In Neil Salkind (Ed.), *Encyclopedia of Research Design.*
Thousand Oaks, CA: Sage. 2010

Barycentric Discriminant Analysis (BADIA)

Hervé Abdi · Lynne J. Williams

1 Introduction

Barycentric discriminant analysis (BADIA) generalizes discriminant analysis and, like discriminant analysis, it is performed when measurements made on some observations are combined to assign these observations or "new" observations to *a-priori* defined categories. For example, BADIA can be used *1)* to assign subjects to a given diagnostic group (*i.e.,* Alzheimer's disease, other dementia, normal aging) on the basis on brain imaging data or psychological tests (here the *a-priori* categories are the clinical groups), *2)* to assign wines to a region of production on the basis of several physical and chemical measurements (here the *a-priori* categories are the regions of production), *3)* to use brain scans taken on a given participant to determine what type of object (*e.g.,* a face, a cat, a chair) was watched by the participant when the scans were taken (here the *a-priori* categories are the types of object), *(4)* to use DNA measurements to predict if a person is at risk for a given health problem (here the *a-priori* categories are the types of health problem).

Hervé Abdi
The University of Texas at Dallas

Lynne J. Williams
The University of Toronto Scarborough

Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

BADIA is more general than standard discriminant analysis because it can be used in cases for which discriminant analysis cannot be used. This is the case, for example, when there are more variables than observations (a case often called the "$N << P$ problem") or when the measurements are qualitative.

BADIA is a class of methods which all rely on the same principle: Each category of interest is represented by the *barycenter* of its observations (*i.e.,* the weighted average; the barycenter is also called the *center of gravity* of the observations of a given category), and a generalized principal component analysis (GPCA) is performed on the category by variable matrix. This analysis gives a set of discriminant factor scores for the categories and another set of factor scores for the variables. The original observations are then projected onto the category factor space, providing a set of factor scores for the observations. The distance of each observation to the set of categories is computed from the factor scores and each observation is assigned to the closest category. The comparison between the *a-priori* and *a-posteriori* category assignments is used to assess the quality of the discriminant procedure. The prediction for the observations which were used to compute the barycenters is called the *fixed effect* prediction. Fixed effect performance is evaluated by counting the number of correct and incorrect assignments and storing these numbers in a confusion matrix. Another index of the performance of the fixed effect model—equivalent to a squared coefficient of correlation—is the ratio

$$R^2 = \frac{\text{category variance}}{\text{category variance} + \text{variance of the observations within category}} .$$

This coefficient is denoted R^2 and is interpreted as the proportion of variance of the observations explained by the categories or as the proportion of the variance explained by the discriminant model. The performance of the fixed effect model can also be represented graphically as a *tolerance* ellipsoid that encompasses a given proportion (say 95%) of the observations. The overlap between the tolerance ellipsoids of two categories is proportional to the number of misclassifications between these two categories.

New observations can also be projected onto the discriminant factor space and they can be assigned to the closest category. When the actual assignment of these observations is not known, the model can be used to *predict* category membership. The model is then called a *random* model (as opposed to the fixed model). An obvious problem, then, is to evaluate the quality of the prediction for new observations. Ideally, the performance of the random effect model is evaluated by counting the number of correct and incorrect classifications for new observations and computing a confusion matrix on these *new* observations. However, it is not always practical or even feasible to obtain new observations and therefore the random effect performance is, in general, evaluated using computational cross-validation techniques such as the *jackknife* (see Abdi & Williams, this volume) or the *bootstrap*. For

example, a jackknife approach (also called "leave one out") can be used by which each observation is taken out of the set, in turn, and predicted from the model built on the other observations. The predicted observations are then projected in the space of the fixed effect discriminant scores. This can also be represented graphically as a *prediction* ellipsoid. A prediction ellipsoid encompasses a given proportion (say 95%) of the *new* observations. The overlap between the prediction ellipsoids of two categories is proportional to the number of misclassifications of *new* observations between these two categories.

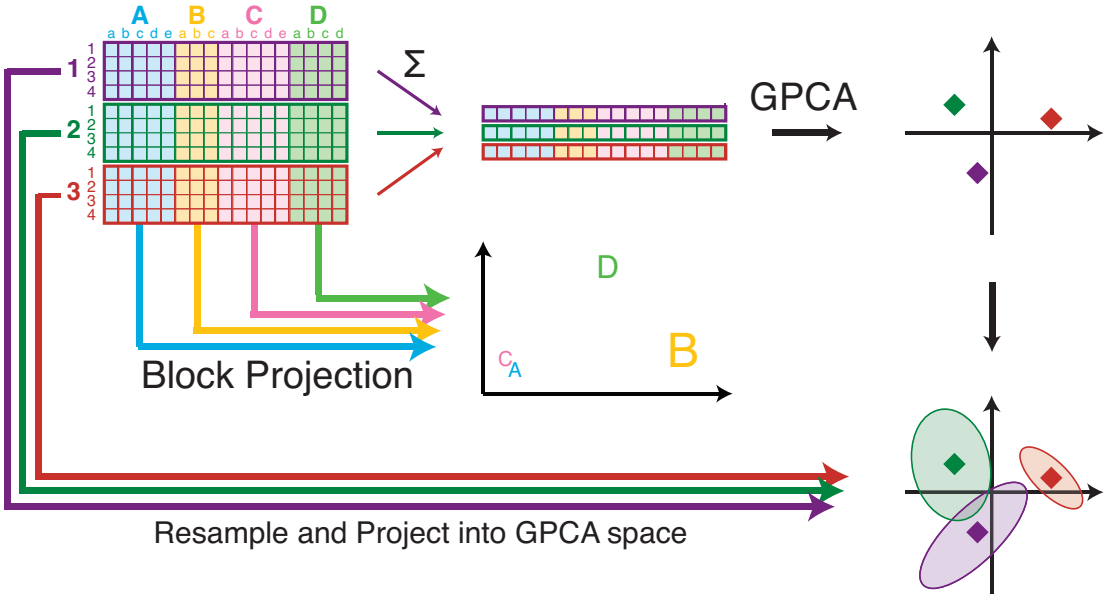
The stability of the discriminant model can be assessed by a cross-validation model such as the bootstrap (see Efron & Tibshirani, 1993). In this procedure, multiple sets of observations are generated by sampling with replacement from the original set of observations and the category barycenters are computed from each of these sets. These barycenters are then projected onto the discriminant factor scores. The variability of the barycenters can be represented graphically as a *confidence* ellipsoid that encompasses a given proportion (say 95%) of the barycenters. When the confidence intervals of two categories do not overlap, this indicates that these two categories are *significantly* different.

In summary, BADIA is a GPCA performed on the category barycenters. Recall that GPCA encompasses different techniques such as, for example, correspondence analysis, biplot, Hellinger distance analysis, discriminant analysis, and canonical variate analysis (see Abdi, 2007; Gittins, 1980; Greenacre, 1984). For each specific type of GPCA, we have a corresponding version of BADIA. For example, when the GPCA used is correspondence analysis, this gives the most well-known version of BADIA: discriminant correspondence analysis (DCA, sometimes also called correspondence discriminant analysis; see Abdi, 2007; Celeux & Nakache, 1994; Leclerc, 1976; Perrière, Lobry, & Thioulouse, 1996; Perrière & Thioulouse, 2003; Saporta & Niang, 2006). Because BADIA is based on GPCA, it can also analyze data tables obtained by the concatenation of blocks (*i.e.,* sub-tables). In this case, the importance (often called the *contribution*) of each block to the overall discrimination can also be evaluated and represented as a graph. A sketch of BADIA is given in Figure 1.

2 Notations

The original data matrix is an I observations by J variables matrix denoted $\mathbf{X}$. Prior to the analysis, the matrix $\mathbf{X}$ can be pre-processed by centering (*i.e.,* subtracting the column mean from each column), by transforming each column into a Z-score, or even by normalizing each row such that the sum of its elements or the

Figure 1: The different steps of BADIA.



sum of its squared elements is equal to one. The observations in $\mathbf{X}$ are partitioned into N *a-priori* categories of interest with I_n being the number of observations of the nth category (and so $\sum_{n}^{N} I_n = I$). The columns of matrix $\mathbf{X}$ can be arranged in K *a priori* blocks (or sub-tables). The number of columns of the kth block are denoted J_k (and so $\sum_{k}^{K} J_k = J$). So, the matrix $\mathbf{X}$ can decomposed into N by K blocks as

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ \vdots \\ n \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & \cdots & k & \cdots & K \end{array} \\ \begin{bmatrix} \mathbf{X}_{1,1} & \cdots & \mathbf{X}_{1,k} & \cdots & \mathbf{X}_{1,K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{X}_{n,1} & \cdots & \mathbf{X}_{n,k} & \cdots & \mathbf{X}_{n,K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{X}_{N,1} & \cdots & \mathbf{X}_{N,k} & \cdots & \mathbf{X}_{N,K} \end{bmatrix} \end{array} \tag{1}$$

2.1 Notations for the categories (rows)

We denote by $\mathbf{Y}$ the I by N design matrix for the categories (*i.e.*, groups) describing the the rows of $\mathbf{X}$: $y_{i,n} = 1$ if row i belongs to category n, $y_{i,n} = 0$ otherwise. We denote by $\mathbf{m}$ the I by 1 vector of *masses* for the rows of $\mathbf{X}$ and by $\mathbf{M}$ the I by I diagonal matrix whose diagonal elements are the elements of $\mathbf{m}$ (*i.e.*, using the diag operator which transforms a vector into a diagonal matrix, we have

$\mathbf{M} = \text{diag}\{\mathbf{m}\}$). We denote by $\mathbf{b}$ the N by 1 vector of masses for the categories describing the rows of $\mathbf{X}$ and by $\mathbf{B}$ the N by N diagonal matrix whose diagonal elements are the elements of $\mathbf{b}$. Masses are positive numbers, it is convenient (but not necessary) to have the sum of the masses equal to one.

2.2 Notations for blocks (columns)

We denote by $\mathbf{Z}$ the J by K design matrix for the blocks from the columns of $\mathbf{X}$: $z_{j,k} = 1$ if column j belongs to block k, $z_{j,k} = 0$ otherwise. We denote by $\mathbf{w}$ the J by 1 vector of *weights* for the columns of $\mathbf{X}$ and by $\mathbf{W}$ the J by J diagonal matrix whose diagonal elements are the elements of $\mathbf{w}$. We denote by $\mathbf{c}$ the K by 1 vector of weights for the blocks of $\mathbf{X}$ and by $\mathbf{C}$ the K by K diagonal matrix whose diagonal elements are the elements of $\mathbf{c}$. Weights are positive numbers, it is convenient (but not necessary) to have the sum of the weights equal to one.

3 BADIA: Barycentric Discriminant Analysis

The first step of BADIA is to compute the barycenter of each of the N categories describing the rows. The barycenter of a category is the weighted average of the rows where the weights are the masses re-scaled such that the sum of the weights for one category is equal to one. The N by J matrix of barycenters is computed as

$$\mathbf{R} = \text{diag}\left\{\mathbf{Y}^{\mathsf{T}}\mathbf{M}\mathbf{1}\right\}^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{M}\mathbf{X} \tag{2}$$

where $\mathbf{1}$ is a conformable vector of 1s (the diagonal matrix $\text{diag}\{\mathbf{Y}\mathbf{M}\mathbf{1}\}^{-1}$ serves to re-scale the masses of the rows such that their sum is equal to one for each category).

3.1 Masses, weights and GPCA

The type of preprocessing of the data and the choice of $\mathbf{B}$ (matrix of masses for the categories) and $\mathbf{W}$ (matrix of weights for the variables) is crucial because this choice determines the specific type of GPCA used. For example, discriminant correspondence analysis will is obtained by transforming rows of $\mathbf{R}$ into relative frequencies, by using the relative frequencies of the barycenters as their masses and by using the inverse of the column frequencies for the weights of the variables. As another example, standard discriminant analysis if obtained when $\mathbf{W}$ is equal to the inverse of the within group variance-covariance matrix.

3.2 GPCA of the barycenter matrix

The $\mathbf{R}$ matrix is then analyzed using a GPCA under the constraints provided by the matrices $\mathbf{B}$ (for the N categories) and $\mathbf{W}$ (for the columns). Specifically, matrix $\mathbf{R}$ is analyzed with the generalized singular value decomposition (see Abdi, 2007a, Greenacre, 1984) as

$$\mathbf{R} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \quad \text{with} \quad \mathbf{P}^{\mathsf{T}}\mathbf{B}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q} = \mathbf{I} \ , \tag{3}$$

where $\boldsymbol{\Delta}$ is the L by L diagonal matrix of the singular values (with L being the number of non-zero singular values), and $\mathbf{P}$ (respectively $\mathbf{Q}$) being the N by L (respectively J by L) matrix of the left (respectively right) generalized singular vectors of $\mathbf{R}$.

3.3 Factor scores

The N by L matrix of factor scores for the categories is obtained as

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta} = \mathbf{R}\mathbf{W}\mathbf{Q} \ . \tag{4}$$

The variance of the columns of $\mathbf{F}$ is given by the square of the corresponding singular values (*i.e.,* the "eigen-value" denoted λ, these are stored in the diagonal matrix $\boldsymbol{\Lambda}$). This can be shown by combining Equations 3 and 4 to give:

$$\mathbf{F}^{\mathsf{T}}\mathbf{B}\mathbf{F} = \boldsymbol{\Delta}\mathbf{P}^{\mathsf{T}}\mathbf{B}\mathbf{P}\boldsymbol{\Delta} = \boldsymbol{\Delta}^2 = \boldsymbol{\Lambda} \ . \tag{5}$$

The I rows of matrix $\mathbf{X}$ can be projected (as "supplementary" or "illustrative" elements) onto the space defined by the factor scores of the barycenters. Note that the matrix $\mathbf{W}\mathbf{Q}$ from Equation 4 is a projection matrix. Therefore, the I by L matrix $\mathbf{H}$ of the factor scores for the rows of $\mathbf{X}$ can be computed as

$$\mathbf{H} = \mathbf{X}\mathbf{W}\mathbf{Q}. \tag{6}$$

These projections are barycentric, because the weighted average of the factor scores of the rows of a category gives the factors scores of the category. This can be shown by first computing the barycenters of the row factor scores as (*cf.* Equation 2) as

$$\overline{\mathbf{H}} = \mathsf{diag}\left\{\mathbf{Y}\mathbf{M}\mathbf{1}\right\}^{-1}\mathbf{Y}\mathbf{M}\mathbf{H} \ , \tag{7}$$

then plugging in Equation 6 and developing. Taking this into account, Equation 4 gives

$$\overline{\mathbf{H}} = \mathsf{diag}\left\{\mathbf{Y}\mathbf{M}\mathbf{1}\right\}^{-1}\mathbf{Y}\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{Q} = \mathbf{R}\mathbf{W}\mathbf{Q} = \mathbf{F} \ . \tag{8}$$

3.4 Loadings

The loadings describe the variables of the barycentric data matrix and are used to identify the variables important for the separation between the groups. As for standard PCA, there are several ways of defining the loadings. The loadings can be defined as the correlation between the columns of matrix $\mathbf{R}$ and the factor scores. They can be defined as the matrix $\mathbf{Q}$ or (as we did in our example) as

$$\mathbf{G} = \mathbf{Q}\boldsymbol{\Delta} \ . \tag{9}$$

4 Quality of the prediction

The performance, or quality of the prediction of a discriminant analysis is assessed by predicting the category membership of the observations and by comparing the predicted with the actual category membership. The pattern of correct and incorrect classifications can be stored in a confusion matrix in which the columns represent the actual categories and the row the predicted categories. At the intersection of a row and a column is the number of observations from the column category assigned to the row category.

The performance of the model can be assessed for the observations used to compute the categories: this is the *fixed effect* model. In addition, the performance of the model can be estimated for *new* observations (*i.e.,* observations not used to compute the model): this is the *random effect* model).

4.1 Fixed effect: Old observations

The *fixed effect* model predicts the category assignment for the observations used to compute the barycenters of the categories. In order to assign an observation to a category, the first step is to compute the distance between this observation and all N categories. Then, the observation is assigned to the closest category. Several possible distances can be chosen, but a natural choice is the Euclidean distance computed in the factor space. If we denote by $\mathbf{h}_i$ the vector of factor scores for the ith observation, and by $\mathbf{f}_n$ the vector of factor scores for the nth category, then the squared Euclidean distance between the ith observation and the nth category is computed as

$$d^2\left(\mathbf{h}_i, \mathbf{f}_n\right) = \left(\mathbf{h}_i - \mathbf{f}_n\right)^{\mathsf{T}} \left(\mathbf{h}_i - \mathbf{f}_n\right) \ . \tag{10}$$

Obviously, other distances are possible (*e.g.,* Mahalanobis distance), but the Euclidean distance has the advantage of being "directly read" on the map.

4.1.1 Tolerance intervals

The quality of the category assignment of the actual observations can be displayed using *tolerance* intervals. A tolerance interval encompasses a given proportion of a sample or a population. When displayed in two dimensions, these intervals have the shape of an ellipse and are called *tolerance ellipsoids*. For BADIA, a category tolerance ellipsoid is plotted on the category factor score map. This ellipsoid is obtained by fitting an ellipse which includes a given percentage (*e.g.,* 95%) of the observations. Tolerance ellipsoids are centered on their categories and the overlap of the tolerance ellipsoids of two categories reflects the proportion of misclassifications between these two categories.

4.2 Random Effect: New observations

The *random effect* model evaluates the quality of the assignment of *new* observations to categories. This estimation is obtained, in general, by using cross validation techniques that partition the data into a *learning set* (used to create the model) and a *testing set* (used to evaluate the model). A convenient variation of this approach is the jackknife (*a.k.a.* "leave one out") approach: Each observation is taken out from the data set, in turn, and then is projected onto the factor space of the remaining observations in order to predict its category membership. For the estimation to be unbiased, the left-out observation should not be used in any way in the analysis. In particular if the data matrix is preprocessed, the left-out observation should not be used in the preprocessing. So, for example, if the columns of the data matrix are transformed into Z scores, the left-out observation should *not* be used to compute the means and standard deviations of the columns of the matrix to be analyzed, but these means and standard deviations will be used to compute the Z-score for the left-out observation.

The assignment of an observation to a category follows the same procedure as for a fixed effect model: the observation is projected onto the category factor scores, and the observation is assigned to the closest category. Specifically, we denote by $\mathbf{X}_{-i}$ the data matrix without the ith observation, and by $\mathbf{x}_i$ the ith observation. If $\mathbf{X}_{-i}$ is preprocessed (*e.g.,* centered and normalized), the pre-processing parameters will be estimated without $\mathbf{x}_i$ (*e.g.,* the mean and standard deviation of $\mathbf{X}_{-i}$ is computed *without* $\mathbf{x}_i$) and $\mathbf{x}_i$ will be pre-processed with the parameters estimated for $\mathbf{X}_{-i}$ (*e.g.,* $\mathbf{x}_i$ will be centered and normalized using the means and standard deviations of the columns of $\mathbf{X}_{-i}$). Then the matrix of barycenters $\mathbf{R}_{-i}$ is computed and its generalized eigendecomposition is obtained as (*cf.* Equation 3):

$$\mathbf{R}_{-i} = \mathbf{P}_{-i}\mathbf{\Delta}_{-i}\mathbf{Q}_{-i}^{\mathsf{T}} \quad \text{with} \quad \mathbf{P}_{-i}^{\mathsf{T}}\mathbf{W}_{-i}\mathbf{P}_{-i} = \mathbf{Q}_{-i}^{\mathsf{T}}\mathbf{B}_{-i}\mathbf{Q}_{-i} = \mathbf{I} \tag{11}$$

(with $\mathbf{B}_{-i}$ and $\mathbf{W}_{-i}$ being the mass and weight matrices for $\mathbf{R}_{-i}$). The matrix of factor scores denoted $\mathbf{F}_{-i}$ is obtained as (*cf.* Equation 4)

$$\mathbf{F}_{-i} = \mathbf{P}_{-i}\mathbf{\Delta}_{-i} = \mathbf{R}_{-i}\mathbf{W}_{-i}\mathbf{Q}_{-i} \ . \tag{12}$$

The jackknifed projection of the ith observation, denoted $\widetilde{\mathbf{h}}_i$ is obtained (*cf.* Equation 6) as

$$\widetilde{\mathbf{h}}_i = \mathbf{x}_i\mathbf{W}_{-i}\mathbf{Q}_{-i} \ . \tag{13}$$

Distances between the ith observation and the N categories can be computed (*cf.* Equation 10) with the factor scores. The observation is then assigned to the closest category.

4.2.1 Prediction intervals

In order to display the quality of the prediction for *new* observations we use *prediction* intervals. In order to compute these intervals, the first step is to project the jackknifed observations onto the original complete factor space. There are several ways to project a jackknifed observation onto the factor score space. Here we proposed a two-step procedure. First, the observation is projected onto the jackknifed space and is reconstructed from its projections. Then, the reconstituted observation is projected onto the full factor score solution. Specifically, a jackknifed observation is reconstituted from its factor scores as (*cf.* Equations 3 and 13):

$$\widetilde{\mathbf{x}}_i = \widetilde{\mathbf{h}}_i\mathbf{Q}_{-i}^{\mathsf{T}} \ . \tag{14}$$

The projection of the jackknifed observation is denoted $\widehat{\mathbf{h}}_i$ and is obtained by projecting $\widetilde{\mathbf{x}}_i$ as a supplementary element in the original solution. Specifically, $\widehat{\mathbf{h}}_i$ is computed as

$$\begin{aligned}
\widehat{\mathbf{h}}_i &= \widetilde{\mathbf{x}}_i\mathbf{W}\mathbf{Q} && (\textit{cf.}\ \text{Equation 4}) \\
&= \widetilde{\mathbf{h}}_i\mathbf{Q}_{-i}^{\mathsf{T}}\mathbf{W}\mathbf{Q} && (\textit{cf.}\ \text{Equation 14}) \\
&= \mathbf{x}_i\mathbf{W}_{-i}\mathbf{Q}_{-i}\mathbf{Q}_{-i}^{\mathsf{T}}\mathbf{W}\mathbf{Q} && (\textit{cf.}\ \text{Equation 13}) \ .
\end{aligned} \tag{15}$$

Prediction ellipsoids are not necessarily centered on their categories (the distance between the center of the ellipse and the category represents the estimation *bias*). Overlap of two predictions intervals directly reflects the proportion of misclassifications for the "new" observations.

5 Quality of the category separation

5.1 R^2 and permutation test

In order to evaluate the quality of the discriminant model, we use a coefficient inspired by the coefficient of correlation. Because BADIA is a barycentric technique, the total *inertia* (*i.e.,* the "variance") of the observations to the grand barycenter (*i.e.,* the barycenter of all categories) can be decomposed into two additive quantities: *1)* the inertia of the observations relative to the barycenter of their own category, and *2)* the inertia of the category barycenters to the grand barycenter.

Specifically, if we denote by $\overline{\mathbf{f}}$ the vector of the coordinates of the grand barycenter (*i.e.,* each component of this vector is the average of the corresponding components of the barycenters), the total inertia, denoted $\mathcal{I}_{\mathsf{Total}}$, is computed as the sum of the squared distances of the observations to the grand barycenter (10):

$$\mathcal{I}_{\mathsf{Total}} = \sum_{i}^{I} m_i d^2\left(\mathbf{h}_i, \overline{\mathbf{f}}\right) = \sum_{i}^{I} m_i \left(\mathbf{h}_i - \overline{\mathbf{f}}\right)^{\mathsf{T}} \left(\mathbf{h}_i - \overline{\mathbf{f}}\right) \ . \tag{16}$$

The inertia of the observations relative to the barycenter of their own category is abbreviated as the "inertia within." It is denoted $\mathcal{I}_{\mathsf{Within}}$ and computed as

$$\mathcal{I}_{\mathsf{Within}} = \sum_{n}^{N} \sum_{i \text{ in category } n} m_i d^2\left(\mathbf{h}_i, \mathbf{f}_n\right) = \sum_{n}^{N} \sum_{i \text{ in category } n} m_i \left(\mathbf{h}_i - \mathbf{f}_n\right)^{\mathsf{T}} \left(\mathbf{h}_i - \mathbf{f}_n\right) \ . \tag{17}$$

The inertia of the barycenters to the grand barycenter is abbreviated as the "inertia between." It is denoted $\mathcal{I}_{\mathsf{Between}}$ and computed as

$$\mathcal{I}_{\mathsf{Between}} = \sum_{i}^{I} b_n \times d^2\left(\mathbf{f}_n, \overline{\mathbf{f}}\right) = \sum_{n}^{N} b_n \times d^2\left(\mathbf{f}_n, \overline{\mathbf{f}}\right) = \sum_{n}^{N} b_n \times \left(\mathbf{f}_n - \overline{\mathbf{f}}\right)^{\mathsf{T}} \left(\mathbf{f}_n - \overline{\mathbf{f}}\right) \ . \tag{18}$$

So the additive decomposition of the inertia can be expressed as

$$\mathcal{I}_{\mathsf{Total}} = \mathcal{I}_{\mathsf{Within}} + \mathcal{I}_{\mathsf{Between}} \ . \tag{19}$$

This decomposition is similar to the familiar decomposition of the sum of squares in the analysis of variance. This suggest that the intensity of the discriminant model can be tested by the ratio of between inertia by the total inertia, as is done in analysis of variance and regression. This ratio is denoted R^2 and it is computed as:

$$R^2 = \frac{\mathcal{I}_{\mathsf{Between}}}{\mathcal{I}_{\mathsf{Total}}} = \frac{\mathcal{I}_{\mathsf{Between}}}{\mathcal{I}_{\mathsf{Between}} + \mathcal{I}_{\mathsf{Within}}} \ . \tag{20}$$

The R^2 ratio takes values between 0 and 1, the closer to one the better the model. The significance of R^2 can be assessed by permutation tests, and confidence intervals can be computed using cross-validation techniques such as the jackknife (see Abdi & Williams, this volume).

5.2 Confidence intervals

The stability of the position of the categories can be displayed using *confidence* intervals. A confidence interval reflects the variability of a population *parameter* or its estimate. In two dimensions, this interval becomes a confidence ellipsoid. The problem of estimating the variability of the position of the categories cannot, in general, be solved analytically and cross-validation techniques need to be used. Specifically, the variability of the position of the categories is estimated by generating *bootstraped* samples from the sample of observations. A bootstraped sample is obtained by sampling *with replacement* from the observations (recall that when sampling with replacement some observations may be absent and some other maybe repeated). The "bootstraped barycenters" obtained from these samples are then projected onto the discriminant factor space and, finally, an ellipse is plotted such that it comprises a given percentage (*e.g.,* 95%) of these bootstraped barycenters. When the confidence intervals of two categories do not overlap, these two categories are "significantly different" at the corresponding alpha level (*e.g.,* $\alpha = .05$).

6 Multiblock analysis

In a multiblock analysis, the blocks can be analyzed by projecting the categories and the observations for each block. As was the case for the categories, these projections are barycentric because the barycenter of the all the blocks gives the coordinates of the whole table.

6.1 Partial projection

Each block can be projected in the common solution. The procedure starts by rewriting Equation 3 in order to show the blocks:

$$\mathbf{R} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Delta}\left[\mathbf{Q}_1, \ldots, \mathbf{Q}_k, \ldots, \mathbf{Q}_K\right]^{\mathsf{T}} \quad , \tag{21}$$

where $\mathbf{Q}_k$ is the kth block (comprising the J_k columns of $\mathbf{Q}$ corresponding to the J_k columns of the kth block). Then, Equation 4 is rewritten to get the projection for the k-th block as

$$\mathbf{F}_k = K\mathbf{X}_k\mathbf{W}_k\mathbf{Q}_k \tag{22}$$

(where $\mathbf{W}_k$ is the weight matrix for the J_k columns of the k-th block).

6.2 Inertia of a block

Recall from Equation 5 that, for a given dimension, the variance of the factor scores of all the J columns of matrix $\mathbf{R}$ is equal to the eigenvalue of this dimension. Because each block comprises a set of columns, the contribution of a block to a dimension can be expressed as the sum of this dimension squared factor scores of the columns of this block. Precisely, the inertia for the kth table and the ℓth dimension is computed as:

$$\mathcal{I}_{\ell,k} = \sum_{j \in J_k} w_j f_{\ell,j}^2 \tag{23}$$

Note that the sum of the inertia of the blocks gives back the total inertia:

$$\lambda_\ell = \sum_k \mathcal{I}_{\ell,k} \ . \tag{24}$$

7 An example: 4 assessors taste 12 wines from 3 regions

It is a common belief that the taste of a wine depends upon its place of origin (*aka* apellation or region). As an (fictitious) illustration, we have sampled 12 wines coming from 3 different French wine regions and asked 4 professional assessors (unaware of the origin of the wines) to rate these wines. Each assessor generated a list of descriptors (up to six) and rated (with a 7-point scale) each wine on these descriptors. The assessors also evaluated an additional French wine from an unknown region with the goal of predicting the origin of the wine from the assessors' ratings. The data are given in Table 1.

Tables 2 and 3 give the factor scores for the analysis and Figure 2 displays them. We can see that the wine regions appear to be well differentiated, with wines from the Loire being characterized as "fruity" and "pleasant", wines from the Rhône being characterized as having a high "alcohol" content, and the wines from Beaujolais being characterized as "tangy" and "tannic."

Table 1: Data from the 3 wine regions example: 12 wines from 3 different wine regions are rated on a 7-point scale on descriptors assigned by each of 4 assessors.

		Assessor 1					Assessor 2				Assessor 3						Assessor 4		
Wine	Region	Wood	Fruit	Sugar	Alcohol	Hedonic	Tannic	Fruity	Phenol	Hedonic	Tangy	Sour	Fruity	Easy	Smooth	Pleasant	Alcohol	Hedonic	Tannic
1	Loire	1	3	2	1	1	2	5	2	2	1	1	3	4	5	4	1	2	1
2	Loire	2	3	3	2	3	2	6	1	3	1	2	4	2	4	5	2	2	2
3	Loire	1	2	2	1	2	2	4	2	4	2	1	5	3	5	6	1	1	3
4	Loire	1	3	3	2	4	1	6	1	3	1	1	4	4	5	5	1	2	1
$\overline{X}$	Loire	1.25	2.75	2.50	1.50	2.50	1.75	5.25	1.50	3.00	1.25	1.25	4.00	3.25	4.75	5.00	1.25	1.75	1.75
1	Rhône	1	2	1	3	3	3	3	4	5	3	1	3	1	1	3	3	5	1
2	Rhône	2	1	1	3	2	5	2	3	4	4	1	4	2	2	2	3	4	2
3	Rhône	3	2	2	2	1	4	4	4	6	3	2	2	1	1	1	2	3	3
4	Rhône	2	3	3	3	4	5	6	5	5	3	1	1	1	2	3	3	3	4
$\overline{X}$	Rhône	2.00	2.00	1.75	2.75	2.50	4.25	3.75	4.00	5.00	3.25	1.25	2.50	1.25	1.50	2.25	2.75	3.75	2.50
1	Beaujolais	3	1	3	1	1	5	4	1	1	4	4	1	2	2	2	3	5	1
2	Beaujolais	2	1	3	1	2	6	4	2	1	4	3	3	3	3	4	1	2	4
3	Beaujolais	3	2	2	2	4	7	4	1	2	5	5	1	2	2	1	1	1	4
4	Beaujolais	3	1	1	1	3	6	1	1	1	4	5	2	1	3	2	1	2	4
$\overline{X}$	Beaujolais	2.75	1.25	2.25	1.25	2.50	6.00	3.25	1.25	1.25	4.25	4.25	1.75	2.00	2.50	2.25	1.50	2.50	3.25
	Mystery	1	3	2	2	1	2	6	1	3	1	1	4	4	5	5	1	2	2

Table 2: Factor scores for the observations, categories and blocks. Jackknifed factor scores for the observation

			Loire Wines				Region	Rhône Wines				Region	Beaujolais Wines				Region	Mystery
Axis	λ	%	1	2	3	4	Loire	1	2	3	4	Rhône	1	2	3	4	Beaujolais	Wine
								Factor Scores for Observations and Regions										
1	0.67	62.91	1.03	0.94	1.10	1.40	1.11	−0.07	−0.44	−0.47	−0.22	−0.30	−0.76	−0.34	−1.13	−1.03	−0.82	1.23
2	0.39	37.09	0.31	0.19	0.16	0.25	0.23	−1.09	−0.62	−0.85	−0.86	−0.86	0.38	0.68	0.66	0.79	0.63	0.27
								Jackknifed Factor Scores for Observations										
1			1.02	0.97	1.09	1.45	–	−0.10	−0.44	−0.44	−0.22	–	−0.70	−0.37	−1.09	−0.96	–	–
2			0.33	0.15	0.09	0.32	–	−1.06	−0.53	−0.81	−0.65	–	0.12	0.63	0.67	0.82	–	–
								Factor Scores for Blocks by Region										
1	Assessor 1		–	–	–	–	0.26	–	–	–	–	−0.04	–	–	–	–	−0.23	–
2	Assessor 1		–	–	–	–	0.02	–	–	–	–	−0.25	–	–	–	–	0.23	–
1	Assessor 2		–	–	–	–	1.36	–	–	–	–	−0.12	–	–	–	–	−1.24	–
2	Assessor 2		–	–	–	–	0.15	–	–	–	–	−1.75	–	–	–	–	1.60	–
1	Assessor 3		–	–	–	–	2.55	–	–	–	–	−0.82	–	–	–	–	−1.72	–
2	Assessor 3		–	–	–	–	0.44	–	–	–	–	−0.97	–	–	–	–	0.53	–
1	Assessor 4		–	–	–	–	0.29	–	–	–	–	−0.21	–	–	–	–	−0.07	–
2	Assessor 4		–	–	–	–	0.30	–	–	–	–	−0.45	–	–	–	–	0.15	–

Table 3: Factor scores for the variables

Axis	λ	%	Wood	Fruit	Sugar	Alcohol	Hedonic	Tannic	Fruity	Phenol	Hedonic	Tangy	Sour	Fruity	Easy	Smooth	Pleasant	Alcohol	Hedonic	Tannic
1	0.67	62.91	−0.59	0.59	0.18	−0.07	0	−1.72	0.85	−0.22	0.34	−1.24	−1.00	0.93	0.66	1.15	1.25	−0.27	−0.49	−0.59
2	0.39	37.09	0.16	−0.16	0.26	−0.65	0	0.28	0.02	−1.22	−1.49	0.09	1.00	−0.07	0.49	0.73	0.33	−0.60	−0.66	0.16

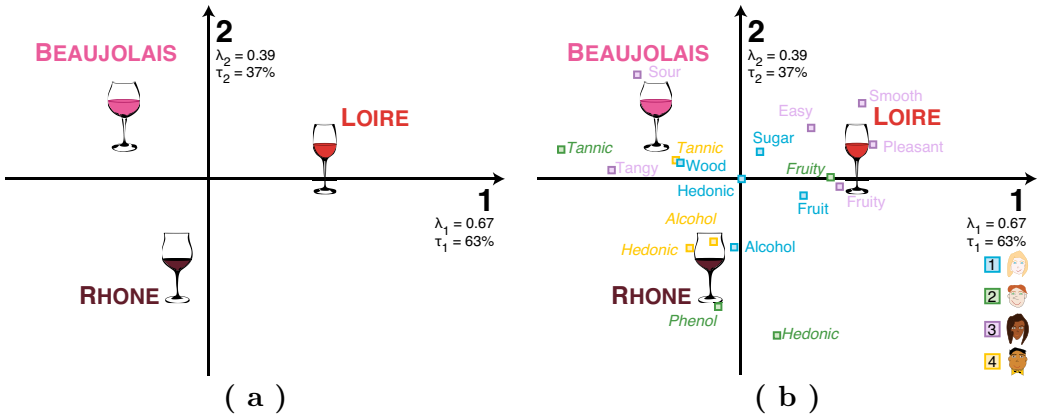


Figure 2: Multiblock Barycentric Discriminant Analysis. Projections on the first 2 dimensions. (a) The N set.Rows of $\mathbf{R}$: wine regions barycenters. (b) The J set. Columns ($i.e.$, descriptors). The wine categories have also been projected for ease of interpretation. Both figures have the same scale.

7.1 Reliability and Stability of the Analysis

7.1.1 R^2

The reliability and the stability of the analysis is evaluated by computing R^2 (see Equation 20). Its large value: R^2 = .94 confirms that the wine regions are well identified. In addition, the p-value of $p < .001$ (obtained from a permutation test using 10,000 permutations) indicates that the discrimination of the 3 wine regions is reliable.

7.1.2 Confidence Intervals

The original $I \times J$ matrix was resampled 1,000 times with replacement and the barycenters for each wine region for each of the 1,000 samples was projected into the original factor space. The 950 closest barycenters were kept and we generated a 95% confidence ellipsoid which comprised these 950 barycenters. These confidence ellipsoids are displayed in Figure 3-a. Because there is no overlap between the 3 wines regions, they are significantly different with $p < .05$.

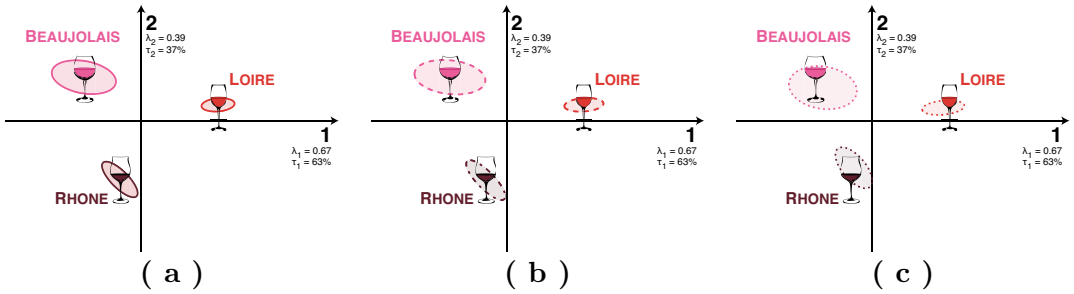


Figure 3: Multiblock Barycentric Discriminant Analysis. Projections on the first 2 dimensions. (a) Confidence Ellipses. (b) Tolerance Ellipses. (c) Prediction Ellipses. All figures are shown on the same scale.

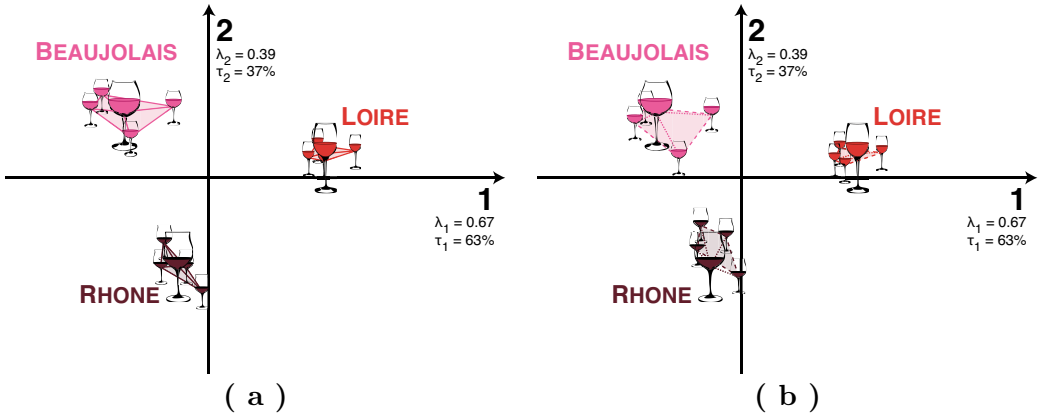


Figure 4: Multiblock Barycentric Discriminant Analysis. Projections on the first 2 dimensions. (a) Fixed Effect Model. The three wine regions and the convex hull for the wines. (b) Random Effect Model. The jackknifed wines have been projected back onto the fixed effect solution. The convex hull shows that the random effect categories have a larger variability and have moved. All figures are shown on the same scale.

7.2 Reliability of the Prediction

7.2.1 Fixed Effect Model

Confusion Matrix. The fixed effect quality of the model is evaluated by the following confusion matrix:

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \tag{25}$$

In this matrix, the rows are the assigned categories and the columns are the real categories. All wines within the sample were correctly classified. The overall quality of the fixed effect model can be computed from the diagonal of the matrix. Here we find that all 12 wines were correctly classified.

The projections of the wines within the sample into the original GPCA space are shown in Figure 4-a. The quality of the model can be evaluated by drawing the convex hull of each category. For the fixed effect model, the center of gravity of the convex hulls are the category barycenters.

Tolerance Intervals. The reliability of the prediction for the fixed effect model can also be displayed graphically as tolerance ellipsoids. These are shown in Figure 3-b. Overlap, between the tolerance ellipsoids, represents the proportion of misclassification of observations within the sample. Because there is no overlap, there were no misclassified wines within the sample.

7.2.2 Random Effect Model

Confusion Matrix. A jackknife procedure was used in order to evaluate the generalization capacity of the analysis to *new* wines. Each wine was taken out of the sample, in turn, and a GPCA was performed on the remaining sample of 11 wines, the left-out wine was then projected onto the discriminant factor space (see Equation 13) and was assigned to its closest category. This gave the following random effect confusion matrix:

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} . \tag{26}$$

The random effect performance is perfect, with all 12 wines correctly assigned.

Prediction Intervals. The projections of the jackknifed wines into the original GPCA space (computed according to Equations 14 and 15) are given in Table 2 and displayed in Figure 4-b. The quality of the model can be illustrated by drawing the convex hull for these observations. All the wines were correctly classified, but note that compared to the fixed effect (*cf.* Figure 4-a), the placement of the convex hull expands and shifts for each of the 3 wine categories.

The reliability of the predictions for the random effect model can also be displayed as prediction ellipsoids. Overlap between ellipsoids represents the proportion of misclassifications of *new* observations. This is shown in Figure 3-c. Because there is no overlap between the prediction ellipsoids, the *new* observations were all correctly classified.

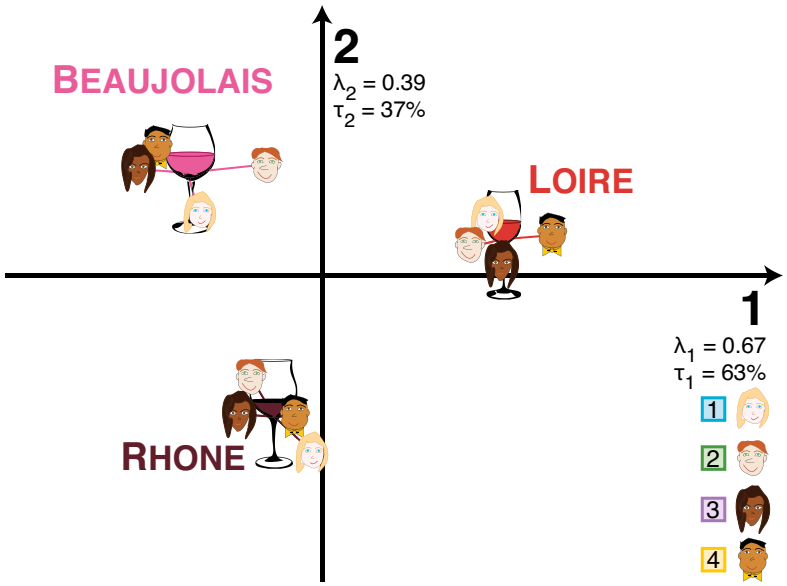


Figure 5: Multiblock Barycentric Discriminant Analysis. Projections on the first 2 dimensions. Position of the assessors relative to the barycenters of the wine regions. The assessors are projected as supplementary elements.

7.3 Multiblock Analysis

In this example, one block corresponds to the set of ratings made by one assessor (*cf.* Table 1).

7.3.1 Partial Projections

To see how each assessor separated the wine regions, we project—as supplementary elements—each assessor's factor scores for the three wine regions (*cf.* Equation 22). These factor scores are given in Table 2 and are displayed in Figure 5.

7.3.2 Inertia of the Blocks

The respective importance of each assessor (*i.e.*, block) is quantified by its associated partial inertia (see Equation 23). The partial inertias are given in Table 4 and are displayed in Figure 6. We can see that assessors 2 and 3 account for most of the inertia in the analysis with assessor 3 determining most of Dimension 1 and assessor 2 determining most of Dimension 2.

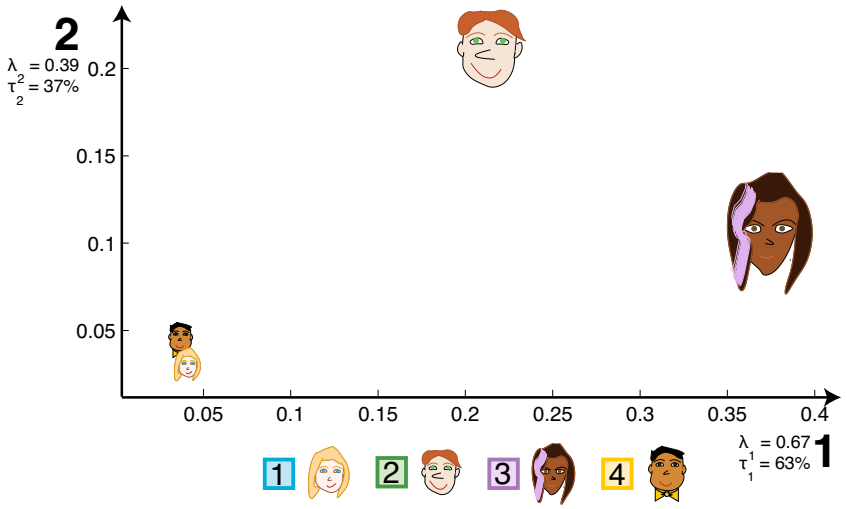


Figure 6: Multiblock Barycentric Discriminant Analysis. Partial Inertia of the blocks (*i.e.*, assessors) on the first 2 dimensions.

Table 4: Partial Inertias of the Blocks

Axis	Assessor 1	Assessor 2	Assessor 3	Assessor 4	$\sum = \lambda$
1	0.0408	0.2139	0.3743	0.0370	0.6660
2	0.0302	0.2113	0.1056	0.0456	0.3927

Related entries

Bootstrap, canonical correlation, correspondence analysis, discriminant analysis, jackknife, matrix algebra, permutation test, principal component analysis,

Further readings

1. Abdi, H. (2003). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage.
2. Abdi, H. (2007a). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 907–912.
3. Abdi, H. (2007a). Discriminant correspondence analysis (DCA). In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 270–275.
4. Abdi, H, Dunlop, J.P., & Williams, L.J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classiffers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage*, **45**, 89–95.
5. Abdi, H., & Williams, L.J. (2010). Jackknife. In N.J. Salkind (Ed.): *Encyclopedia of research design*. Thousand Oaks (CA): Sage.
6. Celeux, P., & Nakache, J.P. (1994). *Analyse discriminante sur variables qualitatives*. Paris: Polytechnica.

7. Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American,* May, 116-130.
8. Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
9. Gittins (1980). *Canonical analysis: A review with applications in ecology*. New York: Springer Verlag.
10. Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
11. Hwang, H., Tomiuk, M.A., Takane, Y. (2009). Correspondence analysis, multiple correspondence analysis and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.). *Handbook of quantitative methods in psychology*. London: Sage Publications. pp.
12. Jackson, E. (2003). *A User's Guide to Principal Components*. New York: Wiley.
13. Jolliffe, I.T. (2002). *Principal Component Analysis*. New York: Springer Verlag.
14. Krzanowski, W.J., & Radley, D. (1989). Nonparametric confidence and tolerance regions in canonical variate analysis. *Biometrics*, **45**, 1163–1173.
15. Leclerc, A. (1976). Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives. *International Statistical Review*, **44**, 241–248.
16. Manly, B.F.J. (1997). *Randomization, Bootstrap, and Monte Carlo methods in biology (2nd Edition)*. New York: Chapman & Hall.
17. Perrière, G., Lobry, J.R., & Thioulouse, J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Bioinformatics*, **12**, 519-524.
18. Perrière, G. & Thioulouse, J. (2003). Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins. *Computer Methods and Programs in Biomedicine*, **70**, 99-105.
19. Saporta, G., & Niang, N. (2006). Correspondence analysis and classification. In M. Greenacre & J. Blasius (Eds.): *Multiple correspondence analysis and related methods*. Boca raton (FL): Chapman & Hall/CRC. pp. 371–392.