance equal to one and each of the K explanatory blocks to $1/K$ (Westerhuis and Coenegracht 1997). The argument `scannf` allows to display the scree plot of eigenvalues to help with the choice of the number of the latent variables to be interpreted (default is `TRUE`). The optional `nf` argument indicates the number of selected dimensions (defaults is 2). The object `res` contains the different outputs of the analysis:

```
R> res

Multiblock principal component analysis with instrumental variables
list of class multiblock
list of class mbpcaiv

$eig: 20 eigen values
44.14 26.33 23.76 19.67 5.364 ...

$call: mbpcaiv(dudiY = dudiY, ktabX = ktabX, scale = TRUE,
    option = "uniform", scannf = FALSE, nf = 5)

$nf: 5 axis saved

   data.frame nrow ncol content
1  $lX        351  20   global components of the explanatory tables
2  $lY        351  5    components of the dependent data table
3  $Tli       1404 5    partial components
4  $Yco       4    5    inertia axes onto co-inertia axis
5  $faX       20   5    loadings to build the global components
6  $bip       4    5    block importances
7  $bipc      4    5    cumulated block importances
8  $vip       20   5    variable importances
9  $vipc      20   5    cumulated variable importances
10 $cov2      4    5    squared covariance between components
other elements: tabX tabY lw X.cw blo rank TL TC Yc1 Tfa Tl1 XYcoef intercept
```

4. Descriptive interpretation tools

Multiblock analyses are descriptive methods that provide an overview of the relationships between variables, blocks and individuals. They can be used to answer questions 1 and 2 (Section 2). The global latent variables $\mathbf{t}$, linear combinations of the explanatory variables, orthogonal by construction, provide the overall graphical displays. The explanatory variables are then depicted by their loadings $\mathbf{w}^*$ where the component is given by $\mathbf{t}^{(h)} = \mathbf{X}\mathbf{w}^{*(h)}$ and $\mathbf{w}^{*(h)} = \Pi_{l=1}^{h-1}[\mathbf{I} - \mathbf{w}^{(l)}({\mathbf{t}^{(l)}}^\top \mathbf{t}^{(l)})^{-1}{\mathbf{t}^{(l)}}^\top \mathbf{X}]\mathbf{w}^{(h)}$ for a given dimension h, as a consequence of the deflation procedure (Wold, Martens, and Wold 1983). The dependent variables are represented by their regression coefficients onto these latent variables ($\mathbf{c}^{(h)} = ({\mathbf{t}^{(h)}}^\top \mathbf{t}^{(h)})^{-1}\mathbf{Y}^\top \mathbf{t}^{(h)}$).

The decomposition of inertia into successive dimensions indicates the quantity of information extracted by the global latent variables $\mathbf{t}$ (element `lX` in `res`). The rank of the analysis is given by the element `rank`. A comprehensive summary of the first dimensions is provided by the `summary` function. For each dimension, the eigenvalues, inertia percentage and cumulated inertia percentage are given. The percentage and the cumulated percentage of the inertia of each dataset, $\mathbf{X}$, $\mathbf{Y}$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$, explained by the global latent variables are also provided (`VarY` and `VarYcum`) for the dependent dataset:

```
R> summary(res)

Multiblock principal component analysis with instrumental variables

Class: multiblock mbpcaiv
Call: mbpcaiv(dudiY = dudiY, ktabX = ktabX, scale = TRUE,
    option = "uniform", scannf = FALSE, nf = 5)

Total inertia: 125.8

Eigenvalues:
    Ax1     Ax2     Ax3     Ax4     Ax5
 44.144  26.332  23.758  19.672   5.364

Projected inertia (%):
    Ax1     Ax2     Ax3     Ax4     Ax5
 35.103  20.939  18.893  15.643   4.265

Cumulative projected inertia (%):
    Ax1   Ax1:2   Ax1:3   Ax1:4   Ax1:5
  35.10   56.04   74.93   90.58   94.84

(Only 5 dimensions (out of 20) are shown)

Inertia explained by the global latent, i.e. res$lX (in %):

dudiY$tab and ktabX:
     varY varYcum varX varXcum
Ax1 41.17    41.2 6.94    6.94
```

```
Ax2 21.49    62.7 7.31   14.25
Ax3 19.15    81.8 5.95   20.20
Ax4 11.07    92.9 5.25   25.45
Ax5  3.02    95.9 5.64   31.08

FarmStructure:
    varXk varXkcum
Ax1  4.68     4.68
Ax2  5.02     9.70
Ax3  2.45    12.15
Ax4  2.67    14.82
Ax5  5.10    19.92

OnFarmHistory:
    varXk varXkcum
Ax1  6.81     6.81
Ax2  9.10    15.91
Ax3  4.49    20.40
Ax4  9.49    29.89
Ax5  4.53    34.42

FlockCharacteristics:
    varXk varXkcum
Ax1 12.38     12.4
Ax2  4.13     16.5
Ax3  8.35     24.9
Ax4  5.22     30.1
Ax5  8.35     38.4

CatchingTranspSlaught:
    varXk varXkcum
Ax1  3.88     3.88
Ax2 11.01    14.88
Ax3  8.51    23.39
Ax4  3.61    26.99
Ax5  4.57    31.56
```

Graphical tools to represent the outputs of the analysis are provided in the new **adegraphics** package (Dray and Siberchicot 2018). The main results are provided by the `plot` method of the '`multiblock`' class. By default, the first two global latent variables **t** (element `lX`) are used but higher order representations can be selected using the arguments `xax` (defaults to 1) and `yax` (defaults to 2).

```
R> library("adegraphics")
R> plotmbpcaiv <- plot(res, xax = 1, yax = 2)
```

Results are illustrated in Figure 2. The first plot (top right corner) depicts the similarities between individuals (i.e., the 351 chicken flocks). The scree plot of eigenvalues is represented

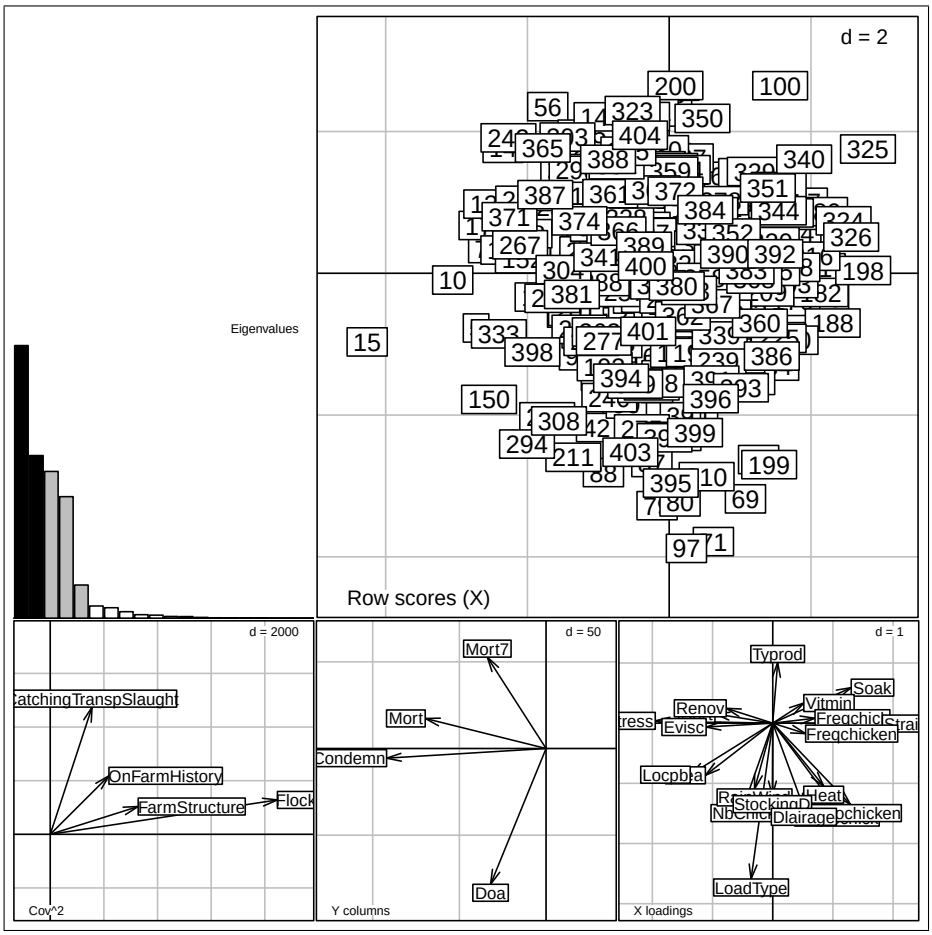


Figure 2: Results of multiblock principal component analysis of the `chickenk` dataset for the first two global latent variables.

in the top left corner. Relationships between blocks are depicted in the third plot (bottom left corner) by the squared covariance between the partial latent variables $\mathbf{t}_k$ (`Tl1`) and $\mathbf{u}$ (`lY`). The fourth plot (bottom middle) depicts the dependent variables by their projection on the latent variables $\mathbf{c}$ (`Yco`). The fifth plot (bottom right corner) depicts simultaneously the 20 explanatory variable loadings $\mathbf{w}^*$ (`faX`). These two last plots are usually interpreted together to identify the relationships between explanatory and dependent variables. Graphical functionalities of **adegraphics** are based on the package **lattice** (Sarkar 2008, 2017). Classes are provided to store simple and multiple graphics as objects. It is thus possible to extract and modify easily a single plot from these multiple graphical outputs. For instance, we updated some aesthetic properties of the plot of the individuals and added colors corresponding to different values of stress during the rearing period.

```
R> class(plotmbpcaiv)
```

```
[1] "ADEgS"
attr(,"package")
[1] "adegraphics"
```

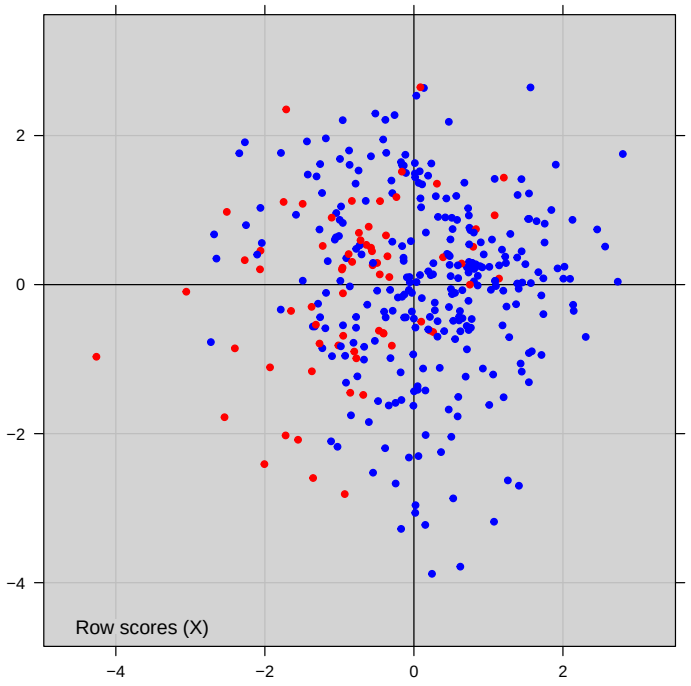


Figure 3: Updated graphical representation of the 351 chicken flocks colored according to stress during the rearing period. Red symbols correspond to stress occurrences (feeding system defection, electrical defect, etc.), blue symbols to the absence of stress.

```
R> mycol <- ifelse(ktabX$FlockCharacteristics$Stress == 0, "blue", "red")
R> update(plotmbpcaiv[[1]], plabel.cex = 0, ppoints.col = mycol,
+    paxes.draw = TRUE, pbackground.col = "lightgrey")
```

5. Selection of the optimal number of latent variables

Multiblock methods can also be used for predictive purposes and the first step is to select the optimal model dimension. The global latent variables can be expressed as linear combinations of $\mathbf{X}$, i.e., $\mathbf{t}^{(h)} = \mathbf{X}\mathbf{w}^{*(h)}$ and the dependent dataset $\mathbf{Y}$ can be split up into the same latent variables such as $\mathbf{Y} = \sum_{l=1}^{h} \mathbf{t}^{(l)}\mathbf{c}^{(l)^\top} + \mathbf{Y}^{(h)}$, $\mathbf{Y}^{(h)}$ being the residual matrix of the model based on h components. This leads to the final model $\mathbf{Y} = \mathbf{X}\sum_{l=1}^{h} \mathbf{w}^{*(l)}\mathbf{c}^{(l)^\top} + \mathbf{Y}^{(h)}$ for all the dimensions $h = (1, \ldots, H)$. The optimal model is obtained by selecting the number of latent variables with a two-fold cross-validation (Stone 1974). The dataset is split in a calibration and a validation sets, this procedure being repeated several times. Among all the models corresponding to the various values of h, the optimal model is retained, as a compromise between a good fitting ability (minimization of the root mean square error of calibration $RMSE_C^{(h)}$) and a good prediction ability (minimization of the root mean square error of validation $RMSE_V^{(h)}$).

We implemented new classes '`randxval`' and '`krandxval`' and methods (`print`, `plot`) to manage the outputs of cross-validation procedures. For multiblock methods, the two-fold cross-

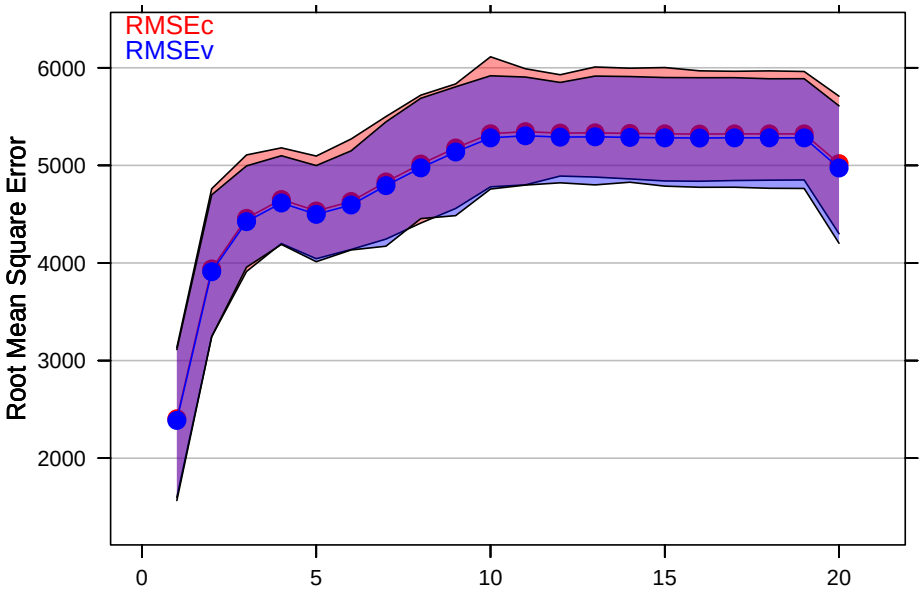


Figure 4: Fitting (in red) and prediction (in blue) abilities of the multiblock model as functions of the number of global latent variables introduced in the model, on the basis of a two-fold cross-validation procedure, for the `chickenk` dataset.

validation is provided by the `testdim` function which has three arguments: the multiblock object, the number of repetitions (`nrepet`) and the lower and upper quantiles to compute (defaults respectively to 0.25 and 0.75) to get confidence intervals. To get reliable results, a minimum number of 100 repetitions is imposed.

```
R> set.seed(123456)
R> testdim.chik <- testdim(res, nrepet = 100)
R> class(testdim.chik)

[1] "krandxval"

R> plot(testdim.chik)
```

Results are summarized in Figure 4 by means and associated confidence intervals for $RMSE_C^{(h)}$ and $RMSE_V^{(h)}$. This plot helps to select the optimal number of latent variables to be introduced in the model by making a trade-off between a stable and a predictive model. In this example, a model with four latent variables both optimizes (i.e., minimizes) fitting and prediction abilities.

6. Interpretation of the optimal model

Lastly, we provide tools to identify, in the optimal model, significant relationships between explanatory and dependent variables at the variable and at the block levels. Bootstrapping simulations are applied to the three main predictive parameters (regression coefficients,

cumulated variable importance index and cumulated block importance index) to provide confidence intervals, computed by the non-Studentized pivotal method (Carpenter and Bithell 2000). The regression coefficients of the optimal model measure the links between each explanatory and dependent variable. A coefficient is considered significant if the bootstrapped 95% confidence interval does not contain the threshold value 0. If the number of dependent variables in $\mathbf{Y}$ is large, the interpretation of these coefficients is difficult. In this case, it is more suitable to measure the contribution of each explanatory variable to the explanation of the whole dependent block $\mathbf{Y}$. The variable importance index (vip) is proposed, derived from the squared explanatory loadings $\mathbf{w}^{*(h)^2}$, weighted according to the associated block importance $a_k^{(h)^2}$ and expressed as a percentage for each dimension h. The cumulated variable importance index (vipc) sums these quantities over all the optimal components under study and weights them according to the amount of the relative importance of each dimension $\lambda^{(h)}$, also expressed as a percentage. Each explanatory variable is considered to be significantly associated with $\mathbf{Y}$ when the 95% confidence interval does not contain the threshold value $1/P$, P being the number of explanatory variables. Lastly, the block importance index (bip) is proposed to assess the contributions of the blocks $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ in the modeling process. It is computed from the coefficients $(a_1^{(h)^2}, \ldots, a_K^{(h)^2})$ which measure the link between $\mathbf{Y}$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$. If the optimal model contains several components, the cumulated block importance index (bipc) is based on the weighted average of the bip indexes, taking as weights the relative importance of each dimension $\lambda^{(h)}$. Both these quantities are expressed as percentages. A block is considered to be significantly associated with the dependent dataset if the 95% confidence interval does not contain the threshold value $1/K$, K being the number of blocks. All the details are given in Bougeard *et al.* (2011).

We implemented new classes '`randboot`' and '`krandboot`' and methods (`print`, `plot`) to manage the outputs of bootstrap simulations. For multiblock methods, the `randboot` method for '`multiblock`' objects can be used. It takes three arguments: the multiblock object, the number of repetitions (`nrepet`) and the optimal number of dimension of the model (`optdim`). By default, the number of repetitions is equal to 199 but a higher number is required to get more stable results. The `randboot` method for '`multiblock`' objects returns a list with 3 elements `XYcoef`, `bipc` and `vipc`. The first element is a list of '`randboot`' objects, the two others are object of the class '`randboot`'.

```
R> set.seed(123456)
R> boot.chik <- randboot(res, nrepet = 199, optdim = 4)
R> class(boot.chik)

[1] "list"

R> names(boot.chik)

[1] "XYcoef" "bipc"   "vipc"

R> class(boot.chik$vipc)

[1] "krandboot"
```

The `plot` function applied to '`randboot`' objects provides graphical summaries of bootstrapped values:

```
R> g1 <- plot(boot.chik$XYcoef$Mort7, main = "Mort7", plot = FALSE)
R> g2 <- plot(boot.chik$XYcoef$Mort, main = "Mort", plot = FALSE)
R> g3 <- plot(boot.chik$XYcoef$Doa, main = "Doa", plot = FALSE)
R> g4 <- plot(boot.chik$XYcoef$Condemn, main = "Condemn", plot = FALSE)
R> ADEgS(list(g1, g2, g3, g4))
R> g5 <- plot(boot.chik$vipc, main = "vipc", plot = FALSE)
R> g6 <- plot(boot.chik$bipc, main = "bipc", plot = FALSE)
R> ADEgS(list(g5, g6))
```

Results are presented in Figure 5. The first four plots illustrate the regression coefficient values and their confidence intervals for all the explanatory variables associated with each of the four dependent ones (and the optimal model involving four dimensions). The last two plots represent the `vipc` and `bipc` values associated with their confidence intervals.

From these results, it follows that each variable in $\mathbf{Y}$ is significantly related to a specific set of explanatory variables. Firstly, the first-week mortality rate (`"Mort7"`) is related to four variables, two of which pertain to the farm structure. The mortality rate during the rest of the rearing (`"Mort"`) is significantly linked with seven variables, four of which pertain to the flock characteristics during the rearing period. The mortality rate during the transport to the slaughterhouse (`"Doa"`) is associated with eight variables, among which four pertain to the catching, transport-lairage conditions, slaughterhouse and inspection features. Finally, the condemnation rate at slaughterhouse (`"Condemn"`) is related to fourteen variables, six of these variables refer to the flock characteristics during the rearing period. Some explanatory variables are specifically related to one variable in $\mathbf{Y}$, e.g., the chick homogeneity (from $\mathbf{X}_2$), whereas others are linked with up to three (out of four) variables in $\mathbf{Y}$, e.g., the genetic strain (from $\mathbf{X}_3$). Therefore, to sort these explanatory variables by a global order of priority thus highlighting their overall contribution to the explanation of the $\mathbf{Y}$ block, the `vipc` values can be used. It turns out that four explanatory variables have a significant impact on the overall losses: the stress occurrence during rearing (`"Stress"`), the type of loading system (`"LoadType"`), the stocking density in transport crates (`"StockingD"`) and the average duration of waiting time on lairage (`"Dlairage"`). Finally, the relative importance of the four production stages in the overall losses explanation highlights the significant importance of the flock characteristics during the rearing period ($\mathbf{X}_3$ block). The interested reader may refer to Bougeard *et al.* (2012) for a detailed interpretation of the results.

7. Conclusion and perspectives

We provide new tools to improve the handling and the statistical analysis of multiblock $(K + 1)$ datasets in the **ade4** package for R. These methods preserve the original structure of the data and provide an adapted framework that combines tools from factorial analysis and regression methods. Traditional graphical outputs are completed by cross-validation and bootstrap procedures to select and interpret the optimal model.

The framework of multiblock methods provide several tools and could be enriched by considering hierarchical-structured design of individuals frequently met in biological surveys (e.g.,

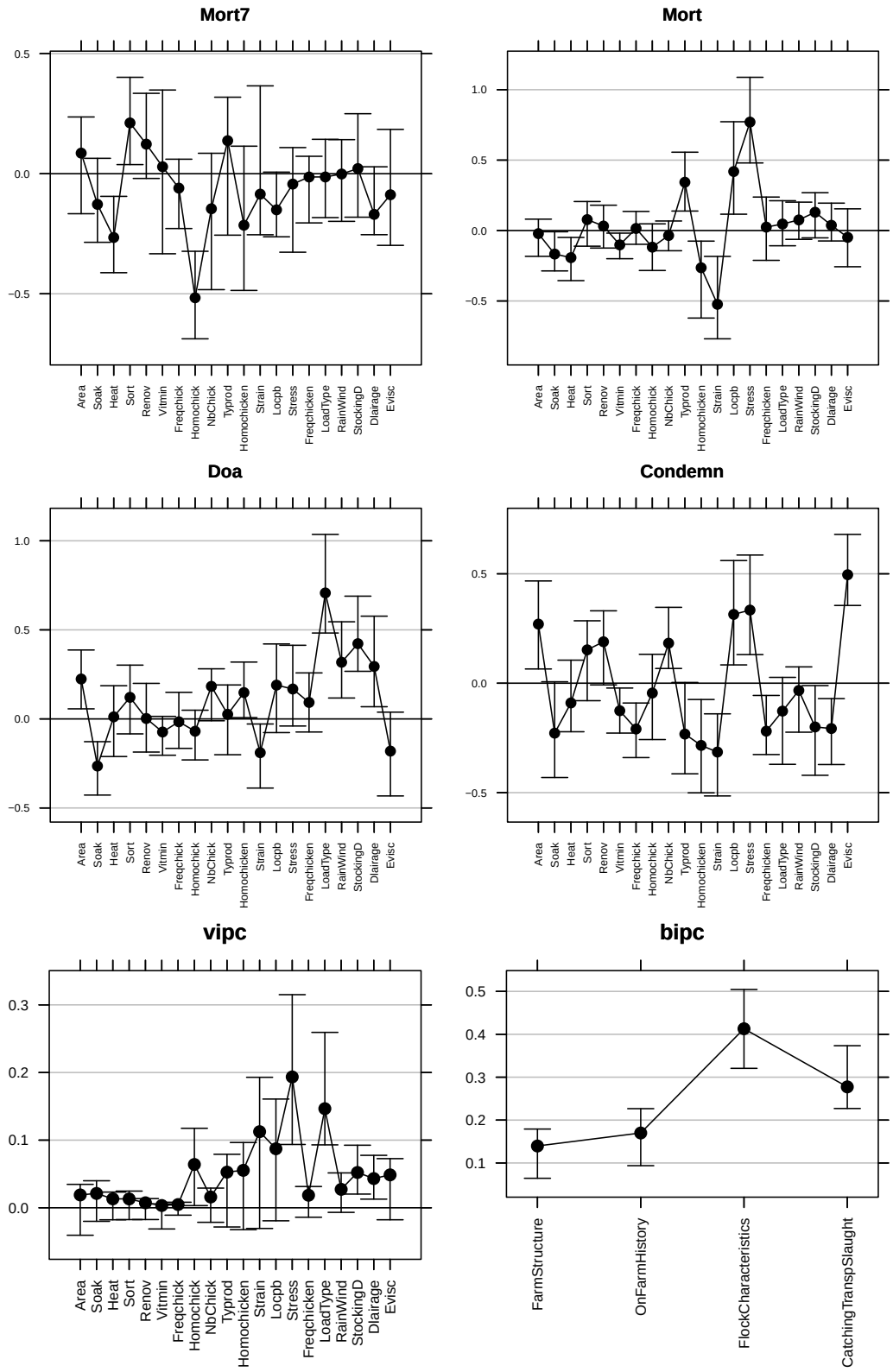


Figure 5: Multiblock predictive plots of the optimal model for the `chickenk` dataset.

individuals partitioned in groups corresponding to different treatments). Multiblock methods can also be directly adapted to the explanation of several dependent datasets $(\mathbf{Y}_1, \ldots, \mathbf{Y}_{K'})$ to handle more complex data structures. We hope that our implementation of multiblock methods and future developments will facilitate the use of these techniques and thus improve the statistical analysis of $(K + 1)$ datasets.

References

Bougeard S, Lupo C, le Bouquin S, Chauvin C, Qannari M (2012). "Multiblock Modelling to Assess the Overall Risk Factors for a Composite Outcome." *Epidemiology and Infection*, **140**(2), 337–347. doi:10.1017/s0950268811000537.

Bougeard S, Qannari M, Rose N (2011). "Multiblock Redundancy Analysis: Interpretation Tools and Application in Epidemiology." *Journal of Chemometrics*, **25**(9), 467–475. doi:10.1002/cem.1392.

Carpenter J, Bithell J (2000). "Bootstrap Confidenc Intervals: When, Which, What? A Practical Guide for Medical Statisticians." *Statistics in Medicine*, **19**(9), 1141–1164. doi:10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f.

Dray S, Dufour AB (2007). "The **ade4** Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, **22**(4), 1–20. doi:10.18637/jss.v022.i04.

Dray S, Dufour AB, Chessel D (2007). "The **ade4** Package — II: Two-Table and K-Table Methods." *R News*, **7**(2), 47–52. URL https://www.R-project.org/doc/Rnews/.

Dray S, Siberchicot A (2018). ***adegraphics**: An S4 **lattice**-Based Package for the Representation of Multivariate Data*. R package version 1.0-12, URL https://CRAN.R-project.org/package=adegraphics.

Hanafi M, Lafosse R (2001). "Généralisation de la régression simple pour analyser la dépendance de K ensembles de variables avec un $K + 1$ième." *Revue de Statistique Appliquée*, **49**, 5–30.

Henseler J (2010). "On the Convergence of the Partial Least Squares Path Modeling Algorithm." *Computational Statstics*, **25**(1), 107–120. doi:10.1007/s00180-009-0164-x.

Kohonen J, Reinikainen SP, Aaljoki K, Perkio A, Vaananen T, Hoskuldsson A (2008). "Multi-Block Methods in Multivariate Process Control." *Journal of Chemometrics*, **22**(3–4), 281–287. doi:10.1002/cem.1120.

Kourti T (2003). "Multivariate Dynamic Data Modeling for Analysis and Statistical Process Control of Batch Processes, Start-Ups and Grade Transitions." *Journal of Chemometrics*, **17**(1), 98–109. doi:10.1002/cem.778.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R Package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18. doi:10.18637/jss.v025.i01.

Lupo C, le Bouquin S, Balaine L, Michel V, Peraste J, Petetin I, Colin P, Chauvin C (2009). "Feasibility of Screening Broiler Chicken Flocks for Risk Markers as an Aid for Meat Inspection." *Epidemiology and Infection*, **137**(8), 1086–1098. doi:10.1017/s095026880900209x.

Måge I, Menichelli E, Naes T (2012). "Preference Mapping by PO-PLS: Separating Common and Unique Information in Several Data Blocks." *Food Quality and Preference*, **24**(1), 8–16. doi:10.1016/j.foodqual.2011.08.003.

Mevik BH, Wehrens R (2007). "The **pls** Package: Principal Component and Partial Least Squares Regression in R." *Journal of Statistical Software*, **18**(2), 1–24. doi:10.18637/jss.v018.i02.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sanchez G, Trinchera L, Russolillo G (2017). *Partial Least Squares Data Analysis Methods*. R package version 0.4-9, URL https://CRAN.R-project.org/package=plspm.

Sarkar D (2008). ***lattice**: Multivariate Data Visualization with R*. Springer-Verlag, New York. doi:10.1007/978-0-387-75969-2. URL http://lmdvr.R-Forge.R-project.org/.

Sarkar D (2017). ***lattice**: Trellis Graphics for R*. R package version 0.20-35, URL https://CRAN.R-project.org/package=lattice.

Stone M (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society B*, **36**(2), 111–147.

The MathWorks, Inc (2015). *MATLAB – The Language of Technical Computing, Version R2015b*. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Van den Berg F (2004). ***Multi-Block Toolbox** for MATLAB (V0.2)*. URL http://www.models.life.ku.dk/MBToolbox/.

Westerhuis JA, Coenegracht PMJ (1997). "Multivariate Modelling of the Pharmaceutical Two-Step Process of Wet Granulation and Tableting with Multiblock Partial Least Squares." *Journal of Chemometrics*, **11**, 379–392. doi:10.1002/(sici)1099-128x(199709/10)11:5<379::aid-cem482>3.3.co;2-\%23.

Westerhuis JA, Kourti T, MacGregor JF (1998). "Analysis of Multiblock and Hierarchical PCA and PLS Model." *Journal of Chemometrics*, **12**(5), 301–321. doi:10.1002/(sici)1099-128x(199809/10)12:5<301::aid-cem515>3.3.co;2-j.

Wold S (1984). "Three PLS Algorithms According to SW." In *Symposium MULDAST (Multivariate Analysis in Science and Technology)*. Umea University, Sweden.

Wold S, Martens H, Wold H (1983). "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method." In *Proceedings of the Conference on Matrix Pencils*. Springer-Verlag, Heidelberg.

Affiliation:

Stéphanie Bougeard
Department of Epidemiology
French Agency for Food, Occupational and Health Safety (Anses)
Zoopole, BP 53, 22440 Ploufragan, France
E-mail: stephanie.bougeard@anses.fr

Stéphane Dray
Laboratoire de Biométrie et Biologie Evolutive (UMR 5558)
CNRS - Université de Lyon 1
43, Boulevard du 11 Novembre 1918
69622 Villeurbanne Cedex, France
E-mail: stephane.dray@univ-lyon1.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2018, Volume 86, Issue 1 *Submitted:* 2015-10-12
doi:10.18637/jss.v086.i01 *Accepted:* 2017-08-24