General overview of methods of analysis of multi-group datasets

Aida Eslami*, El Mostafa Qannari**
Achim Kohler***, Stéphanie Bougeard*

*French Agency for Food, Environmental and Occupational Health Safety,
BP53, F-22440, Ploufragan, France
aida.eslami@anses.fr, stephanie.bougeard@anses.fr,
http://www.anses.fr
** LUNAM University, ONIRIS, Sensometrics and Chemometrics Laboratory,
Nantes, F-44307, France; INRA, Nantes, F-44307, France
elmostafa.qannari@oniris-nantes.fr
http://www.oniris-nantes.fr
*** Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences
and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway
achim.kohler@umb.no

Abstract. Methods of analysis of a dataset where the individuals are partitioned into groups are discussed. These methods encompass known strategies of analysis and a new method called dual generalized Procrustes analysis. The emphasis is put on how the methods used in the context of multi-block data analysis can be adapted to the present context of multi-group setting. The similarities and the differences between the various approaches of analysis are highlighted and illustrated on the basis of three datasets.

1 Introduction

Very often, it occurs that the same J variables are measured on a set of individuals partitioned in M groups. We shall refer to this setting as multi-group datasets. In order to investigate the structure of the data in the groups, principal components analysis (PCA) (Jolliffe, 2002), which is an extensively used tool for the reduction of the dimensionality in multivariate analysis, can be performed on each group separately. Clearly, this strategy of analysis yields a large number of parameters which is likely to lead to an instability problem of the solution because of a lack of sufficient data to accurately estimate all the parameters. Moreover, this strategy of analysis entails a difficulty in interpreting the outcomes and in comparing the results across the groups. It is also possible to perform PCA on the concatenated dataset where the rows refer to the individuals from all the groups. However, in this case the total variance recovered by the principal components mix up both the between and within-group variances.

In order to counteract these problems, several procedures have been proposed using more parsimonious models than separate PCA on the M groups. For instance, common principal

components analysis (Flury, 1984) is defined as a generalization of PCA to the case of multi-group setting. This consists in considering the variance-covariance matrices associated to the groups and seeking common orthogonal vectors of loadings associated with the components in the groups. However, the determination of the common vectors of loadings which is based on maximum likelihood estimation leads to a complex algorithm which is time consuming and whose convergence is not granted.

The aim of this paper is to set up a general framework for the determination of the common vectors of loadings. For this purpose, several strategies are proposed: (i) common principal components analysis, (ii) determination of a common variance-covariance matrix (multi-group principal components analysis, dual multiple factor analysis, dual STATIS), (iii) dual generalized Procrustes analysis, (iv) stepwise determination of common vectors of loadings (between-groups comparison, multi-group PCA retrieved). These strategies typify the main multi-group methods available in the literature. They are compared on the basis of three real datasets.

2 Methods

2.1 Data and notations

Matrices are denoted by upper case bold letters (*e.g.*, $\mathbf{A}$) and column vectors are denoted by lower case bold letters (*e.g.*, $\mathbf{a}$). As stated above, the dataset $\mathbf{X}$ consists in the measurements of J variables on N individuals. Moreover, this dataset is *a priori* divided into M groups $(X_1, \ldots, X_M)$. Each group refers to N_m individuals ($\sum_{m=1}^{M} N_m = N$). We assume that each group $\mathbf{X}_m$ is column centered, therefore the variance-covariance matrix of group m is given by $\mathbf{V}_m = \frac{1}{N_m}\mathbf{X}_m^T\mathbf{X}_m$. Where the superscript ^{T} denotes the matrix transpose operation. The vector $\mathbf{a}^{(h)}$ is the common vector of loadings associated with dimension $h = (1, \ldots, H)$ where H =rank($\mathbf{X}$). The matrix $\mathbf{A}$ is the matrix of common loadings given by $\mathbf{A} = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)}]$. As we will see, it may be useful to exhibit a group vector of loadings $\mathbf{a}_m^{(h)}$ associated with group m and dimension h. The group vector of loadings $\mathbf{a}_m^{(h)}$ is assumed to lie in the space spanned by the rows of matrix $\mathbf{X}_m$. The group component $\mathbf{t}_m^{(h)} = \mathbf{X}_m\mathbf{a}^{(h)}$ is the principal component in group m associated with the common vector of loadings $\mathbf{a}^{(h)}$ $(h = 1, \ldots, H)$. The graphical display in Figure 1 depicts all these elements.

2.2 Common principal components analysis

In order to set up more clearly the aim of the study and introduce the notations that will be used throughout this paper, we find it useful to elaborate on Flury's common principal components analysis, called CPCA (Flury, 1984). Flury's CPCA model is expressed in terms of the variance-covariance matrices associated with the M groups as follows:

$$\mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^T \quad \text{with } \mathbf{A}^T\mathbf{A} = \mathbf{I} \text{ (identity matrix)}, \text{ for } m = (1, \ldots, M) \tag{1}$$

where the matrix $\mathbf{A} = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)}]$ contains the vectors of loadings which are assumed to be common to the various groups and the diagonal matrix $\mathbf{\Lambda}_m$ is supposed to contain the variances associated with the group components $\mathbf{t}_m^{(h)} = \mathbf{X}_m\mathbf{a}^{(h)}$ for dimension $h = (1, \ldots, H)$. Thus, CPCA stipulates that the vector of loadings, assumed to be orthogonal, are common to the

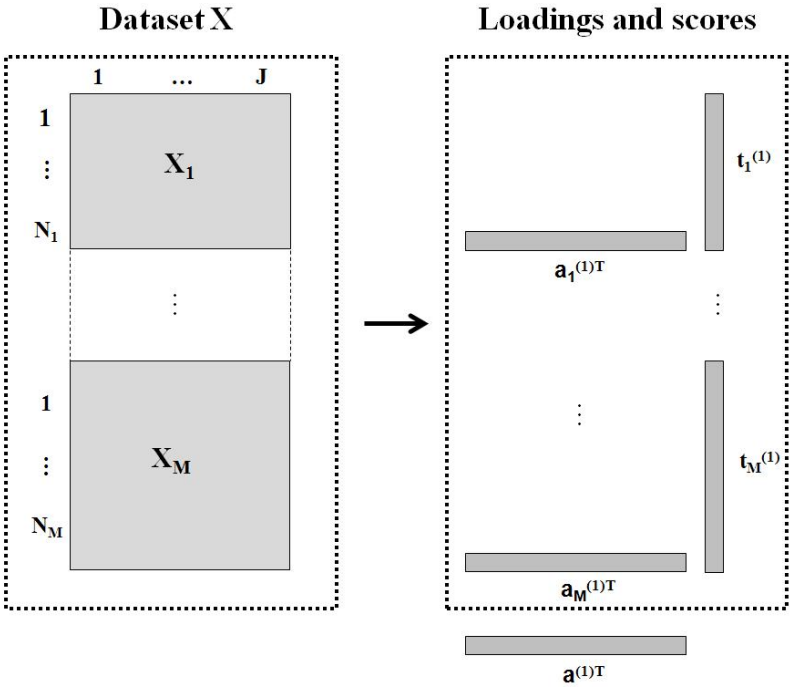

FIG. 1: Graphical display of the common vector of loadings ($\mathbf{a}^{(1)}$), the group vectors of loadings ($\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_M^{(1)}$) and the group components ($\mathbf{t}_1^{(1)}, \ldots, \mathbf{t}_M^{(1)}$).

various groups but the variances associated to the group principal components are specific to each group. For determining the CPCA parameters (namely $\mathbf{A}$ and $\mathbf{\Lambda}_m$), Flury considers a multinormal setting and uses maximum likelihood estimation. This leads to the so-called F-G algorithm (Flury and Gautschi, 1986). The appealing feature of this strategy of analysis is that it makes it possible to set up a hypothesis testing framework. However, the assumption of multinormal setting may be questionable in many situations and, moreover, the algorithm is complex, time consuming and may have some convergence problems. In the following, we discuss simpler alternatives to CPCA method.

2.3 Common variance-covariance matrix

2.3.1 Multi-group principal components analysis

The strategy of analysis, called multi-group principal components analysis (MGPCA) proposed by Krzanowski (1984) is simpler and more straightforward than Flury's CPCA. Indeed, Krzanowski (1984) remarks that if $\mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^T$ stands for all m then it also stands for the following linear combination of the variance-covariance matrices ($\mathbf{V}_1, \ldots, \mathbf{V}_M$):

$$\sum_{m=1}^{M} \frac{N_m}{N}\mathbf{V}_m = \sum_{m=1}^{M} \frac{N_m}{N}\mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^T = \mathbf{A}(\sum_{m=1}^{M} \frac{n_m}{N}\mathbf{\Lambda}_m)\mathbf{A}^T \quad \text{with } \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{2}$$

Therefore, the matrix of common loadings $\mathbf{A}$ could be derived from the eigenanalysis of the matrix $\mathbf{V}_W = \sum_{m=1}^{M} \frac{N_m}{N} \mathbf{V}_m$ which is referred to in the literature as the within groups variance-covariance matrix. From this standpoint the matrix $\mathbf{V}_W$ can be seen as a common variance-covariance matrix to the various groups because it is the closest matrix to $(\mathbf{V}_1, \ldots, \mathbf{V}_M)$, in the sense that it minimizes the following criterion:

$$\min_{\mathbf{V}_c} \sum_{m=1}^{M} N_m \parallel \mathbf{V}_m - \mathbf{V}_c \parallel^2 \tag{3}$$

As a summing up, the strategy of analysis proposed by Krzanowski (1984) consists in computing $\mathbf{V}_W$, the within groups variance-covariance matrix. The common vectors of loadings $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)})$ are the eigenvectors of $\mathbf{V}_W$. The specific variances of group m are determined as $\lambda_m^{(h)} = (\mathbf{a}^{(h)})^T \mathbf{V}_m \mathbf{a}^{(h)}$ for $h = (1, \ldots, H)$.

As a matter of fact, Krzanowski (1984) remarks that if $\mathbf{V}_m = \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^T$ stands for all m, then it stands not only for $\mathbf{V}_W$, but also for any linear combination of $\mathbf{V}_m$ as shown in Eq. (4).

$$\sum_{m=1}^{M} \alpha_m \mathbf{V}_m = \sum_{m=1}^{M} \alpha_m \mathbf{A}\mathbf{\Lambda}_m\mathbf{A}^T = \mathbf{A}(\sum_{m=1}^{M} \alpha_m \mathbf{\Lambda}_m)\mathbf{A}^T \quad \text{with } \alpha_m \geq 0 \tag{4}$$

From this standpoint we may seek a common variance-covariance matrix as a solution to other optimization criteria than the problem stated above in Eq. (3). This is illustrated by means of dual STATIS which is discussed in the next section.

It is worth noting that MGPCA amounts to performing PCA on the matrix $\mathbf{X}$ obtained by stacking the group datasets $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ one above the other. Indeed, it is easy to check that in this case, we are led to the eigenanalysis of matrix $\frac{1}{N}\mathbf{X}^T\mathbf{X} = \sum_{m=1}^{M} \frac{N_m}{N} \mathbf{V}_m = \mathbf{V}_W$. The idea of vertically merging the group datasets and performing a PCA of the matrix thus obtained stands at the root of the method of analysis introduced by Lê et al. (2010) called dual multiple factor analysis.

2.3.2 Dual STATIS

As an alternative criterion to the problem stated in Eq. (3), we propose to seek a common variance-covariance matrix $\mathbf{V}_c$ which is a solution to problem (5):

$$\min_{\mathbf{V}_c, \alpha_1, \ldots, \alpha_M} \sum_{m=1}^{M} \parallel \alpha_m \mathbf{V}_m - \mathbf{V}_c \parallel^2 \quad \text{with} \quad \sum_{m=1}^{M} \alpha_m^2 = 1 \tag{5}$$

This problem is known as dual STATIS (Lavit et al., 1994). Indeed, STATIS is a popular method in the field of multi-block data analysis which seeks a common configuration to several datasets measured on the same individuals. But when these datasets pertain to the same variables instead of the same individuals, it is referred to as dual STATIS (DSTATIS). The solution to problem (5) is given by the compromise variance-covariance matrix $\mathbf{V}_c = \sum_{m=1}^{M} \alpha_m \mathbf{V}_m$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$ is the eigenvector of the matrix $\mathbf{R} = (r_{ik})$ associated with the largest eigenvalue, where r_{ik} =trace$(\mathbf{V}_i\mathbf{V}_k)$ for $(i, k = 1, \ldots, M)$, where 'trace' stands, for a a square matrix, for the sum of the elements on the diagonal. Thereafter, the spectral decomposition of $\mathbf{V}_c$ can be computed as $\mathbf{V}_c = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$, where $\mathbf{A} = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)}]$ will stand for

the matrix of common vectors of loadings and $\mathbf{\Lambda}$ is a diagonal matrix. The specific variances of group m are computed as $\lambda_m^{(h)} = (\mathbf{a}^{(h)})^T \mathbf{V}_m \mathbf{a}^{(h)}$ for dimensions $h = (1, \ldots, H)$.

The interest of DSTATIS over MGPCA is that it takes account of the similarities between the variance-covariance matrices of the groups. In other words, if the structure of one group is different from the other groups, then the contribution of this group to the determination of the common variance-covariance matrix, $\mathbf{V}_c$, will be minimized, *i.e.*, the associated variance-covariance matrix is downweighted because its associated coefficient α_m is relatively small. This is known as an interesting feature of STATIS and DSTATIS (Lavit et al., 1994).

2.4 Dual generalized Procrustes analysis

As stated above, the rationale behind MGPCA and DSTATIS is to compute a variance-covariance matrix common to the various groups. Another way to stipulate the same assumption stems from Property 1. Suppose that $\mathbf{X}_1$ $(n \times J)$ and $\mathbf{X}_2$ $(n \times J)$ are two centered datasets which refer to the same J variables but not necessarily to the same individuals. Without any loss of generality these two datasets are assumed to have the same number of individuals. If this is not the case then the dataset with the smallest number of rows can be augmented with the necessary number of rows containing zeroes. The following property holds.

Property 1 The equality $\mathbf{X}_1^T \mathbf{X}_1 = \mathbf{X}_2^T \mathbf{X}_2$ holds if and only if $\mathbf{X}_1^T = \mathbf{X}_2^T \mathbf{H}$ where $\mathbf{H}$ is an orthogonal matrix.

Indeed if $\mathbf{X}_1^T = \mathbf{X}_2^T \mathbf{H}$ then $\mathbf{X}_1^T \mathbf{X}_1 = \mathbf{X}_2^T \mathbf{H}\mathbf{H}^T \mathbf{X}_2 = \mathbf{X}_2^T \mathbf{X}_2$. Conversely, suppose that $\mathbf{X}_1^T \mathbf{X}_1 = \mathbf{X}_2^T \mathbf{X}_2$, then the singular value decomposition of $\mathbf{X}_1^T$ and $\mathbf{X}_2^T$ can be expressed as $\mathbf{X}_1^T = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Q}_1^T$ and $\mathbf{X}_2^T = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Q}_2^T$, where $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{X}_1^T \mathbf{X}_1 = \mathbf{X}_2^T \mathbf{X}_2$ associated with the eigenvalues in the diagonal matrix $\mathbf{\Lambda}$. It follows that $\mathbf{X}_1^T = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Q}_1^T = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Q}_2^T\mathbf{Q}_2\mathbf{Q}_1^T = \mathbf{X}_2^T\mathbf{H}$, with $\mathbf{H} = \mathbf{Q}_2\mathbf{Q}_1^T$ which is an orthogonal matrix.

It is worth noting that a similar property pertaining to the case where the two datasets at hand refer to the same individuals is proven by Glaçon (1981).

The implication of this property in the multi-group setting is that instead of seeking a common variance-covariance matrix to $\mathbf{V}_m = \frac{1}{N_m}\mathbf{X}_m^T\mathbf{X}_m = (\frac{1}{\sqrt{N_m}}\mathbf{X}_m)^T(\frac{1}{\sqrt{N_m}}\mathbf{X}_m)$, one could look for a dataset that would be some average of groups $\frac{1}{\sqrt{N_m}}\mathbf{X}_m^T$ through orthogonal transforms. This can be achieved by means of generalized Procrustes analysis (GPA) (Gower, 1975). We shall refer to this strategy of analysis as dual GPA (DGPA) as it is based on $\mathbf{X}_m^T$ instead of $\mathbf{X}_m$. Formally, we seek to minimize the following criterion (6):

$$\sum_{m=1}^{M} \| \frac{1}{\sqrt{N_m}}\mathbf{X}_m^T\mathbf{H}_m - \mathbf{C}\|^2 \tag{6}$$

where $\frac{1}{\sqrt{N_m}}\mathbf{X}_m^T$ is orthogonally transformed towards the common matrix $\mathbf{C}$ by $\mathbf{H}_m$, the orthogonal matrix associated with group m. This optimization problem can be solved by one of the several algorithms for GPA (Gower, 1975; Ten Berge, 1977). Once $\mathbf{C}$ is determined, the common vectors of loadings $\mathbf{A} = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)}]$ are calculated as the left singular vectors of $\mathbf{C}$. The specific variances of group m are given as usual by $\lambda_m^{(h)} = (\mathbf{a}^{(h)})^T \mathbf{V}_m \mathbf{a}^{(h)}$.

2.5 Stepwise determination of common vectors of loadings

2.5.1 Between-groups comparison

Instead of determining, in a first step, a variance-covariance matrix common to the various groups and, in a second step, determining the common vectors of loadings through the spectral decomposition of this matrix, we adopt a strategy which consists in directly defining, step by step, the common vectors of loadings. The aim is to seek $\mathbf{a}^{(1)}$, a common vector of loadings for the first dimension ($h = 1$), and group vectors of loadings respectively associated to the various groups, namely $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_M^{(1)})$ so as to maximize a criterion which reflects the proximity (or similarity) of the group vectors of loadings $(\mathbf{a}_1^{(1)}, \ldots, \mathbf{a}_M^{(1)})$ to the common vector of loadings $\mathbf{a}^{(1)}$. For instance, we consider criterion (7):

$$\sum_{m=1}^{M} N_m < \mathbf{a}_m^{(1)}, \mathbf{a}^{(1)} >^2 = \sum_{m=1}^{M} N_m ((\mathbf{a}^{(1)})^T \mathbf{a}_m^{(1)})^2 \quad \text{with} \quad \| \mathbf{a}_m^{(1)} \| = \| \mathbf{a}^{(1)} \| = 1 \tag{7}$$

or equivalently:

$$\sum_{m=1}^{M} N_m \cos^2(\mathbf{a}_m^{(1)}, \mathbf{a}^{(1)}) \tag{8}$$

where $\cos(.,.)$ stands for the cosine between two vectors.

In order to solve this problem, we firstly can recall that a vector of loadings associated with group m is, as stated above, supposed to be a linear combination of the columns of matrix $\mathbf{X}_m^T$: $\mathbf{a}_m^{(1)} = \mathbf{X}_m^T \mathbf{t}_m^{(1)}$, where $\mathbf{t}_m^{(1)}$ is a vector of dimension N_m. Therefore, in criterion (7) (or equivalently (8)), if we assume that $\mathbf{a}^{(1)}$ is fixed, the optimal solution for vector $\mathbf{a}_m^{(1)}$ is given by $\mathbf{a}_m^{(1)} = \frac{\mathbf{P}_m \mathbf{a}^{(1)}}{\|\mathbf{P}_m \mathbf{a}^{(1)}\|}$ where $\mathbf{P}_m = \mathbf{X}_m^T (\mathbf{X}_m \mathbf{X}_m^T)^{-1} \mathbf{X}_m$, the projector upon the subspace spanned by the columns in $\mathbf{X}_m^T$. So far, we assume that $\mathbf{X}_m \mathbf{X}_m^T$ is non-singular but we will discuss below how to proceed if this matrix is singular. If we replace this expression in criterion (7), we are led to maximizing with respect to $\mathbf{a}^{(1)}$ the following expression:

$$\sum_{m=1}^{M} N_m (\mathbf{a}^{(1)})^T \mathbf{P}_m \mathbf{a}^{(1)} = (\mathbf{a}^{(1)})^T \sum_{m=1}^{M} N_m \mathbf{P}_m \mathbf{a}^{(1)} \quad \text{with} \quad \| \mathbf{a}^{(1)} \| = 1 \tag{9}$$

It follows that the optimal solution is given by setting $\mathbf{a}^{(1)}$ equal to the eigenvector of $\sum_{m=1}^{M} N_m \mathbf{P}_m$ associated with the largest eigenvalues.

Subsequent vectors of loadings $(\mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(H)})$ could be sought by considering the same maximization problem and adding constraints of orthogonality of the vector of loadings to be determined at the current stage with those determined at previous stages. As a matter of fact, this leads to setting the common vectors of loadings to the successive eigenvectors of matrix $\sum_{m=1}^{M} n_m \mathbf{P}_m$. It is clear that problem (7) is related to generalized canonical correlation analysis of the M spaces spanned by the columns of $(\mathbf{X}_1^T, \ldots, \mathbf{X}_M^T)$, respectively (Carroll, 1968). The difference of the strategy of analysis followed herein and the usual setting of generalized canonical correlation analysis is that we consider the subspaces spanned by the rows of matrices $\mathbf{X}_m$ $(m = 1, \ldots, M)$ whereas in generalized canonical correlation, we consider the subspaces spanned by the columns.

As stated above, the solution to the optimization problem (7) assumes that the matrices $\mathbf{X}_m\mathbf{X}_m^T$ are invertible. Generally, this assumption does not hold. For instance, for those groups where the number of variables is smaller than the number of individuals, the $(N_m \times N_m)$ matrices $\mathbf{X}_m\mathbf{X}_m^T$ are of rank less than N_m and therefore not invertible. In order to circumvent this problem, we propose to approach by using Eckart and Young (1936) theorem each matrix $\mathbf{X}_m^T$ by a matrix $(\mathbf{X}_m^*)^T$ of rank k ($k < N_m$) such that the matrices $\mathbf{X}_m^*(\mathbf{X}_m^*)^T$ are invertible. Formally, this can be written as $(\mathbf{X}_m^*)^T = \mathbf{L}_m^{(k)}\mathbf{\Lambda}_m^{(k)}(\mathbf{Q}_m^{(k)})^T$ (*i.e.*, singular value decomposition of order k). It follows that $\sum_{m=1}^M N_m(\mathbf{X}_m^*)^T(\mathbf{X}_m^*(\mathbf{X}_m^*)^T)^{-1}\mathbf{X}_m^* = \sum_{m=1}^M N_m\, \mathbf{L}_m^{(k)}(\mathbf{L}_m^{(k)})^T$. This means that apart from the weighting by the size of the various groups, this strategy of analysis amounts to the method of analysis which was proposed by Krzanowski (1979) under the appellation "between-groups comparison of principal components". In the following, we shall refer to this method of analysis as BGC which stands for between-groups comparison.

2.5.2 Multi-group PCA retrieved

We consider the same objective function as in the previous section Eq. (7) but this time we impose other constraints on the parameters involved. Namely we aim at maximizing:

$$\sum_{m=1}^{M} N_m < \mathbf{a}_m^{(1)}, \mathbf{a}^{(1)} >^2 \quad \text{with} \quad \mathbf{a}_m^{(1)} = \mathbf{X}_m^T\mathbf{t}_m^{(1)} \text{ and } \parallel \mathbf{a}^{(1)} \parallel = \parallel \mathbf{t}_m^{(1)} \parallel = 1 \tag{10}$$

This amounts to maximizing under the specified constraints:

$$\sum_{m=1}^{M} N_m((\mathbf{a}^{(1)})^T\mathbf{X}_m^T\mathbf{t}_m^{(1)})^2 \tag{11}$$

For a fixed vector $\mathbf{a}^{(1)}$, the maximum is achieved by setting $\mathbf{t}_m^{(1)} = \frac{\mathbf{X}_m\mathbf{a}^{(1)}}{\|\mathbf{X}_m\mathbf{a}^{(1)}\|}$. Replacing this expression in Eq. (11), we are led to maximizing:

$$\sum_{m=1}^{M} N_m(\mathbf{a}^{(1)})^T\mathbf{X}_m^T\mathbf{X}_m\mathbf{a}^{(1)} = (\mathbf{a}^{(1)})^T \sum_{m=1}^{M} N_m\mathbf{X}_m^T\mathbf{X}_m\mathbf{a}^{(1)} \tag{12}$$

It follows that $\mathbf{a}^{(1)}$ is an eigenvector of matrix $\sum_{m=1}^M N_m\mathbf{X}_m^T\mathbf{X}_m$ associated with the largest eigenvalue. Subsequent common vectors of loadings could be sought by considering the same optimization problem and adding orthogonality constraints of the common vectors of loadings. It follows that we are led to the same solution as multi-group PCA (MGPCA, section 2.3.1).

All these developments make it possible to highlight a striking difference between the method BGC (between-groups comparison) developed in the previous section and MGPCA. It is clear that whereas in MGPCA we aim at recovering the total variance in the various groups, in BGC the main purpose is to find similar patterns no matter whether these patterns are linked to directions of high saliency (*i.e.*, total variance) or not. Indeed, in MGPCA, we aim at maximizing (see Eq. (12)),

$$\sum_{m=1}^{M} N_m(\mathbf{a}^{(1)})^T\mathbf{X}_m^T\mathbf{X}_m\mathbf{a}^{(1)} = \sum_{m=1}^{M} N_m\lambda_m^{(1)} \tag{13}$$

where $\lambda_m^{(1)}$ is the variance of the principal component $\mathbf{t}_m^{(1)} = \mathbf{X}_m\mathbf{a}^{(1)}$. In BGC analysis, it is clear from criterion (8) that the aim is to find common patterns regardless of their importance in terms of the total variance recovered. One can draw a parallel between this situation and generalized canonical correlation (Carroll, 1968) as compared to inter-battery analysis (Tucker, 1958), consensus PCA (Wold et al., 1987) or co-inertia analysis (Hanafi et al., 2011).

2.6 Comparison of methods

The various strategies of determining common vectors of loadings in a multi-group setting can be divided into two families. In the first family, we find CPCA, MGPCA, DSTATIS and DGPA. The second family singles out BGC analysis. The methods in the first family aim at recovering the total variance in the various data blocks associated with the groups whereas BGC analysis aims at unveiling common patterns to the data blocks. This means, in particular, that in this latter method, the first common vector of loadings may not be associated to directions with relatively large total variances. Within the first family of methods, it is clear that MGPCA is the simplest and most straightforward method of determination of the common vectors of loadings. DSTATIS presents the advantage of taking account of the structure of the data in the various groups and downweighting those blocks of data which do not agree with the general pattern. As stated in the introduction CPCA, being based on the assumption of a multinormal setting and maximum likelihood estimation, makes it possible to set up a hypothesis testing framework. However, the assumption of multinormality may not be granted in some situations and, in such situations, CPCA may lead to some discrepancies. Another situation which is worth discussing concerns the case where the variables at hand are not in the same measurement unit. Obviously, a standardization is needed. As stated above, Lê et al. (2010) discussed a method of analysis called dual multiple factor analysis where they advocate to systematically standardize the variables in each dataset $\mathbf{X}_m$ to unit variance. We have already stressed the connection of this method of analysis with MGPCA (section 2.3.1).

The methods of analysis discussed in this paper are compared on the basis of three datasets. We aim at highlighting the similarities and differences between the methods in terms of how similar the common vectors of loadings obtained by the various methods are, and in terms of the total variances in the different groups recovered by the successive components.

For the comparison of the common vectors of loadings between methods, we set up, for each pair of methods, a sequence of indices indexed by the number of components retained in the model. Each index ranges between 0 and 1 and reflects the extent to which the common vectors of loadings obtained by the two methods at hand are similar. More precisely, let $\mathbf{A} = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(H)}]$ and $\mathbf{A}^* = [\mathbf{a}^{*(1)}, \ldots, \mathbf{a}^{*(H)}]$ be the matrices of common loadings associated with two methods to be compared. We will investigate whether these methods lead to similar vectors of loadings up to a given dimension h for $h = (1, \ldots, H)$. For this purpose we consider the sequence of similarity indices given by Eq. (14).

$$S^{(h)} = \frac{1}{h}\sum_{r=1}^{h} |(\mathbf{a}^{(r)})^T\mathbf{a}^{*(r)}| = \frac{1}{h}\sum_{r=1}^{h} |\cos(\mathbf{a}^{(r)}, \mathbf{a}^{*(r)})| \quad \text{for } h = (1, \ldots, H) \tag{14}$$

By considering the absolute value $|(\mathbf{a}^{(r)})^T\mathbf{a}^{*(r)}|$, we take account of the fact that the orientation of a given vector of loadings is arbitrary. The indices $S^{(h)}$ range between 0 and 1, the value 1 being reached in case of perfect agreement of the vectors of loadings up to dimension h.

Contrariwise, this index takes the value 0 if $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(h)})$ are respectively orthogonal to $(\mathbf{a}^{*(1)}, \ldots, \mathbf{a}^{*(h)})$.

The methods are also compared on the basis of the total variances in the various groups recovered by the successive principal components which are given by:

$$\lambda_m^{(h)} = (\mathbf{a}^{(h)})^T \mathbf{V}_m \mathbf{a}^{(h)} \text{ for } h = (1, \ldots, H) \tag{15}$$

or equivalently the percentages of total variance recovered by the principal components:

$$I_m^{(h)} = \frac{\lambda_m^{(h)}}{\text{trace}(\mathbf{V}_m)} \text{ for } h = (1, \ldots, H), \ \ m = (1, \ldots, M) \tag{16}$$

In situations where the aim is to recover as much variation in the dataset as possible, these percentages should be large.

2.7 Application

2.7.1 Iris data

As a first illustration, we consider the iris flower dataset introduced by Fisher (1936). The data are available in the R 'datasets' package. They consist in $(N = 150)$ iris specimens sampled from $(M = 3)$ species: Setosa, Virginica and Versicolor, each group containing $(N_m = 50)$ individuals. Four variables $(J = 4)$ are measured on each flower namely, the length and the width of sepals and petals. Table 1 shows the similarity indices $S^{(h)}$ between all the methods under study for the dimensions $(h = 1$ and $2)$. The results corresponding to the subsequent dimensions $(h = 3$ and $4)$ are not shown herein since they do not show any specific pattern in comparison to the first two dimensions. As a general remark we can see (Table 1) that in spite of the conceptual differences among the methods, the similarity of the results is very high. By considering the average similarity values for the first dimension $(h = 1)$, it can be seen that DSTATIS and CPCA are the methods which give the most similar results in comparison with all other methods. BGC analysis shows the least agreement with the other methods. For the second dimension $(h = 2)$, DGPA and MGPCA appear to be the methods which lead to the most similar results to the other methods. Similarly to the first dimension, BCG analysis is the method with the least agreement with the others.

Table 2 shows that the percentages of total variance recovered by the principal components in the various groups obtained by the different strategies of analysis under study. On the average, these percentages are close to each others although we can note that CPCA and BGC have, on the average, relatively smaller percentages of total variance for the first dimension and relatively larger percentages of variance recovered by the second dimension. For this latter dimension, we can note, in particular, the relatively large percentage of total variance in group Setosa recovered by the principal component obtained by means of CPCA.

2.7.2 Chemical composition of olive oil

The olive oil dataset comes from Forina et al. (1983). The data are available in the R 'pgmm' package. These data concern $(N = 572)$ Italian olive oils, sampled from $(M = 9)$ regions of Italy on which $(J = 8)$ fatty acid variables are measured. Table 3 shows the

The first dimension (h=1)					
	DGPA	DSTATIS	MGPCA	CPCA	BGC
DGPA	1.000				
DSTATIS	0.991	1.000			
MGPCA	0.998	0.997	1.000		
CPCA	0.990	1.000	0.997	1.000	
BGC	0.969	0.983	0.977	0.985	1.000
Average	**0.987**	**0.993**	**0.992**	**0.993**	**0.979**
The first two dimensions ($h = 1$ and 2)					
DGPA	1.000				
DSTATIS	0.976	1.000			
MGPCA	0.992	0.961	1.000		
CPCA	0.940	0.885	0.973	1.000	
BGC	0.934	0.977	0.912	0.833	1.000
Average	**0.961**	**0.950**	**0.960**	**0.908**	**0.914**

TAB. 1: Similarity indices $S^{(h)}$ for dimensions ($h = 1$ and 2). Results from the iris dataset. Abbreviations: dual generalized Procrustes analysis (DGPA), dual STATIS (DSTATIS), multi-group principal components analysis (MGPCA), common principal components analysis (CPCA), between-groups comparison (BGC).

The first dimension (h=1)					
Methods / Groups	DGPA	DSTATIS	MGPCA	CPCA	BGC
Setosa	56.4	47.9	52.5	47.5	45.2
Virginica	76.8	77.6	77.4	77.6	73.7
Versicolor	75.7	78.0	77.1	77.9	76.6
Average	**69.6**	**67.8**	**69.0**	**67.7**	**65.2**
The second dimension (h=2)					
Setosa	26.4	25.7	32.3	40.6	29.0
Virginica	10.0	10.1	9.3	8.8	13.7
Versicolor	13.0	11.9	11.4	8.5	11.5
Average	**16.5**	**15.9**	**17.7**	**19.3**	**18.1**
Cumulated total variance (h=1 and 2)	**86.1**	**83.7**	**86.7**	**87.0**	**83.3**

TAB. 2: Percentages of total variance recovered by the first two principal components. Results from the iris dataset.

similarity coefficients $S^{(h)}$ for the first two dimensions ($h = 1$ and 2). As for the previous case study, the remaining dimensions ($h = 3, \ldots, 8$) follow the same pattern and are not shown herein. The common loadings associated with all the methods under study are highly similar. For the first two dimensions ($h = 1$ and 2), the methods DGPA and MGPCA show, on the average, the highest similarity with the other methods, whereas BGC shows the least similarity with the other methods.

Table 4 shows the percentages of total variance in the various groups recovered by the principal components derived from the various strategies of analysis. The only notable difference between the methods is the relatively smaller percentage of total variance explained by the first principal components obtained by means of BGC analysis. However, this difference is counterbalanced by the second component obtained by BGC analysis which explains a rela-

The first dimension (h=1)					
	DGPA	DSTATIS	MGPCA	CPCA	BGC
DGPA	1.000				
DSTATIS	0.998	1.000			
MGPCA	1.000	0.999	1.000		
CPCA	1.000	0.997	1.000	1.000	
BGC	0.978	0.967	0.976	0.980	1.000
Average	**0.994**	**0.990**	**0.993**	**0.994**	**0.975**
The first two dimensions ($h = 1$ and 2)					
DGPA	1.000				
DSTATIS	0.998	1.000			
MGPCA	0.999	0.999	1.000		
CPCA	1.000	0.996	0.999	1.000	
BGC	0.967	0.960	0.965	0.966	1.000
Average	**0.991**	**0.988**	**0.990**	**0.990**	**0.965**

TAB. 3: Similarity indices $S^{(h)}$ for dimension ($h = 1$ and 2). Results from the olive oil dataset.
Abbreviations: dual generalized Procrustes analysis (DGPA), dual STATIS (DSTATIS), multi-group principal components analysis (MGPCA), common principal components analysis (CPCA), between-groups comparison (BGC).

tively larger amount of total variance than the second components from the other methods of analysis.

2.7.3 Veterinary epidemiological data

The epidemiological dataset comes from Rose et al. (2009) (data not yet made public). The aim of this longitudinal study is to describe a pig disease called Post-weaning Multisystemic Wasting Syndrome (PMWS). These data include ($N = 884,\ N_m \simeq 120$) randomly selected pigs sampled from ($M = 7$) farms on which ($J = 19$) variables are measured. These variables are related to the disease status, the animal performance and the farm structure. The variables being measured on very different scales, they are centered and scaled by group as advocated by Lê et al. (2010). This case study is interesting because unlike the previous two case studies, it shows more discrepancies among the different methods for multi-group data analysis. Table 5 shows the similarity indices between the vectors of loadings derived from the methods. It can be seen that whereas DGPA, DSTATIS and MGPCA seem to be in relatively high agreement, CPCA and BGC analysis show only a fair agreement between them and with the other methods.

The percentages of total variance (Table 6) in the various groups explained by the principal components derived from the different strategies of analysis corroborate the same observation. Indeed, the first family of methods (DGPA, DSTATIS, MGPCA) lead to more or less the same percentages of total variance explained by the first two principal components whereas both the first two principal components derived from CPCA and BGC analysis recover less variation in the various groups. We believe that these differences could be explained by the fact that by considering the correlation matrices in the various groups instead of the variance-covariance matrices, we depart from CPCA normal setting which as stated above assumes a multinormal distribution and derives a solution by maximum likelihood estimation which involves variance-

The first dimension (h=1)					
Methods / Groups	DGPA	DSTATIS	MGPCA	CPCA	BGC
North-Apulia	78.0	76.5	77.7	78.3	79.6
Calabria	82.4	81.1	82.1	82.5	82.3
South-Apulia	79.1	79.0	79.0	78.8	72.9
Sicily	48.2	47.8	48.1	48.2	47.6
Inland-Sardin	87.9	87.7	87.8	88.0	84.2
Coast-Sardini	90.1	90.6	90.2	89.9	84.6
Umbria	82.1	81.9	82.2	82.1	79.0
East-Liguria	28.9	28.8	29.0	28.5	24.5
West-Liguria	59.0	58.1	58.9	59.1	61.0
Average	**70.6**	**70.2**	**70.6**	**70.6**	**68.4**
The second dimension (h=2)					
North-Apulia	14.7	15.6	15.0	14.5	12.8
Calabria	10.7	11.4	10.9	10.5	11.1
South-Apulia	15.6	14.6	15.1	15.9	19.5
Sicily	33.3	32.5	32.9	33.6	31.5
Inland-Sardin	7.0	7.0	6.9	7.1	9.1
Coast-Sardini	7.5	7.5	7.6	7.6	11.5
Umbria	14.7	14.9	14.8	14.4	17.0
East-Liguria	28.8	28.0	28.1	28.4	39.9
West-Liguria	30.1	30.8	30.4	30.0	27.6
Average	**18.1**	**18.0**	**18.0**	**18.0**	**20.0**
Cumulated total variance (h=1 and 2)	**88.7**	**88.8**	**88.6**	**88.6**	**88.4**

TAB. 4: Percentages of total variance recovered by the first two principal components. Results from the olive oil dataset.

covariance matrices. In the case of BGC analysis, we believe that the difference of its outcomes with the other methods stem from the fact that we have highlighted above that is, this method of analysis, unlike MGPCA for instance, does not focus on recovering the variation in the various groups but is concerned by finding common patterns to the various groups.

3 Conclusion and perspectives

The extension of principal components analysis to multi-group setting makes it possible to handle the specificity of complex data in various fields. For the purpose of describing a dataset X with observations *a priori* divided into M groups, several methods are described. Most of them can be seen as adaptation of multi-block methods which are concerned with the analysis of several datasets pertaining to the same individuals to the case where the datasets pertain to the same variables (and not necessarily to the same individuals). From this standpoint, it follows that, yet, several other methods could be adapted to this latter context because the domain of multi-block data analysis has been very fecund these last two or three decades. This transfer of methodology from multi-block data analysis to multi-group data analysis is illustrated by dual generalized Procrustes analysis (DGPA) that we have introduced herein and which turned out to be in high agreement with more known method such as multi-group PCA (MGPCA). Not only multi-block data analysis offers a wide range of methods but it also

The first dimension (h=1)					
	DGPA	DSTATIS	MGPCA	CPCA	BGC
DGPA	1.000				
DSTATIS	0.967	1.000			
MGPCA	0.916	0.984	1.000		
CPCA	0.893	0.817	0.711	1.000	
BGC	0.823	0.678	0.582	0.771	1.000
Average	**0.900**	**0.862**	**0.798**	**0.798**	**0.713**
The first two dimensions ($h = 1$ and 2)					
DGPA	1.000				
DSTATIS	0.963	1.000			
MGPCA	0.909	0.983	1.000		
CPCA	0.646	0.521	0.452	1.000	
BGC	0.428	0.346	0.359	0.630	1.000
Average	**0.737**	**0.703**	**0.676**	**0.562**	**0.441**

TAB. 5: Similarity indices $S^{(h)}$ for dimension ($h = 1$ and 2). Results from the epidemiological dataset.
Abbreviations: dual generalized Procrustes analysis (DGPA), dual STATIS (DSTATIS), multi-group principal components analysis (MGPCA), common principal components analysis (CPCA), between-groups comparison (BGC).

The first dimension (h=1)					
Methods / Groups	DGPA	DSTATIS	MGPCA	CPCA	BGC
Farm1	14.0	13.5	12.8	13.8	12.6
Farm2	22.4	21.6	18.1	24.8	17.2
Farm3	17.9	18.0	17.1	16.0	15.4
Farm4	15.6	19.7	24.2	9.2	11.0
Farm5	8.7	7.3	6.3	9.6	10.8
Farm6	8.1	8.8	9.0	6.4	6.8
Farm7	14.2	13.2	13.3	13.7	16.0
Average	**14.4**	**14.6**	**14.4**	**13.4**	**12.8**
The second dimension (h=2)					
Farm1	13.8	13.9	13.3	10.3	15.0
Farm2	6.4	8.1	11.0	10.7	19.8
Farm3	14.5	13.6	12.9	11.6	9.6
Farm4	20.3	18.4	16.5	9.8	5.1
Farm5	7.7	9.0	10.1	10.0	11.1
Farm6	6.4	5.5	5.4	7.3	6.8
Farm7	14.2	15.0	14.9	7.8	9.2
Average	**11.9**	**11.9**	**12**	**9.6**	**10.9**
Cumulated total variance (h=1 and 2)	**26.3**	**26.5**	**26.4**	**23.0**	**23.7**

TAB. 6: Percentages of total variance recovered by the first two principal components. Results from the epidemiological dataset.

provides visualization and interpretation tools which are helpful to unveil the hidden structure in the data. Tools for diagnostic assessment are also of paramount importance. Adapted to the case of multi-group data analysis, these tools could, for instance, highlight some discrepancies

in the data such as groups of individuals which depart from the assumption of equality of the variance-covariance matrices.

Another interest of the present research work is to offer a bird's eye view of the different methods of analysis of multi-group datasets and highlight their similarities and differences. From this standpoint, it seems that between-groups comparison (BGC) can be singled out because its main concern is not to recover the variation in the various groups but to find common patterns to these groups. We also highlighted the specificity of Flury's common principal components analysis (CPCA). Indeed, this method of analysis being based on the assumption of multinormality may exhibit some discrepancies if the data show a serious departure from normality. The remaining methods, namely DGPA, DSTATIS and MGPCA led to very similar results in the three case studies discussed herein and, judging from their common rationale which is reflected by the objective functions which are optimized, we believe that this should be always the case. Among these methods, a particular emphasis should be put on MGPCA because it is simple and straightforward.

The present research work can be extended in several ways. For instance, we could consider the case of multi-block and multi-group setting where we have more than one dataset which are partitioned into several groups. With this setting in view, we could also investigate other aspects than merely describing the structure of the data. For instance, we could investigate predictive models which take account of the presence of groups in the various multi-block datasets.

Acknowledgements

The authors gratefully acknowledge the Brittany Region for its financial contribution to this research project (ARED funding for PhD).

References

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 67th Annual Convention of the American Psychological Association*, 227–228.

Eckart, C. H. and G. Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika 1*, 211–218.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 179–188.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association 79*, 892–898.

Flury, B. N. and W. Gautschi (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM J. Sci. Stat. Comput. 7*, 169–184.

Forina, M., C. Armanino, S. Lanteri, and E. Tiscornia (1983). *Classification of olive oils from their fatty acid composition*, pp. 189–214. In Martens H. and Russwurm Jr. H., editors, Food Research and Data Analysis. London: Applied Science Publishers.

Glaçon, F. (1981). *Analyse conjointe de plusieurs matrices de données. Comparaison de différentes méthodes*. Phd thesis, University of Grenoble.

Gower, J. (1975). Generalized procrustes analysis. *Psychometrika 40*, 33–51.

Hanafi, M., A. Kohler, and M. Qannari (2011). Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and Intelligent Laboratory Systems 106*, 37–40.

Jolliffe, I. T. (2002). *Principal component analysis*. New York: Springer.

Krzanowski, W. J. (1979). Between-groups comparison of principal components. *Journal of the American statistical Association 74*, 703–707.

Krzanowski, W. J. (1984). Principal component analysis in the. presence of group structure. *Applied Statistics 33*, 164–168.

Lê, S., F. Husson, and J. Pagès (2010). Dmfa: Dual multiple factor analysis. *Communication in Statistics - Theory and Methods 39*, 483–492.

Lavit, C., Y. Escoufier, R. Sabatier, and P. Traissac (1994). The act (statis method). *Computational Statistics and Data Analysis 18*, 97–117.

Rose, N., E. Eveno, B. Grasland, A. C. Nignol, A. Oger, A. Jestin, and F. Madec (2009). Individual risk factors for post-weaning multisystemic wasting syndrome (pmws) in pigs: A hierarchical bayesian survival analysis. *Preventive Veterinary Medicine 90*, 168–179.

Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika 42*, 267–276.

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

Wold, S., S. Hellberg, T. Lundstedt, M. Sjostrom, and H. Wold (1987). Proc. symp. on pls model building: Theory and application. Frankfurt am Main, 1987; also Tech. rep., Department of Organic Chemistry, Umea University.