Version of Record: https://www.sciencedirect.com/science/article/pii/S0169743919303703
Manuscript_b7ef2ec2199b2ecbaef379b2a44d2403

# Unsupervised multiblock data analysis: a unified approach and extensions

Essomanda **Tchandao Mangamana**[(a, b)*], Véronique **Cariou**[a], Evelyne **Vigneau**[a], Romain Lucas **Glèlè Kakaï**[b], El Mostafa **Qannari**[a]

[a] *StatSC, ONIRIS, INRA, 44322, Nantes, France*

[b] *LABEF, University of Abomey-Calavi, 04 BP 1525, Cotonou, Bénin*

[*] Corresponding author: tchanesso@yahoo.fr

**Abstract**

For the analysis of multiblock data, a unified approach of several strategies such as Generalized Canonical Correlation Analysis (GCCA), Multiblock Principal Components Analysis (MB-PCA), Hierarchical Principal Components Analysis (H-PCA) and ComDim is outlined. These methods are based on the determination of global and block components. The unified approach postulates, on the one hand, two link functions that relate the block components to their associated global components and, on the other hand, two summing up expressions to compute the global components from their associated block components. Not only several well-known methods are retrieved but we also introduce a variant of GCCA. More generally, we hint to other possibilities of extensions thus emphasizing the fact that the unified approach, besides being simple, is versatile. We also show how this approach of analysis although basically unsupervised could be adapted to yield a supervised method to be used for a prediction purpose. Illustrations on the basis of simulated and real case studies are discussed.

Keywords: Multiblock data, Generalized Canonical Correlation Analysis, Multiblock Principal Components Analysis, Hierarchical Principal Components Analysis, ComDim

# 1 Introduction

The collection of several blocks of variables has become a common practice to study complex systems in several domains of investigation. For example, in chemometrics, the coupling of different sources of measurements generates a large amount of data which can be arranged into meaningful blocks of variables for the characterization of the same set of samples [1]. In health science, several measurements (e.g., clinical, metabolomic, transcriptomic) can be made to assess the incidence and the prevalence of a disease. In sensory analysis, assessors can be asked to score the intensity of several sensory attributes to characterize a set of products [2]. For the purpose of exploring the structure of these data blocks and investigating their relationships, unsupervised multiblock methods are often used. Since multiblock data analysis has been the focus of many research works these last three decades or so, a plethora of methods dedicated to such a purpose have been proposed and are compared in the literature [3–7]. They include in particular: H-PCA, MB-PCA also called Consensus Principal Components Analysis (CPCA), ComDim, GCCA, etc.

The aim of the paper is manifold: (i) to provide a unified approach that brings several methods of unsupervised data analysis under the same umbrella; (ii) clearly pinpoint the similarities and differences between these strategies of analysis; (iii) open venues for the developments of yet new methods of analysis, as illustrated by the introduction of an original variant of GCCA; (iv) propose optimization criteria that underly the methods of analysis. The interest of the optimization criteria is to help better interpreting the outcomes of the methods. They may also help proving the convergence of iterative algorithms. In the case of multi-start procedures where several solutions are obtained by considering several starting points, the optimization criteria make it possible to compare these solutions and choose the most optimal one. A common feature of the methods considered herein is that they fit under the umbrella of multiblock component analysis [7]. This means that they involve the determination of latent variables or components. A latent variable associated with a dataset is, by definition, a hidden variable that underlies the vector space generated by the variables in this dataset. Very often, this latent variable

is defined as a linear combination of the variables that is determined in an optimal way so as to highlight a specific purpose such as investigating the structure of the dataset, relating this dataset to other datasets, etc. Latent variables may differ in how they are standardized or how the vectors of weights are standardized. This standardization is more or less arbitrary, although some procedures of standardization may be more convenient than others for purposes such as interpretation, graphical displays, prediction, etc.

The paper is organized as follows. We start by giving a general strategy of multiblock data analysis. Then, we present the optimization criteria that underly the various methods. Thereafter, each method is reviewed in turn to specifically highlight its properties. In a subsequent section, we show how an unsupervised method can be adapted to yield a supervised method to be used for a prediction purpose. The multiblock methods are illustrated and compared on the basis of simulated data and real case studies. Finally, we end the paper by a discussion and concluding remarks.

# 2 Theory

## 2.1 A general strategy of multiblock data analysis

### 2.1.1 General considerations

Throughout this paper, we use the well-known Cauchy-Schwartz inequality, which states that: for two vectors $\boldsymbol{x}$ and $\boldsymbol{a}$, $\boldsymbol{x}^\top \boldsymbol{a} \leq ||\boldsymbol{x}||||\boldsymbol{a}||$ and the maximum of the function $\Psi(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{a}$ is achieved if and only if $\boldsymbol{x} = \lambda \boldsymbol{a}$, with $\lambda$, a scalar which can be determined by considering the constraint that is imposed on $\boldsymbol{x}$ (e.g., $||\boldsymbol{x}|| = 1$). Another relationship that we will use is that, for given blocks of variables $\boldsymbol{X}_k$ ($k = 1, 2, ..., K$), $\sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{X}_k^\top = \boldsymbol{X}\boldsymbol{X}^\top$, where $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2|...|\boldsymbol{X}_K]$ (i.e., horizontal concatenation of the blocks of variables).

### 2.1.2 Block and global components

We consider the multiblock setting where we have $K$ blocks of variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_K$, measured on the same $n$ observations and assumed to be column-centered

and pre-scaled so as to have their norms equal to 1. This pre-scaling makes it possible to put all the datasets on the same footing [5].

As stated above the methods of analysis discussed hereinafter belong to the family of multiblock component analysis [7]. They consist in determining global components or latent variables and, associated with each global component, $K$ block components respectively associated with the $K$ blocks of variables. We will focus on how to determine the first order global component and its associated block components. The subsequent components of higher order than 1 are determined following the same strategy of analysis as for the first order after a deflation with respect to the global component [8, 9]. This consists in regressing the variables in the various blocks on the global component that has just been determined and replacing the blocks of variables by the residuals of these regressions. Note that different procedures of deflation such as deflating with respect to the block latent variables could be adopted [4, 10].

### 2.1.3 Relationships between the global and block components

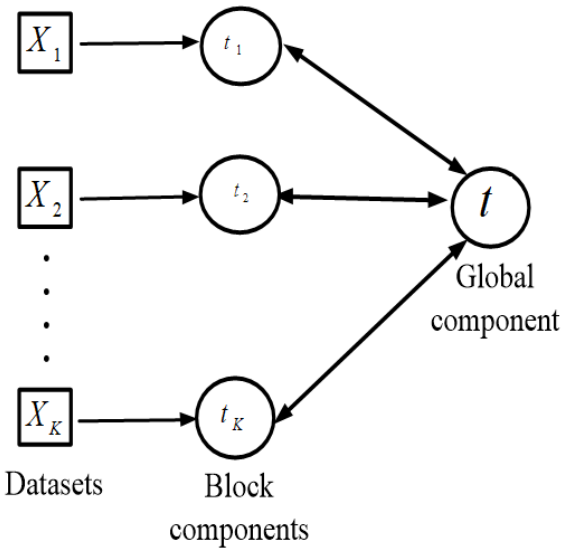



Figure 1: Relationships between the global and block components.

Figure 1 depicts a conceptual scheme which shows the various elements at hand. The connection between the global component $\boldsymbol{t}$, on the one hand, and the block components $\boldsymbol{t}_k$, on the other hand, is a two-way relationship: (i) the block components are the reflexion

of the global component in the space spanned by the variables in the various blocks, (ii) the global component stands as a summary of the block components.

To comply with the first requirement, we shall mainly consider two kinds of relationships between the global component and its associated block components. To comply with the second requirement, we shall consider two strategies of computing a synthetic variable (i.e., $\boldsymbol{t}$) from individual variables (i.e., $\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_K$). These two procedures (i.e., relationships between $\boldsymbol{t}$ and $\boldsymbol{t}_k$ and computation of a synthesis) could be crossed, leading to four methods of analysis.

The first relationship between $\boldsymbol{t}_k$ $(k = 1, 2, ..., K)$ and $\boldsymbol{t}$ postulates that:

$$\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top t = \boldsymbol{P}_k t \tag{1}$$

where $\boldsymbol{P}_k$ is the projector upon the space spanned by the variables in $\boldsymbol{X}_k$. This link function between $\boldsymbol{t}$ and $\boldsymbol{t}_k$ clearly pinpoints the idea behind the fact that $\boldsymbol{t}_k$ is a reflexion of $\boldsymbol{t}$ since, in this case, $\boldsymbol{t}_k$ is the closest variable to $\boldsymbol{t}$ in the space spanned by the variables in $\boldsymbol{X}_k$. However, since this relationship involves the inversion of the matrices $\boldsymbol{X}_k^\top \boldsymbol{X}_k$, we may face a problem of instability in presence of quasi-colinearity among the variables in one or several blocks $\boldsymbol{X}_k$ $(k = 1, 2, ..., K)$ [11–14]. To counteract this problem, we propose a second relationship between $\boldsymbol{t}_k$ $(k = 1, 2, ..., K)$ and $\boldsymbol{t}$, namely:

$$\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top t = \boldsymbol{W}_k t \tag{2}$$

where $\boldsymbol{W}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top$ is the matrix of scalar products between the observations. This expression also reflects a kind of projection of $\boldsymbol{t}$ upon the space spanned by the variables in $\boldsymbol{X}_k$. Indeed, if we denote by $\boldsymbol{x}_{k_1}, \boldsymbol{x}_{k_2}, ..., \boldsymbol{x}_{k_p}$ the $\boldsymbol{X}_k$ - variables, we have: $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t} = \sum_{j=1}^{p} \boldsymbol{x}_{k_j}\boldsymbol{x}_{k_j}^\top t = n\sum_{j=1}^{p} cov(\boldsymbol{x}_{k_j}, \boldsymbol{t})\boldsymbol{x}_{k_j}$. This means that $\boldsymbol{t}_k$ is colinear to the first partial least squares (PLS) regression component of $\boldsymbol{t}$ upon $\boldsymbol{X}_k$.

We may consider a third relationship between $\boldsymbol{t}_k$ and $\boldsymbol{t}$ which consists in setting $\boldsymbol{t}_k = \boldsymbol{X}_k(\gamma I + (1-\gamma)\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$, where $\boldsymbol{I}$ is the identity matrix and $\gamma$ is a tuning parameter comprised between 0 and 1. For $\gamma = 0$, we retrieve the relationship $\boldsymbol{t}_k = \boldsymbol{P}_k\boldsymbol{t}$ and for $\gamma = 1$, we retrieve the relationship $\boldsymbol{t}_k = \boldsymbol{W}_k\boldsymbol{t}$. The introduction of the tuning parameter, $\gamma$, is motivated by a regularization procedure akin to Ridge regression whose

aim is to prevent the problem of quasi-colinearity in a less drastic manner than that which consists in removing the matrices $(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}$ altogether [15, 16]. This kind of regularization will not be pursued any further in this paper except in the discussion section.

Let us discuss two kinds of syntheses that we can operate on the block components $\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_K$ to form the global component $\boldsymbol{t}$. We may state that $\boldsymbol{t}$ is proportional to the average of $\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_K$ or, equivalently, to their sum:

$$\boldsymbol{t} \propto \boldsymbol{t}_1 + \boldsymbol{t}_2 + ... + \boldsymbol{t}_K \tag{3}$$

where the symbol $\boldsymbol{\propto}$ means "proportional to". Alternatively, we may state that $\boldsymbol{t}$ is proportional to the first principal component of $\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_K$:

$$\boldsymbol{t} \propto \alpha_1 \boldsymbol{t}_1 + \alpha_2 \boldsymbol{t}_2 + ... + \alpha_K \boldsymbol{t}_K \tag{4}$$

where $\alpha_k \propto cov(\boldsymbol{t}, \boldsymbol{t}_k)$.

In order for the global latent variable to be close to its block latent variables, we can imagine an iterative process of reciprocal updating between the global latent variable and the block components using alternatively the link functions to compute the block components from the global component and the summing up expressions to compute the global component from the block components. Let us assume that we impose that the global component should be of norm equal to 1. Typically, we shall encounter two kinds of algorithms akin to Non Iterative PArtial Least Squares (NIPALS). The first algorithm is the following:

**Step 0.** Choose randomly $\boldsymbol{t}$ and set $\boldsymbol{t} = \boldsymbol{t}/||\boldsymbol{t}||$

**Step 1.** (Link function) : compute the block components, $\boldsymbol{t_k}$, using one or the other of the relationships considered above (i.e., $\boldsymbol{t_k} = \boldsymbol{P_k t}$ or $\boldsymbol{t_k} = \boldsymbol{W_k t}$)

**Step 2.** (Summing up expression) Update $\boldsymbol{t}$: $\boldsymbol{t} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t_k}$

**Step 3.** Set $\boldsymbol{t} = \boldsymbol{t}/||\boldsymbol{t}||$

**Step 4.** Iterate from Step 1, until convergence.

The second algorithm runs as follows:

**Step 0.** Choose randomly $\boldsymbol{t}$ and set $\boldsymbol{t} = \boldsymbol{t}/||\boldsymbol{t}||$

**Step 1.** (Link function) : compute the block components, $\boldsymbol{t_k}$, using one or the other of the relationships considered above (i.e., $\boldsymbol{t_k} = \boldsymbol{P_k t}$ or $\boldsymbol{t_k} = \boldsymbol{W_k t}$)

**Step 2.** Set $\alpha_k = \boldsymbol{t}^\top \boldsymbol{t_k}$ $(= n\boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t}))$

**Step 3.** (Summing up expression) Update $\boldsymbol{t}$: $\boldsymbol{t} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \alpha_k \boldsymbol{t_k}$

**Step 4.** Set $\boldsymbol{t} = \boldsymbol{t}/||\boldsymbol{t}||$

**Step 5.** Iterate from 1, until convergence.

The convergence of these two iterative algorithms will be assessed in the next section. Table 1 gives a classification of the methods of multiblock data analysis that we shall discuss further in subsequent sections.

Table 1: Classification of some unsupervised methods of multiblock data analysis.

| Link function (Relationship between $\boldsymbol{t}_k$ and $\boldsymbol{t}$) | Summing up expression (Summing up $\boldsymbol{t}_k$ by $\boldsymbol{t}$) | Algorithm | Method of analysis |
|---|---|---|---|
| $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$ | $\boldsymbol{t} \propto \sum_{k=1}^K \boldsymbol{t}_k$ | 1 | GCCA |
| $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$ | $\boldsymbol{t} \propto \sum_{k=1}^K \alpha_k \boldsymbol{t}_k$ | 2 | A new variant of GCCA |
| $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$ | $\boldsymbol{t} \propto \sum_{k=1}^K \boldsymbol{t}_k$ | 1 | MB-PCA |
| $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$ | $\boldsymbol{t} \propto \sum_{k=1}^K \alpha_k \boldsymbol{t}_k$ | 2 | ComDim/H-PCA |

## 2.2 Optimization criteria

We define optimization criteria for the determination of the global latent variable $\boldsymbol{t}$ and, as a byproduct, its associated block latent variables $\boldsymbol{t}_k = \boldsymbol{P}_k\boldsymbol{t}$ or $\boldsymbol{t}_k = \boldsymbol{W}_k\boldsymbol{t}$, according to which choice of the link function has been made. These optimization criteria will come as an echo to the summing up expressions defined above and which allowed us to compute the global components from the block components. Convergence properties of the two algorithms introduced above will also be discussed.

The fact that the global component should be highly related to its associated block components can be reflected by stating that the covariance between the global component and its associated block components should be, on average, as large as possible. This leads us to maximize:

$$\sum_{k=1}^{K} cov(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n}\sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{t}_k = \frac{1}{n}\boldsymbol{t}^\top (\sum_{k=1}^{K} \boldsymbol{t}_k). \tag{5}$$

We choose as a determination constraint $||\boldsymbol{t}|| = 1$. This optimization problem is precisely solved by Algorithm 1. Indeed, for $\boldsymbol{t}$ fixed, $\boldsymbol{t}_k$ are defined by $\boldsymbol{t}_k = \boldsymbol{P}_k\boldsymbol{t}$ or $\boldsymbol{t}_k = \boldsymbol{W}_k t$, as the case applies. For $\boldsymbol{t}_k$ fixed, we can apply the Cauchy-Schwartz inequality to the last member of the equation (5). This leads us to consider $\boldsymbol{t} = \psi \sum_{k=1}^{K} \boldsymbol{t}_k$. The scalar $\psi$ can be computed by considering the determination constraint imposed on $\boldsymbol{t}$, namely $||\boldsymbol{t}|| = 1$.

We can also note that at each updating of the global component, the criterion to be maximized increases. Moreover, we can show that the criterion is upper bounded. For instance, if we consider the case where $\boldsymbol{t}_k = \boldsymbol{W}_k\boldsymbol{t}$, we have:

$$\begin{aligned} \boldsymbol{t}^\top \sum_{k=1}^{K} \boldsymbol{t}_k \leq ||\boldsymbol{t}||\ ||\sum_{k=1}^{K} \boldsymbol{t}_k|| \leq \sum_{k=1}^{K} ||\boldsymbol{t}_k|| &= \sum_{k=1}^{K} \sqrt{\boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}} \\ &\leq \sum_{k=1}^{K} \sqrt{||\boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top||}. \end{aligned} \tag{6}$$

Being an increasing and upper bounded criterion, the algorithm will converge as the number of iterations increases. Thus, the convergence of Algorithm 1 should be understood in the sense that criterion (5) ceases to increase by less than a pre-specified threshold (e.g., $\epsilon = 10^{-8}$). It is also worth noting that in order to avoid that the iterative algorithm gets stuck in local optima, it is recommended to operate a multi-start procedure and eventually choose the solution that corresponds to the largest value of the criterion.

As a matter of fact, we can propose a straightforward solution to the maximization problem (5) which does not necessitate an iterative algorithm. For the case where the link function is given by $\boldsymbol{t}_k = \boldsymbol{W}_k\boldsymbol{t} = \boldsymbol{X}_k\boldsymbol{X}_k^\top\boldsymbol{t}$, we have: $\sum_{k=1}^{K} cov(\boldsymbol{t}_k, \boldsymbol{t}) = \frac{1}{n}\boldsymbol{t}^\top \sum_{k=1}^{K} \boldsymbol{t}_k = \frac{1}{n}\boldsymbol{t}^\top \sum_{k=1}^{K} \boldsymbol{X}_k\boldsymbol{X}_k^\top\boldsymbol{t} = \frac{1}{n}\boldsymbol{t}^\top \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{t}$, where $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2|...|\boldsymbol{X}_K]$. It is known that the maximum of this expression with respect to $\boldsymbol{t}$ is achieved for $\boldsymbol{t}$ equal to the eigenvector of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ associated with the largest eigenvalue. The implication of this finding is that $\boldsymbol{t}$ is the first standardized principal component of $\boldsymbol{X}$. This is a known property of MB-PCA [4]. As for the case where the link function is given by $\boldsymbol{t}_k = \boldsymbol{P}_k\boldsymbol{t} = \boldsymbol{X}_k(\boldsymbol{X}_k^\top\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top\boldsymbol{t}$, we can show following the same pattern of development that $\boldsymbol{t}$ is an eigenvector of $\sum_{k=1}^{K} \boldsymbol{P}_k$ associated with the largest eigenvalue. This is a characteristic property of GCCA [17].

The advantage of the NIPALS-like algorithm over the eigenanalysis solution is that it is faster, can handle very large matrices and can be accomodated in order to cope with missing data [18].

An alternative criterion to express that the global latent variable is highly linked to its associated block components is the following:

$$\sum_{k=1}^{K} cov^2(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n^2}\sum_{k=1}^{K}(\boldsymbol{t}^\top \boldsymbol{t}_k)^2 = \frac{1}{n^2}\sum_{k=1}^{K}\boldsymbol{t}^\top \boldsymbol{t}_k \boldsymbol{t}_k^\top \boldsymbol{t} = \frac{1}{n^2}\boldsymbol{t}^\top(\sum_{k=1}^{K}\boldsymbol{t}_k \boldsymbol{t}_k^\top)\boldsymbol{t} = \frac{1}{n^2}\boldsymbol{t}^\top \boldsymbol{T}\boldsymbol{T}^\top \boldsymbol{t} \quad (7)$$

where $\boldsymbol{T} = [\boldsymbol{t}_1|\boldsymbol{t}_2|...|\boldsymbol{t}_K]$. The solution to this problem entails that $\boldsymbol{t}$ is the first standardized principal component of $\boldsymbol{T}$. From the second member of the formula (7), we can write $\sum_{k=1}^{K} cov^2(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n^2}\sum_{k=1}^{K}\alpha_k \boldsymbol{t}^\top \boldsymbol{t}_k$, where $\alpha_k = \boldsymbol{t}_k^\top \boldsymbol{t}$, which is proportional to the covariance between $\boldsymbol{t}_k$ and $\boldsymbol{t}$. It follows that for a fixed $\boldsymbol{t}$, $\boldsymbol{t}_k = \boldsymbol{W}_k \boldsymbol{t}$ or $\boldsymbol{t}_k = \boldsymbol{P}_k \boldsymbol{t}$ as the case applies and for fixed $\boldsymbol{t}_k$, $\boldsymbol{t} \propto \sum_{k=1}^{K}\alpha_k \boldsymbol{t}_k$, by virtue of Cauchy-Schwartz inequality.

At each updating of $\boldsymbol{t}$ in this iterative algorithm, criterion (7) increases. Moreover, since the criterion is upper bounded, this entails that its algorithm converges.

## 2.3 Methods of multiblock data analysis

To enable us to achieve a straightforward comparison of the various methods of unsupervised data analysis, we will present these methods in summary tables which in a way represent the identity card for each method. Hopefully, the comparison between the methods could easily be made. Additional properties will be discussed with the aim of enhancing the interpretation of the outcomes when applying these methods to specific datasets.

### 2.3.1 GCCA

As early as 1936, Hotelling [19] introduced canonical correlation analysis whose aim is to seek linear combinations of two blocks of variables with maximum correlation. In 1968, Carroll [17] extended this method of analysis to investigate the structure of $K$ blocks of variables. This consists in finding, in a first stage, a global component $\boldsymbol{t}$ and its associated block components $\boldsymbol{t}_k$ so as to maximize $\sum_{k=1}^{K} cor^2(\boldsymbol{t}, \boldsymbol{t}_k)$, where $\boldsymbol{t}_k = \boldsymbol{P}_k \boldsymbol{t}$ and $cor()$ is the

correlation coefficient. Since the correlation coefficient is scale invariant, we can arbitrary choose $||\boldsymbol{t}|| = 1$. The criterion to be maximized is equivalent to: $\sum_{k=1}^{K} \frac{(\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t})^2}{||\boldsymbol{t}||^2 ||\boldsymbol{P}_k \boldsymbol{t}||^2} = \sum_{k=1}^{K} \frac{(\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t})^2}{||\boldsymbol{P}_k \boldsymbol{t}||^2} = \sum_{k=1}^{K} \frac{(\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t})^2}{\boldsymbol{t}^\top \boldsymbol{P}_k^\top \boldsymbol{P}_k \boldsymbol{t}}$. Since $\boldsymbol{P}_k$ is symmetric ($\boldsymbol{P}_k^\top = \boldsymbol{P}_k$ ) and idempotent ($\boldsymbol{P}_k^2 = \boldsymbol{P}_k$ ), it follows that the criterion to be maximized is equivalent to $\sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t} = n \sum_{k=1}^{K} cov(\boldsymbol{t}, \boldsymbol{P}_k \boldsymbol{t})$, which is, in its turn, equivalent to the criterion (5) introduced in section 2.2. The quantity $cov(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n} \boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t}$ which appears in the maximization criterion is also equal to $\frac{1}{n} \boldsymbol{t}^\top \boldsymbol{P}_k^\top \boldsymbol{P}_k \boldsymbol{t} = var(\boldsymbol{P}_k \boldsymbol{t})$, where $var()$ stands for the variance, because as stated above $\boldsymbol{P}_k$ is symmetric and idempotent. Moreover, since we have assumed that $||\boldsymbol{t}|| = 1$, it follows that $\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t} = \frac{var(\boldsymbol{P}_k \boldsymbol{t})}{var(\boldsymbol{t})}$ is the coefficient of determination, $R^2(\boldsymbol{t}/\boldsymbol{X}_k)$ of $\boldsymbol{t}$ with respect to $\boldsymbol{X}_k$. From this stand point, it appears that GCCA seeks the direction in the space, that is, on average, best explained by the blocks of variables $\boldsymbol{X}_k$. The overall importance of the component $\boldsymbol{t}$ can be assessed by $\frac{1}{K} \sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t}$, which highlights how, on average, $\boldsymbol{t}$ is related to the blocks of variables $\boldsymbol{X}_k$. Table 2 gives a summary of the properties of the global component $\boldsymbol{t}$ and its associated block components.

Table 2: Generalized canonical correlation analysis (GCCA).

| | |
|---|---|
| Link function between block and global components | $\boldsymbol{t}_k = \boldsymbol{X}_k (\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^\top \boldsymbol{t} = \boldsymbol{P}_k \boldsymbol{t}$ |
| Summing up expression | $\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{t}_k$ |
| Maximization criterion | $\sum_{k=1}^{K} cov(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n} \sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t}$, with $\|\|\boldsymbol{t}\|\| = 1$ |
| Solution | Algorithm 1 |
| Eigenanalysis solution | $\boldsymbol{t}$ eigenvector of $\sum_{k=1}^{K} \boldsymbol{P}_k$ |

### 2.3.2 GCCA-V: A variant of GCCA

A variant of GCCA is simply obtained by considering the same link function as for GCCA but we choose for the summing up expression the one that states that $\boldsymbol{t}$ is the first principal component of its associated block components. Table 3 sums up this procedure of analysis.

Table 3: A variant of Generalized canonical correlation analysis (GCCA-V).

| | |
|---|---|
| Link function between block and global components | $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t} = \boldsymbol{P}_k \boldsymbol{t}$ |
| Summing up expression | $\boldsymbol{t} \propto \sum_{k=1}^K \alpha_k \boldsymbol{t}_k$, with $\alpha_k = \boldsymbol{t}^\top \boldsymbol{t}_k$ |
| Maximization criterion | $\sum_{k=1}^K cov^2(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n^2}\sum_{k=1}^K \alpha_k \boldsymbol{t}^\top \boldsymbol{P}_k t$, with $\|\|\boldsymbol{t}\|\| = 1$ |
| Solution | Algorithm 2 |
| Eigenanalysis solution | - |

Using similar developments as for GCCA, it follows:
$cov(\boldsymbol{t}_k, \boldsymbol{t}) = \frac{1}{n}\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t} = var(\boldsymbol{P}_k \boldsymbol{t})$ and $\boldsymbol{t}^\top \boldsymbol{P}_k \boldsymbol{t} = R^2(\boldsymbol{t}/\boldsymbol{X}_k)$. This shows that we could compute similar indices as for GCCA to highlight the relative importance of the components.

### 2.3.3 MB-PCA

With the advent of measurement methods that yield datasets where the number of variables is often larger than that of the samples and where, moreover, the variables are highly correlated, MB-PCA also known as Consensus PCA has dethroned GCCA in terms of popularity [14]. Indeed, in these situations GCCA is not applicable unless a regularization procedure is introduced [11–14]. Table 4 sums up the main features of MB-PCA.

Table 4: Multiblock PCA (MB-PCA).

| | |
|---|---|
| Link function between block and global components | $\boldsymbol{t}_k = \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t} = \boldsymbol{W}_k \boldsymbol{t}$ |
| Summing up expression | $\boldsymbol{t} \propto \sum_{k=1}^K \boldsymbol{t}_k$ |
| Maximization criterion | $\sum_{k=1}^K cov(\boldsymbol{t}, \boldsymbol{t}_k) = \frac{1}{n}\sum_{k=1}^K \boldsymbol{t}^\top \boldsymbol{W}_k \boldsymbol{t}$, with $\|\|\boldsymbol{t}\|\| = 1$ |
| Solution | Algorithm 1 |
| Eigenanalysis solution | $\boldsymbol{t}$ eigenvector of $\frac{1}{n}\sum_{k=1}^K \boldsymbol{X}_k \boldsymbol{X}_k^\top = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$<br>($\boldsymbol{t}$ is the first standardized principal component of $\boldsymbol{X}$) |

From the criterion to be maximized, we single out the quantity $\boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t} = n\, cov(\boldsymbol{t}_k, \boldsymbol{t})$,

which reflects the contribution of the block $\boldsymbol{X}_k$ to the determination of the global component $\boldsymbol{t}$. We can also note that since $\boldsymbol{t}$ is assumed to be of length 1, the quantity $\boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}$ reflects the variation in $\boldsymbol{X}_k$ explained by $\boldsymbol{t}$. Moreover, since the norm of $\boldsymbol{X}_k$ was set to 1, the quantity $\boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}$ also expresses the percentage of variation in $\boldsymbol{X}_k$ explained by $\boldsymbol{t}$. The overall importance of $\boldsymbol{t}$ (in percentage) is assessed by $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}$. In this expression, the denominator $K$ corresponds to the total variation in all the blocks: $\sum_{k=1}^{K} trace(\boldsymbol{X}_k \boldsymbol{X}_k^\top) = K$.

For graphical displays, it may be useful to rescale the global component $\boldsymbol{t}$ so that its variance reflects the variation in the various blocks explained by this component. This amounts to considering $\tilde{\boldsymbol{t}} = \boldsymbol{\mu t}$, where $\mu = \sqrt{\sum_{k=1}^{K} cov(\boldsymbol{t}, \boldsymbol{t}_k)} = \sqrt{\frac{1}{n}\boldsymbol{t}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{t}}$. In other words, $\tilde{\boldsymbol{t}}$ corresponds to the non-standardized principal component of $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2|...|\boldsymbol{X}_K]$.

### 2.3.4 ComDim and H-PCA

ComDim stands for an abbreviation of "Common Dimensions". It originated in the context of sensory analysis [5, 20] and was applied to various domains of applications [21–24]. Hanafi et al. [6] showed that it is equivalent to H-PCA which was introduced by Wold et al. [25]. Table 5 sums up how ComDim and H-PCA proceed.

Table 5: ComDim / H-PCA.

| | |
|---|---|
| Link function between block and global components | $\boldsymbol{t}_k = \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t} = \boldsymbol{W}_k \boldsymbol{t}$ |
| Summing up expression | $\boldsymbol{t} \propto \sum_{k=1}^{K} \alpha_k \boldsymbol{t}_k$, with $\alpha_k = \boldsymbol{t}^\top \boldsymbol{t}_k$ |
| Maximization criterion | $\sum_{k=1}^{K} cov^2(\boldsymbol{t}, \boldsymbol{t}_k)$, with $\|\|\boldsymbol{t}\|\| = 1$ |
| Solution | Algorithm 2 |
| Eigenanalysis solution | - |

As for MB-PCA, $\boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}$ which is equal to $\alpha_k$, reflects the contribution of the block $\boldsymbol{X}_k$ to the determination of $\boldsymbol{t}$. It also represents the total variance in $\boldsymbol{X}_k$ explained by $\boldsymbol{t}$. The overall importance of $\boldsymbol{t}$ is assessed by $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{t}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}$. For the graphical displays, we recommend to rescale $\boldsymbol{t}$ by multiplication of the scalar $\sqrt{\sum_{k=1}^{K} cov^2(\boldsymbol{t}, \boldsymbol{t}_k)}$.

It should be noted that there is a version of H-PCA, where the block components are scaled to unit length after each updating [4]. This version of H-PCA seems to have some convergence problems. In any case, we are not concerned by this version.

## 2.4 Comparison of methods

The first key of differentiation between the methods is whether we choose as a link function between the global component, $\boldsymbol{t}$, and the block components $\boldsymbol{t}_k$, the expression $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$ or $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$. The former expression entails that, within each block of variables, the variation in terms of variances and correlations of the variables is obliterated. Therefore, the focus of the method of analysis is to investigate the relationships between the datasets. By contrast, if the latter link function (i.e., $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$) is chosen, then the method of analysis will seek to recover the within block variation as well as investigate the relationships between the blocks of variables. The second key of differentiation is which summing up expression one considers. By choosing the expression $\boldsymbol{t} \propto \sum_{k=1}^{K} \alpha_k \boldsymbol{t}_k$ (i.e., $\boldsymbol{t}$ is the first principal component of $\boldsymbol{t}_k$), the blocks of variables are weighted according to how they agree with each others. In particular, those blocks of variables which do not convey the same information as the other blocks will be down-weighted.

## 2.5 From an unsupervised method to a supervised method

The boundary between unsupervised and supervised methods is thin. For instance, consider the case of GCCA. In this method of analysis, all the datasets at hand play the same role and, from this perspective, GCCA stands as an unsupervised method. Yet, we know that multiple linear regression and linear discriminant analysis, which are supervised methods par excellence, are particular cases of GCCA [26]. The explanation of this seemingly paradoxial finding is that a method of analysis can be conceptually unsupervised, yet, the outcomes of the method of analysis could be used for a supervised purpose. Another method which illustrates this idea is Latent Root Regression [27–29], where a regression model between a univariate variable $\boldsymbol{y}$ and a dataset $\boldsymbol{X}$ is set up by using the

outcomes from a PCA of the dataset $[\boldsymbol{y}|\boldsymbol{X}]$. This idea was extended to the case of a multivariate $\boldsymbol{Y}$ to be predicted by a dataset $\boldsymbol{X}$ [28]. Bougeard et al. [30] introduced a method of analysis called Multiblock Latent Root Regression (MB-LRR) where the aim is to predict a dataset $\boldsymbol{Y}$ from several blocks of variables $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$. The rationale behind MB-LRR is to perform a method of analysis akin to MB-PCA on $\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$, but the global components are constrained to be linear combinations of the variables in the $\boldsymbol{X}_k$ blocks. The procedure of analysis proposed herein is more straightforward and takes advantage of the determination of the block components associated with the blocks of variables. We shall refer to this strategy of analysis as LR-MBPCA, which stands for "Latent Root Multiblock Principal Component Analysis". For the sake of simplicity, we shall restrict ourselves to the case where the multiblock data analysis that is applied to $\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$ is MB-PCA. Obviously, the strategy of analysis can be easily adapted to the other methods.

The general idea is the following. Consider a setting with a dataset $\boldsymbol{Y}$ to be predicted from $K$ datasets $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$. All these datasets are supposed to be column-centered and pre-scaled as discussed below. We advocate performing MB-PCA on $\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$. This yields, in a first stage, a global component $\boldsymbol{t}^{(1)}$ and its associated block components $\boldsymbol{t}_0^{(1)}, \boldsymbol{t}_1^{(1)}, \boldsymbol{t}_2^{(1)}, ..., \boldsymbol{t}_K^{(1)}$ respectively associated with $\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_K}$. A predictive latent component $\boldsymbol{t}_{\boldsymbol{X}}^{(1)} = \boldsymbol{t}_1^{(1)} + \boldsymbol{t}_2^{(1)} + ... + \boldsymbol{t}_K^{(1)}$ is used to predict $\boldsymbol{Y}$ from $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$. Note that, in the expression of $\boldsymbol{t}_{\boldsymbol{X}}^{(1)}$, only the block components associated with the $\boldsymbol{X}_k$ blocks are considered. The same procedure could be applied to the successive global latent variables and their associated block latent variables, yielding new predictive components. These predicitve components are standardized so that their norms are equal to 1.

Two recommendations are of prime importance. The first recommendation is that we advocate performing a deflation with respect to the predictive components $\boldsymbol{t}_{\boldsymbol{X}}^{(1)}, \boldsymbol{t}_{\boldsymbol{X}}^{(2)}, ..., \boldsymbol{t}_{\boldsymbol{X}}^{(A)}$ instead of the global components $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, ..., \boldsymbol{t}^{(A)}$; A being the number of components to be introduced in the prediction model. As a result, the predictive components $\boldsymbol{t}_{\boldsymbol{X}}^{(1)}, \boldsymbol{t}_{\boldsymbol{X}}^{(2)}, ..., \boldsymbol{t}_{\boldsymbol{X}}^{(A)}$ will be orthogonal, which will very likely improve their predictive ability. As stated in section 2.1.2, other deflation strategies could as well be adopted. The one advocated herein

is in line with the deflation that we have adopted up to now (i.e., deflation with respect to the global components). For the second recommendation, we advocate that the $\boldsymbol{X}_k$ blocks of variables should be pre-scaled as previously by dividing them by their respective norms. As for the dataset $\boldsymbol{Y}$, in addition to the division by its norm, we advocate multiplying it by $\sqrt{K}$. By so doing, the total variation in the dataset $\boldsymbol{Y}$ alone will become equal to that of the datasets $\boldsymbol{X}_k$ $(k = 1, 2, ..., K)$ all together. As a consequence, the first component derived from MB-PCA performed on $\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K$ will likely be related to $\boldsymbol{Y}$, carrying along all the connected information from the other blocks of variables. The multiplication of $\boldsymbol{Y}$ by $\sqrt{K}$ should be operated after each deflation with respect to the predictive components in order to ensure that the new component that emerges is highly linked to $\boldsymbol{Y}$.

In order to set up a predictive model, we should note that each block component $\boldsymbol{t}_k^{(1)} = \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{t}^{(1)}$ is, by definition, a linear combination of the variables in $\boldsymbol{X}_k$: $\boldsymbol{t}_k^{(1)} = \boldsymbol{X}_k \boldsymbol{w}_k^{(1)}$, with $\boldsymbol{w}_k^{(1)} = \boldsymbol{X}_k^\top \boldsymbol{t}^{(1)}$. Since the global predictive component $\boldsymbol{t}_{\boldsymbol{X}}^{(1)} = \sum_{k=1}^{K} \boldsymbol{t}_k^{(1)}$, it follows that the global vector of weights, $\boldsymbol{w}^{(1)}$, is formed by the concatenation of the vectors $\boldsymbol{w}_k^{(1)}$. Obviously, these remarks are also valid for the subsequent predictive components, thus leading to the global vector of weights $\boldsymbol{w}^{(2)}$, $\boldsymbol{w}^{(3)}$, .... Each of these vectors is normalized to be of length 1. Let us denote by $\boldsymbol{W}$, the matrix formed by these vectors. Associated with each predictive component $\boldsymbol{t}_{\boldsymbol{X}}^{(h)}$, we compute the vector of loadings $\boldsymbol{p}_{\boldsymbol{X}}^{(h)} = \frac{\boldsymbol{X}^\top \boldsymbol{t}_{\boldsymbol{X}}^{(h)}}{\boldsymbol{t}_{\boldsymbol{X}}^{(h)\top} \boldsymbol{t}_{\boldsymbol{X}}^{(h)}}$. This is the regression coefficient of $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K]$ on $\boldsymbol{t}_{\boldsymbol{X}}^{(h)}$. Let us denote by $\boldsymbol{P}_{\boldsymbol{X}}$, the matrix whose columns are the vectors of $\boldsymbol{p}_{\boldsymbol{X}}^{(h)}$. Similarly, we can compute, at each step $h$, the vector of loadings $\boldsymbol{p}_{\boldsymbol{Y}}^{(h)} = \frac{\boldsymbol{Y}^\top \boldsymbol{t}_{\boldsymbol{X}}^{(h)}}{\boldsymbol{t}_{\boldsymbol{X}}^{(h)\top} \boldsymbol{t}_{\boldsymbol{X}}^{(h)}} \times \frac{1}{K^{\frac{h-1}{2}}}$. The constant $\frac{1}{K^{\frac{h-1}{2}}}$ is introduced to take account of the scaling of the $\boldsymbol{Y}$-variables after each deflation. Let us denote by $\boldsymbol{P}_{\boldsymbol{Y}}$, the matrix containing the vectors $\boldsymbol{p}_{\boldsymbol{Y}}^{(h)}$. It is known that the matrix of weights, $\boldsymbol{W}^*$ which directly refer to the $\boldsymbol{X}$-variables (instead of the deflated $\boldsymbol{X}$-variables) is given by [18,31]:

$$\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{P}_{\boldsymbol{X}}^\top \boldsymbol{W})^{-1} \tag{8}$$

A prediction model to predict $\boldsymbol{Y}$ from $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K]$ is obtained by regressing $\boldsymbol{Y}$ on the $\boldsymbol{X}\boldsymbol{W}^*$, yielding:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \tag{9}$$

where, $\boldsymbol{\beta} = \boldsymbol{W}^* \boldsymbol{P}_{\boldsymbol{Y}}^\top$ is the matrix of the regression coefficients, $\boldsymbol{E}$ is the $\boldsymbol{Y}$-residual matrix.

The appropriate number of components, $A$, to be introduced in the model can be selected by a cross-validation procedure as it is common practice with PLS methods [32, 33].

# 3 Illustrations

A simulation study and a case study are used to illustrate and compare the various unsupervised methods. A third case study is used to illustrate the predictive approach, LR-MBPCA, proposed herein.

## 3.1 A simulation study

The simulated data follow the same pattern as that considered by Westerhuis et al. [4]. Two orthogonal variables $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ and four blocks are considered. All the blocks have fifty observations and five variables. Variables in block $\boldsymbol{X}_1$ are formed by the variable $\boldsymbol{d}_1$ plus twenty percent of random noise: $x_{1j} = \boldsymbol{d}_1 + 0.2\epsilon_{1j}$, where $j = 1, 2, ..., 5$, $\epsilon_{1j} \sim \mathcal{N}(m_1, \sigma_1)$ and $m_1$, $\sigma_1$ are respectively the mean and standard deviation of the variable $\boldsymbol{d}_1$. In block $\boldsymbol{X}_2$, only the first variable is formed by the variable $\boldsymbol{d}_2$ plus twenty percent of random noise; the remaining variables are formed of random noise: $x_{21} = \boldsymbol{d}_2 + 0.2\epsilon_{21}$, where $\epsilon_{21} \sim \mathcal{N}(m_2, \sigma_2)$ and $m_2$, $\sigma_2$ are respectively the mean and standard deviation of the variable $\boldsymbol{d}_2$. $x_{2j} = \epsilon_{2j}$, with $\epsilon_{2j} \sim \mathcal{N}(0, 1)$ and $j = 2, ..., 5$. Blocks $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$ are formed in a similar way as $\boldsymbol{X}_2$. Note that the variable $\boldsymbol{d}_1$ appears five times in $\boldsymbol{X}_1$, whereas, $\boldsymbol{d}_2$ appears in total three times; once in each block $\boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$.

Table 6 gives the percentages of total variance in $\boldsymbol{X}_1$ to $\boldsymbol{X}_4$ recovered by the first two global components computed by means of GCCA, GCCA-V, MB-PCA and ComDim/H-PCA. Not surprisingly, GCCA and GCCA-V seem to be more concerned with the common direction, $\boldsymbol{d}_2$, to the blocks of variables $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$. The first GCCA component is highly correlated with $\boldsymbol{d}_2$ ($r = 0.97$), so is the first GCCA-V component ($r = 0.99$). Contrariwise, the first global component of MB-PCA, on the one hand, and ComDim/H-

PCA, on the other hand are almost completely devoted to recovering the variation in $\boldsymbol{X}_1$ because this block of variables carries one unique information with a substantial weight that is, the variable $\boldsymbol{d}_1$ repeated five times, apart from the added noise. The second global component of MB-PCA, as well as that of ComDim/H-PCA are oriented towards the variable $\boldsymbol{d}_2$ which constitutes a common pattern to $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$. As for the second global components of GCCA and GCCA-V, they both seem to be reflecting noise only. This can be explained by the fact that once the variable $\boldsymbol{d}_2$ is accounted for by the first global component, no common information is left.

It is worth noting that the findings regarding MB-PCA agree with those presented by Westerhuis et al. [4]. However, this is not the case for the findings regarding H-PCA since these authors applied the version of H-PCA where the block components are standardized.

Table 6: Simulated data: Percentages of total variance in blocks $\boldsymbol{X}_1$ to $\boldsymbol{X}_4$ explained by the first two global components and correlations of these global components with variables $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$.

| | | % total variance | | | | | Correlations | |
|---|---|---|---|---|---|---|---|---|
| | | $\boldsymbol{X}_1$ | $\boldsymbol{X}_2$ | $\boldsymbol{X}_3$ | $\boldsymbol{X}_4$ | Global | $\boldsymbol{d}_1$ | $\boldsymbol{d}_2$ |
| GCCA | Dim.1 | 0.28 | **20.52** | **22.64** | **19.98** | 15.85 | 0.03 | **0.97** |
| | Dim.2 | 0.56 | 3.23 | 8.31 | 4.47 | 4.14 | 0.01 | -0.20 |
| GCCA-V | Dim.1 | 0.15 | **21.10** | **22.28** | **21.23** | 16.19 | 0.03 | **0.99** |
| | Dim.2 | 0.07 | 6.29 | 4.96 | 0.40 | 2.93 | 0.03 | -0.01 |
| MB-PCA | Dim.1 | **96.11** | 2.55 | 1.59 | 1.23 | 25.37 | **0.99** | 0.01 |
| | Dim.2 | 0.17 | **21.05** | **23.28** | **22.09** | 16.65 | -0.01 | **0.97** |
| ComDim | Dim.1 | **97.14** | 1.70 | 1.10 | 0.73 | 25.17 | **1.00** | -0.01 |
| | Dim.2 | 0.12 | **20.88** | **23.55** | **22.04** | 16.65 | 0.01 | **0.97** |

## 3.2 Sensory data

The data are extracted from a study which concerns eight American dry-cured ham products differing in aging times [34]. These products were subjected to a sensory evaluation performed by a panel of trained assessors. More precisely, assessors described the

ham flavor (3 variables), aroma (4 variables) and texture (3 variables). Each assessor was instructed to rate, for each product, the intensity of the sensory variables using a 15-point intensity scale, where 0 corresponds to "not detected" and 15 corresponds to "extremely strong". For each ham and each sensory attribute, the data were averaged accross the assessors. The average intensities thus obtained were organized in three blocks of variables respectively corresponding to flavor, aroma and texture. The rows of each block of variables refer to the eight hams and the columns to the sensory attributes. Table 7 shows the total variance explained by the first two global components of GCCA, GCCA-V, MB-PCA and ComDim.

Table 7: Sensory data: Total variance explained by the first two global components.

| | | $\boldsymbol{X}_1$ (Flavor) | $\boldsymbol{X}_2$ (Aroma) | $\boldsymbol{X}_3$ (Texture) | Global |
|---|---|---|---|---|---|
| GCCA | Dim.1 | **48.44** | **40.74** | **30.12** | 39.77 |
| | Dim.2 | 12.68 | 13.77 | 16.33 | 14.26 |
| GCCA-V | Dim.1 | **48.94** | **40.80** | **29.46** | 39.73 |
| | Dim.2 | 12.17 | 13.71 | 16.98 | 14.29 |
| MB-PCA | Dim.1 | **35.49** | **44.51** | **62.70** | 47.57 |
| | Dim.2 | **34.75** | **28.80** | 19.30 | 27.61 |
| ComDim | Dim.1 | **22.58** | **39.38** | **74.66** | 45.54 |
| | Dim.2 | **44.62** | **30.93** | 10.31 | 28.62 |

From Table 7, we can see that the first global component of GCCA and GCCA-V recovers a good proportion of variation in $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and, to a lesser extent, $\boldsymbol{X}_3$.
Table 8, which gives the correlations of these components with the sensory attributes, shows that the first global component associated to GCCA and GCCA-V carries informations from the three blocks of variables related to "Porkcomplex" and "Savory" ($\boldsymbol{X}_1$), "Molasses" and "Caramelized" ($\boldsymbol{X}_3$) and "Mushiness" ($\boldsymbol{X}_3$). The first global component of MB-PCA shows a different pattern since it recovers up to 62.70 % of the total variance in $\boldsymbol{X}_3$ and the smallest recovered variation is that associated with $\boldsymbol{X}_1$. This finding is

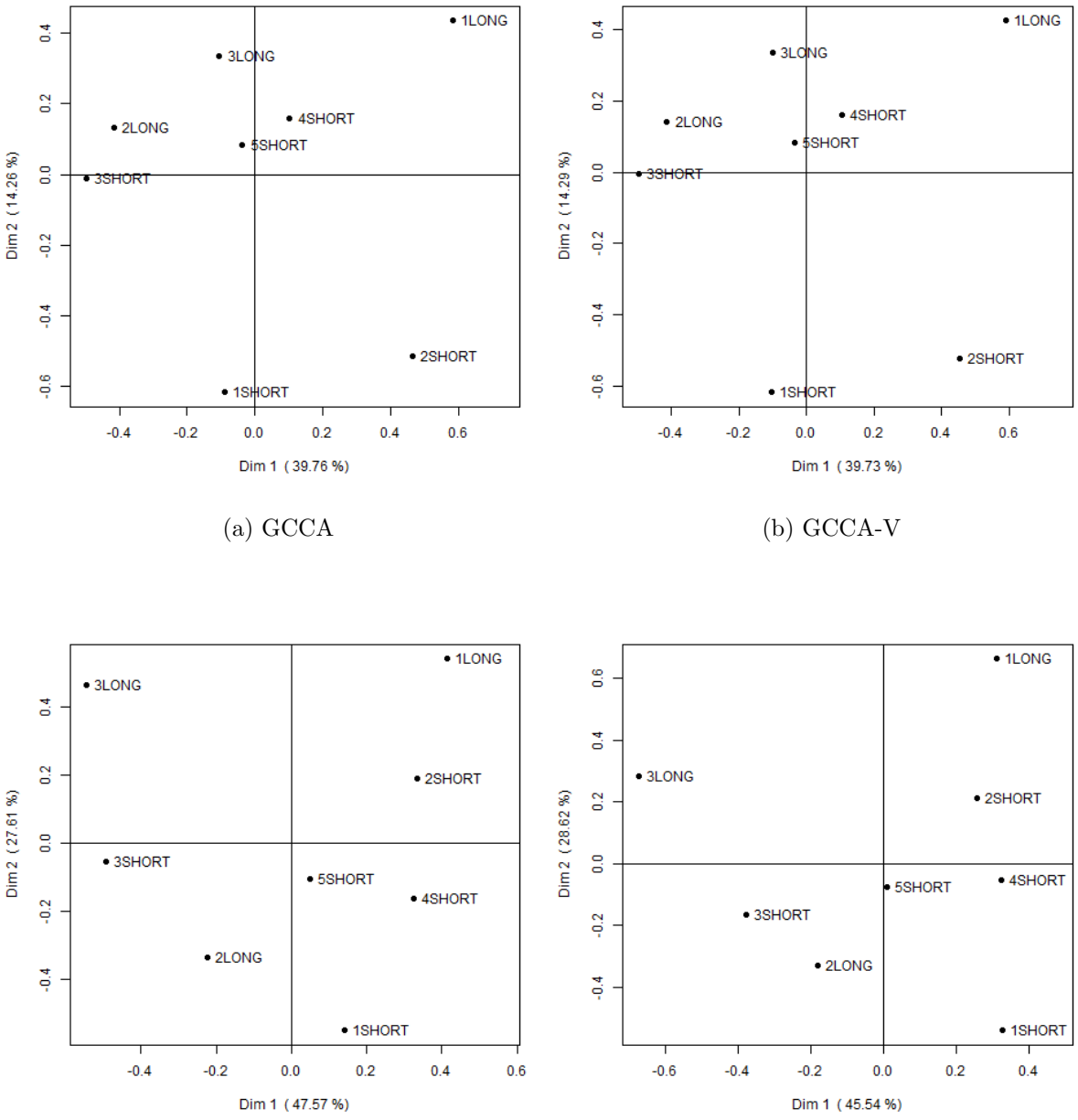
more pronounced with ComDim, since the first global component recovers up to 74.66 % of variation in $\boldsymbol{X}_3$ and only 22.58 % of variation in $\boldsymbol{X}_1$. This can be explained by the fact that the block of variables $\boldsymbol{X}_3$ is formed of variables relatively more correlated with each others than in other blocks of variables. Moreover, the variables of this block are related to the variables "Rancid" and "Earthy" in block $\boldsymbol{X}_2$. By contrast, $\boldsymbol{X}_1$ seems to have been downweighted by ComDim because, on the one hand, the variables in this group are not very correlated to each others and, except the variable "Savory", these variables are not highly related to the variables in the other blocks (data not shown).

Table 8: Correlation between the sensory attributes and the first two global components of GGCA, GCCA-V, MB-PCA and ComDim.

| | GCCA | | GCCA-V | | MB-PCA | | ComDim | |
|---|---|---|---|---|---|---|---|---|
| Variable | Dim. 1 | Dim. 2 | Dim. 1 | Dim. 2 | Dim. 1 | Dim. 2 | Dim. 1 | Dim. 2 |
| Salty | -0.05 | 0.46 | -0.04 | 0.46 | -0.38 | 0.59 | -0.46 | 0.51 |
| Porkcomplex | **-0.75** | -0.34 | **-0.76** | -0.33 | **-0.73** | -0.40 | -0.54 | **-0.60** |
| Savory | **0.94** | 0.23 | **0.95** | 0.21 | **0.63** | **0.73** | 0.43 | **0.85** |
| Rancid | -0.46 | 0.08 | -0.46 | 0.09 | **-0.76** | 0.23 | **-0.73** | -0.01 |
| Molasses | **0.86** | -0.22 | **0.86** | -0.23 | **0.72** | 0.43 | **0.62** | 0.57 |
| Caramelized | **0.81** | 0.40 | **0.82** | 0.39 | 0.56 | **0.71** | 0.37 | **0.84** |
| Earthy | -0.12 | 0.58 | -0.11 | 0.58 | **-0.60** | **0.64** | **-0.72** | 0.46 |
| Dryness | -0.08 | 0.32 | -0.08 | 0.33 | **-0.63** | **0.66** | **-0.80** | 0.45 |
| Juiciness | 0.41 | -0.47 | 0.40 | -0.48 | **0.76** | -0.37 | **0.89** | -0.17 |
| Mushiness | **0.85** | -0.40 | **0.85** | -0.41 | **0.95** | 0.07 | **0.90** | 0.27 |

Figure 2 shows the configurations of the eight hams on the basis of the first two global components associated with the various methods of analysis. The interpretation of these graphical displays can be done using the correlations of the sensory variables with the global components (Table 8). It is clear that, on the one hand, the configurations derived from GCCA and GCCA-V agree with each other and, on the other hand, the configurations from MB-PCA and ComDim bear a high similarity to each other. This confirms the finding that we are in presence of two families of methods. GCCA and

GCCA-V, on the one hand, aim at recovering the information that is common to the blocks of variables and, on the other hand, MB-PCA and ComDim aim at recovering the within and between variation in the blocks.




(a) GCCA

(b) GCCA-V

(c) MB-PCA

(d) ComDim

Figure 2: Configurations of the hams on the first two global components derived from (a) GCCA, (b) GCCA-V, (c) MB-PCA and (d) ComDim.

### 3.3 Potatoes data

For the illustration of the predictive strategy called LR-MBPCA, we consider a case study where the aim is to predict sensory attributes from measurement data. This problem is of high interest in practice since the collection of sensory data is costly and time consuming. Twenty potatoes samples were analyzed after one month of storage and six additional samples were analyzed after eight months of storage. A panel of assessors profiled the texture of these potatoes with respect to nine texture attributes. The sensory data were averaged accross assessors, yielding a dataset $\boldsymbol{Y}$. The block $\boldsymbol{X}_1$ is given by the chemical analysis of the potatoes samples. A second block of variables $\boldsymbol{X}_2$ concerns the uniaxial compression. The third ($\boldsymbol{X}_3$) and fourth ($\boldsymbol{X}_4$) datasets respectively concern the Time Domain-NMR relaxation curves and near infrared (NIR) measurements. More details can be found in [35]. The aim is to investigate the relationships between the sensory attributes $\boldsymbol{Y}$ and the rest of the datasets ($\boldsymbol{X}_1$ to $\boldsymbol{X}_4$). Each dataset was column-centered and pre-scaled so as to have its norm equal to 1. Moreover, the dataset $\boldsymbol{Y}$ was multiplied by 2 in order to have a total variance equal to that of the blocks of variables $\boldsymbol{X}_1$ to $\boldsymbol{X}_4$ all together. By way of assessing the prediction ability of LR-MBPCA, we also performed MB-PLS [4, 36, 37] on the same data. Figure 3 shows the cumulative percentage of variation in $\boldsymbol{Y}$ explained by the first three components from LR-MBPCA and MB-PLS. It turns out that the two methods of analysis lead to results that agree with each others to a large extent. We also performed a leave one out cross validation study. Figure 4 shows the Root Mean Squared Errors of Cross Validation (RMSECV) associated with MB-PLS and LR-MBPCA. Both curves show a typical pattern insofar as RMSECV is concerned, since they decrease implying that the model improves as the number of components increases, then the curves start to increase flagging a problem of overfitting. The two methods seem to have more or the less the same performance. The minimum RMSECV is reached for A=10 components (RMSECV=1.39) for LR-MBPCA and A=11 components (RMSECV=1.33) for MB-PLS.

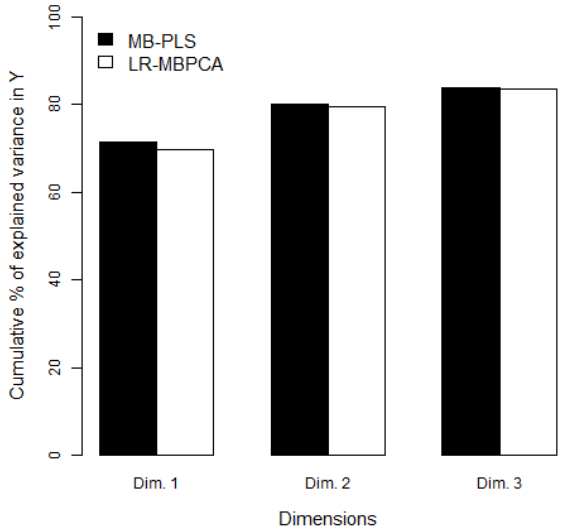



Figure 3: Cumulative percentage of total variance in Y explained by the first three global components of MB-PLS and LR-MBPCA.

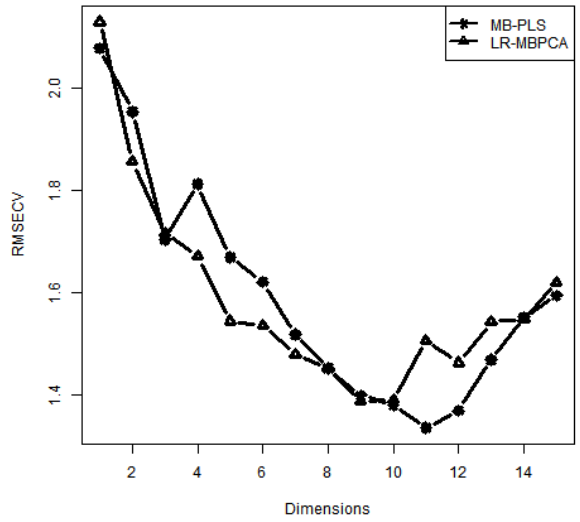



Figure 4: Root mean squared errors obtained by leave one out cross validation for the first fifteen global components of MB-PLS and LR-MBPCA.

# 4 Discussion and concluding remarks

Clearly, the paper sheds light on several unsupervised methods of analysis and highlights their common and differing features. It appears that the first differentiation key between the methods is whether we choose as a link function between the block components

$\boldsymbol{t}_k$ and their association global components, $\boldsymbol{t}$ , the expression $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$ or $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$. The effect of the former expression is to shed off the variation within the blocks and, as a consequence, what will emerge from the analysis is what is common to the blocks of variables no matter whether this is important in terms of variation explained or not. With the link function $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top \boldsymbol{t}$, the methods of analysis will seek to recover the variation in the blocks of variables. We have hinted to an intermediary solution by considering the link function $\boldsymbol{t}_k = \boldsymbol{X}_k(\gamma I + (1-\gamma)\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{t}$. As $\gamma$ increases from 0 to 1, the variances and the correlations among the variables within each block are gradually taken into account and, as a consequence, the method of analysis is likely to realise a better compromise between recovering the common structure to the various blocks of variables and the total variation in these blocks. Another advantage of this link function is that it acts as regularization procedure to counteract the problem of colinearity of the variables within each block [15, 16].

The second key of differentiation between the methods of multiblock data analysis is related to the choice of the summing up expression to compute the global component from its associated block components. Overall, it appears from the case studies that the two choices lead to results that agree with each others to a large extent. However, we believe that the expression based on the first principal component of the block components is likely to offer more possibilities of developments. For instance, instead of postulating that the global component is the first principal component of the block components, we may add the constraint that it should be a sparse principal component. This entails that, for each global component, only a selected number of block components will be included.

We have proposed iterative algorithms to run the multiblock data analyses and for GCCA and MB-PCA, there are also straightforward solutions based on the eigenanalysis of specific matrices. Using the simulated data as well as the sensory data which are used to illustrate the methods of analysis, we have performed a very large number of runs of the iterative algorithms by considering for each run a random starting point. The findings are: (i) not surprisingly, in all the situations the algorithms converged; (ii) the same optimum of the optimization criteria was obtained for all the runs; (iii) this optimum was equal

to that obtained by the eigenanalysis solution when the case applied. Thus, it appears that, for the data considered herein, the algorithms were not sensitive to the starting point. In order to draw a general conclusion regarding this aspect, a large simulation study is needed. For the time being, we recommend performing a multi-start algorithm by considering 30 (say) starting points.

The common feature of the link functions that we have cited, namely $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{X}_k^\top\boldsymbol{t}$, $\boldsymbol{t}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^\top\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top\boldsymbol{t}$ and $\boldsymbol{t}_k = \boldsymbol{X}_k(\gamma\boldsymbol{I} + (1-\gamma)\boldsymbol{X}_k^\top\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top\boldsymbol{t}$ is that they are based on the dot-product kernels $\boldsymbol{X}_k\boldsymbol{X}_k^\top$, $\boldsymbol{X}_k(\boldsymbol{X}_k^\top\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top$ and $\boldsymbol{X}_k(\gamma\boldsymbol{I}+(1-\gamma)\boldsymbol{X}_k^\top\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top$, respectively. This hints to the idea that potentially other dot-product kernels could be used with various purposes such as investigating non-linear relationships among the datasets at hand. Future research will be devoted to these developments.

We have also proposed a supervised strategy of analysis based on the block components associated with the predictive blocks of variables. Besides being very simple, its performance in terms of prediction seems to be very similar to that of MB-PLS. Obviously, this strategy of analysis can easily be adapted to GCCA and ComDim and may offer new extensions by using ideas pertaining to sparse PCA, for instance.

Ongoing research concerns the setting up of a unified approach for supervised methods drawing from ideas developped in this paper.

# Acknowledgements

The first author is very grateful to France overseas cultural and cooperation network of the Embassy of France in Togo and the African Excellence Center in Mathematical Sciences and Applications of the University of Abomey-Calavi (Bénin) for their financial support.

# References

[1] T. Skov, A.H. Honoré, H.M. Jensen, T. Næs, S.B. Engelsen, Chemometrics in foodomics: handling data structures from multiple analytical platforms, TrAC Trends

in Analytical Chemistry. 60 (2014) 71-79.

[2] H.R. Moskowitz, J.H. Beckley, A.V.A. Resurreccion, Sensory and Consumer Research in Food Product Design and Development, 2nd ed, John Wiley & Sons, 2012.

[3] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, Proc. Symp. on PLS Model Building: Theory and Application, Frankfurt am Main, 1987.

[4] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometrics. 12 (5) (1998) 301-321.

[5] E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H. MacFie, Defining the underlying sensory dimensions, Food Qual. Pref. 11 (1-2) (2000) 151-154.

[6] M. Hanafi, A. Kohler, E.M. Qannari, Shedding new light on hierarchical principal component analysis, J Chemometrics. 24 (11-12) (2010) 703-709.

[7] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemometrics. 17 (6) (2003) 323-337.

[8] S. Wold, L. Eriksson, J. Trygg, N. Kettaneh, The PLS method-partial least squares projections to latent structures-and its applications in industrial RDP (research, development, and production), Umea University, 2004.

[9] B.S. Dayal, J.F. MacGregor, Improved PLS algorithms, J. Chemometrics. 11 (1) (1997) 73-85.

[10] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, J. Chemometrics. 15 (2001) 485-493.

[11] I. González, S. Déjean, P.G. Martin, A. Baccini, CCA: An R package to extend canonical correlation analysis, J. Statistical Software. 23 (12) (2008) 1-14.

[12] N.R. Draper, S. Harry, Applied Regression Analysis, 3rd ed., Replika Press, 2005.

[13] H.D. Vinod, Canonical ridge and econometrics of joint production, J. econometrics. 4 (2) (1976) 147-166.

[14] A. Tenenhaus, M. Tenenhaus, Regularized Generalized Canonical Correlation Analysis, Psychometrika. 76 (2) (2011) 257-284. DOI: 10.1007/S11336-011-9206-8

[15] E.M. Qannari, M. Hanafi, A simple continuum regression approach, J. Chemometrics. 19 (2005) 387-392. DOI: 10.1002/cem.942

[16] S. Bougeard, M. Hanafi, E.M. Qannari, Continuum redundancy–PLS regression: A simple continuum approach, Comput. Stat. Data Anal. 52 (2008) 3686-3696.

[17] J.D. Carroll, A generalization of canonical correlation analysis to three or more sets of variables, $76^{th}$ annual convention of the American Psychological Association. (1968) 227-228.

[18] M. Tenenhaus, La régression PLS: Théorie et pratique, Edition TECHNIP, 1998.

[19] H. Hotelling, Relations between two sets variables, Biometrika. 28 (1936) 321-377.

[20] E.M. Qannari, P. Courcoux, E. Vigneau, Common components and specific weights analysis performed on preference data, Food Qual. Pref. 12 (5-7) (2001) 365-368.

[21] G. Mazerolles, M.F. Devaux, E. Dufour, E.M. Qannari, P. Courcoux, Chemometric methods for the coupling of spectroscopic techniques and for the extraction of the relevant information contained in the spectral data tables, Chemometr. Intell. Lab. Syst. 63 (1) (2002) 57-68.

[22] G. Mazerolles, M. Hanafi, E. Dufour, D. Bertrand, E.M. Qannari, Common components and specific weights analysis: a chemometric method for dealing with complexity of food products, Chemometr. Intell. Lab. Syst. 81 (1) (2006) 41-49.

[23] M. Hanafi, G. Mazerolles, E. Dufour, E.M. Qannari, Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques, J. Chemometrics. 20 (5) (2006) 172-183.

[24] J.P. Nielsen, D. Bertrand, E. Micklander, P. Courcoux, L. Munck, Study of NIR spectra, particle size distributions and chemical parameters of wheat flours: a multiway approach, J. Near Infrared Spec. 9 (4) (2001) 275-285.

[25] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometrics. 10 (5-6) (1996) 463-482.

[26] G. Saporta, Probabilités, Analyse des données et statistique. 3è édition révisée, Éditions Technip, 2011.

[27] J.T. Webster, R.F. Gunst, R.L. Mason, Latent root regression analysis, Technometrics. 16 (1974) 513-522.

[28] E. Vigneau, E.M. Qannari, A new algorithm for latent root regression analysis, Comput. Stat. Data Anal. 41 (2002) 231-242.

[29] D. Bertrand, E.M. Qannari, E. Vigneau, Latent root regression analysis: an alternative method to PLS, Chemometr. Intell. Lab. Syst. 58 (2001) 227-234.

[30] S. Bougeard, M. Hanafi, E.M. Qannari, Multiblock latent root regression, Application to epidemiological data, Computational Statistics. 22 (2007) 209-222. DOI 10.1007/s00180-007-0036-1

[31] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

[32] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Stat. Soc. 36 (1974) 111-147.

[33] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS method, In Ruhe A, Kastrom B (eds) Proceedings of conference in matrix pencils, Lecture notes in mathematics, Springer, Heidelberg. (1983) 286-293.

[34] A.J. Pham, M.W. Schilling, W.B. Mikel, J.B. Williams, J.M. Martin, P.C. Coggins, Relationships between sensory descriptors, consumer acceptability and volatile flavor compounds of American dry-cured ham, Meat Science. 80 (2008) 728-737.

[35] A.K. Thybo, I.E. Bechmann, M. Martens, S.B. Engelsen, Prediction of sensory texture of cooked potatoes using uniaxial compression near infrared spectroscopy and low field 1H NMR spectroscopy, LWT Food Science and Technology. 23 (2) (2000) 103-111.

[36] S. Wold, Three PLS algorithms according to SW, In Report from the symposium MULTDAST (multivariate data analysis in science and technology), Umea University, Sweden. (1984) 26-30.

[37] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm forinvestigating complex chemical systems, J. Chemometrics. 3 (1) (1989) 3-20.