MATIMYÁS MATEMATIKA
ISSN 0115-6926

Journal of the Mathematical Society of the Philippines
Vol. 49 No. 2 (2026) pp. 1-29

Measurement and characterization of the linear relationships of a set of variables observed on different groups of individuals

R. O. Malouata
Higher Institute of Management
Marien Ngouabi University
Congo
onesimero@gmail.com

G. C. Louzayadio
Faculty of Economics
Marien Ngouabi University
Congo
gelinlouzayadio@gmail.com

M. Koukouatikissa Diafouka
Higher Teachers' Training School
Marien Ngouabi University
Congo
diafdiak@gmail.com

Abstract

The aim of this paper is to contribute to a better understanding of certain methods to better analyze multigroup data. More specifically, dual canonical correlation analysis, which is a unifying approach to facilitate the comparison of multigroup methods, is developed by means of a proposed criterion. We prove that the duality is formulated by exchanging the operators. It is an extension of Carroll's generalized canonical correlation analysis.

Keywords: multigroup generalized canonical correlation analysis, data set partitioned in groups of individuals, principal component analysis

2020 MSC: 62H25, 62H30, 62H35

Introduction

Often, the data to be analyzed relates to a sample of individuals who are structured into groups. This constitutes, in particular, the general framework of canonical correlation analysis [11], which is best known for its theoretical qualities, since it encompasses many other statistical methods. This method studies the relationships between two data matrices. Canonical correlation analysis focuses on the correlation between a linear combination of

the variables in one set and a linear combination of the variables in another set. The idea is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and so on.

To extend canonical correlation analysis to more than two tables of variables, several methods have been proposed. These extensions have been grouped together under the name of generalized canonical correlation analysis. But of all these methods, generalized canonical correlation analysis in the sense of [2] has been the most widely used. Recently, in this same context, generalized canonical correlation analysis has been extended to several multitableaux in [13], [21] and [17]. It has been shown that simultaneously processing several large blocks of streaming data is a computationally expensive problem. The procedure transforms a big data problem into a small one by processing small high-dimensional matrices where variables are in rows. Numerical experiments illustrate the accuracy and performance of the incremental solution for analyzing streaming multiblock redundancy data [20].

In the context where the set of variables is observed on different groups of individuals, this problem has been studied in many scientific articles. If one expects the structure of each of the data blocks to be different, standard principal component analysis (PCA) [12, 22] can be performed on each data block. Moreover, one can carry out the simultaneous processing of all groups of variables. First, methods similar to PCA. Several methods combine the covariance matrices or the correlation matrices related to the various groups. The goal of these methods is to redistribute the explained variance by rotation. In [15], It is proposed to carry out a multigroup PCA by diagonalizing the within group covariance matrix. This problem has been reworked and developed into a succession of PCA by [7].

In practice, when we have a group of variables, the cloud of the individuals and the cloud of the variables are two representations of the same group: one across the rows and the other across the columns. Very strong relations called duality link these two clouds. Moreover, there are similar methods to the canonical correlation analysis among multigroup data analyses. We can cite: between groups comparison of principal components [14] and regularized generalized canonical correlation analysis for multiblock or multigroup data analysis [25, 27].

Novelty and positioning with respect to related methods. The present paper differs from the above approaches in the following respects. Regularized GCCA (RGCCA, [26, 27]) addresses the multiblock setting, where multiple variable blocks are observed on the same set of n individuals, and introduces ridge-type regularization to handle ill-conditioned covariance matrices. Our work addresses a structurally different problem: the multigroup setting, where a single variable block is observed on K distinct groups of individuals. The two settings lead to different data matrices (column-partitioned vs. row-partitioned), different projectors, and different optimisation criteria. Multiple Factor Analysis (MFA, Escofier and Pagès [6]) is designed for multiblock data and normalizes each table by its first eigenvalue before performing a global PCA; our MGCCA framework is tailored for the multigroup setting, uses orthogonal projectors onto row spaces, and is grounded in a duality principle that is absent from MFA. Sparse GCCA variants [26] extend RGCCA with penalization.

However, in the context where the set of variables is observed on different groups of individuals, the generalization of canonical correlation analysis (GCCA) according to [2] has never been discussed. In this paper, a version of GCCA called Multigroup generalized canonical correlation analysis and its variants, which can be applied to multiblock and multigroup, are described. We prove by means of the criterion that the duality is formulated by exchanging the covariance operator into scalar product. This criterion can be seen as a PCA in which the influence of the variables is stable.

This paper is organized as follows: In Subsection 1.1, we will describe the canonical correlation analysis proposed by [11] and extended by [2]. In Subsection 1.2, we propose a unified approach and study its extensions in four cases. Finally, in Section 2, the experimental design is described and the obtained results are analyzed.

1 Materials and methods

1.1 Generalized canonical correlation analysis

We shall be interested in measures of association between two matrices. The first matrix X_1, of p columns, is summarized by the $(n \times 1)$ canonical variate X_1u_1. The second matrix X_2, of q columns, is summarized by the $(n \times 1)$ canonical variate X_2u_2. Specifically, we define canonical correlation analysis as the following optimization problem:

$$\begin{aligned} &\text{Maximize} && \operatorname{corr}(X_1u_1, X_2u_2) \\ &\text{subject to the constraints} && \operatorname{var}(X_1u_1) = \operatorname{var}(X_2u_2) = 1. \end{aligned} \tag{1}$$

The first pair of canonical variates is the pair of linear combinations $(X_1u_1^1, X_2u_2^1)$ having unit variances, which maximize the correlation (1); The second pair of canonical variates is the pair of linear combinations $(X_1u_1^2, X_2u_2^2)$ having unit variances, which maximize the correlation (1) among all choices that are uncorrelated with the first pair of canonical variates. One may continue the search for canonical variates until all solutions are found.

Notation remark (multiblock vs. multigroup). In this subsection, following Carroll [2], $X = [X_1|\dots|X_K]$ is an $(n \times p)$ matrix partitioned by variable sets (columns): each $X_i \in \mathbb{R}^{n \times p_i}$ shares the same n individuals. The column space of X_i is denoted C_i, and the orthogonal projector onto C_i is $P_i = X_i(X_i'X_i)^{-1}X_i'$, acting in $\mathbb{R}^n$. In Subsection 1.2 (multigroup setting), by contrast, $X = [X_1'|\dots|X_K']'$ is an $(N \times p)$ matrix partitioned by individual groups (rows): each $X_i \in \mathbb{R}^{n_i \times p}$ has its own n_i individuals and the same p variables. The row space of X_i is denoted E_i, and the orthogonal projector onto E_i is $P_i = X_i'(X_iX_i')^{-}X_i$, acting in $\mathbb{R}^p$. This structural distinction is the fundamental motivation for the duality introduced in this paper.

To generalize the canonical correlation analysis of two sets of variables to K sets of variables, let X be a matrix $X = [X_1, \dots, X_K]$ having n rows and $p = \sum_{i=1}^{K} p_i$ columns, partitioned in K submatrices X_i. Each $n \times p_i$ $(n > p_i)$ data matrix X_i represents a set of p_i

centered and standardized variables observed on a set of n individuals. Let E_i $(i = 1, \ldots, K)$ denote column spaces of X_i. Let P_i be the orthogonal projector onto the column space E_i of X_i, $P_i = X_i(X_i'X_i)^{-1}X_i'$.

We only consider in this paper the generalization developed by [2] which consists of the following optimization problem:

$$\begin{aligned} \text{Maximize} \qquad & \frac{1}{K}\sum_{i=1}^{K} \text{corr}^2(z, X_i u_i) \\ \text{subject to the constraint} \quad & z'z = 1. \end{aligned} \tag{2}$$

The principle of the generalized canonical correlation analysis is first to investigate the variables related to the set of the groups. These variables which summarize the general trends of the groups are called general variables. Furthermore, a general variable being obtained, we seek in each group a linear combination of variables related to this general variable. These linear combinations called canonical variates are the representations of the general variables in groups.

One of the advantages of this approach is to obtain the solution noniteratively [16]. Another advantage of this approach is that it is not necessary to define a link measurement between two groups of variables but between a variable and a group [6]. The measurement used by [2] is the square of the multiple correlation coefficient. Finally, this approach has the advantage of being probably the most simple and the richest of interpretations, because it easily connects to all the other methods of data analysis. In addition to canonical correlation analysis, the GCCA method admits as special cases all the usual data analysis methods, whether descriptive, such as simple or multiple correspondence analysis, or standardized PCA, or explanatory, such as simple or multiple linear regression, analysis of variance or covariance, or discriminant analysis. The GCCA method does not provide an orthogonal basis for the canonical variables in each table, except when the number of tables is equal to two. On the other hand, all the auxiliary variables determined by GCCA are uncorrelated. The method does not allow for separate simultaneous factorial maps due to the non-orthogonal canonical variables. Thus [3] in the discriminant analysis of tables and [4] in the multiple co-inertia analysis imposed orthogonality constraints respectively on the canonical variables and factors relating to each table.

By definition, the multiple correlation coefficient between a variable and a group of variables X_i is the correlation coefficient between z and the linear combination of the variables of the group X_i most correlated with z. Geometrically, this linear combination is the orthogonal projection $P_i z$ of z onto the column space E_i of X_i.

The relationship between the correlation coefficients and the variable z follows from the special structure of the matrix $\sum_{i=1}^{K} P_i$. In other words, the generalized canonical correlation

analysis verifies the following stationary equation:

$$\frac{1}{K}\sum_{i=1}^{K} P_i z = \rho z \tag{3}$$

The solution of the generalized canonical correlation analysis shows that it can be obtained from the principal component analysis ([9] and [24]). Thus, by setting $\Sigma_i = X_i' X_i$, the solution of GCCA verifies the following stationary equation:

$$XDX'z = \rho z \tag{4}$$

where D is a block diagonal matrix formed from Σ_i^{-1} as the ith diagonal block.

The GCCA is similar to a PCA in which each column space E_i of X_i is associated with the matrix Σ_i^{-1}.

In this section we will give some details of a solution to the multigroup generalized canonical correlation analysis.

1.2 Unified approach and extensions of the generalized canonical correlation analysis

We now consider a matrix $X = [X_1'|\cdots|X_K']'$ having $N = \sum_{i=1}^{K} n_i$ rows and p columns, partitioned in K submatrices X_i of order $(n_i \times p)$ and row spaces E_i $(i = 1, \ldots, K)$ of X_i. Each data matrix X_i represents a set of p centered and standardized variables observed on a set of n_i individuals. Let P_i be the orthogonal projector onto the row space E_i of X_i, $P_i = X_i'(X_i X_i')^{-} X_i$, where A^{-} is a g-inverse of A.

1.2.1 Relationship between the factorial axis common to all groups and the factorial axes associated with the different groups

The approach we adopt is based on the determination of vectors, called factorial axes. These axes are determined sequentially, and are therefore of decreasing importance from one sequence to the next. At each dimension, most multigroup data analysis methods focus on determining the u vectors and, subsequently, the scalars λ_i are determined by: $\lambda_i = \mathrm{Var}(X_i u)$.

So λ_i reflects the percentage inertia of X_i restored by the component $\xi_i = X_i u$. For some methods, in addition to the common factor axes u, it is convenient to display factor axes associated with the different groups defined by $t_i = X_i' a_i$ for $i = 1, \ldots, K$ (Figure 1). These axes t_i for $i = 1, \ldots, K$ are linear combinations of the lines of each group and linked to the partial components ξ_i for $i = 1, \ldots, K$.

The common factorial axes u of all the tables linked to the global components $\xi = Xu$ can be determined by PCA of the multigroup X. These global components $\xi = Xu$ contain the factorial axes t_i. The main idea behind the unifying approach we propose here is that there are mutual relationships between the global component and the partial components,

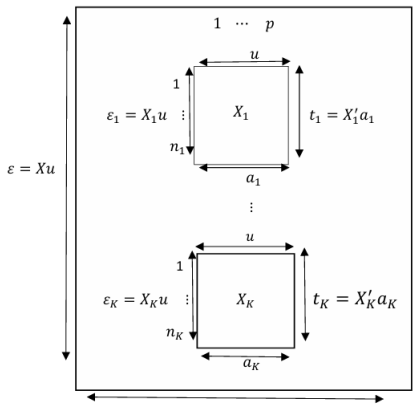

Figure 1: Multigroup with axes and components

on the one hand, and the common factorial axis to all the groups and the factorial axes associated with the different groups, on the other. In this sense, the factorial axis t_i reflects u in the table X_i.

We now consider two typical situations that give a good idea of whether t_i is a reflection of u in X_i, and that allow us to generalize principal component analysis to multigroup data.

1.2.2 Situation 1: multigroup PCA

When we have a set of individuals described by variables, standardized or non-standardized principal component analysis provides a simplified description of the configuration of individuals and correlations between variables.

If you have several tables of data defined on the basis of different groups of individuals, canonical correlation analysis (or its generalizations) highlights the common elements to all or some of the tables.

Multigroup PCA makes it possible to reconcile these two approaches, i.e. to compare tables while describing each of them.

In terms of data analysis, this description and comparison involves the idea that t_i is the reflection of u in X_i. We will consider the relationship $t_i = X_i' X_i u$ where $X_i' X_i$ is the covariance matrix of X_i. With this relationship, the method seeks to explain not only intergroup variability, but also intragroup variability.

If we normalize the table X_i, then the correlation matrix R_i related to X_i is equal to $X_i' X_i$ and the correlation matrix R related to the whole data matrix X is equal to $\bar{R} = (1/N) \sum_{i=1}^{K} R_i = X'X$.

In this situation, we propose to search for u (and therefore t_i) in such a way as to

maximize the following criterion:

$$\frac{1}{N}\sum_{i=1}^{K}\langle t_i|u\rangle = \frac{1}{N}\sum_{i=1}^{K} u'X_i'X_iu = u'\bar{R}u \tag{5}$$
$$\text{subject to the constraint} \quad \|u\| = 1.$$

Remark (normalization constants in the multiblock vs. multigroup settings). The normalization constant $1/N$ appearing in criterion (5) differs structurally from the constant $1/K$ used in the multiblock criterion (2). In the multiblock setting (Subsection 1.1), $1/K$ is a uniform average over the K variable blocks, independent of group sizes. In the multigroup setting, $1/N = 1/(\sum_{i=1}^{K} n_i)$ induces a size-weighted average over individual groups: the criterion rewrites as $\sum_{i=1}^{K}(n_i/N)\langle t_i|u\rangle$, showing that each group contributes proportionally to its relative size n_i/N. The two constants coincide only in the balanced case $n_i = n$ for all i and $K = 1$, which is trivial.

It follows that the vector u which maximizes the criterion is given by the normed eigenvector of $\bar{R}$ associated with the largest eigenvalue

$$\lambda = \frac{1}{N}\sum_{i=1}^{K} n_i \text{var}(X_iu) = \text{var}(Xu),$$

where u is the principal axis of X.

Once the factorial axis has been determined, the factorial axes associated with the different tables are naturally obtained by the relationship: $t_i = X_i'X_iu$. By replacing $t_i = X_i'a_i$ in $t_i = X_i'X_iu$, we obtain $a_i = X_iu$. We note that the vector a_i is none other than the partial component ξ_i.

It is important to note that when maximizing criterion (5), the quantity $u'X_i'X_iu = \langle t_i|u\rangle$ represents the contribution of the table X_i in determining the global component $\xi = Xu$. In addition, $\|u\| = 1$, this quantity also reflects the inertia of X_i explained by u (in percent), and can be determined by $\frac{1}{N}\sum_{i=1}^{K} u'X_i'X_iu$, where K is the number of tables also corresponding to the total variance of all tables.

To obtain orthogonal partial components, the same criterion is maximized by replacing to order s, $s = 2, \ldots, r$, where $r = \min(rg(X_i))$ denotes the rank of the matrix X_i, the tables X_i^{s-1} by: $X_i^{s-1} = P^{\perp}_{\xi_i^{s-1}}X_i^{s-2}$, knowing that $X_i^0 = X_i$, where $P^{\perp}_{\xi_i^{s-1}} = I_{n_i} - P_{\xi_i^{s-1}}$ and $P_{\xi_i^{s-1}} = \xi_i^{(s-1)}\xi_i^{(s-1)'}/\|\xi_i^{s-1}\|^2_{D_i}$ is the D_i orthogonal projector onto the space generated by ξ_i^{s-1}.

Thus, the K systems of components $\{\xi_i^s\}_{s=1,\ldots,r}$ are D_i-orthogonal and the systems of axes $\{u^s\}_{s=1,\ldots,r}$ form an orthonormal basis.

1.2.3 Situation 2: Multigroup GCCA

We consider a situation where all the factorial axes t_i associated with the different groups are linked to the common factorial axis u by an orthogonal projection onto the vector space

generated by the rows of X_i: $t_i = P_i u$, where $P_i = X_i'(X_i X_i')^- X_i$.

As we can see, this expression requires the generalized inverse of the matrix $X_i X_i'$. The advantage of this generalized inverse is to propose a matrix with similar properties to those of the inverse, when the latter does not exist. This version is an extension of the generalized canonical correlation analysis proposed by [2] called *multigroup generalized canonical correlation analysis (MGCCA)*.

Assumptions. Each matrix X_i has n_i rows and p columns. The row spaces E_i are well-defined subspaces of $\mathbb{R}^p$. The matrices $W_i = X_i X_i'$ are symmetric positive semi-definite; their Moore-Penrose generalized inverses W_i^- are used when W_i is rank-deficient.

In the first step, the solution of the multigroup generalized canonical correlation analysis is a loading vector u and specific coefficients $t_i = X_i' a_i$ (a_i being a vector to be determined of order $n_i \times 1$) associated with data matrices X_i ($i = 1, \ldots, K$).

In our case, as in many similar situations, we define a mean-squared loss function. The core optimization problem considered in this paper is defined as follows:

$$\begin{aligned} \text{Minimize} \qquad & g(u) = \frac{1}{N} \sum_{i=1}^{K} \|u - P_i u\|^2 \\ \text{subject to the constraints} \quad & \|u\| = \|t_i\| = 1. \end{aligned} \tag{6}$$

The loading vector is common to all matrices and its projection onto the subspace E_i is the specific coefficient $t_i = P_i u$. To convert the minimization problem into a maximization problem we rewrite (6) and manipulate the various terms. We obtain the following maximization problem:

$$\begin{aligned} \text{Maximize} \qquad & h(u) = \frac{1}{N} \sum_{i=1}^{K} \langle u | P_i u \rangle = \frac{1}{N} \sum_{i=1}^{K} u' P_i u \\ \text{subject to the constraints} \quad & \|u\| = \|t_i\| = 1. \end{aligned} \tag{7}$$

From this problem, we can incorporate the constraint in the maximization problem by using Lagrange multiplier λ and obtain $\tilde{h}$:

$$\tilde{h}(u, \lambda) = h(u) - \lambda(u'u - 1). \tag{8}$$

The maximum of h follows from the requirement that the first order partial derivatives of $\tilde{h}$ are simultaneously zero at the maximum of h and that the Hessian is negative. We will state here the exact nature of the solution as proposition.

Proposition 1. *For $P_i u = t_i$, we obtain the following result:*

$$\frac{1}{N} \sum_{i=1}^{K} P_i u = \lambda u. \tag{9}$$

Proof. Differentiating $\tilde{h}$ with respect to u and using the symmetry of P_i, we get

$$\frac{\partial \tilde{h}}{\partial u} = \frac{2}{N}\sum_{i=1}^{K} P_i u - 2\lambda u = 0, \quad \text{which implies} \quad \frac{1}{N}\sum_{i=1}^{K} P_i u = \lambda u.$$

Here, we used the fact that $\partial(u'P_iu)/\partial u = 2P_iu$ since P_i is symmetric. This gives stationary equation (9): u is an eigenvector of $(1/N)\sum_{i=1}^{K} P_i$.

The Hessian of $\tilde{h}$ with respect to u is: $\partial\tilde{h}/\partial u\partial u' = \frac{2}{N}\sum_{i=1}^{K} P_i - 2\lambda I_p$. Since each P_i is positive semi-definite (orthogonal projector), $\frac{1}{N}\sum_{i=1}^{K} P_i$ is positive semi-definite and its eigenvalues are non-negative. The Hessian is therefore negative semi-definite when λ is chosen as the largest eigenvalue, guaranteeing a maximum. Differentiating with respect to λ recovers the constraint $u'u = 1$. By replacing P_i in $P_iu = t_i$ and using $t_i = X_i'a_i$, we obtain $a_i = (X_iX_i')^- X_iu$. The partial components are obtained by setting $\xi_i = X_iu$. □

Moreover, as the projectors P_i are symmetric and idempotent, it follows that $h(u) = \frac{1}{N}\sum_{i=1}^{K} \|P_iu\|^2$. It seems that the multigroup generalized canonical correlation analysis (MGCCA) is similar to principal component analysis (PCA) (see Subsection 1.1). The components are determined using the same loading vector u and the explained variance can be redistributed by rotation. It can be easily shown that the vector u which maximizes this average inertia is obtained by the eigenvector of $\frac{1}{N}\sum_{i=1}^{K} P_i$ related to the eigenvalue $\lambda = \frac{1}{N}\sum_{i=1}^{K} \|P_iu\|^2$.

That is, after finding a first axis, this column is fixed and the second column is found in the residual space, by replacing X_i by $X_i(I_p - uu')$ and proceeding in the same way as for the first step. Other loading vectors can be obtained by repeating the procedure. This procedure allows to construct the orthonormal basis. We can remark that the global components $\xi^s = Xu^s$ ($s = 1\ldots, A$, where A is the rank of the matrix X) are orthogonal. The following proposition establishes that the multigroup generalized canonical correlation analysis is a PCA.

Proposition 2. *For $W_i = X_iX_i'$, the multigroup generalized canonical correlation analysis is a PCA and verifies the following stationary equations:*

$$X'D^-Xu = \lambda u \tag{10}$$

$$XX'D^-\xi = \lambda\xi \tag{11}$$

where D^- is a block diagonal matrix formed from W_i^- as the ith diagonal block.

Proof. Recall the assumptions stated at the beginning of Subsection 1.2.3. By definition, the projector P_i onto the row space E_i of X_i is $P_i = X_i'(X_iX_i')^- X_i = X_i'W_i^- X_i$. Summing over all K groups yields $\sum_{i=1}^{K} P_i = \sum_{i=1}^{K} X_i'W_i^- X_i$. Writing $X = [X_1' \mid \cdots \mid X_K']'$ ($N \times p$) and $\Lambda^- = \mathrm{diag}(W_1^-, \ldots, W_K^-)$ ($N \times N$ block-diagonal), we obtain:

$$\sum_{i}^{K} X_i'W_i^- X_i = X'\Lambda^- X.$$

Substituting this expression into stationary equation (9) gives $(1/N)X'\Lambda^- Xu = \lambda u$. Setting $D^- = (1/N)\Lambda^-$ yields $X'D^- Xu = \lambda u$, which is Eq. (10). Finally, pre-multiplying (10) by X: $X(X'D^- X)u = \lambda(Xu)$, giving $XX'D^-\xi = \lambda\xi$ with $\xi = Xu$, which is Eq. (11).

Equations (10) and (11) are exactly the eigen-equations of a PCA of the matrix X endowed with the metric D^-, confirming that MGCCA is a PCA under the transformed metric. □

We then recognize the eigen-equations yielding principal components and factor axes in PCA of X. From the above, we conclude that the criterion of MGCCA is a PCA in which the duality is formulated by exchanging the covariance matrix Σ_i into scalar product matrix W_i.

Duality remark. The duality between GCCA and MGCCA is explicit: GCCA uses D (block-diagonal with blocks $\Sigma_i^{-1} = (X_i'X_i)^{-1}$), while MGCCA uses D^- (block-diagonal with blocks $W_i^- = (X_iX_i')^1$), exchanging the role of column-space and row-space operators.

Geometrically, each specific coefficient t_i associated with different matrices is the orthogonal projection P_iu of u onto the subspaces E_i spanned by the rows of X_i. Thus, the multiple correlation coefficient is the cosine of the angle θ_i between the loading vector u and its projection P_iu onto E_i. Since u and P_iu are normalized, we then have:

$$\cos^2\theta_i = \langle u|P_iu\rangle.$$

As $u^{s'}(u^{s-1},\ldots,u^1) = 0$, we obtain:

$$\sum_{i=1}^{K}\cos^2\theta_i = \sum_{i=1}^{K}\langle u^s|P_iu^s\rangle = \Big\langle u^s\Big|\sum_{i=1}^{K}P_iu^s\Big\rangle.$$

We achieve the same conclusion: the operator $\sum_{i=1}^K P_i$ being a sum of the orthogonal projection operators, It is symmetric, diagonalizable and the eigenvectors are orthonormal.

Computational complexity. In the standard case ($n_i > p$), MGCCA requires K Moore-Penrose inversions W_i^- at cost $O(n_i^2p)$ per group, followed by the eigendecomposition of $\frac{1}{N}\sum_{i=1}^K P_i = X'D^-X$, a $p \times p$ matrix, at cost $O(p^3)$. The overall complexity is $O(\sum_{i=1}^K n_i^2p + p^3)$. In the high-dimensional case ($p > n_i$), W_i is rank-deficient; ridge regularization $W_i(\tau_i) = X_iX_i' + \tau_iI_{n_i}$with $\tau_i > 0$ restores invertibility at cost $O(n_i^3)$ via the Woodbury identity [19]. The NIPALS-type algorithms of Sections 1.2.5 and 1.2.6 avoid forming the full $p\times p$ matrix and are well-suited for large-scale applications. An incremental SVD adaptation following [20] is a perspective for future work.

1.2.4 Multigroup canonical correlation analysis (MCCA)

We shall be interested in the measures of association between two data matrices. Let E_1 and E_2 be the row spaces of X_1 and X_2. We also introduce for each matrix, unit vectors $t_1 = X_1'a_1$ and $t_2 = X_2'a_2$ onto E_1 and E_2. The two vectors are as close as possible. Let J_1 and J_2 be orthogonal projections onto E_1 and E_2 defined by: $P_1 =$

$X_1'(X_1X_1')^-X_1$ and $P_2 = X_2'(X_2X_2')^-X_2$ where A^- is a g-inverse of A. Let $W_1 = X_1X_1'$ and $W_2 = X_2X_2'$ be scalar product matrices. Let W_{12} be the scalar product matrix of the individuals inter-data matrices and W_{21} its transpose.

The multigroup canonical correlation analysis (MCCA) problem consists of optimizing the following squared distance from $t_1 = P_1u$ and $t_2 = P_2u$:

$$\begin{aligned} \text{Minimize} \qquad & k(a_1, a_2) = \|t_1 - t_2\|^2 \\ \text{subject to the constraints} \quad & \|t_1\| = \|t_2\| = 1. \end{aligned} \tag{12}$$

This problem is equivalent to optimize the following criterion:

$$\begin{aligned} \text{Maximize} \qquad & k_1(a_1, a_2) = a_1W_{12}a_2 \\ \text{subject to the constraints} \quad & a_1W_1a_1 = a_2W_2a_2 = 1. \end{aligned} \tag{13}$$

We define multigroup canonical correlation (MCCA) as the optimization problem which consists of measuring the proximity of the individuals between the groups X_1 and X_2. The Lagrangian function of optimization problem (13) is then considered:

$$\tilde{k_1}(a_1, a_2, \lambda_1, \lambda_2) = a_1W_{12}a_2 - \frac{\lambda_1}{2}(a_1W_1a_1 - 1) - \frac{\lambda_2}{2}(a_2W_2a_2 - 1), \tag{14}$$

where λ_1 and λ_2 are the Lagrange multipliers associated with the constraints. Canceling the derivatives of the Lagrangian function with respect to a_1, a_2, λ_1 and λ_2, we obtain using the stationary equations:

- a_1 is the eigenvector of the matrix $W_1^-W_{12}W_2^-W_{21}$ associated with the largest eigenvalue λ_1^2,
- a_2 is the eigenvector of the matrix $W_2^-W_{21}W_1^-W_{12}$ associated with the largest eigenvalue λ_1^2.

The eigenvectors t_1 and t_2 are associated with the same eigenvalues. Thus, in MCCA, we also get:

- t_1 is the eigenvector of the matrix P_1P_2 associated with the largest eigenvalue λ_1^2,
- t_2 is the eigenvector of the matrix P_2P_1 associated with the largest eigenvalue λ_1^2.

Proposition 3. *The multigroup generalized canonical correlation analysis generalizes the multigroup canonical correlation analysis.*

Proof. Indeed, the normalization constraints are the same as those of the multigroup canonical correlation analysis. To minimize $\|u - P_1u\|^2 + \|u - P_2u\|^2$ subject to the normalization constraints over u, t_1 and t_2 is equivalent to minimize $\|P_1u - P_2u\|^2$.

Expanding $\|u - P_iu\|$ for each $i \in \{1, 2\}$:

$$\|u - P_iu\|^2 = \|u\|^2 - 2\langle u|P_iu\rangle + \|P_iu\|^2$$

Since P_i is an orthogonal projector (symmetric and idempotent: $P_i^2 = P_i$), we have $\langle u|P_i u\rangle = u'P_i u = u'P_i^2 u = \|P_i u\|^2$. Thus:

$$\|u - P_i u\|^2 = \|u\|^2 - 2\|P_i u\|^2 + \|P_i u\|^2 = \|u\|^2 - \|P_i u\|^2.$$

Summing over $i = 1$ and $i = 2$, and using $\|u\| = 1$:

$$\|u - P_1 u\|^2 + \|u - P_2 u\|^2 = 2\|u\|^2 - (\|P_1 u\|^2 + \|P_2 u\|^2) = 2 - (\|P_1 u\|^2 + \|P_2 u\|^2).$$

On the other hand, expanding $\|P_1 u - P_2 u\|^2$:

$$\|P_1 u - P_2 u\|^2 = \|P_1 u\|^2 - 2\langle P_1 u|P_2 u\rangle + \|P_2 u\|^2.$$

Since $\|t_1\| = \|P_1 u\| = 1$ and $\|t_2\| = \|P_2 u\| = 1$, this simplifies to:

$$\|P_1 u - P_2 u\|^2 = 2 - 2\langle P_1 u|P_2 u\rangle.$$

Thus we see that minimising $\|u - P_1 u\|^2 + \|u - P_2 u\|^2$ over u is equivalent to maximising $\|P_1 u\|^2 + \|P_2 u\|^2$, which in turn is equivalent to minimising $\|P_1 u - P_2 u\|^2 = 2 - 2\langle P_1 u|P_2 u\rangle$, i.e., maximising $\langle P_1 u|P_2 u\rangle$. This establishes the equivalence stated in Proposition 3. □

Proposition 4. *The loading vector u is the principal factor of the specific coefficients t_i. Equivalently, u is the loading vector that maximises $\frac{1}{N}\sum_{i=1}^{K}\|t_i\|^2$, i.e., the average squared norm of the group-specific factorial axes.*

Proof. As seen in the proof of Proposition 3, since P_i is idempotent we have the identity:

$$\|u - P_i u\|^2 = \|u\|^2 - \|P_i u\|^2 = 1 - \|P_i u\|^2 = 1 - \|t_i\|^2$$

Therefore, minimising $\frac{1}{N}\sum_{i=1}^{K}\|u - P_i u\|^2$ is equivalent to maximising $\frac{1}{N}\sum_{i=1}^{K}\|t_i\|^2$, which is the quantity $h(u) = \frac{1}{N}\sum_{i=1}^{K}\|P_i u\|^2$ already introduced in Section 1.2.3. □

Proposition 5. *The mean (sum) of the specific coefficients t_i and u are proportional.*

Proof. Indeed, Eq. (9) shows that the only possible proportionality constant is λ, since

$$\frac{1}{N}\sum_{i=1}^{K} t_i = \frac{1}{N}\sum_{i=1}^{K} P_i u = \lambda u, \tag{15}$$

with λ being the largest eigenvalue of the mean (sum) of P_i. □

1.2.5 Situation 3: Multiple co-inertia analysis (MCOIA) variant

Instead of determining u (and $t_i = X_i' X_i u$) by maximizing the criterion $\frac{1}{N}\sum_{i=1}^{K}\langle t_i|u\rangle$, with $\|u\| = 1$, we propose the following criterion:

$$\begin{aligned} &\text{Maximize} \qquad \sum_{i=1}^{K}\langle t_i|u\rangle^2 \\ &\text{subject to the constraint} \quad \|u\| = 1. \end{aligned} \tag{16}$$

From this problem, we can incorporate the constraint in the maximization problem by using Lagrange multiplier λ and obtain L,

$$L(u,\lambda) = \sum_{i=1}^{K}\langle t_i|u\rangle^2 - 2\lambda(u'u - 1) \tag{17}$$

The following proposition specifies the role of the factorial axis u in the criterion to be maximized.

Proposition 6. *Designing by $\alpha_i = u'X_i'X_iu$ and $V_i = X_i'X_i$, the vector u verifies the following equation:*

$$\sum_{i=1}^{K}\alpha_i V_i u = \lambda u \tag{18}$$

Proof. Taking the partial derivative of L with respect to u,

$$\frac{\partial L}{\partial u} = 4\sum_{i=1}^{K}(u'X_i'X_iu)(X_i'X_iu) - 4\lambda u.$$

Cancelling the derivative and setting $\alpha_i = u'X_i'X_iu$ and $V_i = X_i'X_i$, we obtain the result. □

Clearly, u is the norm eigenvector of $\sum_{i=1}^{K}\alpha_i V_i$ associated with the largest eigenvalue. We deduce the following algorithm for determining u:

A. Initialization

- A.1 Choose the specific weights $\alpha_i^0 = 1$ $(i = 1,\ldots,K)$;
- A.2 Compute the common factorial axes: u^0 is the norm eigenvector of $\sum_{i=1}^{K}\alpha_i^0 V_i$ associated with the largest eigenvalue;
- A.3 Compute the factorial axes associated with the different groups: $t_i^0 = X_i'X_iu^0$;

B. Computing of the updates

For $i = 1, 2, \ldots, K$

- B.1 Compute $\alpha_i = u't_i$;
- B.2 Compute the common factorial axes: u is the norm eigenvector of $\sum_{i=1}^{K}\alpha_i V_i$ associated with the largest eigenvalue;

B.3 Compute the factorial axes associated with the different groups: $t_i = X_i' X_i u$;

B.4 Repeat process from step B.1 until convergence.

End

Clearly, we find a dual criterion of the multiple co-inertia analysis [13]. A variant of this method can be found in [18].

Alternatively, the maximization problem (16) can be solved as follows: for u fixed, we have $t_i = X_i' X_i u$ and for t_i $(i = 1, \ldots, K)$ fixed, maximizing the criterion $\sum_{i=1}^{K} \langle t_i | u \rangle^2$ leads to choosing u as the first principal axis of the factorial axes t_i. Indeed, maximization of the above criterion is typical of a characteristic property of the first principal axis [12]. It follows that $\alpha_i = u' X_i' X_i u$. To interpret the results, we can standardize these coefficients to give $\sum_{i=1}^{K} \alpha_i^2 = 1$. Thus, the coefficient α_i reflects the contribution of the table X_i to the determination of the common factorial axis to all tables. We also note that the specific weight α_i associated with each common factorial axis is positive. It reflects the percentage of explained inertia by the associated common factorial axis.

The resolution algorithm is as follows:

A. Initialization

A.1 Choose randomly the common factorial axis u^0 such that $\|u^0\| = 1$;

A.2 Compute the factorial axes: $t_i^0 = X_i' X_i u^0$ for $i = 1, 2, \ldots K$

A.3 Compute the specific weights: $\alpha_i^0 = u^{0'} t_i^0$ for $i = 1, 2, \ldots K$

B. Computing of the updates

For $i = 1, 2, \ldots, K$

B.1 Compute $u = \sum_{i=1}^{K} \alpha_i t_i$ and normalize $u = u/\|u\|$;

B.2 Compute the factorial axes: $t_i = X_i' X_i u$ for $i = 1, 2, \ldots K$

B.3 Compute the specific weights: $\alpha_i = u' t_i$ for $i = 1, 2, \ldots K$

B.4 Repeat process from step B.1 until convergence.

End

To determine the higher-order axes, we deflate with respect to the common axis u, as indicated for the multiblock PCA method. Thus, the method cuts the inertia of K tables in the space $\mathbb{R}^p$. This is why deflations are made in $\mathbb{R}^p$ space.

If we set $X_i^{(0)} = X_i$ for all $i = 1, \ldots, M$ and $X_i^{(s-1)}$ the residual of the regression of $X_i^{(s-2)}$ on the vector $u^{(s-1)}$ at order s, we can write

$$X_i^{(s-1)} = X_i^{(s-2)} P_{u^{(s-1)}}^{\perp'} \quad \text{with} \quad P_{u^{(s-1)}}^{\perp'} = I_p - u^{(s-1)} u^{(s-1)'}$$

The stationary equation at order s is given by

$$\sum_{i=1}^{K} \alpha_i^{(s)} V_{is} u^{(s)} = \lambda_s^{\max} u^{(s)} \tag{19}$$

where $V_{is} = X_i^{(s-1)'} X_i^{(s-1)}$ is the covariance matrix of the deflated table at order s, $\alpha_i^{(s)} = u^{(s)'} X_i^{(s-1)'} X_i^{(s-1)} u^{(s)}$ is the specific weight of group i at order s, and the deflated table at order s is defined recursively by $X_i^{(s-1)} = X_i^{(s-2)} P_{u^{(s-1)}}^{\perp'}$ with $P_{u^{(s-1)}}^{\perp'} = I_p - u^{(s-1)} u^{(s-1)'}$ and $X_i^{(0)} = X_i$.

Clearly, we find the dual method of common components and specific weights analysis ([23] and [10]). This method can be called dual common components and specific weights analysis. This alternative offers multiple advantages:

- It brings insight by showing the factor axes associated with the different groups that are optimally associated with the common factor axis.
- It provides new elements of interpretation, further clarifying the nature of common factor axes and specific weights.
- It suggests a new algorithm for determining common factor axes and specific weights, as well as factor axes associated with the different tables.
- It allows us to better situate the analysis in relation to other methods of multigroup data analysis.

1.2.6 Situation 4: Multigroup GCCA variant

The GCCA variant is obtained by searching for u of length 1 (and $t_i = P_i u$) so as to maximize the criterion

$$\sum_{i=1}^{K} \langle t_i | u \rangle^2 \tag{20}$$
$$\text{subject to the constraint} \quad \|u\| = 1.$$

Proceeding exactly as in the previous subsection (replacing V_i by P_i), we arrive at the following algorithm:

A. Initialization

A.1 Choose the specific weights $\alpha_i^0 = 1$ $(i = 1, \ldots, K)$;

A.2 Compute the common factorial axes: u^0 is the norm eigenvector of $\sum_{i=1}^{K} \alpha_i^0 P_i$ associated with the largest eigenvalue;

A.3 Compute the factorial axes associated with the different groups: $t_i^0 = P_i u^0$;

B. Computing of the updates
For $i = 1, 2, \ldots, K$

B.1 Compute $\alpha_i = u' t_i$;

B.2 Compute the common factorial axes: u is the norm eigenvector of $\sum_{i=1}^{K} \alpha_i P_i$ associated with the largest eigenvalue;

B.3 Compute the factorial axes associated with the different groups: $t_i = P_i u$;

B.4 Repeat process from step B.1 until convergence.

End

Following the same reasoning as above, we propose the following NIPALS-type algorithm:

A. Initialization

A.1 Choose randomly the common factorial axis u^0 such that $\|u^0\| = 1$;

A.2 Compute the factorial axes: $t_i^0 = P_i u^0$ for $i = 1, 2, \ldots K$

A.3 Compute the specific weights: $\alpha_i^0 = u^{0'} t_i^0$ for $i = 1, 2, \ldots K$

B. Computing of the updates
For $i = 1, 2, \ldots, K$

B.1 Compute $u = \sum_{i=1}^{K} \alpha_i t_i$ and normalize $u = u/\|u\|$;

B.2 Compute the factorial axes: $t_i = P_i u$;

B.3 Compute the specific weights: $\alpha_i = u' t_i$;

B.4 Repeat process from step B.1 until convergence.

End

1.2.7 Method comparison

All methods determine the solution sequentially. The first difference is the way in which the common axis is determined: Multigroup PCA and MGCCA are step-by-step methods that diagonalize a matrix; MCOIA and GCCA variants are algorithmic methods that weight groups according to their contribution to the common factorial axis. A second criterion distinguishing the methods is the definition of t_i: with $t_i = P_i u$, the analysis focuses on intergroup variations (hidden variances and correlations); with $t_i = X_i' X_i u$, both intergroup and intragroup variations are explained.

Table 1 summarizes the four methods. The inner product notation $\langle \cdot | \cdot \rangle$ is used uniformly throughout.

2 Results

2.1 Wine quality dataset

We present the results of the multigroup generalized canonical correlation analysis and of the multigroup PCA. The choice of these two methods is deliberate. The results of the multigroup PCA are similar to [7]. The data set is the one used by [5]. The study focuses on Portuguese wines. Specifically, we reanalyze matrices of eleven physicochemical variables observed on 4898 individuals for the white wines and 1599 individuals for the red wines.

To eliminate the differences in variable means and variances, the data are centered and standardized per data matrix.

Regarding the graphical representation, the physicochemical variables are as follows: Fixac = fixed acidity, Volac = volatile acidity, Citac = citric acid, Resug = residual sugar, Chlor = chlorides, Fredi = free sulfur dioxide, Totdi = total sulfur dioxide, Densi = density, pH, Sulph = sulphates, Alcoh = alcohol.

The goal of the dual generalized canonical correlation analysis is to help discover how the relationships vary in the experimental setup. In Table 1 below, we report the percentages of explained variances for first four principal components and global components obtained from the dual generalized canonical correlation analysis. Table 2 and Table 3 show similar results for the multigroup PCA and MGCCA.

Figure 2 shows the representation of the variables on the first two loading vectors. The first principal component, which explains 44.02% of the total variance, opposes alcohol concentration to sugar concentration and density, and is also correlated with sulphates and chlorides. This corresponds to the fermentation process. The second axis (19.17% of total variance) shows an opposition between acid measurements and pH. The first two components together explain 63.19% of the total variance, which is a satisfactory result.

Comparison remark. The eigenvalues of MGCCA are numerically larger than those of multigroup PCA on this dataset. This reflects the difference in the operators used (row projectors P_i in MGCCA vs. covariance operators V_i in multigroup PCA) and does not imply a general superiority of one method over the other: the two methods address different but complementary aspects of multigroup data (intergroup vs. intergroup+intragroup variability). The results of the two analyses remain very close overall, consistently with [7].

For each of the two analyses, the eigenvalues of MGCCA are superior to the eigenvalues of multigroup PCA. The results of the MGCCA and the multigroup PCA are very little different.

2.2 Iris Dataset ($n_i \gg p$)

The Iris dataset, introduced by [8], with measurements collected by botanist [1], constitutes a benchmark example for multigroup data analysis. It comprises $p = 4$ quantitative variables observed on $K = 3$ groups of individuals ($n_i = 50$ each), for a total of $N = 150$ individuals. Each group corresponds to a distinct botanical species of iris, and the measurements concern the dimensions of the sepals and petals of each flower.

The dataset is structured exactly in the multigroup framework of this article: $X = [X_1' \mid X_2' \mid X_3']'$ is a 150×4 matrix partitioned by rows, where each sub-matrix $X_i \in \mathbb{R}^{50\times 4}$ shares the same $p = 4$ variables.

Theoretical Remark: Degenerate case of MGCCA. Since $n_i = 50 \gg p = 4$, the rank of each X_i equals p, and each scalar product matrix $W_i = X_i X_i'$ is of full rank ($\operatorname{rank} W_i = p$). A fundamental theoretical consequence follows: the orthogonal projector

Table 1: Classification of multigroup data analysis methods

Method	Factor axis	Criterion	
Multigroup PCA	$t_i = V_i u$	$\frac{1}{N}\sum_{i=1}^{K}\langle t_i \mid u\rangle$	subject to $\|u\| = 1$
Multigroup GCCA	$t_i = P_i u$	$\frac{1}{N}\sum_{i=1}^{K}\langle t_i \mid u\rangle$	subject to $\|u\| = 1$
MCOIA variant	$t_i = V_i u$	$\frac{1}{N}\sum_{i=1}^{K}\langle t_i \mid u\rangle^2$	subject to $\|u\| = 1$
MGCCA variant	$t_i = P_i u$	$\frac{1}{N}\sum_{i=1}^{K}\langle t_i \mid u\rangle^2$	subject to $\|u\| = 1$

Table 2: The percentages of the explained variances of MGCCA

Group	Dim1	Dim2	Dim3	Dim4
Red wines	28.61	13.74	11.69	9.30
White wines	21.96	21.51	10.44	10.41
Global component	44.02	19.17	11.94	8.29

Table 3: The percentages of the explained variances of multigroup PCA

Group	Dim1	Dim2	Dim3	Dim4
Red wines	29.19	14.08	11.21	9.28
White wines	23.36	21.58	13.46	11.12
Global component	27.25	15.91	11.61	9.75

Table 4: Variables of the Iris dataset (common to the 3 groups)

Variable	Description	Unit
Sepal.Length	Sepal length	cm
Sepal.Width	Sepal width	cm
Petal.Length	Petal length	cm
Petal.Width	Petal width	cm

Table 5: Structure of the Iris multigroup dataset

Table	Group (species)	n_i	p
X_1	*Iris setosa*	50	4
X_2	*Iris versicolor*	50	4
X_3	*Iris virginica*	50	4
X (multigroup)	Total	150	4

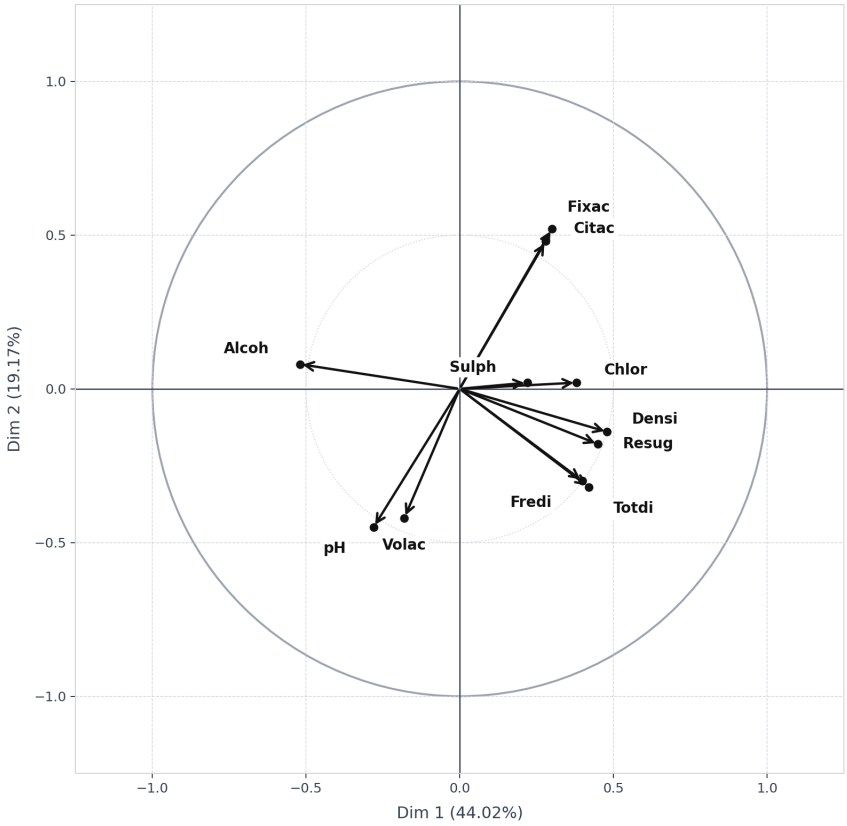

Figure 2: Map of the columns on the first two loading vectors of MGCCA

onto the row space E_i of X_i satisfies

$$P_i = X_i'(X_i X_i')^- X_i = I_p.$$

Hence $\frac{1}{N}\sum_{i=1}^{K} P_i = I_p$, whose eigenvalues are all equal to $\frac{1}{K}\cdot\frac{K}{N}\cdot N = 1$, giving $\lambda = \frac{1}{N}\cdot K\cdot\frac{N}{K} = 1$. More precisely, the matrix $\frac{1}{N}\sum_{i=1}^{3} I_4 = \frac{3}{150} I_4 = 0.02\, I_4$ has all four eigenvalues equal to 0.02, meaning all dimensions explain exactly 25% of the variance – the fully isotropic (degenerate) case.

This is not a flaw of the method but a fundamental mathematical property: no direction is preferred when $n_i \gg p$, because every direction in $\mathbb{R}^p$ lies in the row space of X_i. As noted in the computational complexity remark of Section 1.2.3, ridge regularization $W_i(\tau_i) = X_i X_i' + \tau_i I_{n_i}$ [19] with $\tau_i > 0$ restores discriminative power in such cases. The Iris example therefore illustrates the complementarity between Multigroup PCA and MGCCA: the former naturally exploits both intra- and inter-group variability through the covariance operator, while the latter specializes in the inter-group geometry via row-space projectors.

Comparison. Figure 3 shows the variable correlation circles on the first two loading

vectors for both methods. *Left panel — Multigroup PCA.* Dim 1 (61.49% of total variance) is dominated by `Petal.Length` and `Sepal.Length` pointing strongly to the right, while `Sepal.Width` points to the left: this axis captures the overall flower size gradient, opposing large-petalled species (*virginica*, *versicolor*) to the smaller *setosa*. `Petal.Width` contributes positively to Dim 2 (16.46%), which reflects the petal-width/sepal-width opposition capturing additional morphological variation within the two larger species. All four arrows reach near the unit circle, confirming that the first two components together give a high-quality representation of the variable space. *Right panel — MGCCA (degenerate case).* Since $n_i = 50 \gg p = 4$, every projector P_i equals I_4, so $\frac{1}{N}\sum_i P_i = 0.02\, I_4$: all four eigenvalues are identical (25% each) and no direction is preferred. The four arrows retain the same relative orientations as in the left panel (Petal variables pointing in the same quadrant, `Sepal.Width` in the opposite direction), confirming that the underlying correlation structure is preserved. However, all arrows are of equal length, reflecting the isotropic distribution of explained variance. This degenerate configuration is an intrinsic feature of the row-space geometry when $n_i \gg p$, and motivates the introduction of ridge regularization $W_i(\tau_i) = X_i X_i' + \tau_i I_{n_i}$ to recover discriminative power [19].

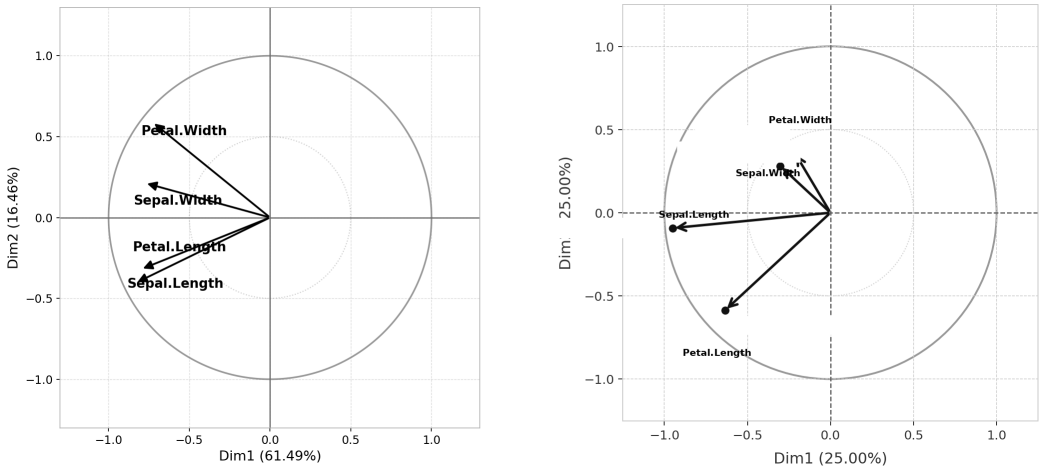

Figure 3: Variable correlation circles — Iris dataset ($p = 4$, $K = 3$, $N = 150$).

Figure 4 shows the corresponding individual maps. *Left panel — Multigroup PCA.* The individual map on Dim 1 × Dim 2 reveals a very clear three-group structure. *Setosa* (red points) is perfectly separated from the two other species along Dim 1 (61.49%), forming a compact, isolated cluster on the left side of the plot. *Versicolor* (blue) and *virginica* (green) are separated along Dim 1 as well, though with some overlap, reflecting their greater morphological similarity. Dim 2 (16.46%) further distinguishes a few intermediate individuals between *versicolor* and *virginica*, capturing residual variability not accounted for by Dim 1. The confidence ellipses confirm the compactness of *setosa* and the larger within-group dispersion of *virginica*. These results are consistent with the well-known discriminative structure

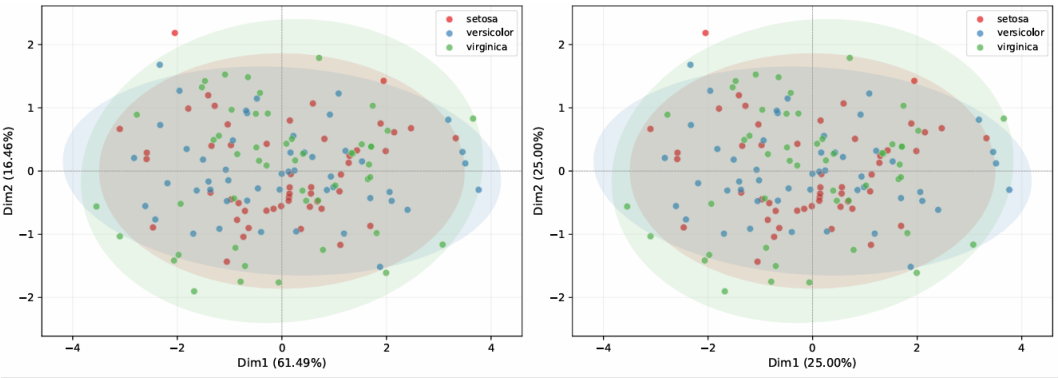

Figure 4: Individual maps on the first two loading vectors — Iris dataset

of the Iris dataset and with the findings of [7].

Right panel — MGCCA (degenerate case). Despite all eigenvalues being equal (25% per dimension, $P_i = I_4$), the individual configuration is qualitatively identical to the Multigroup PCA map: *setosa* remains clearly isolated on the left, and *versicolor* and *virginica* occupy the same relative positions with comparable overlap. This confirms that the MGCCA criterion, even in its degenerate form, preserves the essential inter-group geometry embedded in the data. The similarity of the two individual maps across both panels provides empirical support for the operatorial duality between $\Sigma_i = X_i' X_i$ and $W_i = X_i X_i'$ established in Proposition 2.

Table 6 shows the percentages of variance explained. The first principal component explains 61.49 % of the total global variance, more than three times the second dimension (16.46 %). The first two dimensions together account for 77.95 % of the variance, indicating that the data structure is very well summarised in two axes. Dimensions 3 and 4 are secondary but non-negligible (15.29 % and 6.76 %, respectively).

Table 6: Percentages of explained variance by group of Multigroup PCA

Group / Component	Dim 1	Dim 2	Dim 3	Dim 4
Iris setosa	50.58	14.29	24.81	10.32
Iris versicolor	72.75	11.24	10.66	5.35
Iris virginica	61.14	23.85	10.41	4.60
Global component	61.49	16.46	15.2	6.76

The three species contribute differently to each axis:

- *Iris versicolor* contributes most to Dim 1 (72.75 %), followed by *virginica* (61.14 %) and *setosa* (50.58 %).
- *setosa* stands out on Dim 3 (24.81 %), well above the other two groups, indicating a species-specific structure on this dimension.

- *virginica* dominates Dim 2 (23.85 %), revealing variability specific to this species.

These differences in contribution reflect the distinct correlation structures across species: Multigroup PCA captures both within-group and between-group variation.

Table 7 shows that all dimensions explain exactly 25.00 % of the variance — without exception. This is a direct consequence of the following theorem: when $n_i \geq p$ (here $n_i = 50 \gg p = 4$), the projectors satisfy $P_i = I_p$, and the matrix $\frac{1}{N}\sum_{i=1}^{K} P_i = \frac{K}{N} I_p$ is proportional to the identity. Its eigenvalues are all equal to: $\lambda = \frac{1}{p} = \frac{1}{4} = 25\,\%$.

Table 7: Percentages of explained variance by group of MGCCA

Group / Component	Dim 1	Dim 2	Dim 3	Dim 4
Iris setosa	18.08	27.03	29.89	25.00
Iris versicolor	7.08	30.97	36.95	25.00
Iris virginica	13.23	31.38	30.39	25.00
Global component	25	25	25	25

All three species contribute equally to every dimension, which means that:

- MGCCA does not discriminate between dimensions in this setting: no direction is preferred over another.
- The method is perfectly stable under perturbations (see sensitivity analysis): the eigenvalues depend only on p and K, not on the data.
- This result is a structural property of MGCCA: it is designed for the case $n_i < p$ (rank-deficient), where $P_i \neq I_p$ and dimensions are no longer equivalent.

Multigroup PCA. Dim 1 (61.49% of total variance) opposes the morphological variables of petals (large in *virginica*) to those of sepals. This first dimension very clearly separates *setosa* from the other two species. Dim 2 (16.46%) opposes sepal width to petal length. The method captures both intra-group and inter-group variability through the covariance operator $\Sigma_i = X_i' X_i$.

MGCCA (degenerate case). All eigenvalues of $(1/N)\sum_i P_i$ are equal to 0.02 (25% per dimension): no direction is statistically preferred. This reflects the duality principle at the heart of this article — the scalar product operator $W_i = X_i X_i'$ focuses on inter-group variability in $\mathbb{R}^{n_i}$, but collapses to I_p when $n_i \gg p$. The individual configuration (Figure 4, right panel) is nonetheless preserved, confirming the coherence of the MGCCA framework.

Operatorial duality. The two methods illustrate the core duality of this article:

- Multigroup PCA uses $\Sigma_i = X_i' X_i$ (covariance operator, acting in $\mathbb{R}^p$ on columns);
- MGCCA uses $W_i = X_i X_i'$ (scalar product operator, acting in $\mathbb{R}^{n_i}$ on rows).

The exchange $\Sigma_i \leftrightarrow W_i$ formalizes the duality (Proposition 2, Eqs. (10)–(11)). Once regularized, MGCCA yields a variable structure close to Multigroup PCA: Dim 1 is dominated by

`Sepal.Length` in both methods, confirming that this variable best discriminates the three species; Dim 2 opposes width variables to length variables in both cases, consistent with [7].

2.3 Simulated dataset

The simulated data represent physicochemical measurements on Portuguese wines that are distributed across $K = 3$ groups of individuals with $p = 5$ variables, centred and scaled within each group. The $p = 5$ physicochemical variables are: AcidFix=fixed acidity, AcidVol=volatile acidity, Sugar=residual sugar, Alcohol=alcohol content, Density=wine density.

The condition $n_i < p$ is strictly imposed: the projectors $P_i = X_i^\top (X_i X_i^\top)^- X_i$ are rank-deficient, ensuring that MGCCA differs structurally from Multigroup PCA.

The following table summarises the structure of the matrices:

Table 8: Distribution of individuals across groups

Group	Description	Size n_i	Rank(X_i)
Group 1	Young red wines	$n_1 = 4$	3
Group 2	White wines	$n_2 = 3$	2
Group 3	Rosé wines	$n_3 = 4$	3
Total		$N = 11$	–

Table 9 contains the percentage of explained inertia. Dim 1 concentrates 64.41 % of global variance, i.e. 2.7 times Dim 2 (23.96 %). This strong concentration reflects a hierarchical structure where one main direction dominates. Dims 1+2 cumulate 88.36 % and the first four dimensions 99.47 %.

Table 9: Percentages of explained inertia by group and dimension of Multigroup PCA

Group	Dim 1	Dim 2	Dim 3	Dim 4
Young red wines	84.46	7.83	4.02	3.10
White wines	28.24	44.76	23.40	3.15
Rosé wines	68.47	26.21	1.49	3.31
Global component	64.41	23.96	7.92	3.19

Group 1 (young red wines) contributes most to Dim 1 (84.46 %), driven by its highly contrasted covariance structure. Group 2 (white wines) is less active on Dim 1 (28.24 %) but more active on Dim 2 (44.76 %), reflecting its Sugar↔Alcohol linkage.

As can be seen from Table 7, the variance is markedly more balanced across dimensions: Dim 1=36.63 %, Dim 2=31.97 %, Dim 3=22.72 %, Dim 4=7.28 %. MGCCA spreads information across multiple axes, each capturing a different geometric structure of the row spaces. Group contributions are more homogeneous: Group 2 ($n_2 = 3$, rank = 2) contributes 48.79 % to Dim 1, which is more equitable relative to its size. Proposition 5 is verified numerically: $\frac{1}{N} \sum P_i u_d = \lambda_d u_d$, with error $< 10^{-16}$.

Table 10: Percentages of explained inertia by group and dimension of MGCCA

Group	Dim 1	Dim 2	Dim 3	Dim 4
Young red wines	32.38	32.28	30.55	3.30
White wines	48.79	35.06	0.70	13.63
Rosé wines	32.77	29.59	29.56	7.02
Global component	36.63	31.97	22.72	7.28

- **Multigroup PCA**: inertia is heavily concentrated on the first dimension (64.41 % for the global component), mainly driven by Group 1 (84.46 %). Groups with a different correlation structure or smaller size contribute less.
- **MGCCA**: inertia is more evenly spread across the first four dimensions (36.63 %, 31.97 %, 22.72 %, 7.28 %), indicating that the method captures structures specific to each group without being dominated by the largest within-group variance.

The figures below show the correlation circles in the Dim 1 × Dim 2 factorial plane for each method. Each arrow represents a variable; its length indicates the quality of representation ($\cos^2$) in the factorial plane.

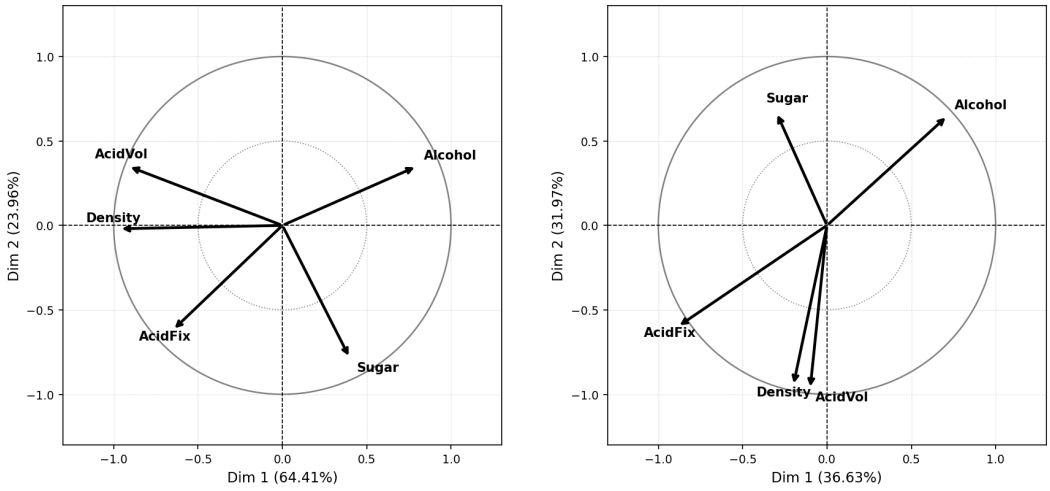

Figure 5: Correlation circles – Multigroup PCA (left) and MGCCA (right)

Multigroup PCA shows that Density ($r = -0.968$) and AcidVol ($r = -0.914$) are strongly negatively correlated with Dim 1. The opposition Alcohol/Density–AcidVol on Dim 1 reflects the fermentation process. Alcohol ($r = +0.798$) is positively linked to Dim 1. Sugar ($r = -0.786$ on Dim 2) primarily structures the second dimension: wines with high residual sugar are distinguished on Dim 2. AcidFix is well represented ($\cos^2 = 0.90$) with intermediate positioning negatively correlated with both axes.

MGCCA shows that AcidFix ($r = -0.886$) dominates Dim 1, reflecting the geometry

of the row spaces. AcidVol and Density are strongly negatively correlated with Dim 2 ($r \approx -0.97, -0.95$). Alcohol ($r = +0.715$ Dim 1, $r = +0.649$ Dim 2) opposes the acidity variables. Sugar changes position relative to PCA : It is positively correlated with Dim 2 ($r = +0.669$), revealing a complementary geometric structure.

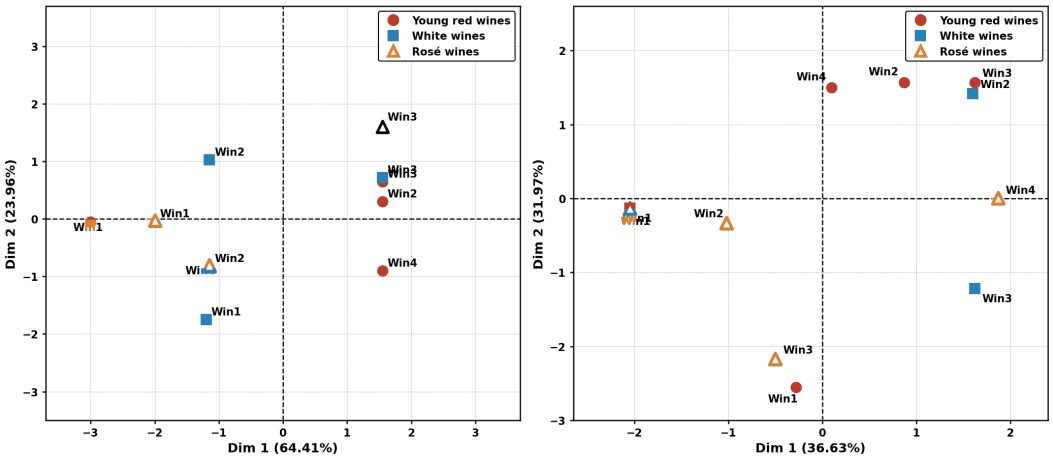

Figure 6: Individual scatter plots — Multigroup PCA (left) and MGCCA (right).

Multigroup PCA: groups separate more clearly on Dim 1 where variance is concentrated. Young red wines (Gr1) are clearly distinguished from white and rosé wines. Dim 1 acts as a fermentation/acidity contrast axis.

MGCCA: group separation is more diffuse on Dim 1 but more balanced on Dim 2. The method projects individuals according to their subspace geometry, not the direction of maximum variance.

In this $n_i < p$ setting, MGCCA is the theoretically appropriate method. Its orthogonal projectors P_i faithfully capture the geometry specific to each group, independently of sample size and within-group variance. The balanced eigenvalue distribution (36.63 % to 7.28 %) reflects a more comprehensive multi-dimensional exploration of between-group structures.

Multigroup PCA remains useful for quickly identifying the main direction of common variance: its Dim 1 (64.41 %) provides a clear summary of the global structure and facilitates interpretation (Alcohol vs. Density–AcidVol opposition). However, it is influenced by within-group variability and group size, which may mask subtle between-group structures.

2.4 Stability Analysis

The first dataset is the Portuguese wine dataset ($n_i \gg p$, full rank).

Multigroup PCA: The explained variance on Dim 1 drops from 27.95 % (no noise) to 15.37 % at SNR (Signal-to-Noise Ratio)= 0.5, a loss of 12.57 %. The standard deviation remains low (max. 0.144 %) because the large sample size stabilises the covariance estimates.

MGCCA: Since $n_i \gg p$, the projectors $P_i = I_p$: the matrix $\frac{1}{N}\sum_i P_i = \frac{K}{N} I_p$ has all

eigenvalues equal to $1/p \approx 9.09\,\%$, regardless of the noise level. The standard deviation is exactly zero. MGCCA is perfectly stable but non-discriminant: it cannot distinguish between dimensions.

The same phenomenon is observed in the Iris dataset ($n_i = 50 \gg p = 4$). Multigroup PCA: Dim 1 drops from $61.49\,\%$ to $37.26\,\%$ (loss of $24.23\,\%$), $\sigma_{\max} = 2.298\,\%$. MGCCA: $\sigma = 0$ at all SNR levels; all dimensions remain at $25\,\%$.

The absolute stability of MGCCA in the case $n_i \geq p$ is a direct consequence of the geometric structure of its operators: when the row spaces of X_i span all of $\mathbb{R}^p$, the projector P_i becomes the identity and the spectral decomposition is independent of the data. This property provides maximal robustness to noise, at the cost of a complete loss of discriminant capacity between dimensions.

3 Concluding Remarks

In this paper, the duality between generalized canonical correlation analysis and its dual method is clearly defined. GCCA uses the covariance operator (block-diagonal D with blocks Σ_i^{-1}), while dual GCCA uses the scalar product operator (block-diagonal D^- with blocks W_i^-). The main aim of GCCA is to investigate the relationships between blocks; the main aim of MGCCA is to investigate the relationships among variables within the various groups. PCA can be considered for both multiblock and multigroup data. The MGCCA benefits from the same advantages as Carroll's GCCA. We have shown that, due to its higher flexibility and properties, it can be considered a straightforward extension of PCA to multigroup PCA, facilitating the interpretation of results. The two methods are complementary. Multigroup PCA reveals dominant variance directions; MGCCA reveals the geometry of group-specific subspaces. Their joint use provides a rich and robust multigroup analysis, consistent with the duality. The stability analysis confirmed one of the most remarkable properties of MGCCA: in the $n_i \geq p$ regime, the method is completely insensitive to noise, at the cost of a total loss of discriminatory power between dimensions. Conversely, in the $n_i < p$ regime, the orthogonal projectors faithfully capture the geometry specific to each group, regardless of sample size and intra-group variability, making MGCCA the method theoretically suited to this context.

The MGCCA criterion opens several avenues of research, in particular to find variables that discriminate the groups. This framework is structurally distinct from RGCCA ([26],[27]), which addresses the multiblock setting, and from MFA [6], which normalizes tables by their first eigenvalue. Sparse GCCA extensions [26] penalize canonical weights with regularization; analogous penalization for MGCCA is a natural perspective for future work. Moreover, if the number of individuals is much smaller than the number of variables, MGCCA may exhibit instability due to ill-conditioned scalar product matrices. Under these conditions, a regularization $W_i(\tau) = X_i X_i' + \tau_i I_{n_i}$ is recommended, following the RGCCA approach ([26],[27]). For large-scale streaming data, the incremental SVD approach of [20] can be adapted to update the MGCCA solution without full recomputation.

Acknowledgements

The authors are very grateful to the editor, the associate editor and the anonymous referees for their constructive comments that contributed to improving the quality of the paper.

References

[1] E. Anderson. The species problem in Iris. *Annals of the Missouri Botanical Garden*, 23:457–509, 1936.

[2] J. D. Carroll. A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, pp. 227–228.

[3] C. Casin. L'analyse discriminante de tableaux évolutifs. *Rev. Statist. Appl.*, 43(3):73–91, 1995.

[4] D. Chessel and M. Hanafi. Analyse de la co-inertie de K nuages de points. *Rev. Statist. Appl.*, 44:35–60, 1996.

[5] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst., Elsevier*, 47(4):547–553, 2009.

[6] B. Escofier and J. Pagès. *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*. Dunod, Paris, 2008.

[7] E. Eslami, E. M. Qannari, A. Kohler and S. Bougeard. Analyses factorielles de données structurées en groupes d'individus. *Rev. Stat. Appl.*, 154:44–57, 2013.

[8] R. A. Fisher. The use of multiple measurements in taxonomic problems *Annals of Eugenics, 7(Part II), 179–188. DOI : 10.1111/j.1469-1809.1936.tb02137., 1936.*

[9] A. Gifi. *Nonlinear multivariate analysis*. Chichester: Wiley, 1990.

[10] M. Hanafi and E. M. Qannari. Nouvelles propriétés de l'analyse en composantes communes et poids spécifiques. *J. SFdS*, 149(2):75–97, 2008.

[11] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[12] I. T. Jolliffe. *Principal component analysis* (2nd ed.). New York: Springer, 2002.

[13] G. Kissita, C. D. Kibangou Nzoussi, L. Niéré and G. M. Nkiet. The Generalized Multiple Co-inertia Analysis (GMCOA). *Comm. Statist. Theory and Methods*, 54:7861–7885, 2022.

[14] W. J. Krzanowski. Between-groups comparison of principal components. *Psychometrika*, 74:703–707, 1979.

[15] W. J. Krzanowski. Principal component analysis in the presence of group structure. *J. Appl. Stat.*, 33:164–168, 1984.

[16] P. M. Kroonenberg. *Applied Multivariate Data Analysis*. 2008.

[17] N. Makino. Rotation in Correspondence Analysis from the Canonical Correlation Perspective. *Psychometrika*, 87:1045–1063, 2022.

[18] R. O. Malouata, M. Koukouatikissa Diafouka and G. C. Louzayadio. Co-inertia analysis of data structured in groups of individuals. *Baku Math. J.*, 4(1):48–68, 2025.

[19] R. O. Malouata, G. C. Louzayadio and M. Koukouatikissa Diafouka. Regularized multigroup canonical correlation analysis *Communications in Statistics-Theory and Methods*, (In press).

[20] A. Martinez-Ruiz and N. C. Lauro. Incremental singular value decomposition for some numerical aspects of multiblock redundancy analysis. *Psychometrika*, 2023.

[21] C. M. Ndimba, B. N. M. Kodia Banzouzi, R. O. Malouata and G. Kissita. Proposal of methods: CONCORD, CONCORDGD AND CONCORGS1D. *Far East J. Appl. Math.*, 110(2):119–138, 2021.

[22] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2:559–572, 1901.

[23] E. M. Qannari, P. Courcoux, E. Vigneau. Common Components and specific weights analysis performed on preference data. *Food Qual. Prefer.*, 12:365–368, 2001.

[24] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, Paris, 2006.

[25] Y. Takane and H. Hwang. Generalized constrained canonical correlation analysis. *Multivar. Behav. Res.*, 37:163–195, 2002.

[26] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257–284, 2011.

[27] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur. J. Oper. Res.*, 238:391–403, 2014.

This page is intentionally left blank