Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v6n2p211

Interpolative and Predictive Biplots applied to Co-Inertia Analysis
By Amenta

October 14, 2013

Electronic Journal of Applied Statistical Analysis
Vol. 06, Issue 02, 2013, 211-223
DOI: 10.1285/i20705948v6n2p211

Interpolative and Predictive Biplots applied to Co-Inertia Analysis

Pietro Amenta*

Department of Law, Economics, Management and Quantitative Methods. University of Sannio. Piazza Arechi II, Palazzo De Simone, 82100, Benevento, Italy.

October 14, 2013

Aim of this paper is to introduce Gower and Hand's biplot procedures within the Co-Inertia Analysis. Predictive biplots can be considered a good choice for interpretation purposes while interpretative ones are very useful to classify new samples not used for the construction of the main co-inertia axes.

keywords: Co-inertia Analysis, Interpolative Biplots, Predictive Biplots.

1 Introduction

In applied or theoretical contexts we often have to deal with the study of numerical data tables obtained in experimental applications. The study of these tables often requires the use of multivariate analyses in order to investigate the relationships between the two data sets. In literature, in order to study symmetrical interdependence relationships, several techniques, originated from Canonical Correlation Analysis (Hotelling, 1936) (hereafter CCA), or from Tucker's Inter-Battery Analysis Tucker (1958) and Co-Inertia Analysis (Dray et al., 2003) (hereafter COA) and their generalizations, have been proposed. For instance, in the field of near infrared spectroscopy, the potential of near infrared spectroscopy to measure the main inorganic components of seawater as salt-manufacturing materials, was investigated with CCA (Chen et al., 2003). However, CCA could create highly correlated linear combinations but not necessarily the most explicative ones. COA, based on the covariance criterion, has been proposed to improve the Correlation Analysis. COA and CCA should not be considered as competing techniques due to their different goals (Cherry, 1996, 1997).

*Corresponding author: amenta@unisannio.it.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

COA is very popular in several fields. Tucker (1958) introduced this method in the social sciences, with the name of Inter-Battery Factor Analysis, in order to find common factors in two batteries of tests presented to the same group of statistical units. It has been popularized in the atmospheric sciences, where it is well known as Singular Value Decomposition Analysis, by Prohaska (1976), Bretherton et al. (1992) and Wallace et al. (1992). It is also very famous in ecology by the papers of Chessel and Mercier (1993), and Dolédec and Chessel (1994). It was also used in Sensometrics in analyzing relationships between instrumental and sensorial measurements (Sabatier et al., 1992).

However, for inexperienced users of Statistics, the factorial graphical outputs of COA are sometimes difficult to conceptualize because interpretations are carried out based on the analysis of synthetic variables. The procedures presented by Gower and Hand (1996), divided into predictive and interpolative biplots, seem to be meaningful and straightforward, owing to the ability of allowing the establishment of direct relationships between the observed data structures and the original variables with original units of measurement.

Aim of this paper is to introduce the Gower and Hand's biplot procedures within COA. We highlight that this approach could be also a comprehensive scheme for applying Interpolative and Predictive Biplots in a single algebraic framework. In fact, if we consider the different nature and coding of data matrices, and using diverse kinds of metrics, then a variety of existing biplots, applied to other methods, are realized by this approach. This is due to the natural links of several multivariate coupling methods with COA. Two examples of these procedures with well-known data sets are given.

2 A summary of Co-Inertia Analysis

The main goal of collecting the values taken by n statistical units on p variables in a table of order $(n \times p)$, is generally the comparison both of the statistical units and the variables. For the former comparison, if we choose to compute a distance between the statistical units, then we have to define a $(p \times p)$ symmetric positive definite matrix $\mathbf{Q}$ defining the scale of the several variables. For the latter, if we perform the comparison between the variables by a linear correlation coefficient, then we have to define a positive diagonal matrix $\mathbf{D}$ collecting the weights of the statistical units. Then we can consider the notation of the statistical study (triplet) ($\mathbf{X}$, $\mathbf{Q}$, $\mathbf{D}$) (Tenenhaus and Young, 1985) to describe the data and their use. This notation allows to present the factorial methods in a single theoretical framework by using suitable choices for the metrics $\mathbf{Q}$ and $\mathbf{D}$.

Let $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ be a statistical study associated with the matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}^T$ of order $n \times p$, collecting a set of p quantitative/qualitative variables observed on n statistical units where $\mathbf{D} = diag(d_1, ..., d_n)$ specifies the (diagonal) weights metric in the vectorial space $\Re^n$ of variables with $\sum_{i=1}^{n} d_i = 1$. Without loss of generality, we can assume uniform weights ($d_i = 1/n$) for $\mathbf{D}$ in this paper. $\mathbf{Q_X}$ is a $(p \times p)$ non-negative definite (hereafter *nnd*) matrix defining the metric measuring the distance between the data vectors $\mathbf{x}_j$, $\mathbf{x}_k$ of two statistical units j, k in $\Re^p$ given by $(\mathbf{x}_j - \mathbf{x}_k)^T \mathbf{Q_X} (\mathbf{x}_j - \mathbf{x}_k)$. $\mathbf{Q_X}$ can assume several forms:

1. if the coordinates that summarize a data matrix are mutually orthogonal and have comparable scales, then $\mathbf{Q_X}$ can be set to the identity matrix $\mathbf{I}_p$ in $\Re^p$ (essentially the Euclidean metric);

2. when the variables of $\mathbf{X}$ are measured on incomparable scales, it is common to transform the data into standard scores before analyzing them; this is equivalent to using the inverse of the diagonal matrix of the reciprocals of the standard deviation of each variable as $\mathbf{Q_X}$;

3. finally, if the variables are not highly correlated, we may set $\mathbf{Q_X} = (\mathbf{X}^T\mathbf{X})^{-1}$ (essentially the Mahalanobis metric), otherwise any generalized inverse of $\mathbf{X}^T\mathbf{X}$ can be used.

Let $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$ be the statistical study associated with the matrix $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}^T$ of order $(n \times q)$, collecting a second set of q (quantitative/qualitative) variables observed on the same n statistical units where $\mathbf{Q_Y}$ is the $(q \times q)$ *nnd* metric of the statistical units in $\Re^q$. We assume that both sets of variables are mean centred with respect to $\mathbf{D}$, i.e. the weighted mean value of each column in $\mathbf{X}$ and $\mathbf{Y}$ is set to zero ($\mathbf{1}_n^T\mathbf{DX} = \mathbf{0}$ with $\mathbf{1}_n$ unitary column vector).

In order to study the common geometry of the statistical triplets $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ and $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$, Co-Inertia Analysis (Chessel and Mercier, 1993) seeks linear combinations of the data $\mathbf{t}_i = \mathbf{XQ_X}\mathbf{w}_i$ and $\mathbf{u}_i = \mathbf{YQ_Y}\mathbf{c}_i$ ($i = 1, ..., s$, $s = \min(p, q)$) with the maximum covariance

$$\max_{\mathbf{w}_i\mathbf{c}_i} cov^2(\mathbf{t}_i, \mathbf{u}_i) = \max_{\mathbf{w}_i\mathbf{c}_i} (\mathbf{w}_i^T\mathbf{Q_X}\mathbf{X}^T\mathbf{DYQ_Y}\mathbf{c}_i)^2$$

such that the unknown weight vectors $\mathbf{w}_i$ and $\mathbf{c}_j$ ($i \neq j$) satisfy the constraints $\mathbf{w}_i^T\mathbf{Q_X}\mathbf{w}_i = 1$, $\mathbf{c}_i^T\mathbf{Q_Y}\mathbf{c}_i = 1$, and $\mathbf{w}_i^T\mathbf{Q_X}\mathbf{w}_j = \mathbf{c}_i^T\mathbf{Q_Y}\mathbf{c}_j = 0$ (that is $\mathbf{W}^T\mathbf{Q_X}\mathbf{W} = \mathbf{I}$ and $\mathbf{C}^T\mathbf{Q_Y}\mathbf{C} = \mathbf{I}$).

The COA$(\mathbf{X}, \mathbf{Y})_{\mathbf{Q_X}, \mathbf{Q_Y}}$ criterion can be also written as $cov^2(\mathbf{t}_i, \mathbf{u}_i) = cor^2(\mathbf{t}_i, \mathbf{u}_i) \times var(\mathbf{t}_i) \times var(\mathbf{u}_i)$ where $cor^2(\mathbf{t}_i, \mathbf{u}_i)$ is the square cosinus of the angle between $\mathbf{t}_i$ and $\mathbf{u}_i$, with $var(\mathbf{t}_i) = \mathbf{w}_i^T\mathbf{Q_X}\mathbf{X}^T\mathbf{DXQ_X}\mathbf{w}_i$ and $var(\mathbf{u}_i) = \mathbf{c}_i^T\mathbf{Q_Y}\mathbf{Y}^T\mathbf{DYQ_Y}\mathbf{c}_i$.

The criterion $cov^2(\mathbf{t}_i, \mathbf{u}_i)$ measures the proximity between the variables of both sets, so maximizing it we maximize the correlation between the components $cor^2(\mathbf{t}_i, \mathbf{u}_i)$ and their respective variances also.

From a computational point of view, COA$(\mathbf{X}, \mathbf{Y})_{\mathbf{Q_X}, \mathbf{Q_Y}}$ amounts also to the (Generalized) Singular Value Decomposition of the matrix $\mathbf{X}^T\mathbf{DY} = \mathbf{W}\Delta\mathbf{C}^T$, where the columns of $\mathbf{W}$ and $\mathbf{C}$ are constrained to be orthonormalized with respect to $\mathbf{Q_X}$ and $\mathbf{Q_Y}$ (that is $\mathbf{W}^T\mathbf{Q_X}\mathbf{W} = \mathbf{I}$ and $\mathbf{C}^T\mathbf{Q_Y}\mathbf{C} = \mathbf{I}$), respectively, and Δ is a diagonal and positive definite matrix containing the generalized singular values, ordered from largest to smallest.

This method is also defined by the analysis of the statistical triplets $(\mathbf{Y}^T\mathbf{DX}, \mathbf{Q_X}, \mathbf{Q_Y})$. The pairs of axes $\mathbf{w}_i$ and $\mathbf{c}_j$ are then obtained by the eigenvectors $\mathbf{g}_{(\mathbf{X})}$ and $\mathbf{g}_{(\mathbf{Y})}$ associated to the decomposition of the operators $\mathbf{Q}_X^{1/2}\mathbf{X}^T\mathbf{DYQ_Y}\mathbf{Y}^T\mathbf{DXQ}_X^{1/2}$ $(p < q)$ or $\mathbf{Q}_Y^{1/2}\mathbf{Y}^T\mathbf{DXQ_X}\mathbf{X}^T\mathbf{DYQ}_Y^{1/2}$ $(q < p)$, respectively, linked to the same maximum eigenvalue $\lambda = (\mathbf{w}_i^T\mathbf{Q_X}\mathbf{X}^T\mathbf{DYQ_Y}\mathbf{c}_i)^2$ where $\sqrt{\lambda}$ is the covariance between $\mathbf{t}_i$ and $\mathbf{u}_i$. After diagonalization s principal axes are preserved. Finally, weight vectors $\mathbf{w}_i$ and $\mathbf{c}_i$ are given by $\mathbf{w}_i = \mathbf{Q_X}^{-1/2}\mathbf{g}_{(\mathbf{X})_i}$ and $\mathbf{c}_i = \mathbf{Q_Y}^{-1/2}\mathbf{g}_{(\mathbf{Y})_i}$, respectively.

Co-Inertia Analysis is then a symmetric coupling metric-based method that provides a decomposition of the co-inertia criterion $tr(\mathbf{Q}_X^{1/2}\mathbf{X}^T\mathbf{D}\mathbf{Y}\mathbf{Q}_\mathbf{Y}\mathbf{Y}^T\mathbf{D}\mathbf{X}\mathbf{Q}_X^{1/2}) = \sum_{s=}^{p}\lambda_s$ on a set of orthogonal vectors where $tr()$ is the trace operator of a square matrix.

It is easy to show that if we set $\mathbf{Q_X} = \mathbf{I}_p$, $\mathbf{Q_Y} = \mathbf{I}_q$ and $\mathbf{D} = \mathbf{I}_n$ then Tucker's approach $\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{I}_p,\mathbf{I}_q}$ and Undeflated PLS (Burnham et al., 1996) lead to the same results, and first solutions of $\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{I}_p,\mathbf{I}_q}$ and PLS Regression (Höskuldsson, 1988) are equal. In this sense, Co-Inertia Analysis can be easily considered like a metric generalization of the Tucker's approach. For deeper COA features and its links with other multivariate coupling methods, see Dolédec and Chessel (1994), and Dray et al. (2003), respectively.

3 Interpolative and Predictive Coa-Inertia Biplots

$\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ graphical outputs can be enriched by using the additional relevant information given by the biplots (Gower and Hand, 1996). They are the multivariate analogue of scatter plots. Similar to scatter plots, biplots are useful for giving a graphical description of the data, for detecting patterns that possibly lead to more formal analyses, and for displaying results found by more formal methods of analysis. Both samples and measured variables are often represented in two-dimensional representations.

The conceptual distinction between interpolation and prediction can be easily highlighted. If we have a point with given values of the variables in a conventional Cartesian coordinate system for two quantitative variables, then its interpolative representation is given as vector-sum, while if we have a point then their associated values of the variables are given by orthogonal projection (Figure 1).

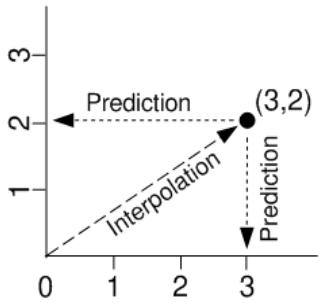


Figure 1: Cartesian axes. Interpolation and prediction.

Predictive biplots can be used to carry out interpretations in relation to initial values and variables instead of latent variables, without loosing the benefits of the multivariate modulation. The main advantage of predictive biplots is then due to an easier interpretation of the most important data structures overcoming the need for understanding and interpreting synthetic variables. Once multivariate graphs are produced, the interpolative biplots can be used in routine laboratory practice to position new statistical units in the graph by hand, using the vector sum method or the complete parallelograms method

for routine classification purposes also.

3.1 Interpolative Co-Inertia Biplot

COA$(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ biplot procedure consists of several steps. To achieve the objectives of biplots, a COA biplot axis for each main variable is drawn in the graphs. Due to the symmetry between the properties concerning $\mathbf{X}$ and $\mathbf{Y}$ we refer all the formulas to the weight vector $\mathbf{w}_i$ avoiding a double explanation. Let $\mathbf{w}_i = \mathbf{Q}_\mathbf{X}^{-1/2}\mathbf{g}_{(\mathbf{X})_i}$ be the COA weight vectors such that $\mathbf{w}_i^T\mathbf{Q_X}\mathbf{w}_i = 1$. The COA components are then given by $\mathbf{t}_i = \mathbf{XQ_X}\mathbf{w}_i$ and biplots graphs are built by plotting scores $\mathbf{t}_s$ vs $\mathbf{t}_{s'}$ $(s \neq s')$. We can assume $s = 1$ and $s' = 2$.

A convenient choice of the scale values is then performed, as well as the coordinates of each scale value in the multidimensional subspaces are computed. These steps are obtained according to the interpolative or the predictive objectives. In the Interpolative COA biplot, for each variable $\mathbf{X}_k$ $(k = 1, ..., p)$, an auxiliary column vector $\mathbf{e}_i$ is created with p elements, with 0 everywhere except for a 1 in position k. A column vector μ_k, with a set of m convenient values defining a scale compatible with the observed initial values, is also constructed.

Scale markers for all variables are then standardized and projected onto the factorial plane of the latent variables $[\mathbf{t}_s, \mathbf{t}_{s'}]$ as $\mu_k\mathbf{e}_k^T\mathbf{W}^{(s,s')}$ with $\mathbf{W}^{(s,s')} = [\mathbf{w}_s, \mathbf{w}_{s'}]$. These points are added on the graph labelling them with the original scale values. Afterwards we link these points by a straight line to represent the axis of the variable $\mathbf{X}_k$ equipped with a suitable measurement scale.

Finally, we position new additional statistical units in the interpolative graph by hand, using the vector sum method or the complete parallelograms method (Gower and Hand, 1996).

3.2 Predictive Co-Inertia Biplot

Since prediction is the inverse of interpolation, the projection of the scale markers in the plane $[\mathbf{t}_s, \mathbf{t}_{s'}]$ is obtained by taking into account the inverse of $\mathbf{W}$. The matrix $\mathbf{W}_{(s,s')}$ is then obtained selecting the columns s and s' of $\mathbf{W}^{-1}$. In addition, we have to consider an adjustment factor necessary to ensure back-projection to the right place onto the variable axes. This happens because in COA biplots the axes for prediction and interpolation lie in different directions (Gower and Hand, 1996). In addition, the inclusion of the adjustment factors necessary to carry out predictions to the right places over the variable axes, changes the magnitude of the gaps between scale markers. Scale markers for the predictive COA$(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ biplot will be then given by $\mu_k\mathbf{e}_k^T\mathbf{W}_{(s,s')}^T \times (\mathbf{e}_k^T\mathbf{W}_{(s,s')}^T\mathbf{W}_{(s,s')}\mathbf{e}_k)^{-1}$.

All these points are then labeled with the initial scale values and linked by a straight line to represent the variable's axis equipped with a measurement scale. Finally, original predictive values are read drawing orthogonal projections from points to variable axes.

When the variable markers are projected onto the factorial plane $[\mathbf{t}_s, \mathbf{t}_{s'}]$, some of them will fit well in the graph while others could appear with not-suitable representations (too

long or too short vectors). Moreover, variables with high predictive power could not have the same effects for interpolative purposes and vice versa. So a suitable variable selection for biplot construction is obligatory. Several trials suggested that it is right to start by building the predictive biplots without taking into account all the too long and too short vectors. Afterwards, the subset of variables to be used is selected by checking the accuracy of a small number of predicted values with the original ones.

A similar procedure can be repeated for the interpolative biplots by checking the proximity of a small set of interpolated points with the corresponding real coordinates. A crossed check of both the selected subsets of variables could lead to an optimal selection for conjoint interpolative and predictive purposes.

We highlight that this approach could be also a comprehensive scheme for applying Interpolative and Predictive Biplots in a single algebraic framework. In fact, if we consider the different nature and coding of $\mathbf{X}$ and $\mathbf{Y}$, and various choices of $\mathbf{Q_X}$ and $\mathbf{Q_Y}$, then a variety of existing biplots are realized by this approach. This is due to the natural links of several methods with $\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ (Dray et al., 2003). For instance, let $\hat{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}$ and $\hat{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{D}\mathbf{Y})^{-1}$ be pre-treatments of the matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. If we perform a Co-Inertia Analysis of the matrices $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ with the metrics $\mathbf{Q_X} = \mathbf{X}^T\mathbf{D}\mathbf{X}$ and $\mathbf{Q_Y} = \mathbf{Y}^T\mathbf{D}\mathbf{Y}$, and apply the suggested approach, then we obtain the Canonical Correlation Biplots suggested by Gower and Hand (1996), and Alves and Oliveira (2003).

Moreover, If $\mathbf{Y}$ is a qualitative variable with q categories then $\mathbf{U}$ is the $\mathbf{D}$-uncentered binary indicator matrix related to the complete disjunctive coding of this variable. In this case, $\mathbf{D_Y} = \mathbf{U}^T\mathbf{D}\mathbf{U}$ is the $q \times q$ diagonal matrix with the relative frequencies of each category. If we perform then a Co-Inertia Analysis of the matrices $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}} = \mathbf{U}\mathbf{D}_\mathbf{Y}^{-1}$ with the metrics $\mathbf{Q_X} = \mathbf{X}^T\mathbf{D}\mathbf{X}$ and $\mathbf{Q_Y} = \mathbf{D_Y}$, and apply the suggested approach, we obtain the Discriminant Biplots proposed by Gower and Hand (1996).

Finally, $\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ scores $\mathbf{t}_i$ and $\mathbf{u}_i$ result to be not $\mathbf{D}$ orthogonal in$\Re^n$. The Wold's two-block "Mode A" Partial Least Squares (Wold, 1985) overcomes this remark by adding orthogonality constraints $\mathbf{t}_i^T\mathbf{D}\mathbf{t}_{i'} = 0$ and $\mathbf{u}_i^T\mathbf{D}\mathbf{u}_{i'} = 0$ (with $i \neq i'$) to the original $\mathrm{COA}(\mathbf{X},\mathbf{Y})_{\mathbf{Q_X},\mathbf{Q_Y}}$ criteria, getting orthogonal scores and weights simultaneously. It is evident that we can apply these interpolative and predictive biplot procedures to Wold's two-block "Mode A" Partial Least Squares analogously.

4 Examples

4.1 Distribution of a butterfly

We consider a well-know genetic data set (McKechnie et al., 1975; Manly, 2005) to present COA-Biplots, as the first example of application. The data concern four environmental variables (altitude, annual precipitation, and the minimum and maximum temperatures) and four genetic variables (percentage frequencies for different phosphoglucose-isomerase [Pgi] genes as determined by the technique of electrophoresis) regarding 16 colonies of the butterfly *Euphydryas editha* in California and Oregon. The study of the relationships between the environmental and genetic variables could indicate the adap-

tation of the *Euphydryas editha* to the local environments. The environmental variables have been treated as **X** variables and the gene frequencies as the **Y** variables.

We report only the COA biplot outputs because the full COA study about the relationships between the environmental and genetic variables, is not the main goal of this paper. See Figure 2 for the main COA graphics outputs. For a deeper analysis of this data set see also Manly (2005).

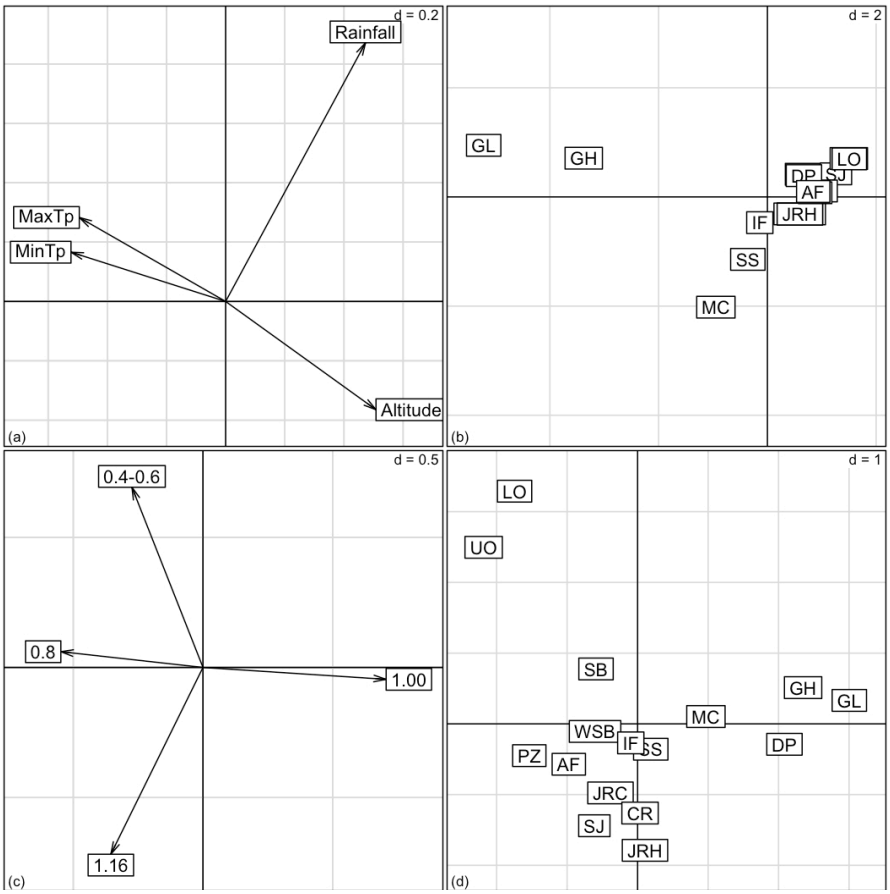


Figure 2: Main COA graphical outputs: Environmental (a) and genetic variables (c); Colonies of the butterfly with respect to environmental (b) and genetic variables (d).

In order to integrate the COA graphics with the biplot procedures, a suitable variable selection for biplot construction has been developed and a subset of variables with high predictive and interpretative power has been selected. This subset is formed for the interactive graph by the genetic variables 0.4-0.6, 0.8 and 1.16, while for the predictive graph no-variable selection has been performed.

Figure 3 shows how a predictive biplot can be used to read off from the graph the original values for the environmental variables for the units "GL", "GH" and "SJ" (Fig-

ure 3.a) and for the genetic variables "0.4-0.6", "0.8" and "1.16" for the units "GL", "GH", "LO" and "SJ" (Figure 3.b) overcoming the need for understanding and interpreting synthetic variables. It is evident how the predictive biplot is a complementary graph to the standard plot giving additional relevant information to carry out final interpretations.

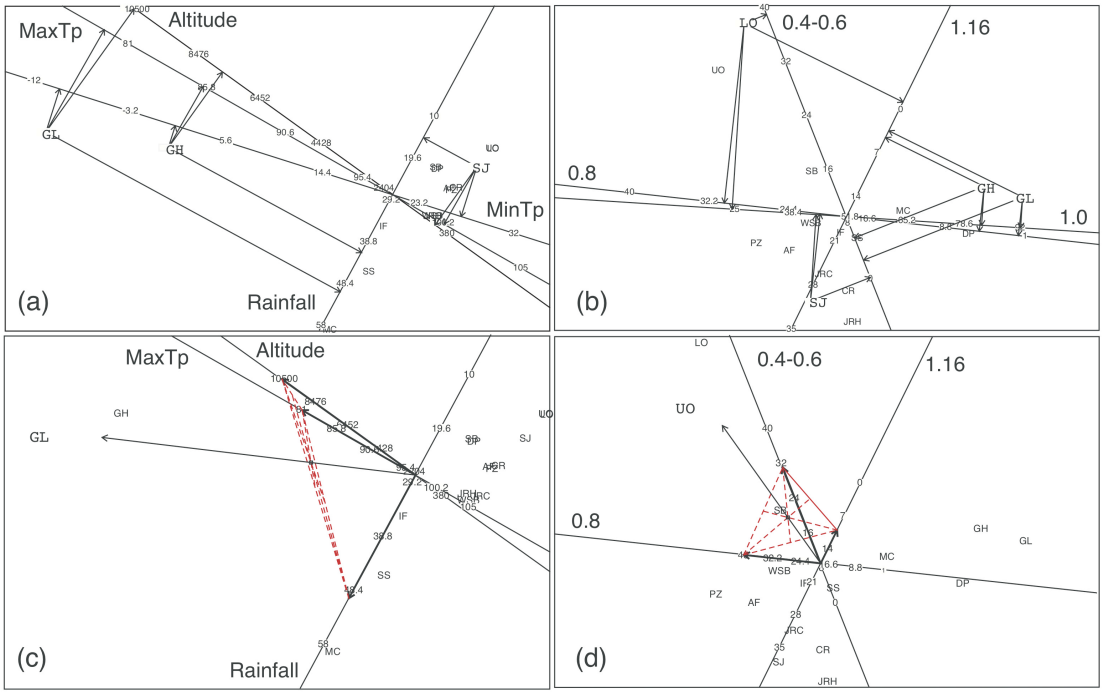


Figure 3: Predictive biplots based on COA of environmental (a) and genetic variables (b). Interpolative biplots based on COA of environmental (c) and genetic variables (d).

In addition, Figure 3.c and Figure 3.d show how given samples "GL" and "UO", respectively, are interpolated in the two-dimensional space by the vector sum method with a good approximation (exact positions are indicated by labels). We highlight how the COA interpolative biplot can be also considered as a complementary graph for routine classification purposes in routine laboratory practice. Aim of this graphical tool is to position new statistical units in the graph already available by hand without pretreatment of the values or calculations (e.g. positioning of the supplementary points in Principal Component Analysis). This can be very interesting if we want to position new points in a published graph without having the intermediate calculus and/or the full results.

Finally, Figure 4 shows the biplots of a Co-Inertia Analysis of the matrices $\mathbf{X}$ and $\mathbf{Y}$ with the metrics $\mathbf{Q_X} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{Q_Y} = \mathbf{Y}^T\mathbf{Y}$ for the same data set. They are equivalent to the Canonical Correlation Biplots of the matrices $\mathbf{X}$ and $\mathbf{Y}$ as proposed

by Gower and Hand (1996).

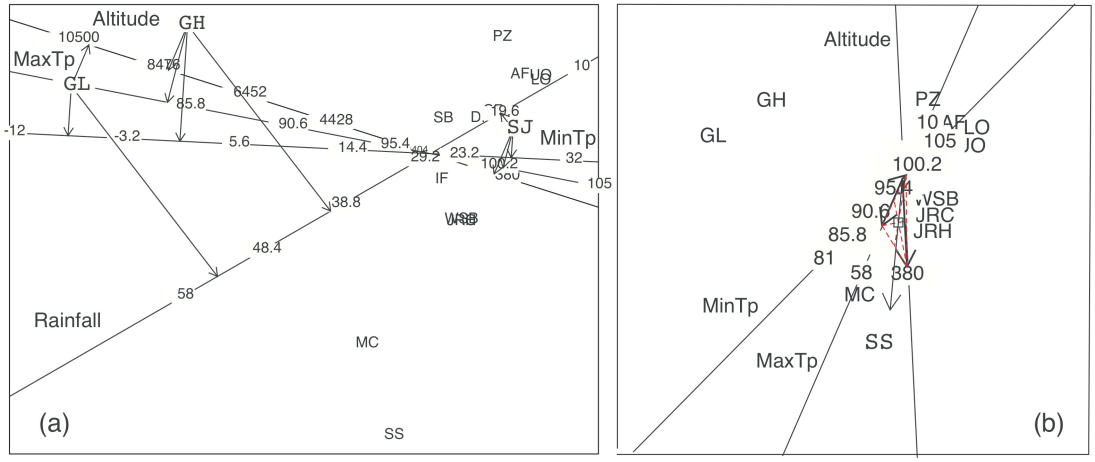


Figure 4: Predictive biplots based on CCA of environmental variables (a). Interpolative biplots based on CCA of environmental variables (b).

4.2 The thymallus thymallus data

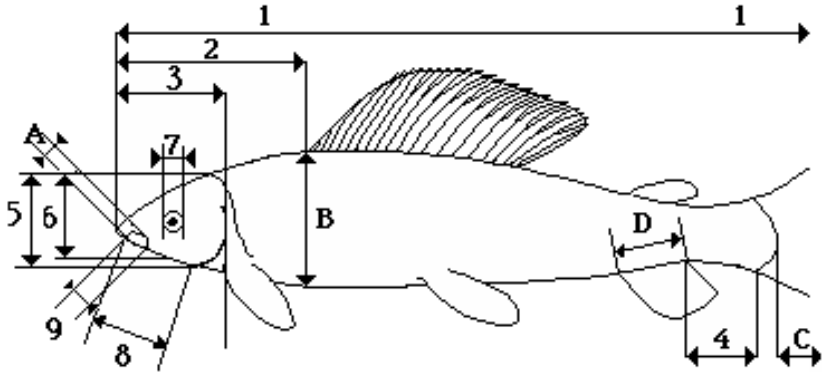
The Thymallus thymallus data (Persat, 1978) is a well-known data-set in Morphometrics. This discipline is the branch of mathematics studying the metrical and statistical properties of shapes and shape changes of geometric objects. It refers then to the quantitative analysis of form, commonly performed on organisms, a concept that encompasses size and shape. Morphometric analyses are useful in studying several aspects: the analysis of fossil record, the impact of mutations on shape, developmental changes in form, main factors that affect shape, and so on. It is evident that Discriminant Analysis plays an important role in this discipline.

The Thymallus thymallus belongs to the Salmoneide species and it is present in the higher part of the side basin of the Rhone in France even if we can find them in other rivers of the districts (Persat, 1978; Sabatier et al., 2003). Data are free available with the statistical software ADE 4 (Thioulouse et al., 1997).

Sabatier et al. (2003) used this data-set by applying Linear Discriminant Analysis (hereafter LDA) to discriminate 5 geographically different populations of thymallus thymallus (named Ain upstream, Bienne, Loue, Ain downstream, and Loire, respectively) with respect to 13 morphometric variables (sizes and shapes) and 120 statistical units. Figure 5 shows the codes of the morphometric variables. We report only the LDA biplot outputs because the full LDA study is not the main goal of this paper. See Figures 6, 7 and 8 for the main LDA biplot graphics outputs. Moreover, for a deeper LDA analysis of this data set, see also Chessel (2008).

To integrate the LDA graphics with the biplot procedures, we performed a Co-

Inertia Analysis of the matrices $\mathbf{X}$ and $\mathbf{Y} = \mathbf{U}\mathbf{D}_{\mathbf{Y}}^{-1}$ with the metrics $\mathbf{Q}_{\mathbf{X}} = \mathbf{X}^T\mathbf{D}\mathbf{X}$ and $\mathbf{Q}_{\mathbf{Y}} = \mathbf{D}_{\mathbf{Y}} = \mathbf{U}^T\mathbf{D}\mathbf{U}$, where $\mathbf{U}$ is the $\mathbf{D}$-uncentered binary indicator matrix related to the complete disjunctive coding of the 5 geographically different populations. Axes are then given by the diagonalization of the matrix $\mathbf{B}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}$, where $\mathbf{B} = (\mathbf{X}^T\mathbf{D}\mathbf{U}\mathbf{D}_{\mathbf{Y}}^{-1})\mathbf{D}_{\mathbf{Y}}(\mathbf{D}_{\mathbf{Y}}^{-1}\mathbf{U}^T\mathbf{D}\mathbf{X})$ and $\mathbf{X}^T\mathbf{D}\mathbf{X}$ are the between groups variance and variance-covariance (or correlation) matrices, respectively. COA outputs are then equivalent to LDA ones.



Label	Variable	Label	Variable
1	Total length	8	Jaw length
2	Out dorsal distance	9	Jawbone length
3	Head length	A	Jawbone width
4	Caudal anal distance	B	Maximum height of the body
5	Orbital height	C	Caudal fin length
6	Occipital length	D	Anal fin length
7	Eye Diameter		

Figure 5: The thymallus thymallus data. Codes of the morphometric variables.

A suitable variable selection for the biplot construction has been then developed and a subset of variables with high predictive and interpretative power has been selected. This subset is formed by the morphometric variables: out dorsal distance (code 2 in Figure 5), head length (code 3) and caudal anal distance (code 4).

Figure 6 shows how a predictive biplot is used to read off from the graph the initial values of variables head length and caudal anal distance for the five mean groups (G1 to G5) directly, overcoming the need for understanding and interpreting synthetic variables.

Likewise we can read off from the graph (Figure 7) (left hand) the initial values of the selected variables for single points. It is evident how the predictive biplot is a complementary graph to the standard plot (Figure 6, units and stargraph plots) giving additional relevant information to carry out final interpretations.

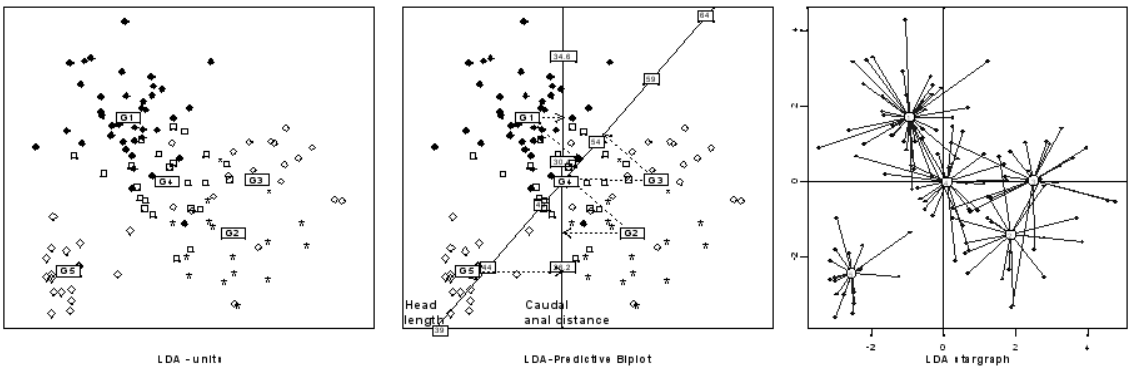


Figure 6: Thymallus data. LDA Predictive Biplot and LDA Stargraph.

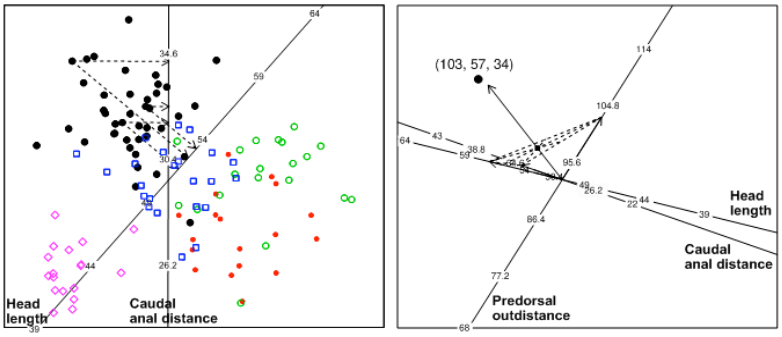


Figure 7: Thymallus data. LDA Predictive and Interpolative Biplots.

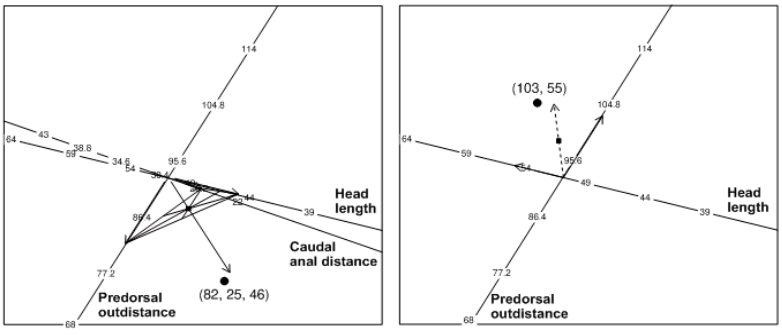


Figure 8: Thymallus data. LDA Interpolative Biplots.

Finally, Figure 7 (right hand) and Figure 8 show how given samples with coordinates $(103, 57, 34)$, $(82, 25, 46)$, and $(103, 55)$, respectively, are interpolated in the two-

dimensional space by the vector sum method with a good approximation (exact positions are indicated by full dots).

References

Alves, R. and Oliveira, M. (2003). Interpolative biplots applied to principal component analysis and canonical correlation analysis. *Journal of Chemometrics*, 17:594–602.

Bretherton, C., Smith, C., and Wallace, J. (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, 5:541–560.

Burnham, A., Viveros, R., and MacGregor, J. (1996). Frameworks for latent variable multivariate regression. *Journal of Chemometrics*, 10(31-45).

Chen, J., Matsunaga, R., Ishikawa, K., and H., Z. (2003). Main inorganic component measurement of seawater using near-infrared spectroscopy. *Applied Spectroscopy*, 57(11):1399–1406.

Cherry, S. (1996). Singular value decomposition analysis and canonical correlation analysis. *Journal of Climate*, 9:2003–2009.

Cherry, S. (1997). Some comments on singular value decomposition analysis. *Journal of Climate*, 10(7):1759–1761.

Chessel, D. (2008). Extensions de l'analyse discriminante linéaire.

Chessel, D. and Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environement. In Lebreton, J. and B., A., editors, *Biométrie et Environnement*, pages 15–44. Masson, Paris.

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

Dray, S., Chessel, D., and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological tables. *Ecology*, 84:3078–3089.

Gower, J. and Hand, D. (1996). *Biplots*. Chapman and Hall: London.

Höskuldsson, A. (1988). Partial least square regression methods. *Journal of Chemometrics*, 2:211–228.

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

Manly, B. (2005). *Multivariate Statistical Methods. A primer*. Chapman and Hall/Crc.

McKechnie, S., Ehrlich, P., and White, R. (1975). Population genetics of euphydryas butterflies. i. genetic variation and the neutrality hypothesis. *Genetics*, 81:571–594.

Persat, H. (1978). Ecologie de l'ombre commun. *Bulletin de la Revue Francaise de Pisciculture*, 26:11–20.

Prohaska, J. (1976). A technique for analyzing the linear relationships between two meteorological fields. *Monthly Weather Review*, 104:1345–1353.

Sabatier, R., Chessel, D., and Maury, L. (1992). Comment mesurer la concordance entre jugements subjectifs et observations multivariables pour des produits alimentaires. In *Agro-Industrie et méthodes statistiques. Compte-rendu des 3èmes journées*

européennes. Montpellier 30/11-1/12, 1992, pages 103–106. Association pour la Statistique et ses Utilisations, Paris.

Sabatier, R., Vivien, M., and Amenta, P. (2003). Two approaches for discriminant partial least squares. In Schader, M., Gaul, W., and Vichi, M., editors, *Between Data Science and Applied Data Analysis*, pages 100–108. Springer-Verlag, Berlin.

Tenenhaus, M. and Young, F. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling and other methods for quantifying categorial multivariate data. *Psychometrika*, 50:90–104.

Thioulouse, J., Chessel, D., Dolédec, S., and Olivier, J. (1997). Ade-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7:75–83.

Tucker, L. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136.

Wallace, J., Smith, C., and Bretherton, C. (1992). Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *Journal of Climate*, 5:561–576.