THÈSE DE DOCTORAT DE

ONIRIS
ET
L'UNIVERSITE D'ABOMEY-CALAVI (UAC)

Ecole doctorale ONIRIS : ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation

Ecole doctorale UAC : Ecole doctorale des Sciences Agronomiques et de l'Eau

Spécialité Oniris : *Statistique / Modélisation en écologie, géosciences, agronomie et alimentation*
Spécialité UAC : *Biométrie*

UAC: Thèse N° 174

Par
TCHANDAO MANGAMANA Essomanda

Analyse des données multiblocs: approche unifiée et développement de nouvelles méthodes

Thèse présentée et soutenue à Nantes, le 30 septembre 2021

Unité de recherche Oniris : Statistique, Sensométrie et Chimiométrie
Unité de recherche UAC : Laboratoire de Biomathématiques et d'Estimations Forestières

Rapporteuses avant soutenance :

Ndèye NIANG, Maître de Conférences, Conservatoire National des Arts et Métiers, France
Atinuke ADEBANJI, Professeure Titulaire, Kwame Nkrumah University Of Science and Technology, Ghana

Composition du Jury :

Président : Mahouton Norbert HOUNKONNOU, Professeur Titulaire, Université d'Abomey-Calavi, Bénin
Examinatrice : Evelyne VIGNEAU, Professeure, Oniris, France
Rapporteuse 1 : Ndèye NIANG, Maître de Conférences, Conservatoire National des Arts et Métiers, France
Rapporteuse 2 : Atinuke ADEBANJI, Professeure Titulaire, Kwame Nkrumah University Of Science and Technology, Ghana
Dir. de thèse : El Mostafa QANNARI, Professeur, Oniris, France
Co-dir. de thèse : Romain GLELE KAKAÏ, Professeur Titulaire, Université d'Abomey-Calavi, Bénin

Remerciements

La tenue effective de cette thèse de doctorat n'aurait été possible sans l'implication de certaines institutions et personnes.

Merci à l'Ambassade de France au Togo et au Centre d'Excellence Africain en Sciences Mathématiques, Informatique et Applications de l'Université d'Abomey-Calavi (Bénin) pour leurs soutiens financiers au cours de cette thèse de doctorat.

Ma profonde gratitude va à l'endroit de mes directeurs de thèse. Notamment, à mon directeur de thèse de France, le Professeur El Mostafa Qannari. Merci pour son engagement dans la réalisation de cette thèse, pour sa clémence, pour sa promptitude dans toutes mes sollicitations. Que Dieu demeure toujours avec lui et lui procure tous ce dont il a besoin.

A mon directeur de thèse du Bénin, le Professeur Romain Glèlè Kakaï, merci pour sa disponibilité sans cesse renouvelée, pour son implication dans cette thèse, pour sa patience. Que Dieu soit avec lui et soit son guide quotidien.

Un sincère merci va également à l'endroit des membres de mon comité de suivi individuel d'Oniris: Professeure Lise Bellanger et Professeur Achim Kohler et de mon comité de thèse du Bénin: Professeur Marcel Senou et Dr Jonas Doumate. Merci pour vos orientations pour la bonne marche de cette thèse.

Merci à Evelyne Vigneau et à Véronique Cariou pour leur collaboration.
Merci aux membres de l'Unité de Statistique, Sensométrie et Chimiométrie pour leur accueil au sein de l'Unité, plus particulièrement à Mohamed Hanafi, je lui dis un sincère merci pour son assistance et encouragement.
Merci également aux membres du Laboratoire de Biomathématiques et d'Estimations Forestières.
Merci au Président de l'Université de Kara et à son personnel.
Merci à Bruno Enagnon Lokonon et à toute sa famille pour leur hospitalité.
Merci à mes parents pour leur soutien.
Je bénis Dieu pour avoir été à mes côtés chaque jour qu'il fait et pour le dévouement de toutes ces personnes et institutions qu'il a mis à ma disposition pour m'accompagner dans la réalisation de ce projet de thèse de doctorat.
Merci à la Vierge Marie pour son intercession pour moi auprès de son fils, mon Seigneur Jésus-Christ.

Table des matières

Abréviations

ACP	Analyse en Composantes Principales
PCA	Principal Components Analysis
CPCA	Consensus Principal Components Analysis
H-PCA	Hierarchical Principal Components Analysis
MB-PCA	Multiblock Principal Components Analysis
ACC	Analyse Canonique des Corrélations
CCA	Canonical Correlation Analysis
ACCG	Analyse Canonique des Corrélations Généralisées
GCCA	Generalized Canonical Correlation Analysis
ACCG-V	Analyse Canonique des Corrélations Généralisées Pondérées
GCCA-V	Weighted Generalized Canonical Correlation Analysis
ACCGR	Analyse Canonique des Corrélations Généralisées Régularisées
RGCCA	Regularized Generalized Canonical Correlation Analysis
RA	Redundancy Analysis
MB-RA	Multiblock Redundancy Analysis
MB-WRA	Multiblock Weighted Redundancy Analysis
ComDim	Dimensions Communes / Common Dimensions
P-ComDim:	Predictive Common Dimensions
Path-ComDim:	Path Common Dimensions

ANOVA-ComDim:	ANalysis Of VAriance Common Dimensions
ComDim-PCA:	Common Dimensions Principal Components Analysis
ComDim-Quali:	Common Dimensions for Qualitative variables
Sparse ComDim:	Sparse Common Dimensions
LRR	Régression par Analyse des Valeurs Latentes Latent Root Regression
MB-LRR	Régression par Analyse des Valeurs Latentes Multiblocs Multiblock Latent Root Regression
LR-MBPCA	Latent Root Multiblock Principal Components Analysis
PLS	Régression au sens des Moindres Carrés Partiels Partial Least Squares regression
H-PLS	Hierarchical Partial Least Squares regression
MB-HPLS	Multiblock Hierarchical Partial Least Squares regression
MB-PLS	Régression au sens des Moindres Carrés Partiels Multiblocs Multiblock Partial Least Squares regression
MB-WCov	Multiblock Weighted Covariate analysis
PLS-PM:	Partial Least Squares regression Path Modeling
ACM	Analyse des Correspondances Multiples
MCA	Multiple Correspondence Analysis
cov()/var()/cor()	Covariance / Variance / Corrélation
LOO	Leave One Out
NIPALS	Non Iterative Partial Least Squares
RMSEP	Root Mean Squared Errors of Prediction
RMSECV	Root Mean Squared Errors of Cross Validation

Liste des tableaux

Liste des figures

1

Résumé en Anglais

1.1 Context

With the advent of technology, data from different sources are often collected in several domains of application, with the aim of acquiring a better and more accurate knowledge on some phenomena of interest. For example, in sensometrics, for the characterization and optimization of food products, sensory and physico-chemical measurements can be made and related to consumer preferences. Or, we may postulate more complex relationships between these measurements. For instance, physico-chemical data may be assumed to have an influence on the sensory and preference data. Moreover, we may assume the sensory data to have an influence on the preference data. In health science, clinical, metabolomic and transcriptomic measurements can be made and related to variables reflecting a disease. In ecology, the environmental and spatial situation of different sites can be used to explain the abundance of some species in those sites. These data are often structured into meaningful blocks of variables, called "multiblock data". We assume that all the blocks of variables are mea-

sured on the same individuals, but the variables within these blocks can be different. Since the collection of multiblock data has increased these last decades, the number of statistical methods devoted to these data has also increased. These methods can be divided into two families: "unsupervised" and "supervised" methods. The former methods aim at exploring the structure of different blocks of variables and investigating the relationships between them. Examples of such methods are MB-PCA, H-PCA, ComDim, GCCA, etc. As regards the supervised methods, they aim at predicting a block of variables from other blocks of variables (e.g., MB-PLS, H-PLS, P-ComDim, etc.) or more generally, analyzing several blocks of variables taking account of a network of relationships that exist among them (e.g., PLS-PM, Path-ComDim, RGCCA, etc.).

1.2 Problems and objectives

With these families of methods, practitioners may be confused in choosing the suitable family of methods, namely, unsupervised or supervised. Moreover, with this plethora of methods within each family, they may be misled in choosing the suitable method to analyze their data. Therefore, one of the objectives of this thesis, is to set up a unified approach for some methods of both unsupervised and supervised families. The purpose of such unification is to better compare the methods of analysis by clearly pinpointing the similarities and differences between them.

Furthermore, it is noticed that some existing methods are too restrictive. For example, all the blocks of variables can be constrained to have the

same importance in the analysis. In order to alleviate these restrictions, we propose to relax them by proposing new unsupervised and supervised strategies of multiblock data analysis. Obviously, the purpose of such developments is to improve the results interpretation.

1.3 Unsupervised multiblock methods: a unified approach and extensions.

Three unsupervised multiblock methods have been compared. These methods are GCCA, MB-PCA and ComDim/H-PCA. Moreover, we introduce the weighted version for GCCA, called "GCCA-V". The common feature of all these methods is that they are based on the determination of global and block components associated with the various blocks of variables. Moreover, they are based on clear optimization criteria. However, the differing features among them are the way the global component is reflected on each block of variables to give the block component and the way the global component is formed from its block components. Regarding the expressions of the block components, we clearly note that we face two groups of methods. The first group embrasses GCCA and its weighted version GCCA-V. Both methods aim at assessing the common traits existing between the considered blocks of variables. By contrast, the methods of the second group (ComDim/H-PCA and MB-PCA) aim not only to assess the common traits between the blocks of variables at investigation, but also to recover the variance within each block of variables. For the expression of the global components, we have also two groups of methods. The first

group concerns GCCA and MB-PCA while the second group is formed by GCCA-V and ComDim/H-PCA. As regards the former methods, blocks of variables are given the same importance in the determination of the global components. Contrariwise, for the latter methods, the blocks of variables are weighted according to how they agree with each others. We have also proposed a new strategy of analysis pertaining to the family of supervised methods by adapting an unsupervised method to be used for a prediction purpose. Illustrations on the basis of simulated and real case studies are discussed. More details regarding these aspects can be found in chapter 3.

1.4 A general strategy for setting up supervised methods of multiblock data analysis.

Within the framework of multiblock data analysis, a unified approach of supervised methods is discussed. It encompasses multiblock redundancy analysis (MB-RA) and multiblock partial least squares (MB-PLS) regression. Moreover, we develop new supervised strategies of multiblock data analysis, which can be seen as variants of one or the other of these two methods. They are respectively referred to as multiblock weighted redundancy analysis (MB-WRA) and multiblock weighted covariate analysis (MB-WCov). The four methods are based on the determination of latent variables associated with the various blocks of variables. They are derived from clear optimization criteria whose aim is to maximize either the sum of the covariances or the sum of squared covariances between the latent variable associated with the response block of variables and the block la-

tent variables associated with the various explanatory blocks of variables. We also propose indices to help better interpreting the outcomes of the analyses. The methods are illustrated and compared based on simulated and real datasets. More details are presented in chapter 4.

1.5 New developments around ComDim.

ComDim is an unsupervised method whose aim is to analyze simultaneously several blocks of quantitative variables measured on the same individuals. Variants of this method are proposed. The first variant concerns the particular case where each block is reduced to a single variable. This variant, referred to as ComDim-PCA, is illustrated and compared to Principal Components Analysis. The second variant is related to the analysis of qualitative variables. The proposed variant is called ComDim-Quali and is illustrated and compared to Multiple Correspondence Analysis.
One of the particularities of ComDim is to provide specific weights that denote the importance of each block of variables in the determination of the global components. However, it very often appears that not all blocks of variables are useful to the determination of these global components. Therefore, we propose to set the weights of the useless blocks of variables to zero in the analysis. This leads to a new strategy of analysis that we refer to as Sparse ComDim. In the same vein, we propose the sparse versions to all the variants described above. These aspects are more investigated in chapter 5.

1.6 New developments around MB-WCov.

MB-WCov is a supervised method introduced to predict one dataset by $\boldsymbol{K}$ datasets. This method is a straightforward extension of ComDim to the analysis of "$\boldsymbol{K}+\mathbf{1}$" datasets. MB-WCov is presented using a new formulation which shows an interesting property in relation with the prediction. Moreover, this method is applied to particular cases where each dataset is reduced to a single variable and to the case where the predicted dataset is univariate / multivariate, while the predictive datasets are reduced to a single variable. The sparse version of MB-WCov is also introduced, together with the sparse versions of its particular cases. All these methods are illustrated using real datasets. More details on these aspects are presented in chapter 6.

1.7 Conclusions and perspectives.

In the frame of this thesis, several existing methods have been discussed and new other methods have been developed to help practitioners analyzing their data. These developments include both unsupervised and supervised methods of multiblock data analysis.

Regarding the unsupervised methods, we have compared them by clearly indicating their similarities and differences. In the same vein as ComDim and for the purpose of relaxing the constraint imposed on GCCA regarding the importance of each block of variables in the computation of the global components, we have proposed GCCA-V to allow each block of variables

to have a different weight in the analysis. Furthermore, we have shown how an unsupervised method can be adapted to yield a supervised method to be used for a prediction purpose.

We have also compared the supervised methods, MB-RA and MB-PLS regression and proposed their weighted versions. More developments are thereafter made on ComDim and MB-WCov to allow them to be applied to some particular cases.

A package R, called "MBAnalysis" has been developed to help practitioners implement some methods discussed in this work.

This research work sketches interesting extensions as persepectives. For example, it will be interesting to explore our general strategy of unifying methods in the framework of path-modeling (i.e., the analysis of several blocks of variables linked with arrows, reflecting a chain of influence for instance). We have discussed two ways of projecting latent variables on blocks of variables. Moreover, we have hinted to an intermediary solution, which bridges the two previous ways of projecting latent variables. This suggests that other dot-product kernels could be used with various purposes such as investigating non-linear relationships among the blocks of variables.

2

Introduction générale

Avec le développement de la technologie, on assiste de nos jours à une prolifération des données dans plusieurs domaines d'application et de la recherche. Ces données, provenant généralement de différentes sources, sont collectées dans le but de mieux étudier, comprendre et décrire des phénomènes d'intérêt. Par exemple, pour la caractérisation et l'optimisation des produits alimentaires, on pourrait utiliser plusieurs types de mesures (données sensorielles, physico-chimiques, etc.) et les relier à des données de qualité (par exemple, les préférences des consommateurs). On pourrait aussi postuler l'existence d'autres liens de causalité plus complexes entre ces données. Par exemple, on pourrait supposer que les données physico-chimiques ont une influence sur les données sensorielles et de préférence et que les données sensorielles ont une influence sur les données de préférence. Dans le domaine de la santé, des mesures cliniques, métabolomiques, transcritomiques, etc. pourraient être relevées et mises en relation avec des variables reflétant, par exemple, l'expression d'une maladie. En écologie, nous pouvons nous intéresser à l'exploration des relations entre l'abondance de certaines espèces dans différents sites d'une part et, d'autre part, les vari-

ables décrivant ces sites (environnement, biodiversité, situation spatiale, etc.). Ces données sont généralement structurées en tableaux de données. Nous nous plaçons particulièrement dans le cas où tous les tableaux de données portent sur les mêmes individus mais les variables peuvent être différentes d'un tableau à un autre. On parle alors des données multi-blocs (appariées par individus). Du fait de l'intérêt particulier que les chercheurs portent à ces données au cours des deux ou trois dernières décennies, on recense une multitude de méthodes statistiques proposées pour leur analyse. Ces méthodes sont regroupées en deux familles: la famille des méthodes "non supervisées", qui sont à vocation exploratoire et celle des méthodes "supervisées", qui sont à vocation prédictive. Comme exemple de méthodes non supervisées, nous avons, entre autres, l'Analyse Canonique des Corrélations Généralisées "ACCG" [1,2], l'Analyse en Composantes Principales multiblocs "ACP multiblocs" [3], aussi connue sous le nom de l'Analyse en Composantes Principales consensuelle [3], l'Analyse en Composantes Principales hiérarchique [3] et la méthode ComDim [4,5]. Quant à la famille des méthodes supervisées, elle est subdivisée en deux sous-familles. La première sous-famille regroupe les méthodes qui ont pour but de prédire un tableau de données "réponse" (ou "à prédire") à partir de plusieurs tableaux de données "explicatifs" (ou "prédictifs"). Parmi ces méthodes, nous pouvons citer la régression des moindres carrés partiels multiblocs "MB-PLS" [3,6], la régression des moindres carrés partiels hiérarchique "MB-HPLS" [3], l'analyse des redondances multiblocs [7–9] et la méthode P-ComDim [10]. La deuxième sous-famille des méthodes supervisées concerne celles qui ont pour but d'analyser simultanément plusieurs

tableaux de données en tenant compte d'un graphe de causalité qui existe entre eux. Parmi ces méthodes, nous citons en particulier la méthode Partial Least Squares Path Modeling "PLS-PM" [11], Regularized Generalized Canonical Correlation Analysis "RGCCA" [12] et Path-ComDim [13].

L'objectif général de ce travail est de contribuer à une meilleure compréhension de certaines méthodes existantes pour mieux analyser les données multiblocs. Plus particulièrement, nous allons:

- Unifier les méthodes pour en faciliter la comparaison.
- Proposer de nouvelles méthodes non supervisées d'analyse des données multiblocs.
- Proposer de nouvelles méthodes supervisées pour l'analyse de deux à plusieurs tableaux de données.
- Développer un package R pour la mise en œuvre de certaines méthodes non supervisées et supervisées d'analyse des données multiblocs.

La suite du travail sera organisée comme suit. Dans le chapitre 3, nous aborderons l'approche unifiée des méthodes non supervisées, nous proposerons également une nouvelle méthode non supervisée puis nous montrerons comment étendre une méthode non supervisée en vue d'obtenir une méthode supervisée. Dans le chapitre 4, nous comparerons dans un cadre unifié, certaines méthodes supervisées. Le chapitre 5 sera consacré à la proposition de variantes pour la méthode non supervisée ComDim. Dans le même ordre d'idées, des variantes de la méthode MB-WCov seront proposées dans le chapitre 6. Une conclusion générale ainsi que des perspectives seront données dans le dernier chapitre.

3

Méthodes non supervisées d'analyse des données multiblocs: une approche unifiée et extensions

3.1 Introduction

Comme cela est indiqué dans l'introduction générale, l'importance de la collecte des données multiblocs appariées par individus n'est plus à démontrer. Elle est plus qu'une nécessité pour une meilleure compréhension et investigation de plusieurs problèmes dans tous les domaines d'application ou de recherche. Pour l'analyse de ces données, l'on se sert souvent des méthodes multiblocs exploratoires ou prédictives en fonction de l'objectif de l'étude.

Dans ce chapitre, nous allons dans un premier temps, nous intéresser aux méthodes non supervisées. Plus particulièrement, nous proposons une démarche unifiée qui regroupe plusieurs méthodes non supervisées à savoir: l'ACCG, l'ACP multiblocs et la méthode ComDim/H-PCA. Ce développement nous permettra d'introduire une version pondérée de l'ACCG. Dans un deuxième temps, nous allons nous intéresser aux méthodes su-

pervisées en montrant comment l'on pourrait adapter une méthode non supervisée pour en faire une méthode supervisée.

Les différentes méthodes sont comparées sur la base d'un jeu de données portant sur la caractérisation de variétés de pommes de terre.

3.2 Méthodes

3.2.1 Relations entre la composante globale et ses composantes partielles

Considérons $\boldsymbol{K}$ tableaux (ou blocs de variables) $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, ..., $\boldsymbol{X_K}$ portant sur les mêmes $\boldsymbol{n}$ individus. Nous supposons que ces tableaux sont centrés, réduits (si nécessaire) et pré-traités de sorte que la norme de chaque tableau soit égale à 1. Ce pré-traitement a pour but de permettre à tous les tableaux d'avoir a priori la même importance dans l'analyse.

L'approche que nous adoptons est basée sur la détermination des composantes, appelées aussi variables latentes [14]. Ces variables sont déterminées de manière séquentielle et, de ce fait, revêtent une importance décroissante d'une séquence à une autre. A chaque dimension, nous distinguons deux types de composantes (figure 3.1):

- des composantes partielles ou de bloc: à chaque tableau $\boldsymbol{X_k}$, nous associons une composante $\boldsymbol{t_k}$ qui est une combinaison linéaire des variables de ce tableau;

- une composante globale $\boldsymbol{t}$ qui est censée faire une synthèse des composantes partielles $\boldsymbol{t_k}$.

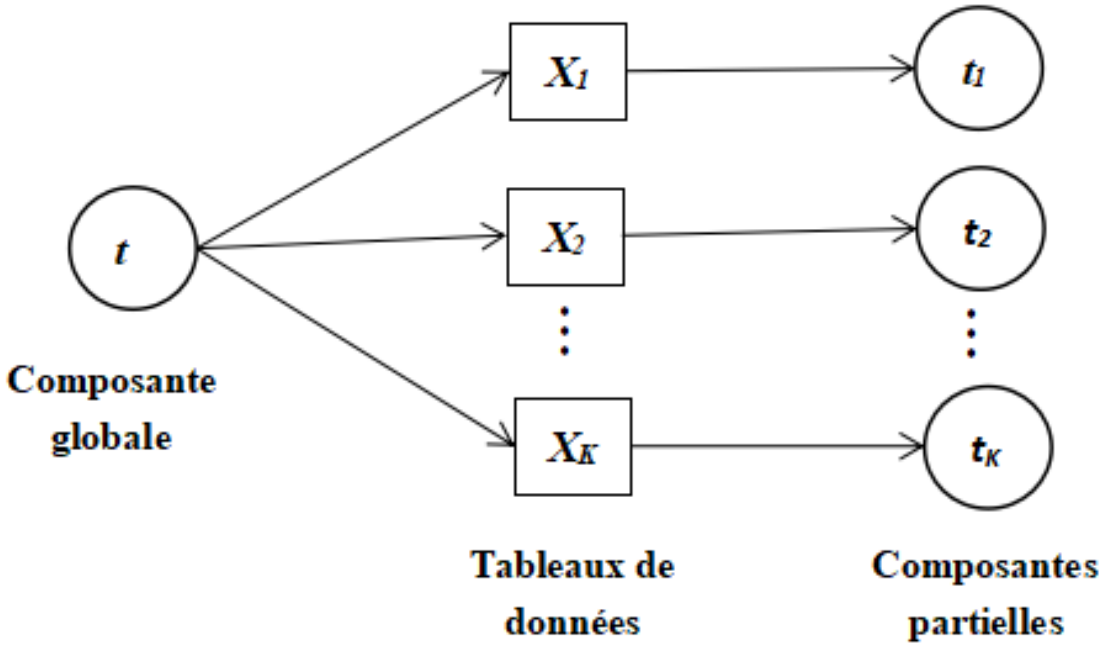


Fig. 3.1: Types de composantes (composantes partielles et composante globale).

L'idée majeure qui sous-tend l'approche unificatrice que nous proposons ici est qu'une méthode d'analyse dépend des relations mutuelles que nous pourrions supposer entre la composante globale, d'un côté, et les composantes partielles de l'autre côté. Naturellement, ces relations devraient formaliser l'idée que, d'un côté, la composante $\boldsymbol{t_k}$, associée au tableau $\boldsymbol{X_k}$, devrait être le reflet de la composante globale $\boldsymbol{t}$ dans le tableau $\boldsymbol{X_k}$ et, d'un autre côté, la composante globale $\boldsymbol{t}$ devrait opérer une synthèse des composantes partielles $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$.

L'idée que $\boldsymbol{t_k}$ devrait être le reflet de $\boldsymbol{t}$ dans $\boldsymbol{X_k}$ peut se traduire de deux manières. Premièrement, $\boldsymbol{t_k = P_k t}$, où $\boldsymbol{P_k = X_k(X_k^\top X_k)^{-1} X_k^\top}$ est le projecteur orthogonal sur l'espace vectoriel engendré par les variables de $\boldsymbol{X_k}$. Comme nous pouvons le constater, cette expression nécessite l'inversion de la matrice $\boldsymbol{X_k^\top X_k}$. L'avantage de cette inversion est que les variances des variables et les corrélations entre les variables à l'intérieur

de chaque tableau $\boldsymbol{X_k}$ sont masquées. Par conséquent, ce qui émergera de l'analyse portera uniquement sur ce qui est commun aux différents tableaux. Cependant, l'inconvénient majeur de l'inversion d'une telle matrice est qu'en présence de variables fortement corrélées, nous risquons d'avoir des modèles instables [15–18]. Ceci nous amène à considérer une deuxième relation pour traduire l'idée que la composante $\boldsymbol{t_k}$ est le reflet de la composante globale $\boldsymbol{t}$ sur le tableau $\boldsymbol{X_k}$: $\boldsymbol{t_k} = \boldsymbol{W_k t}$, où $\boldsymbol{W_k} = \boldsymbol{X_k X_k^\top}$ est la matrice de produit scalaire entre les individus du tableau $\boldsymbol{X_k}$. Avec cette relation, la méthode cherchera non seulement à expliquer la variation inter-tableaux mais aussi la variation intra-tableaux.

L'idée de concevoir la composante globale, $\boldsymbol{t}$, comme une synthèse des composantes partielles $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ pourrait être formalisée de plusieurs manières; conduisant chaque fois à une méthode d'analyse de tableaux multiblocs. Nous pouvons postuler que $\boldsymbol{t}$ est proportionnelle à la moyenne des composantes partielles $\boldsymbol{t_k}$, ou de manière équivalente, à leur somme: $\boldsymbol{t \propto \sum_{k=1}^{K} t_k}$, où le symbole $\propto$ signifie "proportionnelle à". Alternativement, nous pouvons postuler que $\boldsymbol{t}$ est proportionnelle à une combinaison linéaire des composantes partielles $\boldsymbol{t_k}$ (par exemple, la première composante principale): $\boldsymbol{t \propto \sum_{k=1}^{K} \alpha_k t_k}$. Dans les deux situations, le coefficient de proportionnalité sera déterminé grâce à la contrainte de détermination qui sera imposée à $\boldsymbol{t}$ (par exemple, $\boldsymbol{||t|| = 1}$).

En croisant les deux options à savoir, d'un côté, comment les composantes $\boldsymbol{t_k}$ sont déduites à partir de $\boldsymbol{t}$ (c'est-à-dire, $\boldsymbol{t_k = P_k t}$ ou $\boldsymbol{t_k = W_k t}$) et, d'un autre côté, comment $\boldsymbol{t}$ est reconstituée à partir des $\boldsymbol{t_k}$ (c'est-à-dire, $\boldsymbol{t \propto \sum_{k=1}^{K} t_k}$ ou $\boldsymbol{t \propto \sum_{k=1}^{K} \alpha_k t_k}$), nous obtenons quatre

méthodes d'analyse qui sont consignées dans le tableau 3.1. Dans ce tableau, nous indiquons également des algorithmes de résolution qui, intuitivement, émergent de la relation mutuelle entre la composante globale et ses composantes partielles. Pour ces algorithmes, nous avons choisi comme contrainte de détermination $||\boldsymbol{t}|| = \mathbf{1}$.

Tab. 3.1: Classification des méthodes non supervisées d'analyse des données multiblocs.

$\boldsymbol{W_k} = \boldsymbol{X_k}\boldsymbol{X_k}^\top$ et $\boldsymbol{P_k} = \boldsymbol{X_k}(\boldsymbol{X_k}^\top\boldsymbol{X_k})^{-1}\boldsymbol{X_k}^\top$

Méthode	$\boldsymbol{t_k}$ reflet de $\boldsymbol{t}$	$\boldsymbol{t}$ synthèse des $\boldsymbol{t_k}$	Algorithme
A	$\boldsymbol{t_k} = \boldsymbol{W_k t}$	$\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{t_k}$	0. Choix aléatoire de $\boldsymbol{t}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 1. $\boldsymbol{t_k} = \boldsymbol{W_k t}$; 2. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{t_k}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 3. Réitérer à partir de 1.
B	$\boldsymbol{t_k} = \boldsymbol{P_k t}$	$\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{t_k}$	0. Choix aléatoire de $\boldsymbol{t}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 1. $\boldsymbol{t_k} = \boldsymbol{P_k t}$; 2. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{t_k}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 3. Réitérer à partir de 1.
C	$\boldsymbol{t_k} = \boldsymbol{W_k t}$	$\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{\alpha_k t_k}$ $\boldsymbol{\alpha_k} = \boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t})$	0. Choix aléatoire de $\boldsymbol{t}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 1. $\boldsymbol{t_k} = \boldsymbol{W_k t}$; 2. $\boldsymbol{\alpha_k} = \boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t})$; 3. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\alpha_k t_k}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 4. Réitérer à partir de 1.
D	$\boldsymbol{t_k} = \boldsymbol{P_k t}$	$\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{\alpha_k t_k}$ $\boldsymbol{\alpha_k} = \boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t})$	0. Choix aléatoire de $\boldsymbol{t}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 1. $\boldsymbol{t_k} = \boldsymbol{P_k t}$; 2. $\boldsymbol{\alpha_k} = \boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t})$; 3. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\alpha_k t_k}$ puis $\boldsymbol{t} = \boldsymbol{t}/\|\|\boldsymbol{t}\|\|$; 4. Réitérer à partir de 1.

Les algorithmes proposés seront étayés et leur convergence démontrée

dans la section suivante. Nous montrerons aussi que les méthodes A, B et C se rattachent respectivement à l'ACP multiblocs, l'ACCG et à la méthode ComDim. Quant à la méthode D, elle définit une nouvelle stratégie d'analyse qui apparaît comme une variante de l'ACCG.

3.2.2 Critères d'optimisation et algorithmes

Un critère d'optimisation est un trait caractéristique d'une méthode d'analyse. Il revêt une importance capitale dans la compréhension de cette méthode. Il permet de mieux clarifier l'objectif de la méthode, de déterminer l'algorithme de résolution, de prouver la convergence de l'algorithme dans le cas d'une procédure itérative.

Ci-dessus, nous avons postulé l'existence de liens étroits entre la composante globale et ses composantes partielles. L'objectif est donc de maximiser ces liens.

ACP multiblocs

Dans un premier temps, nous postulons que $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$. Nous proposons de déterminer $\boldsymbol{t}$ (et, par conséquent, $\boldsymbol{t_k}$) de manière à maximiser le critère suivant:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{cov(t_k, t)} \tag{3.1}$$

Nous imposons la contrainte de détermination $\boldsymbol{||t|| = 1}$.

Nous avons:

$$\sum_{k=1}^{K} cov(t_k, t) = \frac{1}{n}\sum_{k=1}^{K} t_k^\top t = \frac{1}{n}\sum_{k=1}^{K} t^\top X_k X_k^\top t$$
$$= t^\top \left(\frac{1}{n}\sum_{k=1}^{K} X_k X_k^\top\right) t = t^\top \left(\frac{1}{n} X X^\top\right) t \qquad (3.2)$$

où $X = [X_1|X_2|...|X_K]$ est le tableau obtenu par concaténation horizontale des tableaux X_k $(k = 1, 2, ..., K)$.

Il s'ensuit que le vecteur t qui maximise le critère considéré est donné par le vecteur propre normé de $\frac{1}{n} X X^\top$ associé à la plus grande valeur propre. En d'autres termes, t est la première composante principale du tableau X. Cette propriété est caractéristique de l'ACP multiblocs (MB-PCA) [3].

Une fois la composante globale déterminée, les composantes partielles sont naturellement données par $t_k = X_k X_k^\top t$. La composante globale t étant un vecteur propre normé de $\frac{1}{n}\sum_{k=1}^{K} X_k X_k^\top = \frac{1}{n} X X^\top$, nous pouvons proposer l'algorithme de résolution de type NIPALS, en se basant sur l'idée des puissances itérées pour la détermination d'un vecteur propre.

0. Choix de manière aléatoire de la composante globale t et sa normalisation $(t = t/||t||)$;

1. Détermination des composantes partielles, $t_k = X_k X_k^\top t$ $(k = 1, 2, ..., K)$;

2. Mise à jour de la composante globale t: $t = \sum_{k=1}^{K} t_k$ et sa normalisation $(t = t/||t||)$;

3. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Naturellement, cet algorithme est convergent car à chaque itération le critère à maximiser croît. Comme, par ailleurs, ce critère est majoré, nous en déduisons que la suite de valeurs du critère déterminées au cours des itérations est convergente.

Pour la détermination des composantes globales et des composantes partielles suivantes, nous procédons à une déflation par rapport à la composante globale $\boldsymbol{t}$. De manière précise, ceci consiste à régresser chaque variable des tableaux $\boldsymbol{X_k}$ sur la composante globale $\boldsymbol{t}$ et de considérer les résidus de cette régression à la place de la variable elle-même.

Il est important de souligner qu'à partir du critère de maximisation (3.2), la quantité $\boldsymbol{t^\top X_k X_k^\top t = n \times cov(t_k, t)}$ représente la contribution du tableau $\boldsymbol{X_k}$ dans la détermination de la composante globale $\boldsymbol{t}$. Nous pouvons remarquer que, comme $\boldsymbol{t}$ est supposée de longueur 1, cette quantité reflète aussi l'inertie de $\boldsymbol{X_k}$ expliquée par $\boldsymbol{t}$. De plus, la norme de $\boldsymbol{X_k}$ étant de 1, la quantité $\boldsymbol{t^\top X_k X_k^\top t}$ représente aussi le pourcentage de variation de $\boldsymbol{X_k}$ expliquée par $\boldsymbol{t}$. L'importance globale de $\boldsymbol{t}$ (en pourcentage), peut être déterminée par $\frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t^\top X_k X_k^\top t}$, où $\boldsymbol{K}$, le nombre de tableaux, correspond également à la variation totale de tous les tableaux (c'est-à-dire, $\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{trace(X_k X_k^\top) = K}$).

Pour les représentations graphiques, il convient de redimensionner la composante globale $\boldsymbol{t}$ de sorte que sa variance reflète la variation totale de tous les tableaux expliquée par cette composante. Ce qui revient à considérer $\boldsymbol{\tilde{t} = \mu t}$, où $\boldsymbol{\mu = \sqrt{\sum_{k=1}^{K} cov(t_k, t)} = \sqrt{\sum_{k=1}^{K} \frac{1}{n} t^\top X_k X_k^\top t}}$. En d'autres termes, $\boldsymbol{\tilde{t}}$ correspond à la première composante principale non standardisée de $\boldsymbol{X = [X_1 | X_2 | ... | X_K]}$.

Méthode B: ACCG

Pour la méthode B, nous postulons que $\boldsymbol{t_k = P_k t}$. Nous cherchons une composante $\boldsymbol{t}$ de longueur 1 et, par conséquent, des composantes partielles $\boldsymbol{t_k = P_k t}$, de sorte à maximiser:

$$\begin{aligned}\sum_{k=1}^{K} \boldsymbol{cov}(\boldsymbol{t_k}, \boldsymbol{t}) &= \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{k=1}^{K} \boldsymbol{t_k^\top t} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{k=1}^{K} \boldsymbol{t^\top P_k t} \\ &= \boldsymbol{t^\top} \left(\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{k=1}^{K} \boldsymbol{P_k} \right) \boldsymbol{t} \end{aligned} \tag{3.3}$$

Le vecteur, $\boldsymbol{t}$, qui maximise ce critère est donné par le premier vecteur propre normé de $\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{P_k}$ associé à la plus grande valeur propre.

L'algorithme itératif associé à la résolution du critère (3.3) est le suivant:

0. Choix de manière aléatoire de la composante globale $\boldsymbol{t}$ et sa normalisation ($\boldsymbol{t = t/||t||}$);

1. Détermination des composantes partielles, $\boldsymbol{t_k = P_k t}$ $\boldsymbol{(k = 1, 2, ..., K)}$;

2. Mise à jour de la composante globale $\boldsymbol{t}$: $\boldsymbol{t = \sum_{k=1}^{K} t_k}$ et sa normalisation ($\boldsymbol{t = t/||t||}$);

3. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

La détermination des composantes successives est faite après une déflation de tous les tableaux de données par rapport à la composante globale $\boldsymbol{t}$.

Nous pouvons montrer que le critère (3.3) que nous avons présenté ci-dessus, est équivalent au critère de Carroll [2] qui consiste à maximiser

$\sum_{k=1}^{K} \boldsymbol{cor^2(t_k, t)}$, avec $\boldsymbol{t_k = P_k t}$ et $\boldsymbol{cor()}$ est le coefficient de corrélation. En effet, du fait que le coefficient de corrélation soit invariant par changement d'échelle, nous pouvons choisir arbitrairement $\boldsymbol{||t|| = 1}$. Nous avons:

$$\sum_{k=1}^{K} \boldsymbol{cor^2(P_k t, t)} = \sum_{k=1}^{K} \frac{(t^\top \boldsymbol{P_k} t)^2}{||t||^2 ||\boldsymbol{P_k} t||^2} = \sum_{k=1}^{K} \frac{(t^\top \boldsymbol{P_k} t)^2}{||\boldsymbol{P_k} t||^2} = \sum_{k=1}^{K} \frac{(t^\top \boldsymbol{P_k} t)^2}{t^\top \boldsymbol{P_k}^\top \boldsymbol{P_k} t} \tag{3.4}$$

Puisque $\boldsymbol{P_k}$ est symétrique ($\boldsymbol{P_k^\top = P_k}$) et idempotent ($\boldsymbol{P_k^2 = P_k}$), il s'ensuit que le critère à maximiser est équivalent à:

$$\sum_{k=1}^{K} t^\top \boldsymbol{P_k} t = n \sum_{k=1}^{K} \boldsymbol{cov(P_k t, t)} = n \sum_{k=1}^{K} \boldsymbol{cov(t_k, t)} \tag{3.5}$$

Nous pouvons aussi noter que:

$$\boldsymbol{cov(t_k, t)} = \frac{1}{n} t^\top \boldsymbol{P_k} t = \frac{1}{n} t^\top \boldsymbol{P_k}^\top \boldsymbol{P_k} t = \boldsymbol{var(P_k t)} \tag{3.6}$$

où $\boldsymbol{var()}$ désigne la variance.

De plus, puisque nous avons supposé que $\boldsymbol{||t|| = 1}$, nous avons:

$$t^\top \boldsymbol{P_k}^\top t = \frac{\boldsymbol{var(P_k t)}}{\boldsymbol{var(t)}} \tag{3.7}$$

qui est le coefficient de détermination de $\boldsymbol{t}$ par rapport à $\boldsymbol{X_k}$.

Méthode C: ComDim

Au lieu de déterminer $\boldsymbol{t}$ (et $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$) en maximisant le critère $\sum_{k=1}^{K} \boldsymbol{cov(t_k, t)}$, avec $\boldsymbol{||t|| = 1}$, nous proposons de maximiser le critère:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)} \qquad (\boldsymbol{||t|| = 1}) \tag{3.8}$$

L'expression de Lagrange associée à ce problème de maximisation est la suivante:

$$\boldsymbol{L(t)} = \sum_{k=1}^{K} \boldsymbol{(t^\top X_k X_k^\top t)^2 - 2\lambda(t^\top t - 1)} \tag{3.9}$$

où $\boldsymbol{-2\lambda}$ est le multiplicateur de Lagrange.

La dérivée de cette expression par rapport à $\boldsymbol{t}$ conduit à:

$$\boldsymbol{L'(t)} = 4 \sum_{k=1}^{K} \boldsymbol{(t^\top X_k X_k^\top t) X_k X_k^\top t - 4\lambda t} \tag{3.10}$$

En désignant par $\boldsymbol{\alpha_k = t^\top X_k X_k^\top t}$ et par $\boldsymbol{W_k = X_k X_k^\top}$, il s'ensuit que le point stationnaire (qui annule $\boldsymbol{L'(t)}$) vérifie l'équation:

$$\sum_{k=1}^{K} \boldsymbol{\alpha_k W_k t = \lambda t} \tag{3.11}$$

De là, il est clair que $\boldsymbol{t}$ est le vecteur propre normé de $\sum_{k=1}^{K} \boldsymbol{\alpha_k W_k}$ associé à la plus grande valeur propre. Nous déduisons l'algorithme suivant pour la détermination de $\boldsymbol{t}$:

0. Initialisation des poids: $\boldsymbol{\alpha_k = 1}$ $\boldsymbol{(k = 1, 2, ..., K)}$;

1. Calcul de la composante globale: $\boldsymbol{t}$ est le vecteur propre normé de

$\sum_{k=1}^{K} \boldsymbol{\alpha_k W_k}$ associé à la plus grande valeur propre;

2. Calcul des composantes partielles: $\boldsymbol{t_k = W_k t}$;

3. Mise à jour des poids: $\boldsymbol{\alpha_k = t^\top t_k}$;

4. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Alternativement, le problème de maximisation (3.8) peut se résoudre de la façon suivante: pour $\boldsymbol{t}$ fixée, nous avons $\boldsymbol{t_k = X_k X_k^\top t}$ et pour $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ fixées, la maximisation du critère $\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)}$ conduit à choisir $\boldsymbol{t}$ comme étant la première composante principale des variables $\boldsymbol{t_k}$. En effet, la maximisation du critère considéré est typique d'une propriété caractéristique de la première composante principale [19]. Il s'ensuit que $\boldsymbol{\alpha_k \propto cov(t_k, t)}$. Pour l'interprétation des résultats, nous pourrons standardiser ces coefficients de manière à avoir $\sum_{k=1}^{K} \boldsymbol{\alpha_k^2 = 1}$. Ainsi, le coefficient $\boldsymbol{\alpha_k}$ reflète la contribution du tableau $\boldsymbol{X_k}$ à la détermination de la composante globale.

Il est clair que nous retrouvons la méthode ComDim. L'algorithme de résolution de type NIPALS est le suivant:

0. Choix de manière aléatoire de la composante globale $\boldsymbol{t}$ et sa normalisation $(\boldsymbol{t = t/||t||})$;

1. Détermination des composantes partielles: $\boldsymbol{t_k = W_k t}$ $\boldsymbol{(k = 1, 2, ..., K)}$;

2. Calcul des poids $\boldsymbol{\alpha_k = t^\top t_k}$;

3. Mise à jour de la composante globale $\boldsymbol{t}$: $\boldsymbol{t = \sum_{k=1}^{K} \alpha_k t_k}$ et sa normalisation $(\boldsymbol{t = t/||t||})$;

4. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Pour la détermination des composantes d'ordre supérieur, nous procédons à une déflation par rapport à la composante globale $\boldsymbol{t}$, comme cela est indiqué pour la méthode ACP multiblocs.

Tout comme dans le cas de l'ACP multiblocs, $\boldsymbol{\alpha_k = t^\top X_k X_k^\top t}$ reflète la contribution du bloc de variables $\boldsymbol{X_k}$ dans la détermination de la composante globale $\boldsymbol{t}$. Cela représente aussi l'inertie de $\boldsymbol{X_k}$ expliquée par $\boldsymbol{t}$. L'importance globale de $\boldsymbol{t}$ peut être déterminée comme étant égale à $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{t^\top X_k X_k^\top t}$.

Pour des représentations graphiques, nous recommandons de redimensionner la composante globale $\boldsymbol{t}$ en la multipliant par le scalaire $\sqrt{\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)}}$. De la sorte, la variable latente $\boldsymbol{t}$ apparaît comme la composante principale non standardisée des composantes partielles $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$.

Méthode D: Variante de l'ACCG

La variante de l'ACCG s'obtient en cherchant $\boldsymbol{t}$ de longueur 1 (et $\boldsymbol{t_k = P_k t}$) de manière à maximiser:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)} \tag{3.12}$$

En procédant exactement comme dans la section précédente (en remplaçant $\boldsymbol{W_k}$ par $\boldsymbol{P_k}$), nous aboutissons à l'algorithme suivant:

0. Initialisation des poids: $\boldsymbol{\alpha_k = 1}$ $\boldsymbol{(k = 1, 2, ..., K)}$;

1. Calcul de la composante globale: $\boldsymbol{t}$ est le vecteur propre normé de $\sum_{k=1}^{K} \boldsymbol{\alpha_k P_k}$ associé à la plus grande valeur propre;

2. Calcul des composantes partielles: $\boldsymbol{t_k = P_k t}$;

3. Mise à jour des poids: $\boldsymbol{\alpha_k = t^\top t_k}$;

4. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

De même, nous pouvons, en suivant le même raisonnement que précédemment, proposer l'algorithme de type NIPALS suivant:

0. Choix de manière aléatoire de la composante globale $\boldsymbol{t}$ et sa normalisation ($\boldsymbol{t = t/||t||}$);

1. Détermination des composantes partielles: $\boldsymbol{t_k = P_k t\ (k = 1, 2, ..., K)}$;

2. Calcul des poids $\boldsymbol{\alpha_k = t^\top t_k}$;

3. Mise à jour de la composante globale $\boldsymbol{t}$: $\boldsymbol{t = \sum_{k=1}^{K} \alpha_k t_k}$ et sa normalisation ($\boldsymbol{t = t/||t||}$);

4. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

En utilisant le même développement que l'ACCG, il vient que:

$$\boldsymbol{cov(t_k, t) = \frac{1}{n} t^\top P_k t = var(P_k t)} \tag{3.13}$$

$\boldsymbol{t^\top P_k t}$ est le coefficient de détermination de $\boldsymbol{X_k}$ par rapport à $\boldsymbol{t}$. D'autres indices similaires peuvent être calculés pour déterminer l'importance relative des composantes.

3.2.3 Comparaison des méthodes non supervisées

Un premier critère qui permet de distinguer les méthodes non supervisées discutées ci-dessus est la manière dont les composantes partielles

$\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ sont déterminées à partir de la composante globale $\boldsymbol{t}$: $\boldsymbol{t_k = P_k t}$ ou $\boldsymbol{t_k = W_k t}$. Avec la première expression, les variances des variables ainsi que les corrélations des variables sont masquées. Du coup, la méthode d'analyse se focalisera sur l'explication des variations inter-blocs. Par contre, pour la deuxième expression, la méthode d'analyse cherchera à expliquer aussi bien les variations inter-blocs que les variations intra-blocs.

Une deuxième différence entre les méthodes non supervisées est la manière dont la composante globale est déterminée à partir de ses composantes partielles: $\boldsymbol{t \propto \sum_{k=1}^{K} t_k}$ ou $\boldsymbol{t \propto \sum_{k=1}^{K} \alpha_k t_k}$ (c'est-à-dire, $\boldsymbol{t}$ est la première composante principale des $\boldsymbol{t_k}$). Avec la dernière expression, les blocs de variables sont pondérées en fonction de leurs contributions à la détermination de la composante globale. Tous ces aspects seront illustrés grâce à des données simulées.

3.2.4 Proposition d'une méthode supervisée à partir d'une méthode non supervisée

Généralement, une méthode d'analyse des données multiblocs est conçue pour être soit une méthode non supervisée, soit une méthode supervisée. Cependant, on note la possibilité d'étendre une méthode non supervisée pour obtenir une méthode supervisée. C'est le cas, par exemple, de la méthode ComDim (méthode non supervisée) et de la méthode P-ComDim [10] (méthode supervisée) ou encore Path-ComDim [13]. Il est également possible d'utiliser les résultats d'une méthode non supervisée pour élaborer une méthode supervisée. Un exemple illustratif de cette idée est la régression par analyse des valeurs latentes (Latent Root Regression, LRR) [20, 21].

Cette méthode est obtenue en établissant un modèle de prédiction en régressant une variable $\boldsymbol{y}$ sur les composantes obtenues à partir d'une ACP du tableau concaténé $[\boldsymbol{y}|\boldsymbol{X}]$. Par la suite, cette méthode a été étendue dans le cas où la variable "réponse" est multivariée [22] puis au cas multiblocs où l'on souhaite prédire un tableau "réponse" $\boldsymbol{Y}$ à partir de $\boldsymbol{K}$ tableaux prédictifs $\boldsymbol{X_1, X_2, ..., X_K}$ [22]. Cette dernière méthode est connue sous le nom de la régression par analyse des valeurs latentes multiblocs et consiste à appliquer l'ACP multiblocs sur les tableaux de données $\boldsymbol{Y, X_1, X_2, ..., X_K}$. Cependant, les composantes globales sont contraintes d'être une combinaison linéaire des variables de $\boldsymbol{X_k}$. La méthode que nous proposons s'appelle Latent Root Multiblock Principal Components Analysis "LR-MBPCA". Elle est plus directe que la régression par analyse des valeurs latentes multiblocs et permet de déterminer les composantes partielles associées à chaque tableau.

Nous disposons des tableaux de données $\boldsymbol{Y, X_1, X_2, ..., X_K}$ et nous visons à prédire $\boldsymbol{Y}$ à partir de $\boldsymbol{X_1, X_2, ..., X_K}$. Le principe de la méthode LR-MBPCA est le suivant.

- Centrer, réduire (si nécessaire) et normer tous les tableaux de données $\boldsymbol{X_k}$ de manière à avoir leur norme égale à 1. De plus, nous recommandons de multiplier le tableau $\boldsymbol{Y}$ par la racine carrée du nombre de tableaux prédictifs, afin que l'inertie du tableau $\boldsymbol{Y}$ soit à elle seule égale à l'inertie totale des tableaux $\boldsymbol{X_1, X_2, ..., X_K}$.

- Appliquer l'ACP multiblocs sur les tableaux $\boldsymbol{Y, X_1, X_2, ..., X_K}$. En désignant le tableau $\boldsymbol{Y}$ par $\boldsymbol{X_0}$, nous obtenons ainsi pour une dimen-

sion h donnée, les composantes partielles $\boldsymbol{t_k^{(h)} = X_k X_k^\top t^{(h)},\ k = 0, 1, ..., K}$ respectivement associées à $\boldsymbol{X_k,\ k = 0, 1, ..., K}$ et la composante globale $\boldsymbol{t^{(h)} = \sum_{k=0}^{K} t_k^{(h)}}$, avec $\boldsymbol{||t^{(h)}|| = 1}$.

- Déterminer la composante latente prédictive: $\boldsymbol{t_X^{(h)} = \sum_{k=1}^{K} t_k^{(h)}}$ et la normer (c'est-à-dire, $\boldsymbol{t_X^{(h)} = t_X^{(h)} / ||t_X^{(h)}||}$).

- Pour déterminer les composantes latentes prédictives d'ordre supérieur à h, nous recommandons d'effectuer une déflation de tous les tableaux par rapport à la composante latente prédictive $\boldsymbol{t_X^{(h)}}$ au lieu de la composante globale $\boldsymbol{t^{(h)}}$ pour rendre orthogonales les composantes latentes prédictives et améliorer la prédiction. Après chaque déflation, nous recommandons aussi de multiplier le tableau $\boldsymbol{Y}$ par la racine carrée du nombre de tableaux prédictifs.

- Établir enfin un modèle de prédiction, en régressant les variables du tableau $\boldsymbol{Y}$ sur les composantes latentes prédictives déterminées.

De manière plus concrète, on peut noter que chaque composante partielle $\boldsymbol{t_k^{(h)} = X_k X_k^\top t^{(h)}}$ est par définition, une combinaison linéaire des variables de $\boldsymbol{X_k}$: $\boldsymbol{t_k^{(h)} = X_k w_k^{(h)}}$, avec $\boldsymbol{w_k^{(h)} = X_k^\top t^{(h)}}$. Puisque $\boldsymbol{t_X^{(h)} = \sum_{k=1}^{K} t_k^{(h)}}$, il vient que le vecteur de poids, $\boldsymbol{w^{(h)}}$, est une concaténation des vecteurs $\boldsymbol{w_k^{(h)}}$. Par suite, nous normons $\boldsymbol{w^{(h)}}$ et nous notons par $\boldsymbol{W}$, la matrice formée par ces vecteurs de poids. Le vecteur de loadings associé à chaque variable latente prédictive $\boldsymbol{t_X^{(h)}}$, est $\boldsymbol{p_X^{(h)} = \frac{X^\top t_X^{(h)}}{t_X^{(h)\top} t_X^{(h)}}}$, qui n'est rien d'autre que le coefficient de régression de $\boldsymbol{X = [X_1|X_2|...|X_K]}$ sur $\boldsymbol{t_X^{(h)}}$. Notons par $\boldsymbol{P_X}$, la matrice dont les colonnes sont les vecteurs

de $\boldsymbol{p}_{\boldsymbol{X}}^{(h)}$. De manière analogue, le vecteur de loadings associé à $\boldsymbol{Y}$ est $\boldsymbol{p}_{\boldsymbol{Y}}^{(h)} = \frac{\boldsymbol{Y}^\top \boldsymbol{t}_{\boldsymbol{X}}^{(h)}}{\boldsymbol{t}_{\boldsymbol{X}}^{(h)\top} \boldsymbol{t}_{\boldsymbol{X}}^{(h)}} \times \frac{\mathbf{1}}{\boldsymbol{K}^{\frac{h-1}{2}}}$, où la constante $\frac{\mathbf{1}}{\boldsymbol{K}^{\frac{h-1}{2}}}$ est introduite du fait de la multiplication de $\boldsymbol{Y}$ par $\sqrt{\boldsymbol{K}}$ après chaque déflation. Notons par $\boldsymbol{P}_{\boldsymbol{Y}}$, la matrice contenant les vecteurs $\boldsymbol{p}_{\boldsymbol{Y}}^{(h)}$. Nous savons que la matrice des poids, $\boldsymbol{W}^*$ qui est directement exprimée en fonction des variables de $\boldsymbol{X}$ (au lieu des valeurs de $\boldsymbol{X}$ obtenues après déflation) est donnée par: [23, 24]

$$\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{P}_{\boldsymbol{X}}^\top \boldsymbol{W})^{-1} \tag{3.14}$$

Le modèle pour prédire $\boldsymbol{Y}$ à partir de $\boldsymbol{X} = [\boldsymbol{X_1}|\boldsymbol{X_2}|...|\boldsymbol{X_K}]$ est obtenu en régressant $\boldsymbol{Y}$ sur $\boldsymbol{X}\boldsymbol{W}^*$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \tag{3.15}$$

où, $\boldsymbol{\beta} = \boldsymbol{W}^* \boldsymbol{P}_{\boldsymbol{Y}}^\top$ est la matrice des coefficients de régression et $\boldsymbol{E}$, la matrice résiduelle.

Le nombre de composantes à inclure dans le modèle est déterminé à partir de la technique de cross-validation, comme il est de coutume pour la régression PLS [25, 26].

3.3 Illustrations

Les méthodes non supervisées d'analyse des données multiblocs sont illustrées et comparées sur la base de données simulées et de données réelles. Un troisième jeu de données a servi d'illustration pour la nouvelle approche prédictive, LR-MBPCA, que nous avons proposée.

3.3.1 Simulation des données

Les données que nous simulons sont similaires à celles de Westerhuis et al. [3]. Nous considérons deux variables orthogonales $\boldsymbol{d_1}$, $\boldsymbol{d_2}$, quatre blocs de variables correspondant à cinquante individus et cinq variables, chacun. Les variables du bloc $\boldsymbol{X_1}$ sont formées de $\boldsymbol{d_1}$. Puis, nous ajoutons à chaque variable 20% de bruit: $\boldsymbol{x_{1j} = d_1 + 0.2\epsilon_{1j}}$, avec $\boldsymbol{j = 1, 2, ..., 5}$, $\boldsymbol{\epsilon_{1j} \sim \mathcal{N}(m_1, \sigma_1)}$, $\boldsymbol{m_1}$ et $\boldsymbol{\sigma_1}$ étant respectivement la moyenne et l'écart-type de la variable $\boldsymbol{d_1}$. Le bloc $\boldsymbol{X_2}$ est formé de $\boldsymbol{d_2}$ comme première variable, à laquelle est rajoutée 20% de bruit. Le reste des variables ne relève que du bruit: $\boldsymbol{x_{21} = d_2 + 0.2\epsilon_{21}}$, $\boldsymbol{x_{2j} = \epsilon_{2j}}$, avec $\boldsymbol{\epsilon_{21} \sim \mathcal{N}(m_2, \sigma_2)}$, $\boldsymbol{m_2}$ et $\boldsymbol{\sigma_2}$ étant respectivement la moyenne et l'écart-type de la variable $\boldsymbol{d_2}$ et $\boldsymbol{\epsilon_{2j} \sim \mathcal{N}(0, 1)}$. Les blocs $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$ sont formés de manière analogue à $\boldsymbol{X_2}$. Tous les tableaux $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ sont centrés, réduits et normés. Par la suite, nous avons appliqué l'ACCG, l'ACCG-V, l'ACP multiblocs et la méthode ComDim.

Le Tableau 3.2 présente les pourcentages d'inerties de $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ expliquées par les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs et la méthode ComDim. De ce tableau, il ressort que la première composante globale de l'ACCG et l'ACCG-V est très corrélée avec la variable $\boldsymbol{d_2}$ car l'information concernant cette variable est partagée par les tableaux $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$. Par contre, la première composante globale de l'ACP multiblocs, d'une part, et celle de la méthode ComDim, d'autre part, est plus concernée par la variable $\boldsymbol{d_1}$ qui définit le bloc de variables $\boldsymbol{X_1}$ (corrélation de $\boldsymbol{r = 0.99}$ et $\boldsymbol{r = 1}$, respectivement). Quant à la deuxième

composante globale de l'ACP multiblocs et de la méthode ComDim, elle est orientée vers la variable $\boldsymbol{d_2}$ ($\boldsymbol{r = 0.97}$). La deuxième composante globale de l'ACCG et l'ACCG-V, ne reflète que du bruit car presque toute la variabilité contenue dans la variable $\boldsymbol{d_2}$, commune aux tableaux $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$, a été déjà expliquée par la première composante globale.

Les résultats obtenus à partir de l'ACP multiblocs s'accordent bien avec ceux présentés par Westerhuis et al. [3]. Cependant, ce n'est pas le cas pour les résultats de l'ACP hiérarchique qui est équivalente à ComDim [27] puisque ces auteurs utilisent la version de l'ACP hiérarchique où les composantes partielles sont normées; ce qui pose de sérieux problèmes de convergence.

Cette étude de simulation confirme bien le fait que l'ACCG et l'ACCG-V cherchent à expliquer les variabilités inter-blocs (c'est-à-dire, ce qui est commun aux différents tableaux de données). Quant à l'ACP multiblocs et la méthode ComDim, elles cherchent à expliquer les variabilités intra-blocs et les variabilités inter-blocs.

Tab. 3.2: Données simulées: pourcentages d'inerties des blocs $\boldsymbol{X}_1$ à $\boldsymbol{X}_4$ expliquées par les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs, ComDim et corrélations de ces composantes globales avec les variables $\boldsymbol{d}_1$ et $\boldsymbol{d}_2$.

		% d'inerties restituées					Corrélations	
		$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	Globale	$\boldsymbol{d}_1$	$\boldsymbol{d}_2$
ACCG	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
ACCG-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
ACP multiblocs	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**

3.3.2 Données sensorielles "jambon"

Un panel d'experts en évaluation sensorielle (juges) a évalué la flaveur, l'arôme et la texture de huit types de jambons américains cuits après séchage, en se servant d'un profil conventionnel (c'est-à-dire, les variables sensorielles sont communes à tous les juges) [28]. Pour chaque type de jambon, les juges ont attribué une note allant de 0 à 15 qui exprime l'intensité perçue pour chaque variable sensorielle. Par exemple, une note de 0 signifie que la variable n'est pas détectée et une note de 15, signifie qu'elle est extrêmement détectée. Les notes moyennes attribuées par les juges sont consignées dans trois tableaux: le premier tableau, $\boldsymbol{X_1}$ (3 variables), porte sur la flaveur des jambons, le deuxième tableau, $\boldsymbol{X_2}$ (4 variables) concerne l'arôme et le dernier tableau, $\boldsymbol{X_3}$ (3 variables), la texture. En lignes de

ces tableaux, nous avons les huit types de jambons et en colonnes, les variables sensorielles. Les tableaux $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$ sont soumis à l'ACCG, l'ACCG-V, l'ACP multiblocs et à la méthode ComDim.

Le tableau 3.3 présente les pourcentages d'inerties expliquées par les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs et la méthode ComDim. Nous pouvons noter que la première composante globale de l'ACCG et de l'ACCG-V recouvre une bonne proportion de la variabilité contenue dans les tableaux $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et recouvre un peu moins la variabilité du tableau, $\boldsymbol{X_3}$. Ce qui pourrait s'expliquer par le fait que la première composante globale associée à l'ACCG et à l'ACCG-V contient les informations communes aux trois blocs de variables: "Porkcomplex" et "Savory" ($\boldsymbol{X_1}$), "Molasses" et "Caramelized" ($\boldsymbol{X_2}$) et "Mushiness" ($\boldsymbol{X_3}$) (voir le tableau 3.4).

Tab. 3.3: Données sensorielles: pourcentages d'inerties de $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$ expliquées par les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs et de la méthode ComDim.

		$\boldsymbol{X}_1$ (Flaveur)	$\boldsymbol{X}_2$ (Arôme)	$\boldsymbol{X}_3$ (Texture)	Globale
ACCG	Dim.1	**48.44**	**40.74**	**30.12**	39.77
	Dim.2	12.68	13.77	16.33	14.26
ACCG-V	Dim.1	**48.94**	**40.80**	**29.46**	39.73
	Dim.2	12.17	13.71	16.98	14.29
ACP multiblocs	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

Quant à la première composante globale de l'ACP multiblocs, elle présente un différent aspect, puisqu'elle recouvre jusqu'à 62.70% de l'inertie contenue dans le tableau $\boldsymbol{X_3}$. Pour cette composante, la plus faible variabilité expliquée est associée au tableau $\boldsymbol{X_1}$. Ce résultat est beaucoup plus prononcé avec la méthode ComDim, puisque sa première composante globale recouvre jusqu'à 74.66% de la variabilité du tableau $\boldsymbol{X_3}$ et seulement 22.58% de la variabilité du tableau $\boldsymbol{X_1}$. Pour expliquer cela, nous présentons dans le tableau 3.5, les corrélations entre variables sensorielles. Ce tableau nous indique que le bloc de variables $\boldsymbol{X_3}$ est formé des variables relativement plus corrélées entre elles que les variables d'autres blocs. De plus, les variables de ce bloc sont corrélées avec les variables "Rancid" et "Earthy" du bloc $\boldsymbol{X_2}$. La méthode ComDim semble attribuer un faible

poids au tableau $\boldsymbol{X_1}$, puisque, d'une part, les variables dans ce tableau ne sont pas très corrélées les unes aux autres, et, à l'exception de la variable "Savory", ces variables ne sont pas non plus fortement corrélées avec les variables d'autres blocs (voir le tableau 3.5).

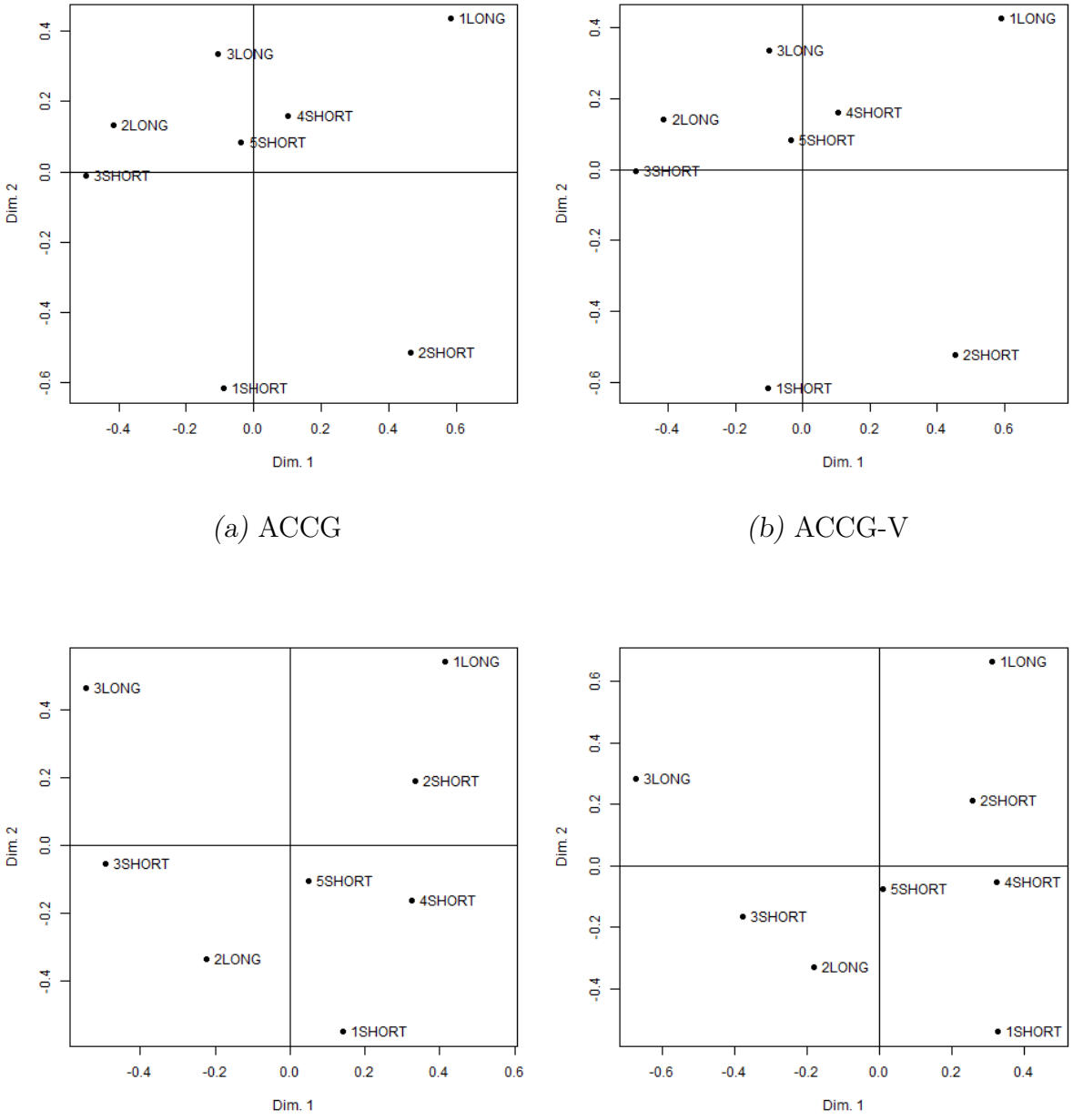
Tab. 3.4: Corrélations entre les variables sensorielles et les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs et de ComDim.

Variables	ACCG		ACCG-V		ACP multiblocs		ComDim	
	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
Pork.	**-0.75**	-0.34	**-0.76**	-0.33	**-0.73**	-0.40	-0.54	**-0.60**
Sav.	**0.94**	0.23	**0.95**	0.21	**0.63**	**0.73**	0.43	**0.85**
Ranc.	-0.46	0.08	-0.46	0.09	**-0.76**	0.23	**-0.73**	-0.01
Mol.	**0.86**	-0.22	**0.86**	-0.23	**0.72**	0.43	**0.62**	0.57
Caram.	**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
Dry.	-0.08	0.32	-0.08	0.33	**-0.63**	**0.66**	**-0.80**	0.45
Juic.	0.41	-0.47	0.40	-0.48	**0.76**	-0.37	**0.89**	-0.17
Mus.	**0.85**	-0.40	**0.85**	-0.41	**0.95**	0.07	**0.90**	0.27

Tab. 3.5: Corrélations entre les variables sensorielles.

	Salty	Pork.	Sav.	Ranc.	Mol.	Caram.	Earthy	Dry.	Juic.	Mus.
Salty	1									
Pork.	0.13	1								
Sav.	0.13	**-0.65**	1							
Ranc.	0.32	**0.62**	-0.30	1						
Mol.	0.09	**-0.53**	**0.76**	-0.34	1					
Caram.	0.17	**-0.75**	**0.85**	-0.20	**0.54**	1				
Earthy	0.43	0.21	0.15	**0.78**	-0.19	0.18	1			
Dry.	**0.54**	0.22	0.11	**0.56**	-0.09	-0.01	**0.72**	1		
Juic.	-0.35	-0.19	0.25	-0.49	**0.57**	0.12	**-0.60**	**-0.80**	1	
Mus.	-0.39	**-0.62**	**0.67**	**-0.62**	**0.76**	**0.58**	**-0.54**	**-0.50**	**0.69**	1

La figure 3.2 représente les configurations des huit jambons dans le plan formé par les deux premières composantes globales associées aux différentes méthodes non supervisées. L'interprétation de ces figures peut mieux se faire en se basant sur les corrélations entre les variables sensorielles et les deux premières composantes globales de l'ACCG, l'ACCG-V, l'ACP multiblocs et de la méthode ComDim (tableau 3.4). Il est clair que d'une part, les configurations obtenues à partir de l'ACCG et l'ACCG-V s'accordent entre elles et, d'autre part, nous notons une très grande similarité entre les configurations de l'ACP multiblocs et de la méthode ComDim. Cela confirme bien que nous sommes en présence de deux familles de méthodes. L'ACCG et l'ACCG-V, d'une part, qui ont pour but d'expliquer les variations inter-blocs, et, d'autre part, l'ACP multiblocs et ComDim, qui ont pour but d'expliquer les variations intra-blocs et inter-blocs.



(a) ACCG

(b) ACCG-V

(c) ACP multiblocs

(d) ComDim

Fig. 3.2: Configurations des jambons dans le plan formé par les deux premières composantes globales des méthodes (a) ACCG, (b) ACCG-V, (c) ACP multiblocs et (d) ComDim.

3.3.3 *Données de "pommes de terre"*

Afin d'évaluer la performance de la nouvelle approche, LR-MBPCA, en termes de prédiction et la comparer à celle de la régression MB-PLS [3,29], nous avons utilisé une étude de cas dans laquelle il s'agit de prédire des attributs sensoriels des variétés de pommes de terre à partir des données obtenues en effectuant des mesures physico-chimiques sur ces pommes de terre. Ce problème revêt un intérêt primordial en pratique car la collecte des données sensorielles est très coûteuse et nécessite beaucoup de temps.

Vingt variétés de pommes de terre ont été analysées après un mois de stockage et six autres variétés, après huit mois de stockage. Plusieurs juges ont ensuite évalué leurs textures à l'aide de neuf attributs. La moyenne des notes attribuées par tous les juges a donné un bloc de variables que nous notons par $\boldsymbol{Y}$. Le bloc $\boldsymbol{X}_1$ est donné par les analyses chimiques des variétés de pommes de terre. Un deuxième bloc de variables $\boldsymbol{X}_2$ concerne la compression uni-axiale. Le troisième bloc ($\boldsymbol{X}_3$) est relatif aux courbes de relaxation et le quatrième bloc de variables ($\boldsymbol{X}_4$) concerne les mesures de spectroscopie proche infrarouge (NIR). Pour plus de précisions sur ces données, le lecteur peut se référer à Thybo et al. [30].

Tous ces blocs de variables sont ensuite centrés, réduits puis normés. De plus, le bloc de variables $\boldsymbol{Y}$ est multiplié par 2 afin qu'il ait une inertie égale à celle des blocs $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ tous ensemble. Nous appliquons sur ces données la méthode LR-MBPCA puis la régression MB-PLS. La figure 3.3 présente les pourcentages d'inerties cumulées de $\boldsymbol{Y}$ expliquées par les trois premières composantes globales de LR-MBPCA et la régression MB-PLS.

De cette figure, il apparaît clairement que les deux méthodes aboutissent pratiquement aux mêmes résultats. Nous avons aussi appliqué la technique de cross-validation appelée "leave-one-out" pour évaluer la qualité de prédiction de ces deux méthodes.

La figure 3.4 présente les erreurs de prédiction (Root Mean Squared Errors of Prediction, RMSEP) associée à ces deux méthodes. Nous voyons à nouveau que les deux méthodes semblent avoir des performances similaires, bien que nous puissions remarquer que pour la régression MB-PLS, l'erreur croît légèrement à la composante 4 avant de décroître; ce phénomène n'étant pas observé pour LR-MBPCA.

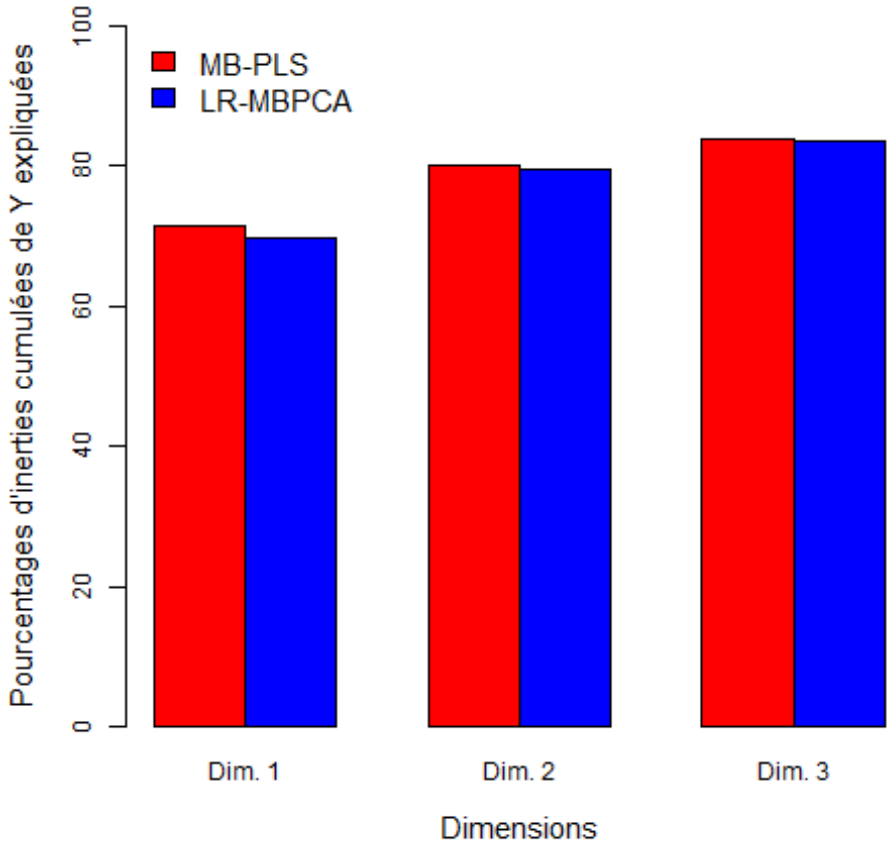


Fig. 3.3: Pourcentages d'inerties cumulées de $\boldsymbol{Y}$ expliquées par les trois premières composantes globales de la régression MB-PLS et LR-MBPCA.

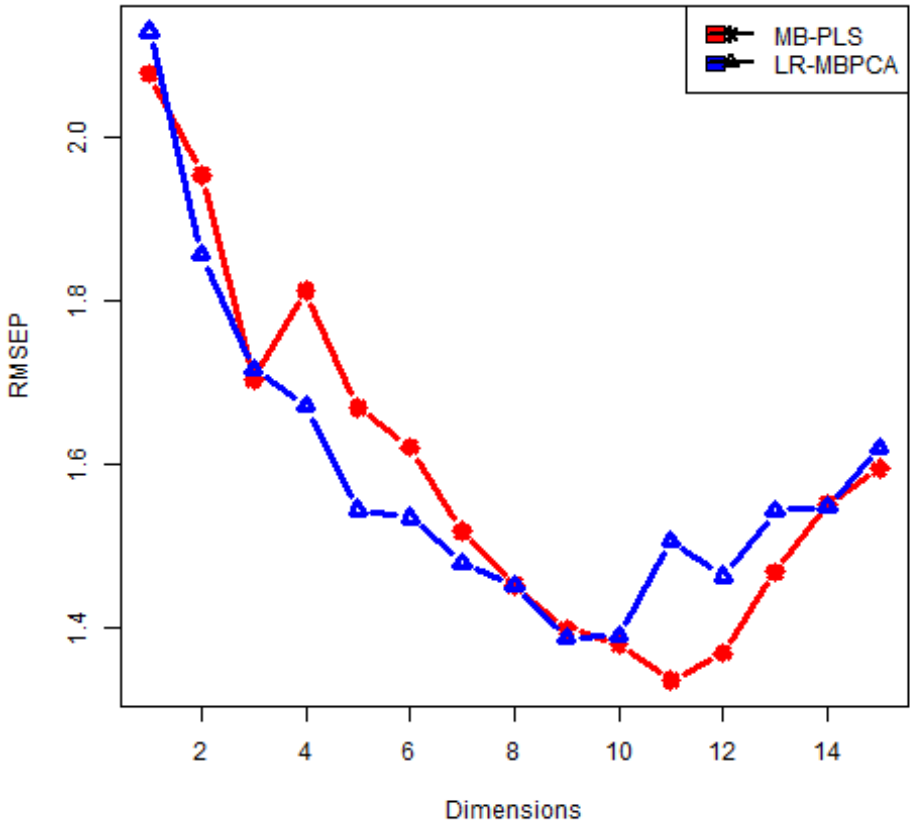


Fig. 3.4: Erreurs de prédiction pour les quinze premières composantes globales de la régression MB-PLS et LR-MBPCA.

3.4 Discussion et conclusion

Dans ce chapitre, nous avons comparé quelques méthodes non supervisées en faisant clairement ressortir les similitudes et les différences entre elles. La première différence fondamentale entre les méthodes considérées est la manière dont la composante globale, $\boldsymbol{t}$, est reflétée dans chaque bloc de variables $\boldsymbol{X_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ pour donner les composantes partielles $\boldsymbol{t_k}$, c'est-à-dire, $\boldsymbol{t_k = X_k(X_k^\top X_k)^{-1} X_k^\top t}$ ou $\boldsymbol{t_k = X_k X_k^\top t}$. En choisissant la première expression, on occulte non seulement les variances des variables mais aussi les corrélations entre les variables du bloc considéré.

Ainsi, la méthode d'analyse focalisera uniquement sur l'explication des traits communs aux différents blocs de variables, peu importe si cela est important en termes d'inertie expliquée ou pas. Avec la deuxième expression $\boldsymbol{t_k} = \boldsymbol{X_k}\boldsymbol{X_k}^\top \boldsymbol{t}$, la méthode d'analyse cherchera à restituer aussi bien la variabilité inter-blocs que la variabilité intra-blocs.

La deuxième différence fondamentale entre les méthodes est la synthèse faite par la composante globale à partir de ses composantes partielles. C'est-à-dire, l'expression $\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{t_k}$ (somme des composantes partielles) ou $\boldsymbol{t} \propto \sum_{k=1}^{K} \boldsymbol{\alpha_k t_k}$ (première composante principale des composantes partielles). Cette dernière expression offre plus de possibilités de développements de méthodes que la première car, on pourra utiliser, par exemple, l'idée de composantes partielles "sparse". En d'autres termes, pour une composante globale donnée, seules les composantes partielles fortement liées à elle seront considérées. Ce point sera développé dans le chapitre 5.

Nous avons proposé des algorithmes itératifs pour exécuter les méthodes non supervisées d'analyse de données multiblocs appariées par individus. De plus, pour l'ACCG et l'ACP multiblocs, il existe aussi une solution beaucoup plus directe (solution basée sur l'analyse des vecteurs propres), qui ne nécessite pas un algorithme itératif. A partir des données simulées, nous avons exécuté plusieurs fois ces algorithmes itératifs en considérant pour chaque exécution, un vecteur aléatoire pour l'initialisation. De cette investigation, il résulte que: (i) dans toutes les situations, les algorithmes ont convergé; (ii) le même optimum est obtenu pour toutes les exécutions; (iii) et, le cas échéant, cet optimum est égal à celui obtenu à partir de la solution directe. Pour les données considérées, les algorithmes itératifs

n'étaient pas sensibles au choix du vecteur initial.

Nous avons aussi proposé une méthode supervisée basée sur les composantes partielles associées aux blocs de variables prédictifs. Outre sa simplicité, sa performance en termes de prédiction semble être similaire à celle de la régression MB-PLS. Évidemment, cette stratégie d'analyse peut facilement être adaptée à l'ACCG ou à la méthode ComDim et pourrait offrir d'autres extensions en s'inspirant des idées de l'ACP "sparse" [31,32], par exemple.

4

Une stratégie générale pour unifier les méthodes supervisées d'analyse des données multiblocs

4.1 Introduction

Rappelons qu'une méthode supervisée d'analyse des données multiblocs est une méthode utilisée pour relier des tableaux de données portant sur les mêmes individus dans le but d'explorer les relations entre, d'un côté, un ou plusieurs tableau(x) réponse(s) et d'un autre côté, un ou de plusieurs tableau(x) dit(s) "prédictif(s)". Comme exemple de méthodes, nous pouvons citer la méthode multiblock redundancy analysis "MB-RA" [7–9], la régression multiblock partial least squares "MB-PLS" [3, 6, 33–35], la méthode Hierachical PLS "H-PLS" [3] et la méthode P-ComDim [10].

L'objectif de ce chapitre est de: (i) proposer une approche unifiée pour MB-RA et la régression MB-PLS, (ii) définir une variante de chacune des deux méthodes, (iii) faire ressortir les points communs et les points divergents entre toutes ces méthodes, (iv) proposer des indices qui pourraient servir à une meilleure interprétation des résultats d'analyse.

Ce chapitre est structuré comme suit. Dans la section suivante, nous proposons une présentation originale de la méthode Redundancy Analysis "RA" [36, 37] et celle de la régression PLS [33]. Ensuite, sous deux angles d'attaque, nous allons étendre ces deux méthodes supervisées dans le cadre multiblocs. La première extension de la méthode RA nous conduit à la méthode MB-RA [7–9] et la deuxième extension définit une variante de la méthode MB-RA que nous désignons par "Multiblock Weighted Redundancy Analysis" (MB-WRA). De même, la première extension de la régression PLS nous conduit à la régression MB-PLS [3, 6, 33–35] et la deuxième extension définit une variante de MB-PLS, que nous désignons par "Multiblock Weighted Covariate analysis" (MB-WCov). Les différentes méthodes sont illustrées et comparées sur la base des données simulées et des données réelles. Enfin, une conclusion clôt le chapitre.

4.2 Méthodes

4.2.1 Relations entre deux tableaux de données:

Redundancy analysis

Soit deux tableaux de données $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ et $\boldsymbol{Y}$ $(\boldsymbol{n} \times \boldsymbol{q})$ portant sur les mêmes individus. Ces tableaux sont supposés être centrés et réduits (si nécessaire). L'objectif de l'étude est d'explorer les relations entre les tableaux $\boldsymbol{Y}$ et $\boldsymbol{X}$.

Nous désignons par $\boldsymbol{u}$, la variable latente associée au tableau $\boldsymbol{Y}$. La variable latente $\boldsymbol{u}$ est définie par une combinaison linéaire des variables de $\boldsymbol{Y}$, sous contrainte que le vecteur des loadings, $\boldsymbol{\nu}$, soit de longueur

égale à 1. De manière formelle, nous avons $\boldsymbol{u} = \boldsymbol{Y\nu}$, avec $||\boldsymbol{\nu}|| = \mathbf{1}$. La projection de $\boldsymbol{u}$ dans l'espace engendré par les variables de $\boldsymbol{X}$ est donnée par $\boldsymbol{t} = \boldsymbol{P_X u}$, où $\boldsymbol{P_X} = \boldsymbol{X(X^\top X)^{-1}X^\top}$ est le projecteur orthogonal dans l'espace formé par les variables de $\boldsymbol{X}$. Nous cherchons $\boldsymbol{u}$, et, par conséquent, $\boldsymbol{t}$, de manière à maximiser le critère:

$$\boldsymbol{cov(u,t)} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{u^\top t} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{\nu^\top Y^\top P_X Y\nu} \tag{4.1}$$

Par ce critère, nous cherchons une variable latente $\boldsymbol{u}$ dans l'espace engendré par les variables de $\boldsymbol{Y}$ et une variable latente $\boldsymbol{t}$ dans l'espace engendré par les variables de $\boldsymbol{X}$, de sorte que les deux variables latentes soient les plus liées possibles.

Il est clair que le vecteur $\boldsymbol{\nu}$ qui maximise ce critère est donné par le vecteur propre de $\boldsymbol{Y^\top P_X Y}$ associé à la plus grande valeur propre. Un algorithme de type NIPALS pour trouver la solution du critère de maximisation précédent est le suivant:

0. Choix de manière aléatoire du vecteur $\boldsymbol{\nu}$ et sa normalisation ($\boldsymbol{\nu} = \boldsymbol{\nu}/||\boldsymbol{\nu}||$);

1. Détermination de la variable latente associée à $\boldsymbol{Y}$: $\boldsymbol{u = Y\nu}$;

2. Détermination de la variable latente associée à $\boldsymbol{X}$: $\boldsymbol{t = P_X u}$;

3. Mise à jour du vecteur $\boldsymbol{\nu}$: $\boldsymbol{\nu = Y^\top t/||Y^\top t||}$;

4. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence. La convergence est atteinte lorsque la différence du critère à maximiser à deux itérations successives est inférieure à un seuil $\boldsymbol{\epsilon}$, pré-défini par

l'utilisateur (par exemple, $\boldsymbol{\epsilon = 10^{-8}}$).

Cette procédure itérative étant semblable à l'algorithme des puissances itérées pour la détermination d'un vecteur propre associé à la plus grande valeur propre, nous pouvons être rassurés quant à la convergence de cet algorithme.

Par construction, la variable $\boldsymbol{t}$ est une combinaison linéaire des variables de $\boldsymbol{X}$: $\boldsymbol{t = Xw^*}$ avec $\boldsymbol{w^* = (X^\top X)^{-1} X^\top u}$. Le vecteur $\boldsymbol{w^*}$ pourrait être normalisé à 1 ($\boldsymbol{w = w^*/||w^*||}$) pour des raisons d'interprétation.

Étant donné que le projecteur $\boldsymbol{P_X}$ est idempotent ($\boldsymbol{P_X^2 = P_X}$) et symétrique ($\boldsymbol{P_X^\top = P_X}$), nous avons:

$$\boldsymbol{Y^\top P_X Y = Y^\top P_X P_X Y = Y^\top P_X^\top P_X Y = (P_X Y)^\top (P_X Y)} \quad (4.2)$$

Il s'ensuit alors que $\boldsymbol{trace(Y^\top P_X Y) = n \sum_{j=1}^{q} var(\hat{y}_j)}$, où $\boldsymbol{\hat{y}_j = P_X y_j}$ est la projection de la $\boldsymbol{j^{ème}}$ variable, $\boldsymbol{y_j}$ de $\boldsymbol{Y}$ dans l'espace formé par les variables de $\boldsymbol{X}$. En d'autres termes, $\boldsymbol{trace(Y^\top P_X Y)}$ reflète l'inertie totale de $\boldsymbol{Y}$ expliquée par $\boldsymbol{X}$. De même, nous avons:

$$\boldsymbol{\nu^\top Y^\top P_X Y \nu = u^\top P_X u = (P_X u)^\top (P_X u) = n \times var(P_X u)} \quad (4.3)$$

Ainsi, nous pouvons calculer un indice donnant la variation de $\boldsymbol{P_X Y}$ expliquée par la variable latente $\boldsymbol{t}$. Il vaut:

$$\boldsymbol{I = \frac{\nu^\top Y^\top P_X Y \nu}{trace(Y^\top P_X Y)} = \frac{var(P_X u)}{\sum_{j=1}^{q} var(P_X y_j)} = \frac{var(t)}{\sum_{j=1}^{q} var(P_X y_j)}} \quad (4.4)$$

Cet indice varie entre 0 et 1.

Notons également que le cas particulier où le tableau de données $\boldsymbol{Y}$ est constitué d'une seule variable $\boldsymbol{y}$, nous conduit à la régression linéaire multiple.

Des variables latentes d'ordre supérieur à 1 peuvent être déterminées suivant une déflation des tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$ par rapport aux composantes, $\boldsymbol{t} = \boldsymbol{P_X u}$, associées à $\boldsymbol{X}$. De manière concrète, la déflation du tableau $\boldsymbol{X}$ est donnée par: $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{tt}^\top)\boldsymbol{X}$ et celle du tableau $\boldsymbol{Y}$ par: $\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{tt}^\top)\boldsymbol{Y}$, avec $\boldsymbol{I}$, la matrice identité. L'intérêt de cette déflation est de soustraire de chaque tableau, l'information qui est déjà extraite par les variables latentes précédentes. En conséquence, les variables latentes successives associées à $\boldsymbol{X}$ sont orthogonales et l'interprétation des résultats en est facilitée.

Cette déflation nous amène à déterminer d'une part, les variables latentes $\boldsymbol{u^{(1)}, u^{(2)}, ...}$ associées à $\boldsymbol{Y}$ et d'autre part, les variables latentes $\boldsymbol{t^{(1)}, t^{(2)}, ...}$ associées à $\boldsymbol{X}$. A chaque dimension, $\boldsymbol{h}$, nous pouvons calculer l'indice:

$$\boldsymbol{I_h} = \frac{\boldsymbol{var(t^{(h)})}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{var(P_X y_j)}} \tag{4.5}$$

Ces indices peuvent être représentés graphiquement en fonction du nombre de variables latentes. Cela permet de les interpréter de manière similaire au diagramme en éboulis de l'Analyse en Composantes Principales (ACP) avec les pourcentages de variances expliquées par les composantes principales successives [19]. De ce fait, ce graphique nous donne une indication sur le nombre de variables latentes à retenir pour l'analyse.

Un modèle de prédiction des variables de $\boldsymbol{Y}$ à partir de celles de $\boldsymbol{X}$ peut

être établi en régressant $\boldsymbol{Y}$ sur les variables latentes $\boldsymbol{t^{(1)}, t^{(2)}, ..., t^{(A)}}$. Le nombre de variables latentes, $\boldsymbol{A}$, à introduire dans le modèle de prédiction peut être déterminé à partir de la technique de cross-validation [25].

Le projecteur, $\boldsymbol{P_X}$, nécessite l'inversion de la matrice $\boldsymbol{X^\top X}$. Comme nous l'avons indiqué dans le chapitre précédent, l'avantage de cette inversion est que les variances des variables de $\boldsymbol{X}$ ainsi que leurs corrélations sont masquées. Par conséquent, l'analyse se focalisera sur l'explication de la variabilité de $\boldsymbol{Y}$. Cependant, l'inconvénient majeur de l'inversion d'une telle matrice est que cela risque de poser de sérieux problèmes en cas de quasi-colinéarité entre les variables de $\boldsymbol{X}$ [15–18].

Régression PLS2

Nous considérons les mêmes hypothèses et notations que dans la section précédente. Afin de contourner la difficulté liée au problème de quasi-colinéarité, nous considérons, à la place du projecteur $\boldsymbol{P_X = X(X^\top X)^{-1}X^\top}$, l'opérateur $\boldsymbol{W_X = XX^\top}$. Ainsi, la matrice $\boldsymbol{(X^\top X)^{-1}}$ est remplacée par la matrice identité.

A la variable $\boldsymbol{u = Y\nu}$ $\boldsymbol{(||\nu|| = 1)}$, nous associons la variable, $\boldsymbol{t}$, de $\boldsymbol{X}$, définie par $\boldsymbol{t = W_X u}$. Par la suite, nous cherchons $\boldsymbol{u}$ (et, par conséquent, $\boldsymbol{t}$) de manière à maximiser le critère:

$$\boldsymbol{cov(u,t) = \frac{1}{n} u^\top t = \frac{1}{n} \nu^\top Y^\top X X^\top Y \nu}. \tag{4.6}$$

L'idée sous-jacente de ce critère de maximisation est la même que celle de la méthode RA, c'est-à-dire, chercher une variable latente $\boldsymbol{u}$ dans l'espace

de $\boldsymbol{Y}$ et une variable latente $\boldsymbol{t}$ dans l'espace de $\boldsymbol{X}$, de sorte que les deux variables latentes soient les plus liées possibles.

Il vient que, le vecteur $\boldsymbol{\nu}$, qui maximise cette forme quadratique est donné par le vecteur propre de $\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}$ associé à la plus grande valeur propre. Il est clair que nous retrouvons la régression PLS [34]. Un algorithme de type NIPALS pour la résolution du critère de maximisation ci-dessus est donné par:

0. Choix de manière aléatoire du vecteur $\boldsymbol{\nu}$ et sa normalisation ($\boldsymbol{\nu} = \boldsymbol{\nu}/||\boldsymbol{\nu}||$);

1. Détermination de la variable latente associée à $\boldsymbol{Y}$: $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{\nu}$;

2. Détermination de la variable latente associée à $\boldsymbol{X}$: $\boldsymbol{t} = \boldsymbol{W_X}\boldsymbol{u}$;

3. Mise à jour du vecteur $\boldsymbol{\nu}$: $\boldsymbol{\nu} = \boldsymbol{Y}^\top \boldsymbol{t}/||\boldsymbol{Y}^\top \boldsymbol{t}||$;

4. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence (c'est-à-dire, jusqu'à ce que la différence du critère de maximisation à deux itérations successives soit inférieure à un seuil, $\boldsymbol{\epsilon}$, pré-défini par l'utilisateur, par exemple, $\boldsymbol{\epsilon} = \mathbf{10^{-8}}$).

La convergence de cet algorithme est assurée par le fait qu'à chaque étape, le critère de maximisation croît. De plus, étant donné que ce critère est majoré, il s'ensuit que la suite de valeurs de ce critère générées au cours des itérations, converge.

Comme la régression PLS2 est liée à la décomposition en valeurs propres et vecteurs propres de la matrice $\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}$, nous pouvons remarquer que $\boldsymbol{trace}(\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}) = \boldsymbol{trace}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Y}^\top)$. En nous servant des

relations $\boldsymbol{XX^\top} = \sum_{i=1}^{p} \boldsymbol{x_i x_i^\top}$ et $\boldsymbol{YY^\top} = \sum_{j=1}^{q} \boldsymbol{y_j y_j^\top}$, avec $\boldsymbol{x_i}$ et $\boldsymbol{y_j}$ respectivement, la $\boldsymbol{i^{ème}}$ variable de $\boldsymbol{X}$ et la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{Y}$, nous avons:

$$\boldsymbol{trace(Y^\top XX^\top Y) = n^2 \sum_{i=1}^{p} \sum_{j=1}^{q} cov^2(x_i, y_j)} \tag{4.7}$$

De ce fait, la quantité $\boldsymbol{trace(Y^\top XX^\top Y)}$ reflète la force de la liaison entre les variables de $\boldsymbol{X}$ et celles de $\boldsymbol{Y}$. Cette quantité, introduite par Robert et Escouffier [38, 39], est intimement liée au coefficient RV, qui est largement utilisé en Sensométrie et Chimiométrie [40–45]. De même, nous avons:

$$\boldsymbol{\nu^\top Y^\top XX^\top Y \nu = n^2 \sum_{i=1}^{p} cov^2(x_i, u)} \tag{4.8}$$

Nous pouvons calculer l'indice:

$$\begin{aligned} \boldsymbol{I} &= \frac{\boldsymbol{\nu^\top Y^\top XX^\top Y \nu}}{\boldsymbol{trace(Y^\top XX^\top Y)}} = \frac{\boldsymbol{n \times cov(t, u)}}{\boldsymbol{trace(Y^\top XX^\top Y)}} \\ &= \frac{\sum_{i=1}^{p} \boldsymbol{cov^2(x_i, u)}}{\sum_{i=1}^{p} \sum_{j=1}^{q} \boldsymbol{cov^2(x_i, y_j)}} = \frac{\boldsymbol{\lambda_1}}{\boldsymbol{trace(Y^\top XX^\top Y)}} \end{aligned} \tag{4.9}$$

avec $\boldsymbol{\lambda_1}$ la plus grande valeur propre de $\boldsymbol{Y^\top XX^\top Y}$.
Cet indice reflète la proportion de covariation (c'est-à-dire, $\boldsymbol{trace(Y^\top XX^\top Y)}$) expliquée par $\boldsymbol{t}$ (et $\boldsymbol{u}$).

Les variables latentes successives peuvent être déterminées après une déflation de $\boldsymbol{X}$ et de $\boldsymbol{Y}$ par rapport aux variables latentes associées à $\boldsymbol{X}$ qui sont déterminées de manière séquentielle. Nous obtenons ainsi les variables latentes $\boldsymbol{u^{(1)}, u^{(2)}, ...}$ et $\boldsymbol{t^{(1)}, t^{(2)}}$, etc. A chaque dimension, $\boldsymbol{h}$,

nous pouvons calculer l'indice:

$$\boldsymbol{I_h = \frac{\sum_{i=1}^{p} cov^2(x_i, u^{(h)})}{\sum_{i=1}^{p}\sum_{j=1}^{q} cov^2(x_i, y_j)}} \tag{4.10}$$

Nous pourrons ensuite faire une représentation graphique de ces indices en fonction de $\boldsymbol{h}$, de manière similaire à un diagramme en éboulis. L'intérêt de cette représentation graphique est de fournir un nouvel indicateur pour décider du nombre de variables latentes à retenir pour l'analyse.

4.2.2 Relations entre plusieurs tableaux

Considérons le cas multiblocs où nous disposons d'un tableau $\boldsymbol{Y}$ $\boldsymbol{(n \times q)}$ à prédire à partir de $\boldsymbol{K}$ tableaux prédictifs $\boldsymbol{X_1, X_2, ..., X_K}$; chacun de dimension $\boldsymbol{n \times p_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$. Tous ces tableaux portent sur les mêmes individus et sont supposés être centrés, réduits (si nécessaire) et normés de manière à avoir leur norme égale à 1. De manière plus précise, cela consiste à diviser chaque tableau de données $\boldsymbol{X_k}$ par sa norme $||\boldsymbol{X_k}|| = \sqrt{\boldsymbol{trace}(\boldsymbol{X_k^\top X_k})}$.

La figure 4.1 précise les relations qui existent entre les différentes composantes dans le cadre supervisé.

Pour chaque dimension, nous partons du tableau à prédire, $\boldsymbol{Y}$, auquel nous lui associons une variable latente (ou composante), $\boldsymbol{u}$: $\boldsymbol{u = Y\nu}$ avec $||\boldsymbol{\nu}|| = \boldsymbol{1}$. Cette variable latente est ensuite projetée dans l'espace formé par les variables de $\boldsymbol{X_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$, donnant ainsi les composantes partielles $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$. Enfin une synthèse des composantes partielles est faite pour donner la composante globale $\boldsymbol{t}$.

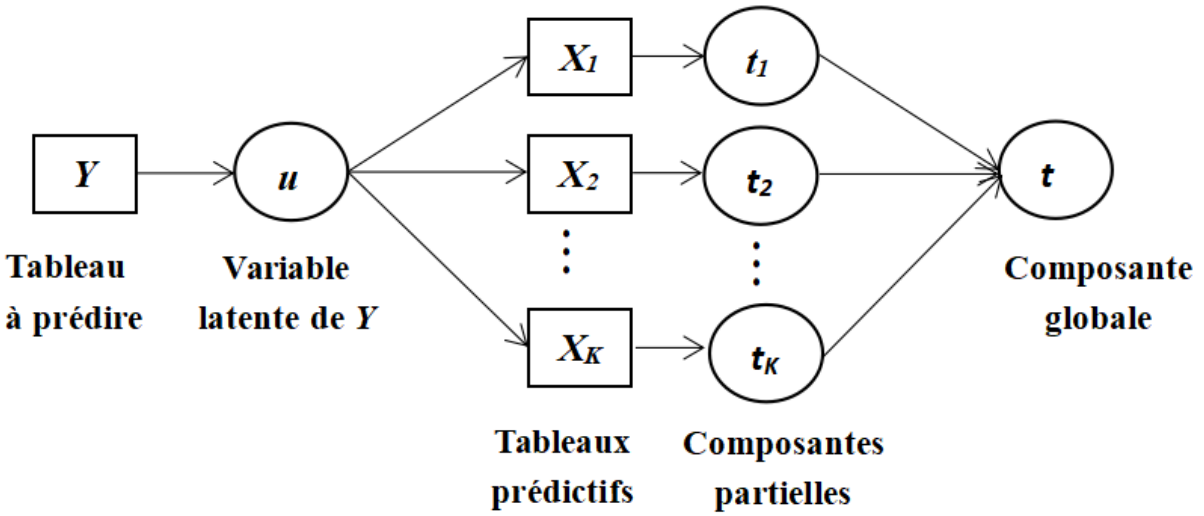


Fig. 4.1: Relation entre les composantes partielles et composantes globales dans le cas supervisé.

Multiblock Redundancy Analysis

Notons par $\boldsymbol{P_k} = \boldsymbol{X_k}(\boldsymbol{X_k^\top X_k})^{-1}\boldsymbol{X_k^\top}$, le projecteur orthogonal sur l'espace engendré par les variables du tableau $\boldsymbol{X_k}$ et par $\boldsymbol{W_k} = \boldsymbol{X_k X_k^\top}$, la matrice de produit scalaire entre les individus de $\boldsymbol{X_k}$.

Partant de la variable latente $\boldsymbol{u} = \boldsymbol{Y\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$ associée au tableau $\boldsymbol{Y}$, nous considérons sa projection orthogonale dans l'espace engendré par les variables de $\boldsymbol{X_k}$. Nous obtenons ainsi, $\boldsymbol{t_k} = \boldsymbol{P_k u}$. Cette variable définit la composante partielle associée au tableau $\boldsymbol{X_k}$ $(\boldsymbol{k} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{K})$. Par la suite, nous cherchons $\boldsymbol{u}$, (et, par conséquent, $\boldsymbol{t_k}$), de sorte à maximiser:

$$\sum_{k=1}^{K} \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t_k}) = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{u}^\top \sum_{k=1}^{K} \boldsymbol{t_k} = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \left(\sum_{k=1}^{K} \boldsymbol{P_k} \right) \boldsymbol{Y\nu}. \qquad (4.11)$$

L'idée sous-jacente à ce problème de maximisation est de chercher une variable latente $\boldsymbol{u}$ dans l'espace de $\boldsymbol{Y}$ qui soit la plus proche possible des variables $\boldsymbol{t_k}$ dans les espaces $\boldsymbol{X_k}$ $(\boldsymbol{k} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{K})$.

Ainsi, le vecteur $\boldsymbol{\nu}$, qui maximise cette forme quadratique est donné par le vecteur propre de $\boldsymbol{Y}^\top \left(\sum_{k=1}^{K} \boldsymbol{P_k}\right) \boldsymbol{Y}$ associé à la plus grande valeur propre.

Nous pouvons remarquer que le critère de maximisation pourrait encore s'écrire de la manière suivante:

$$\sum_{k=1}^{K} \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t_k}) = \boldsymbol{cov}\left(\boldsymbol{u}, \sum_{k=1}^{K} \boldsymbol{t_k}\right) = \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t}) \tag{4.12}$$

où $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{t_k}$ est la composante globale.

L'algorithme de type NIPALS qui permet de résoudre ce problème de maximisation est:

0. Choix de manière aléatoire du vecteur $\boldsymbol{\nu}$ et sa normalisation ($\boldsymbol{\nu} = \boldsymbol{\nu}/||\boldsymbol{\nu}||$);

1. Calcul de la composante associée à $\boldsymbol{Y}$: $\boldsymbol{u} = \boldsymbol{Y\nu}$ et celle associée à $\boldsymbol{X_k}$: $\boldsymbol{t_k} = \boldsymbol{P_k u}$;

2. Calcul de la composante globale: $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{t_k}$;

3. Mise à jour du vecteur $\boldsymbol{\nu}$: $\boldsymbol{\nu} = \boldsymbol{Y}^\top \boldsymbol{t}/||\boldsymbol{Y}^\top \boldsymbol{t}||$;

4. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence.

Notons que, la composante partielle $\boldsymbol{t_k} = \boldsymbol{X_k}(\boldsymbol{X_k}^\top \boldsymbol{X_k})^{-1} \boldsymbol{X_k}^\top \boldsymbol{u}$ est une combinaison linéaire des variables de $\boldsymbol{X_k}$: $\boldsymbol{t_k} = \boldsymbol{X_k w_k}$, avec $\boldsymbol{w_k} = (\boldsymbol{X_k}^\top \boldsymbol{X_k})^{-1} \boldsymbol{X_k}^\top \boldsymbol{u}$. Pour des raisons d'interprétation des résultats, ce vecteur pourrait être standardisé de sorte à avoir une longueur de 1, c'est-à-dire $\boldsymbol{t_k} = \boldsymbol{X_k w_k^*}$, avec $\boldsymbol{w_k^*} = \boldsymbol{w_k}/||\boldsymbol{w_k}||$. De la même manière, la composante

globale $\boldsymbol{t}$ peut être écrite comme $\boldsymbol{t = Xw}$, avec $\boldsymbol{X = [X_1|X_2|...|X_K]}$ et $\boldsymbol{w = (w_1, w_2, ..., w_K)^\top}$. A nouveau, pour des raisons d'interprétation des résultats, nous pouvons standardiser ce dernier vecteur ($\boldsymbol{w^* = w/||w||}$).

Les composantes globales et partielles d'ordre supérieur à 1 sont obtenues après la déflation de tous les tableaux par rapport aux composantes globales associées à $\boldsymbol{X}$ et qui sont déterminées de manière séquentielle. Ainsi, nous sommes amenés à déterminer les variables latentes associées à $\boldsymbol{Y}$: $\boldsymbol{u^{(1)}, u^{(2)}, ...}$, et, parallèlement, celles associées à $\boldsymbol{X_k}$: $\boldsymbol{t_k^{(1)},\ t_k^{(2)}, ...\ (k = 1, 2, ..., K)}$ ainsi que les composantes globales $\boldsymbol{t^{(1)},\ t^{(2)}}$, etc.

Tout comme dans le cas de la méthode RA, nous pouvons calculer à chaque dimension, $\boldsymbol{h}$, l'indice:

$$\begin{aligned} \boldsymbol{I_h} &= \frac{\sum_{k=1}^{K} \boldsymbol{u^{(h)\top} P_k u^{(h)}}}{\sum_{k=1}^{K} \boldsymbol{trace(Y^\top P_k Y)}} = \frac{\sum_{k=1}^{K} \boldsymbol{var(P_k u^{(h)})}}{\sum_{k=1}^{K} \sum_{j=1}^{q} \boldsymbol{var(P_k y_j)}} \\ &= \frac{\sum_{k=1}^{K} \boldsymbol{var(t_k^{(h)})}}{\sum_{k=1}^{K} \sum_{j=1}^{q} \boldsymbol{var(P_k y_j)}} \end{aligned} \tag{4.13}$$

Cet indice reflète la variation de $\boldsymbol{P_k Y\ (k = 1, 2, ..., K)}$ expliquée par $\boldsymbol{t^{(h)}}$. Ces indices peuvent ensuite être utilisés pour faire une représentation graphique similaire à un diagramme en éboulis pour choisir le nombre de composantes à retenir pour l'analyse.

A chaque dimension, $\boldsymbol{h}$, la quantité:

$$\boldsymbol{cont_k^{(h)}} = \frac{\boldsymbol{var(P_k u^{(h)})}}{\sum_{l=1}^{K} \boldsymbol{var(P_l u^{(h)})}} = \frac{\boldsymbol{var(t_k^{(h)})}}{\sum_{l=1}^{K} \boldsymbol{var(t_l^{(h)})}} \tag{4.14}$$

peut être calculée pour évaluer la contribution du tableau $\boldsymbol{X_k}$ à la détermination des composantes $\boldsymbol{u^{(h)}}$ et $\boldsymbol{t^{(h)}}$.

Multiblock Weighted Redundancy Analysis (MB-WRA)

Considérons toujours les mêmes notations que dans les sections précédentes. De même, nous effectuons le centrage et les mêmes pré-traitements des différents tableaux de données.

Nous considérons une variante du critère de maximisation précédent. Au lieu de maximiser la quantité $\sum_{k=1}^{K} \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{P_k u})$, nous proposons de chercher un vecteur $\boldsymbol{\nu}$ de longueur 1, qui maximise la quantité:

$$\sum_{k=1}^{K} \boldsymbol{cov}^{\mathbf{2}}(\boldsymbol{u}, \boldsymbol{P_k u}) = \frac{\mathbf{1}}{\boldsymbol{n}^{\mathbf{2}}} \sum_{k=1}^{K} (\boldsymbol{\nu}^{\top} \boldsymbol{Y}^{\top} \boldsymbol{P_k Y \nu})^{\mathbf{2}} \tag{4.15}$$

Il est clair que l'idée sous-jacente de ce problème est exactement la même que précédemment, à savoir, chercher une direction dans l'espace de $\boldsymbol{Y}$ qui soit la plus proche possible de celles des espaces de $\boldsymbol{X_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$.

Pour la résolution de ce critère de maximisation, nous utilisons la méthode de Lagrange. Ainsi, l'expression de Lagrange associée à ce critère est:

$$\sum_{k=1}^{K} (\boldsymbol{\nu}^{\top} \boldsymbol{Y}^{\top} \boldsymbol{P_k Y \nu})^{\mathbf{2}} - \mathbf{2}\boldsymbol{\mu}(\boldsymbol{\nu}^{\top} \boldsymbol{\nu} - \mathbf{1}) \tag{4.16}$$

où $-\mathbf{2}\boldsymbol{\mu}$ est le multiplicateur de Lagrange lié à la contrainte $||\boldsymbol{\nu}|| = \mathbf{1}$ ou de manière équivalente $||\boldsymbol{\nu}||^{\mathbf{2}} = \boldsymbol{\nu}^{\top} \boldsymbol{\nu} = \mathbf{1}$.

En dérivant cette expression par rapport à $\boldsymbol{\nu}$ et en annulant cette dérivée, nous obtenons:

$$\mathbf{4} \sum_{k=1}^{K} (\boldsymbol{\nu}^{\top} \boldsymbol{Y}^{\top} \boldsymbol{P_k Y \nu}) \boldsymbol{Y}^{\top} \boldsymbol{P_k Y \nu} - \mathbf{4}\boldsymbol{\mu}\boldsymbol{\nu} = \mathbf{0}. \tag{4.17}$$

En désignant par $\boldsymbol{\lambda_k}$ la quantité $\boldsymbol{\lambda_k = \nu^\top Y^\top P_k Y \nu}$, nous avons:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k Y^\top P_k Y \nu = \mu \nu}. \tag{4.18}$$

En multipliant les deux membres de l'égalité par $\boldsymbol{\nu^\top}$ et en utilisant la contrainte $\boldsymbol{\nu^\top \nu = 1}$, nous obtenons $\boldsymbol{\mu = \sum_{k=1}^{K} \lambda_k^2}$. De là, nous pouvons déduire le point stationnaire qui est donné par: $\boldsymbol{\nu = \frac{\sum_{k=1}^{K} \lambda_k Y^\top P_k Y \nu}{\sum_{k=1}^{K} \lambda_k^2}}$.
Cette résolution nous suggère l'algorithme itératif suivant:

0. Choisir aléatoirement le vecteur $\boldsymbol{\nu}$ et le standardiser $\boldsymbol{\nu = \nu / ||\nu||}$;

1. $\boldsymbol{\lambda_k = \nu^\top Y^\top P_k Y \nu}$;

2. $\boldsymbol{\nu = \sum_{k=1}^{K} \lambda_k Y^\top P_k Y \nu / \sum_{k=1}^{K} \lambda_k^2}$;

3. $\boldsymbol{\nu = \nu / ||\nu||}$;

4. Réitérer le même processus à partir de l'étape 1, jusqu'à convergence.

Nous pouvons montrer qu'à chaque itération, le critère que nous cherchons à maximiser croît et comme, par ailleurs, il est majoré, nous en déduisons que la suite générée par le critère au cours des itérations est convergente. Pour plus de détails sur la démonstration de la convergence, nous renvoyons à l'annexe en fin de chapitre.

Nous avons:

$$\begin{aligned} \sum_{k=1}^{K} \boldsymbol{cov^2(u, t_k)} &= \boldsymbol{n} \sum_{k=1}^{K} \boldsymbol{\lambda_k cov(u, t_k)} = \boldsymbol{n} \times \boldsymbol{cov}\left(\boldsymbol{u}, \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}\right) \\ &= \boldsymbol{n} \times \boldsymbol{cov(u, t)} \end{aligned} \tag{4.19}$$

La variable $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}$ représente la variable latente globale. C'est une combinaison linéaire des composantes partielles $\boldsymbol{t_k} = \boldsymbol{P_k u}$ associées aux tableaux $\boldsymbol{X_k}$. Plus précisément, puisque $\boldsymbol{\lambda_k} = \boldsymbol{u}^\top \boldsymbol{t_k} = \boldsymbol{n} \times \boldsymbol{cov(u, t_k)}$, il vient que $\boldsymbol{t}$ est proportionnelle à la première composante PLS de $\boldsymbol{u}$ sur $\boldsymbol{t_1, t_2, ..., t_K}$. Toutes ces remarques suggèrent un algorithme alternatif pour la résolution du problème de maximisation énoncé plus haut. En effet, ce critère peut également s'exprimer sous la forme:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(u, t_k)} = \boldsymbol{n} \sum_{k=1}^{K} \boldsymbol{\lambda_k cov(u, t_k)} = \boldsymbol{\nu}^\top \left(\sum_{k=1}^{K} \boldsymbol{\lambda_k Y}^\top \boldsymbol{P_k Y} \right) \boldsymbol{\nu}. \tag{4.20}$$

Ainsi, il vient que, pour des valeurs fixées de $\boldsymbol{\lambda_k}$, le vecteur optimal, $\boldsymbol{\nu}$, est donné par le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k Y}^\top \boldsymbol{P_k Y}$ associé à la plus grande valeur propre. De même, pour $\boldsymbol{\nu}$ fixé, $\boldsymbol{\lambda_k}$ est donné par: $\boldsymbol{\lambda_k} = \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \boldsymbol{P_k Y \nu}$. L'algorithme associé à cette résolution est le suivant:

0. Initialisation des poids $\boldsymbol{\lambda_k}$ (exemple, $\boldsymbol{\lambda_k = 1}$ pour $\boldsymbol{k = 1, 2, ..., K}$);

1. $\boldsymbol{\nu}$ est le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k Y}^\top \boldsymbol{P_k Y}$ associé à la plus grande valeur propre;

2. $\boldsymbol{\lambda_k = \nu}^\top \boldsymbol{Y}^\top \boldsymbol{P_k Y \nu}$;

3. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence. C'est-à-dire, jusqu'à ce que la différence entre les valeurs du critère de maximisation à deux itérations successives est inférieure à un seuil, $\boldsymbol{\epsilon}$, pré-défini par l'utilisateur (par exemple, $\boldsymbol{\epsilon = 10^{-8}}$).

Des composantes latentes d'ordre supérieur à 1 peuvent être obtenues

en poursuivant la même démarche après déflation de tous les tableaux par rapport aux composantes globales, $\boldsymbol{t}$, associées aux tableaux prédictifs.

Notons par $\boldsymbol{t^{(h)}}$, la composante globale pour la dimension $\boldsymbol{h}$ et par $\boldsymbol{t_1^{(h)}, t_2^{(h)}, ..., t_K^{(h)}}$, ses composantes partielles, pour la même dimension. De même, notons par $\boldsymbol{u^{(1)}, u^{(2)}, ..., u^{(h)}}$, les variables latentes successives de $\boldsymbol{Y}$. Enfin, notons par $\boldsymbol{\lambda_k^{(h)} = n \times cov(u^{(h)}, t_k^{(h)})}$, le poids du tableau $\boldsymbol{X_k}$, pour la dimension $\boldsymbol{h}$. Pour aider à l'interprétation des résultats, nous pouvons calculer les indices ci-après:

$$\boldsymbol{I_h = \frac{\sum_{k=1}^{K} \lambda_k^{(h)}}{\sum_{k=1}^{K} trace(Y^\top P_k Y)}} \tag{4.21}$$

pour déterminer l'importance de la composante globale $\boldsymbol{t^{(h)}}$ dans l'explication de la covariation entre $\boldsymbol{Y}$ et $\boldsymbol{X_1, X_2, ..., X_K}$;

$$\boldsymbol{cont_k^{(h)} = \frac{\lambda_k^{(h)}}{\sum_{l=1}^{K} \lambda_l^{(h)}}} \tag{4.22}$$

pour refléter la contribution du tableau $\boldsymbol{X_k}$ à la détermination des variables latentes $\boldsymbol{t^{(h)}}$ et $\boldsymbol{u^{(h)}}$.

Multiblock PLS regression

Afin de contourner le problème de colinéarité lié à l'inversion des matrices $\boldsymbol{X_k^\top X_k}$ dans l'expression de $\boldsymbol{P_k}$, nous proposons, tout comme dans le cas de la régression PLS2, de remplacer les matrices $\boldsymbol{(X_k^\top X_k)^{-1}}$ par la matrice identité et ainsi, de considérer les opérateurs $\boldsymbol{W_k = X_k X_k^\top}$, à la place des projecteurs $\boldsymbol{P_k = X_k (X_k^\top X_k)^{-1} X_k^\top}$.

Partant de la variable latente $\boldsymbol{u} = \boldsymbol{Y\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$ associée à $\boldsymbol{Y}$, nous définissons la composante partielle associée au tableau $\boldsymbol{X_k}$ par $\boldsymbol{t_k} = \boldsymbol{W_k u}$. Par la suite, nous cherchons $\boldsymbol{u}$ (et, par conséquent, $\boldsymbol{t_k}$) de sorte à maximiser le critère:

$$\begin{aligned}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t_k}) &= \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{u}^\top \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t_k} = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \left(\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{W_k} \right) \boldsymbol{Y\nu} \\ &= \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{\nu}^\top \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{Y}^\top \boldsymbol{X_k} \boldsymbol{X_k}^\top \boldsymbol{Y\nu} = \boldsymbol{n\nu}^\top \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{V_{Yk}} \boldsymbol{V_{kY}} \boldsymbol{\nu} \end{aligned} \quad (4.23)$$

où $\boldsymbol{V_{kY}} = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{X_k}^\top \boldsymbol{Y}$ est la matrice de covariance entre $\boldsymbol{X_k}$ et $\boldsymbol{Y}$ et $\boldsymbol{V_{Yk}} = \boldsymbol{V_{kY}}^\top$, la matrice de covariance entre $\boldsymbol{Y}$ et $\boldsymbol{X_k}$.

Ainsi, le vecteur optimal, $\boldsymbol{\nu}$, est donné par le vecteur propre de $\boldsymbol{V_{Yk}} \boldsymbol{V_{kY}}$ associé à la plus grande valeur propre. Par conséquent, nous sommes conduits à la régression MB-PLS.

Nous avons:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t_k}) = \boldsymbol{cov}\left(\boldsymbol{u}, \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t_k} \right) = \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t}) \quad (4.24)$$

avec $\boldsymbol{t} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t_k}$, la variable latente globale.
Un algorithme de type NIPALS qui permet de trouver la solution du critère de maximisation ci-dessus est le suivant:

0. Choix de manière aléatoire de $\boldsymbol{\nu}$, avec $||\boldsymbol{\nu}|| = \mathbf{1}$;

1. $\boldsymbol{u} = \boldsymbol{Y\nu}$: variable latente associée à $\boldsymbol{Y}$;

2. $\boldsymbol{t_k} = \boldsymbol{X_k} \boldsymbol{X_k}^\top \boldsymbol{u}$: variable latente associée à $\boldsymbol{X_k}$;

3. $\boldsymbol{t = \sum_{k=1}^{K} t_k}$: variable latente globale;

4. $\boldsymbol{\nu = Y^\top t / ||Y^\top t||}$;

5. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence.

Il est clair qu'à chaque étape de l'algorithme, le critère de maximisation croît, et comme par ailleurs, il est majoré, nous déduisons que l'algorithme converge.

Nous pouvons aussi noter que:

$$\sum_{k=1}^{K} \boldsymbol{cov(u, t_k)} = \frac{1}{n} \boldsymbol{\nu^\top Y^\top} \sum_{k=1}^{K} \boldsymbol{X_k X_k^\top Y \nu} = \frac{1}{n} \boldsymbol{\nu^\top Y^\top X X^\top Y \nu} \quad (4.25)$$

avec $\boldsymbol{X = [X_1|X_2|...|X_K]}$, le tableau obtenu en concaténant horizontalement les tableaux $\boldsymbol{X_1, X_2, ..., X_K}$. Cela signifie que le vecteur optimal $\boldsymbol{\nu}$ est le vecteur propre de la matrice $\boldsymbol{Y^\top X X^\top Y}$ associé à la plus grande valeur propre. En d'autres termes, la solution obtenue à partir de la régression MB-PLS est la même que celle obtenue à partir de la régression PLS2 de $\boldsymbol{Y}$ sur $\boldsymbol{X}$ [3].

Des variables latentes d'ordre supérieur à 1 peuvent être déterminées après déflation comme exposé dans les sections précédentes. En désignant, comme précédemment, par $\boldsymbol{u^{(h)}}$, $\boldsymbol{t^{(h)}}$ et $\boldsymbol{t_k^{(h)}}$, les variables latentes associées respectivement à $\boldsymbol{Y}$, $\boldsymbol{X}$ et $\boldsymbol{X_k}$ ($\boldsymbol{h = 1, 2, ..., H}$ et $\boldsymbol{k = 1, 2, ..., K}$), des indices pourraient être calculés pour mieux interpréter les résultats. Nous avons, d'une part:

$$\boldsymbol{I_h} = \frac{\boldsymbol{\lambda^{(h)}}}{\sum_{k=1}^{K} \boldsymbol{trace(Y^\top X_k X_k^\top Y)}} = \frac{\boldsymbol{\lambda^{(h)}}}{\boldsymbol{trace(Y^\top X X^\top Y)}} \quad (4.26)$$

avec $\boldsymbol{\lambda^{(h)} = u^{(h)\top} t^{(h)} = n \times cov(u^{(h)},\ t^{(h)})}$.

$\boldsymbol{I_h}$ reflète la proportion de covariation entre $\boldsymbol{Y}$ et $\boldsymbol{X_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ expliquée par la composante globale $\boldsymbol{t^{(h)}}$. D'autre part, nous avons:

$$\boldsymbol{cont_k^{(h)} = \frac{\lambda_k^{(h)}}{\sum_{l=1}^{K} \lambda_l^{(h)}}} \tag{4.27}$$

qui reflète la contribution du tableau $\boldsymbol{X_k}$ pour la détermination de la composante globale $\boldsymbol{t^{(h)}}$.

Multiblock Weighted Covariate analysis (MB-WCov)

Nous poursuivons la même démarche en adoptant les mêmes notations que la section précédente. Au lieu du critère de maximisation ayant permis d'introduire la régression MB-PLS, à savoir, $\boldsymbol{\sum_{k=1}^{K} cov(u, t_k)}$, nous considérons le critère $\boldsymbol{\sum_{k=1}^{K} cov^2(u, t_k)}$. L'idée sous-jacente de ces deux critères est d'explorer la covariation entre, d'une part, $\boldsymbol{Y}$ et, d'autre part, les tableaux $\boldsymbol{X_k}$. Ce dernier critère de maximisation introduit une nouvelle stratégie d'analyse des données que nous désignons par Multiblock Weighted Covariate analysis (MB-WCov). Son intérêt est de donner explicitement des poids aux différents tableaux $\boldsymbol{X_k}$ qui reflètent leurs contributions à la détermination de la composante globale.

En partant de la variable latente $\boldsymbol{u = Y\nu}$ $\boldsymbol{(||\nu|| = 1)}$ associée à $\boldsymbol{Y}$, nous déterminons les variables $\boldsymbol{t_k = W_k u}$, qui constituent les composantes partielles associées aux tableaux $\boldsymbol{X_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$. Par la suite, nous

cherchons un vecteur $\boldsymbol{u}$ (et, par conséquent, $\boldsymbol{t_k}$) qui maximise le critère:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(u, t_k)} = \frac{\mathbf{1}}{\boldsymbol{n^2}} \sum_{k=1}^{K} (\boldsymbol{\nu^\top Y^\top W_k Y \nu})^{\mathbf{2}} \tag{4.28}$$

Soit l'expression de Lagrange de ce problème de maximisation:

$$\sum_{k=1}^{K} (\boldsymbol{\nu^\top Y^\top W_k Y \nu})^{\mathbf{2}} - \mathbf{2}\boldsymbol{\mu}(\boldsymbol{\nu^\top \nu} - \mathbf{1}) \tag{4.29}$$

avec $-\mathbf{2}\boldsymbol{\mu}$, le multiplicateur de Lagrange associé à la contrainte $||\boldsymbol{\nu}|| = \mathbf{1}$. En dérivant cette expression par rapport à $\boldsymbol{\nu}$ et en annulant cette dérivée, nous obtenons $\mathbf{4}\sum_{k=1}^{K}(\boldsymbol{\nu^\top Y^\top W_k Y \nu})\boldsymbol{Y^\top W_k Y \nu} - \mathbf{4}\boldsymbol{\mu\nu} = \mathbf{0}$. En notant par $\boldsymbol{\lambda_k}$, la quantité $\boldsymbol{\lambda_k} = \boldsymbol{\nu^\top Y^\top W_k Y \nu}$, nous avons $\sum_{k=1}^{K} \boldsymbol{\lambda_k Y^\top W_k Y \nu} = \boldsymbol{\mu\nu}$. En multipliant les deux membres de cette dernière égalité par $\boldsymbol{\nu^\top}$ et en appliquant la contrainte $\boldsymbol{\nu^\top \nu} = \mathbf{1}$, nous obtenons $\boldsymbol{\mu} = \sum_{k=1}^{K} \boldsymbol{\lambda_k^2}$. Ainsi, le point stationnaire est donné par $\boldsymbol{\nu} = \frac{\sum_{k=1}^{K} \boldsymbol{\lambda_k Y^\top W_k Y \nu}}{\sum_{k=1}^{K} \boldsymbol{\lambda_k^2}}$. L'algorithme itératif permettant de résoudre le problème de maximisation ci-dessus est le suivant:

0. Choix de manière aléatoire du vecteur $\boldsymbol{\nu}$ ($\boldsymbol{\nu} = \boldsymbol{\nu}/||\boldsymbol{\nu}||$);

1. $\boldsymbol{\lambda_k} = \boldsymbol{\nu^\top Y^\top W_k Y \nu}$;

2. $\boldsymbol{\nu} = \sum_{k=1}^{K} \boldsymbol{\lambda_k Y^\top W_k Y \nu} / \sum_{k=1}^{K} \boldsymbol{\lambda_k^2}$;

3. $\boldsymbol{\nu} = \boldsymbol{\nu}/||\boldsymbol{\nu}||$;

4. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence.

La convergence de cet algorithme peut être démontrée en poursuivant des développements très similaires à ceux de la méthode MB-WRA donnés

en annexe. Il suffit de remplacer les matrices $\boldsymbol{P_k} = \boldsymbol{X_k}(\boldsymbol{X_k}^\top \boldsymbol{X_k})^{-1}\boldsymbol{X_k}^\top$ par $\boldsymbol{W_k} = \boldsymbol{X_k}\boldsymbol{X_k}^\top$.

Le critère $\sum_{k=1}^{K} \boldsymbol{cov^2(u, t_k)}$ peut s'écrire:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k cov(u, t_k)} = \boldsymbol{cov}\left(\boldsymbol{u}, \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}\right). \tag{4.30}$$

Ceci permet de définir la variable latente globale $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}$. Nous pouvons remarquer que $\boldsymbol{t}$ est proportionnelle à la première composante PLS de $\boldsymbol{u}$ sur $\boldsymbol{t_1, t_2, ..., t_K}$. De là, nous pouvons en déduire un algorithme alternatif. En effet, pour les valeurs de $\boldsymbol{\lambda_k}$ fixées, le vecteur optimal, $\boldsymbol{\nu}$, est donné par le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k Y}^\top \boldsymbol{W_k Y}$ associé à la plus grande valeur propre. Inversement, pour $\boldsymbol{\nu}$ fixé, $\boldsymbol{\lambda_k}$ est donné par: $\boldsymbol{\lambda_k} = \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \boldsymbol{W_k Y \nu}$. En résumé, l'algorithme alternatif pour la résolution du critère de maximisation de MB-WCov est le suivant:

0. Initialisation des poids $\boldsymbol{\lambda_k}$ (exemple, $\boldsymbol{\lambda_k = 1}$ pour $\boldsymbol{k = 1, 2, ..., K}$);

1. $\boldsymbol{\nu}$ est le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k Y}^\top \boldsymbol{W_k Y}$ associé à la plus grande valeur propre;

2. Mise à jour des poids: $\boldsymbol{\lambda_k} = \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \boldsymbol{W_k Y \nu}$;

3. Réitération de la procédure à partir de l'éatpe 1, jusqu'à convergence.

Les variables latentes d'ordre supérieur à 1, peuvent être déterminées après une déflation de tous les tableaux par rapport aux composantes globales précédentes.

Les mêmes indices que ceux de la régression MB-PLS peuvent être calculés pour une meilleur interprétation des résultats d'analyse.

4.2.3 Comparaison des méthodes supervisées

Le tableau 4.1 présente un aperçu général des différentes méthodes supervisées étudiées.

Tab. 4.1: Aperçu général des méthodes supervisées d'analyse des données multiblocs.

$$\boldsymbol{u = Y\nu}\ (||\boldsymbol{\nu}|| = \mathbf{1});\ \boldsymbol{P_k = X_k(X_k^\top X_k)^{-1}X_k^\top};\ \boldsymbol{W_k = X_kX_k^\top}$$

Méthode	Composante partielle	Critère de maximisation	Composante globale	Premier algorithme	Algorithme NIPALS
MB-RA	$\boldsymbol{t_k = P_ku}$	$\sum_{k=1}^{K} \boldsymbol{cov(u,t_k)}$	$\boldsymbol{t = \sum_{k=1}^{K} t_k}$	$\boldsymbol{\nu}$- 1er vecteur propre de $\boldsymbol{Y^\top\left(\sum_{k=1}^{K} P_k\right)Y}$.	0. Initialiser $\boldsymbol{\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$; 1. $\boldsymbol{u = Y\nu}$; $\boldsymbol{t_k = P_ku}$; 2. $\boldsymbol{t = \sum_{k=1}^{K} t_k}$; 3. $\boldsymbol{\nu = Y^\top t/||Y^\top t||}$; 4. Réitérer à partir de 1.
MB-WRA	$\boldsymbol{t_k = P_ku}$	$\sum_{k=1}^{K} \boldsymbol{cov^2(u,t_k)}$	$\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$ avec $\boldsymbol{\lambda_k = u^\top t_k}$	0. $\boldsymbol{\lambda_k = 1}$; 1. $\boldsymbol{\nu}$- 1er vecteur propre de $\boldsymbol{Y^\top\left(\sum_{k=1}^{K} \lambda_k P_k\right)Y}$; 2. $\boldsymbol{\lambda_k = \nu^\top Y^\top P_k Y\nu}$; 3. Réitérer à partir de 1.	0.Initialiser $\boldsymbol{\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$; 1. $\boldsymbol{u = Y\nu}$; $\boldsymbol{t_k = P_ku}$; 2. $\boldsymbol{\lambda_k = u^\top t_k}$; 3. $\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$; 4. $\boldsymbol{\nu = Y^\top t/||Y^\top t||}$; 5. Réitérer à partir de 1.
MB-PLS	$\boldsymbol{t_k = W_ku}$	$\sum_{k=1}^{K} \boldsymbol{cov(u,t_k)}$	$\boldsymbol{t = \sum_{k=1}^{K} t_k}$	$\boldsymbol{\nu}$- 1er vecteur propre de $\boldsymbol{Y^\top\left(\sum_{k=1}^{K} W_k\right)Y}$.	0.Initialiser $\boldsymbol{\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$; 1. $\boldsymbol{u = Y\nu}$ $\boldsymbol{t_k = W_ku}$; 2. $\boldsymbol{t = \sum_{k=1}^{K} t_k}$; 3. $\boldsymbol{\nu = Y^\top t/||Y^\top t||}$; 4. Réitérer à partir de 1.
MB-WCov	$\boldsymbol{t_k = W_ku}$	$\sum_{k=1}^{K} \boldsymbol{cov^2(u,t_k)}$	$\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$ avec $\boldsymbol{\lambda_k = u^\top t_k}$	0. $\boldsymbol{\lambda_k = 1}$; 1. $\boldsymbol{\nu}$- 1er vecteur propre de $\boldsymbol{Y^\top\left(\sum_{k=1}^{K} \lambda_k W_k\right)Y}$; 2. $\boldsymbol{\lambda_k = \nu^\top Y^\top W_k Y\nu}$; 3. Réitérer à partir de 1.	0. Initialiser $\boldsymbol{\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$; 1. $\boldsymbol{u = Y\nu}$; $\boldsymbol{t_k = W_ku}$; 2. $\boldsymbol{\lambda_k = u^\top t_k}$; 3. $\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$; 4. $\boldsymbol{\nu = Y^\top t/||Y^\top t||}$; 5. Réitérer à partir de 1.

Il est clair que les quatre méthodes supervisées d'analyse des données

multiblocs présentées dans ce chapitre se différencient selon deux clefs. La première clef concerne la relation entre la composante $\boldsymbol{u}$, dans l'espace de $\boldsymbol{Y}$, et sa "projection" dans chacun des espaces engendrés par les variables de $\boldsymbol{X_k}$, c'est-à-dire, $\boldsymbol{t_k}$. De manière précise, deux options s'offrent à nous: (i) $\boldsymbol{t_k} = \boldsymbol{X_k}(\boldsymbol{X_k}^\top \boldsymbol{X_k})^{-1}\boldsymbol{X_k}^\top \boldsymbol{u}$ qui représente vraiment la projection de $\boldsymbol{u}$ sur l'espace engendré par les variables de $\boldsymbol{X_k}$, (ii) $\boldsymbol{t_k} = \boldsymbol{X_k}\boldsymbol{X_k}^\top \boldsymbol{u}$ qui représente l'avantage de contourner les problèmes de colinéarité qui peuvent survenir du fait de l'inversion des matrices $\boldsymbol{X_k}^\top \boldsymbol{X_k}$. Il est clair que l'option (i) oriente vers les méthodes d'analyse qui s'apparente à la redundancy analysis (c'est-à-dire, MB-RA et MB-WRA), alors que l'option (ii) oriente vers des méthodes qui s'apparentent à la régression PLS (c'est-à-dire, MB-PLS et MB-WCov).

La deuxième clef de différentiation des méthodes est la relation entre la composante globale, $\boldsymbol{t}$, et les composantes partielles, $\boldsymbol{t_k}$. Cette relation découle directement des critères de maximisation pour la détermination des variables latentes. Là encore, deux options s'offrent à nous. (i) Pour le critère basé sur la somme des covariances entre $\boldsymbol{t_k}$ et $\boldsymbol{u}$, nous avons $\boldsymbol{t}$ qui est proportionnelle à la somme des composantes partielles: $\boldsymbol{t} \propto \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{t_k}$, (ii) Pour le critère basé sur la somme des carrés de la covariance entre $\boldsymbol{t_k}$ et $\boldsymbol{u}$, nous avons $\boldsymbol{t}$ qui est égale à une combinaison linéaire des composantes partielles: $\boldsymbol{t} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k t_k}$ (avec $\boldsymbol{\lambda_k} = \boldsymbol{u}^\top \boldsymbol{t_k}$).

4.3 Illustrations

Les méthodes supervisées évoquées dans ce chapitre sont illustrées et comparées à partir d'une étude de simulation et de données réelles.

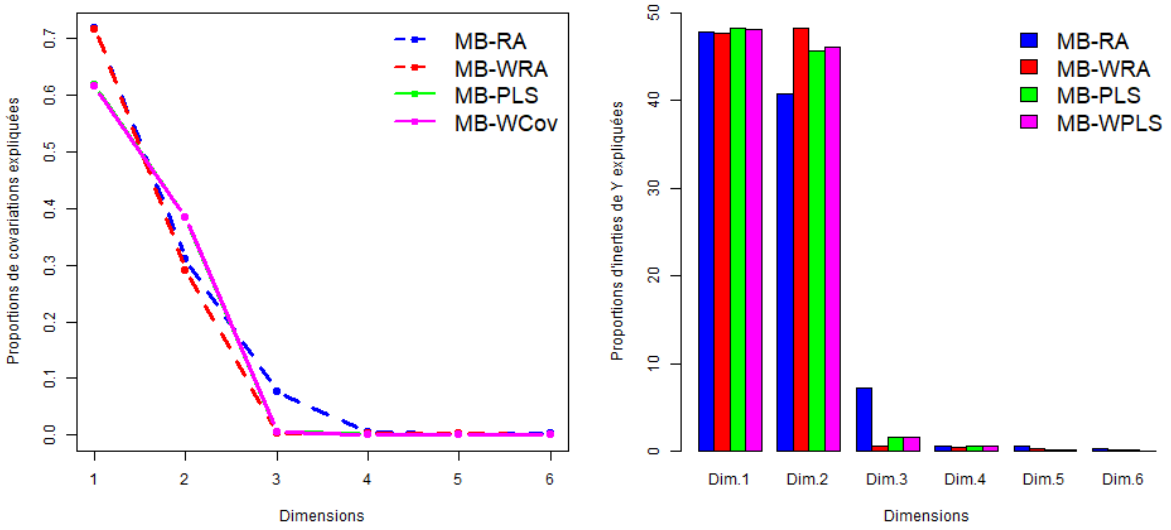
4.3.1 Étude de simulation

L'étude de simulation est, dans une large mesure, similaire à celle proposée par Westerhuis et al. [3]. Nous considérons deux variables orthogonales $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$, quatre tableaux prédictifs $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ et un tableau réponse (ou à prédire), $\boldsymbol{Y}$. Tous ces tableaux portent sur cinquante observations et sont définis comme suit: $\boldsymbol{X_1} = [\boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}]$, $\boldsymbol{X_2} = \boldsymbol{X_3} = \boldsymbol{X_4} = [\boldsymbol{d_2}, \boldsymbol{randn(4)}]$ et $\boldsymbol{Y} = [\boldsymbol{d_1}, \boldsymbol{d_2}]$, où randn(4) désigne quatre variables aléatoires normalement distribuées. Dans chaque tableau, nous rajoutons 20% de bruit aux variables $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$. Après avoir pré-traité ces tableaux, nous leur appliquons les méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov pour prédire $\boldsymbol{Y}$ à partir de $\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}$ et $\boldsymbol{X_4}$.

La figure 4.2a présente les proportions de covariations entre les tableaux prédictifs, $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$, et le tableau réponse, $\boldsymbol{Y}$, expliquées par les six premières variables latentes obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov. Comme nous l'avions indiqué plus haut, cette figure peut être interprétée de la même manière qu'un diagramme en éboulis et permet de déterminer le nombre de variables latentes à retenir pour l'analyse. De cette figure, nous pouvons noter que toutes les courbes décroissent entre la deuxième et la troisième composante et se stabilisent à partir de la troisième composante, exceptée pour la méthode MB-RA pour

laquelle la stabilisation est faite à partir de la quatrième composante. Cela suggère de retenir les trois premières composantes.

La figure 4.2b présente les proportions d'inerties de $\boldsymbol{Y}$ expliquées par les six premières composantes obtenues à partir des différentes méthodes d'analyse. De cette figure, nous pouvons tirer les mêmes conclusions que pour la figure 4.2a.



(a) Covariation *(b)* Inertie totale

Fig. 4.2: Données simulées: (a) Proportions de covariations entre le tableau réponse et les tableaux prédictifs, expliquées par les six premières composantes obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov et (b) Proportions d'inerties de $\mathbf{Y}$ expliquées par les six premières composantes obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov.

Le tableau 4.2 donne les corrélations entre les composantes globales $\boldsymbol{t}^{(1)}$ et $\boldsymbol{t}^{(2)}$ avec les variables $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$. Il donne également les contributions des tableaux $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ à la détermination de ces composantes globales.

La première composante, $\boldsymbol{t}^{(1)}$, obtenue à partir de MB-RA ou MB-WRA est très fortement corrélée à la variable $\boldsymbol{d_2}$ et la seconde composante, $\boldsymbol{t}^{(2)}$,

est très fortement corrélée à la variable $\boldsymbol{d_1}$. Quant à la régression MB-PLS et MB-WCov, ces composantes présentent une tendance totalement inverse, en ce sens que leurs premières composantes respectives sont très fortement corrélées à la variable $\boldsymbol{d_1}$ et leurs deuxièmes composantes respectives, à la variable $\boldsymbol{d_2}$. Ceci s'explique par le fait que puisque les méthodes MB-RA et MB-WRA ne tiennent pas compte de la variation (c'est-à-dire, variances et corrélations) intra-tableaux, c'est la variable $\boldsymbol{d_2}$ qui est favorisée parce qu'elle apparaît dans $\boldsymbol{Y}$, d'une part, et, $\boldsymbol{X_2, X_3}$ et $\boldsymbol{X_4}$, d'autre part. Elle est donc le trait commun à tous les tableaux, sauf $\boldsymbol{X_1}$. Par contre, la variable $\boldsymbol{d_1}$ apparaît dans le tableau $\boldsymbol{X_1}$ cinq fois. Mais étant donné que MB-RA et MB-WRA ne tiennent pas compte de la variation intra-tableaux, elle ne sera comptabilisée qu'une seule fois pour ces deux méthodes. En revanche, elle sera privilégiée par la régression MB-PLS et MB-WCov parce que, ces deux méthodes tiennent compte non seulement des relations du tableau $\boldsymbol{Y}$ avec les tableaux $\boldsymbol{X_k}$ mais également des variations au sein des tableaux prédictifs $\boldsymbol{X_k}$.

Pour ce qui est des contributions des tableaux $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ à la détermination de la composante globale $\boldsymbol{t^{(1)}}$ (tableau 4.2), nous pouvons noter, de manière logique, que cette composante est déterminée plus ou moins équitablement par les tableaux $\boldsymbol{X_2, X_3}$ et $\boldsymbol{X_4}$ pour les méthodes MB-RA et MB-WRA. Avec les méthodes MB-PLS et MB-WCov, on note que la première composante est essentiellement déterminée par le tableau $\boldsymbol{X_1}$. Des conclusions similaires peuvent être tirées pour la deuxième composante $\boldsymbol{t^{(2)}}$. Toutes ces conclusions corroborent bien le principe de chaque méthode, à savoir: MB-RA et MB-WRA ne tiennent pas compte de la variation intra-tableaux,

alors que MB-PLS et MB-WCov tiennent compte aussi bien de la variation intra-tableaux que la variation inter-tableaux (c'est-à-dire, la relation entre chaque tableau prédictif et le tableau réponse).

Tab. 4.2: Données simulées: corrélations entre les composantes globales $\boldsymbol{t^{(1)}}$, $\boldsymbol{t^{(2)}}$ et les variables $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$ et les contributions des différents tableaux à la détermination des composantes globales $\boldsymbol{t^{(1)}}$ et $\boldsymbol{t^{(2)}}$.

		Corrélations		Contributions			
		$\boldsymbol{d_1}$	$\boldsymbol{d_2}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$
MB-RA	$\boldsymbol{t^{(1)}}$	0.04	**-0.99**	0.02	**0.32**	**0.33**	**0.33**
	$\boldsymbol{t^{(2)}}$	**-0.90**	-0.06	**0.78**	0.10	0.04	0.08
MB-WRA	$\boldsymbol{t^{(1)}}$	-0.01	**0.99**	0.02	**0.32**	**0.33**	**0.33**
	$\boldsymbol{t^{(2)}}$	**0.99**	0.01	**0.84**	0.06	0.02	0.08
MB-PLS	$\boldsymbol{t^{(1)}}$	**-0.99**	0.09	**0.96**	0.01	0.01	0.02
	$\boldsymbol{t^{(2)}}$	0.09	**0.98**	0.00	**0.34**	**0.33**	**0.33**
MB-WCov	$\boldsymbol{t^{(1)}}$	**-1.00**	0.01	**0.97**	0.01	0	0.02
	$\boldsymbol{t^{(2)}}$	-0.01	**-0.98**	0	**0.34**	**0.33**	**0.33**

Afin d'étudier la performance des quatre méthodes supervisées en termes de prédiction, nous avons subdivisé les données en deux parties. Le premier jeu de données concerne les données d'étalonnage (ou d'apprentissage) avec trente observations et le second jeu, les données de validation, avec vingt observations. Les données d'étalonnage ont servi à établir les modèles de prédiction et les données de validation ont servi à valider ces modèles. Ainsi, nous avons confronté les prédictions faites aux valeurs observées du tableau réponse $\boldsymbol{Y}$ à l'aide de l'indicateur RMSEP (Root Mean Squared

Errors of Prediction), c'est-à-dire la racine carrée de la moyenne des erreurs de prédictions pour chaque méthode et pour chaque composante introduit dans le modèle [25]. La figure 4.3 présente l'évolution de ces erreurs en fonction du nombre de composantes introduits dans le modèle. De cette figure, on note que les quatre méthodes présentent plus ou moins la même allure. De manière plus précise, on note une décroissance nette du RMSEP lorsqu'on introduit la deuxième composante dans le modèle. Puis, une décroissance légère lorsqu'on introduit la troisième composante. Par suite, les différentes courbes tendent à se stabiliser ou à croître légèrement. La plus petite valeur du RMSEP est obtenue avec la méthode MB-WRA sur la troisième composante.

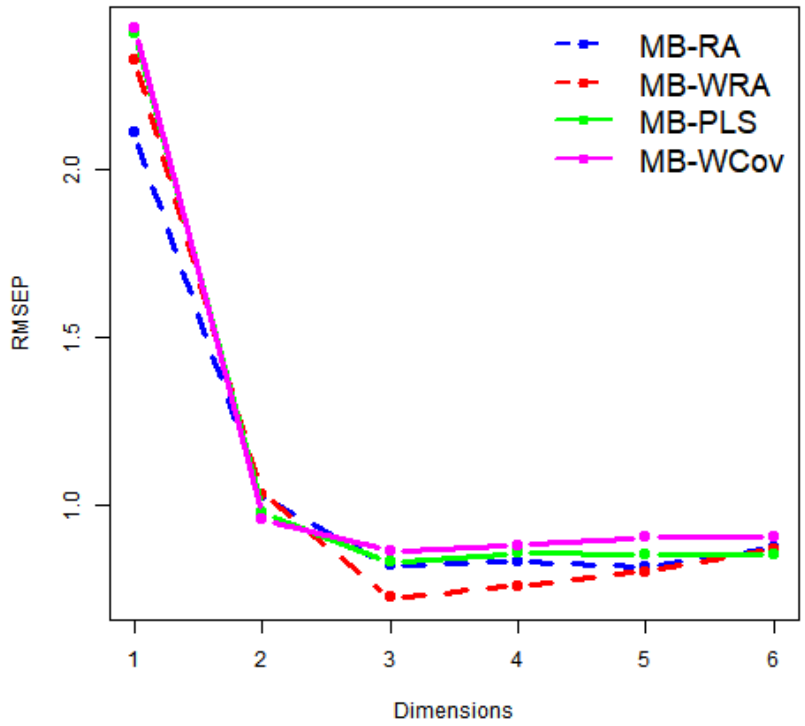


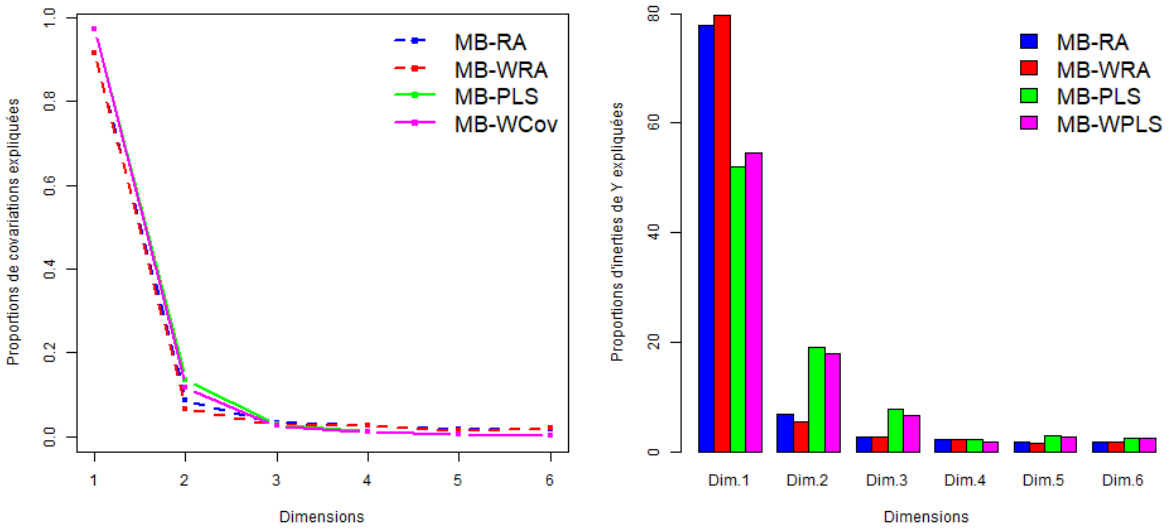
Fig. 4.3: Données simulées: RMSEP obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov en fonction du nombre de composantes introduits dans le modèle.

4.3.2 *Données réelles: données de "pommes de terre"*

Les données utilisées pour illustrer et comparer les quatre méthodes supervisées sont les données de pommes de terre qui sont davantage détaillées dans Thybo et al. [30]. L'objectif est de prédire des attributs sensoriels, $\boldsymbol{Y}$ (9 variables) concernant vingt-six variétés de pommes de terre, à partir des mesures chimiques, $\boldsymbol{X_1}$ (14 variables), et des mesures de compression uni-axiale, $\boldsymbol{X_2}$ (6 variables). Après avoir pré-traité ces tableaux, nous les avons soumis aux méthodes MB-RA, MB-PLS, MB-WRA et MB-WCov.

La figure 4.4a présente les proportions de covariations entre les tableaux prédictifs et le tableau réponse, expliquées par les six premières variables latentes obtenues à partir des différentes méthodes. Il apparaît que, toutes les courbes décroissent et se stabilisent à partir de la troisième composante. Ainsi, l'analyse suggère de retenir les trois premières composantes.
Pour la première composante, MB-RA et sa variante MB-WRA sont les méthodes qui expliquent la plus grande proportion de covariation entre le tableau réponse, $\boldsymbol{Y}$ et les tableaux prédictifs $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$. Quant à la régression MB-PLS et MB-WCov, elles présentent plus ou moins le même comportement.

Les proportions d'inerties du tableau réponse, $\boldsymbol{Y}$, expliquées par les six premières variables latentes obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov sont présentées dans la figure 4.4b. Tout comme dans le cas de la figure 4.4a, nous notons une décroissance de la proportion d'inertie expliquée de $\boldsymbol{Y}$, avec une stabilisation à partir de la quatrième composante.



(a) Covariations *(b)* Inerties

Fig. 4.4: Données de pommes de terre: (a) Proportions de covariations entre les tableaux prédictifs et le tableau réponse, expliquées par les six premières variables latentes obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov et (b) Proportions d'inerties du tableau $\mathbf{Y}$, expliquées par les six premières variables latentes des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov.

Les contributions de chaque tableau prédictif pour la détermination des variables latentes $\boldsymbol{t}^{(1)}$ et $\boldsymbol{t}^{(2)}$ obtenues à partir des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov sont présentées dans le tableau 4.3. Globalement, on note que pour toutes les méthodes, le tableau relatif aux données chimiques ($\boldsymbol{X_1}$) contribue le plus à la détermination des deux premières variables latentes.

Tab. 4.3: Données de pommes de terre: contributions des tableaux $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$ à la détermination des variables latentes $\boldsymbol{t^{(1)}}$ et $\boldsymbol{t^{(2)}}$.

	MB-RA		MB-WRA		MB-PLS		MB-WCov	
	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
$\boldsymbol{t^{(1)}}$	**0.59**	0.41	**0.59**	0.41	**0.55**	0.45	**0.56**	0.44
$\boldsymbol{t^{(2)}}$	**0.55**	0.45	**0.56**	0.44	**0.67**	0.33	**0.67**	0.33

La figure 4.5 présente les valeurs du critère RMSECV (Root Mean Squared Errors of Cross-Validation) obtenues par la technique de cross-validation leave-one-out (LOO) en fonction du nombre de variables latentes des différentes méthodes. Il apparaît que les valeurs de RMSECV associées à la régression MB-PLS et MB-WCov décroissent en fonction du nombre de variables latentes introduits dans le modèle de prédiction et se stabilisent à partir de la troisième variable latente. De plus, ces deux méthodes d'analyse semblent avoir la même performance. Quant aux méthodes MB-RA et MB-WRA, les conclusions qu'on peut tirer sont très différentes. En effet, les valeurs de RMSECV croissent en fonction du nombre de variables latentes introduits dans le modèle de prédiction. Avec un modèle basé uniquement sur la première variable latente, les méthodes MB-RA et MB-WRA présentent des valeurs du RMSECV plus petites que celles obtenues à partir des méthodes MB-PLS et MB-WCov. Cependant, lorsqu'on introduit de nouvelles variables latentes dans le modèle, la performance en termes de prédiction de MB-RA et MB-WRA se détériore de manière alarmante. Ce dernier résultat est typique des méthodes qui sont vulnérables au problème de multicolinéarité.

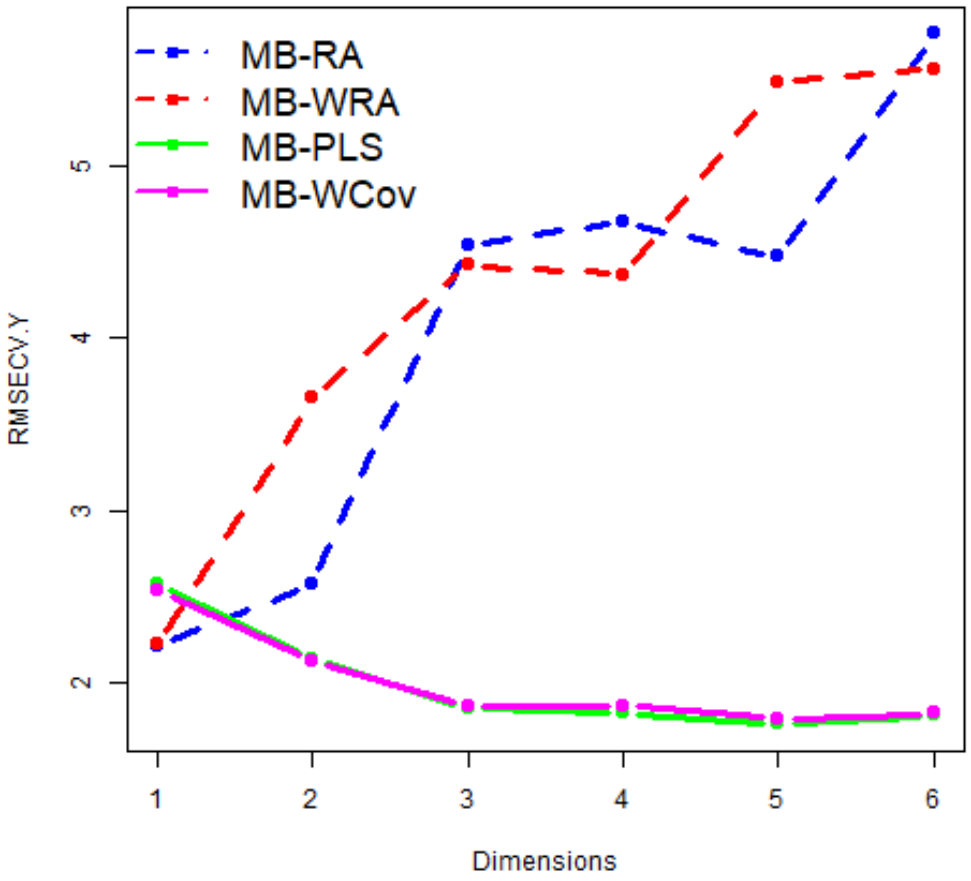


Fig. 4.5: Données de pommes de terre: RMSECV obtenues à partir de la procédure de cross-validation LOO pour les six premières variables latentes des méthodes MB-RA, MB-WRA, MB-PLS et MB-WCov.

4.4 Discussion et conclusion

Dans ce chapitre, nous avons proposé une approche unifiée pour deux méthodes supervisées, notamment, MB-RA et la régression MB-PLS. Nous avons aussi proposé deux nouvelles approches pour prédire un tableau réponse à partir de plusieurs tableaux prédictifs. Les quatre méthodes d'analyse sont issues des critères de maximisation qui sont clairs. Elles peuvent se distinguer selon deux clefs. La première clef est la relation entre

la variable latente $\boldsymbol{u}$ de $\boldsymbol{Y}$ et les variables latentes partielles $\boldsymbol{t_k}$ de $\boldsymbol{X_k}$. Nous avons considéré ici deux types de relations: (i) $\boldsymbol{t_k = X_k(X_k^\top X_k)^{-1} X_k^\top u}$ et (ii) $\boldsymbol{t_k = X_k X_k^\top u}$. En choisissant la relation (i), les variances des variables et les corrélations des variables de $\boldsymbol{X_k}$ sont masquées. Par conséquent, les méthodes d'analyse concernées par cette relation (MB-RA et MB-WRA) cherchent uniquement à restituer l'inertie de $\boldsymbol{Y}$. Cependant, ces méthodes sont très sensibles au problème de quasi-colinéarité entre les variables prédictives. Avec la relation (ii), nous retrouvons les méthodes MB-PLS et MB-WCov, qui ne sont pas vulnérables au problème de quasi-colinéarité. Ces méthodes restituent aussi bien la variabilité intra-tableaux que la variabilité inter-tableaux (c'est-à-dire, la relation entre $\boldsymbol{Y}$ et $\boldsymbol{X_k}$, $\boldsymbol{k = 1, 2, ..., K}$).

La deuxième clef permettant de distinguer les quatre méthodes supervisées est la relation que nous postulons entre la composante globale $\boldsymbol{t}$ et ses composantes partielles $\boldsymbol{t_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$. Nous avons également considéré deux types de relations: (i) la composante globale $\boldsymbol{t}$ est égale à la somme de ses composantes partielles $\boldsymbol{t_k}$ ($\boldsymbol{t = \sum_{k=1}^{K} t_k}$) et (ii) la composante globale $\boldsymbol{t}$ est une combinaison linéaire de ses composantes partielles $\boldsymbol{t_k}$ ($\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$, avec $\boldsymbol{\lambda_k = u^\top t_k}$). Cette dernière relation signifie que $\boldsymbol{t}$ est la première composante PLS de $\boldsymbol{u}$ sur $\boldsymbol{t_1, t_2, ..., t_K}$. La relation (i) nous conduit à MB-RA et à la régression MB-PLS. Avec la relation (ii), nous avons les méthodes MB-WRA et MB-WCov. Pour ces deux dernières méthodes, un poids spécifique, $\boldsymbol{\lambda_k}$, est attribué à chaque tableau $\boldsymbol{X_k}$ et reflète l'importance qu'il revêt dans la détermination des variables latentes $\boldsymbol{t}$ et $\boldsymbol{u}$.

Nous avons déjà évoqué que la relation $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda_k} \boldsymbol{t_k}$, avec $\boldsymbol{\lambda_k} = \boldsymbol{u}^\top \boldsymbol{t_k}$ signifie que $\boldsymbol{t}$ est la première composante PLS de $\boldsymbol{u}$ sur $\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_K}$. De ce point de vue, MB-WCov présente des similarités avec la méthode Multiblock Hierarchical PLS "MB-HPLS" [23] qui présente la même propriété. Cependant, comme cette dernière méthode n'est basée sur aucun critère d'optimisation précis, son algorithme souffre de problèmes de convergence [3, 46].

Le fait que pour MB-WRA et MB-WCov, la composante globale $\boldsymbol{t}$ soit la première composante PLS de $\boldsymbol{u}$ sur $\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_K}$, suggère d'intéressants développements. En effet, au lieu d'utiliser la régression PLS usuelle, on pourrait utiliser la régression PLS éparse. Ce qui signifie qu'à chaque étape, les tableaux qui ne contribuent pas significativement à la détermination de la composante globale de cette étape, seront écartés. Par conséquent, nous obtenons des modèles parcimonieux qui sont faciles à interpréter sans pour autant affecter leur pouvoir prédictif. Cet aspect sera abordé dans le chapitre 6.

Afin de contourner le problème de quasi-colinéarité qu'on rencontre souvent avec les projecteurs $\boldsymbol{P_k} = \boldsymbol{X_k}(\boldsymbol{X_k}^\top \boldsymbol{X_k})^{-1} \boldsymbol{X_k}^\top$, nous avons utilisé les opérateurs $\boldsymbol{W_k} = \boldsymbol{X_k} \boldsymbol{X_k}^\top$. Ceci revient, en définitive, à substituer les matrices $\boldsymbol{X_k}^\top \boldsymbol{X_k}$ par une matrice identité. Cependant, nous pourrions adopter une substitution graduelle en utilisant les opérateurs $\boldsymbol{P_{k\gamma}} = \boldsymbol{X_k} \left[\boldsymbol{\gamma I} + \boldsymbol{(1-\gamma)} \boldsymbol{X_k}^\top \boldsymbol{X_k}\right]^{-1} \boldsymbol{X_k}^\top$. Dans cette expression, $\boldsymbol{\gamma}$ est un paramètre de régularisation, compris entre 0 et 1. Nous obtenons ainsi une approche continuum dont les points extrêmes (c'est-à-dire, $\boldsymbol{\gamma = 0}$ et $\boldsymbol{\gamma = 1}$) sont les méthodes discutées dans ce chapitre. En pratique, le paramètre de

régularisation peut être déterminé conjointement avec le nombre de variables latentes à introduire dans le modèle à l'aide de la technique de cross-validation [47].

Nous avons aussi proposé des indices pour mieux interpréter les résultats des analyses. Parmi ceux-ci, nous avons un indice qui donne la proportion de covariation qui est expliquée par les variables latentes. L'évolution de cet indice peut servir à choisir le nombre de composantes à retenir pour l'analyse. Nous avons également un indice qui nous renseigne sur l'importance de chaque tableau prédictif dans la détermination de la composante globale.

Annexe du Chapitre 4

L'objectif de cette annexe est de prouver la convergence de l'algorithme associé au critère de maximisation de MB-WRA.

La maximisation par rapport à $\boldsymbol{\nu}$ ($||\boldsymbol{\nu}|| = \mathbf{1}$) du critère:

$$\boldsymbol{C(\nu)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} (\boldsymbol{\nu}^\top \boldsymbol{Y}^\top \boldsymbol{P_k} \boldsymbol{Y} \boldsymbol{\nu})^{\mathbf{2}} \tag{.31}$$

nous a conduit à l'algorithme suivant:

0. Initialisation de $\boldsymbol{\nu}$, avec $||\boldsymbol{\nu}|| = \mathbf{1}$;

1. $\boldsymbol{\nu} = \frac{\sum_{k=1}^{K} \boldsymbol{\lambda_k} \boldsymbol{A_k}}{\sum_l \boldsymbol{\lambda_l^2}} \boldsymbol{\nu}$, avec $\boldsymbol{A_k} = \boldsymbol{Y}^\top \boldsymbol{P_k} \boldsymbol{Y}$ et $\boldsymbol{\lambda_k} = \boldsymbol{\nu}^\top \boldsymbol{A_k} \boldsymbol{\nu}$;

2. $\boldsymbol{\nu} = \boldsymbol{\nu} / ||\boldsymbol{\nu}||$;

3. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence.

Cet algorithme génère une séquence de nombres réels positifs:

$$\boldsymbol{C(n)} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} (\boldsymbol{\nu_n}^\top \boldsymbol{A_k} \boldsymbol{\nu_n})^{\mathbf{2}} \tag{.32}$$

où $\boldsymbol{\nu_n}$ est le vecteur, $\boldsymbol{\nu}$, obtenu à la $\boldsymbol{n^{ème}}$ itération.

Nous montrons que cette séquence forme une suite croissante lorsque $\boldsymbol{n}$ croît. Étant donné que le critère de maximisation $\boldsymbol{C_\nu}$ est majoré par $\sum_{k=1}^{K} ||\boldsymbol{Y}^\top \boldsymbol{P_k} \boldsymbol{Y}||^2$, nous déduisons que l'algorithme convergence.

Notons:

$$\boldsymbol{G(n)} = \frac{\sum_{k=1}^{K} \boldsymbol{\lambda_k(n)} \boldsymbol{A_k}}{\sum_{k=1}^{K} [\boldsymbol{\lambda_k(n)}]^2} \tag{.33}$$

avec $\boldsymbol{\lambda_k(n)} = \boldsymbol{\nu_n^\top} \boldsymbol{A_k} \boldsymbol{\nu_n}$.

Nous avons:

$$\boldsymbol{\nu_{n+1}} = \frac{\boldsymbol{G(n)\nu_n}}{||\boldsymbol{G(n)\nu_n}||} \tag{.34}$$

Nous cherchons donc à montrer que:

$$\boldsymbol{C(n)} \leq \boldsymbol{C(n+1)} \tag{.35}$$

Nous avons la propriété suivante:

$$\boldsymbol{\nu_n^\top G(n) \nu_n} \leq \boldsymbol{\nu_{n+1}^\top G(n) \nu_n} \tag{.36}$$

Cette propriété est facilement démontrable en remarquant par le biais de l'inégalité de Cauchy-Schwarz, que le maximum (par rapport à $\boldsymbol{x}$, $||\boldsymbol{x}|| = \boldsymbol{1}$) de la fonction $\boldsymbol{x}^\top \boldsymbol{G(n)\nu_n}$ est obtenue pour:

$$\boldsymbol{x} = \frac{\boldsymbol{G(n)\nu_n}}{||\boldsymbol{G(n)\nu_n}||} = \boldsymbol{\nu_{n+1}} \tag{.37}$$

En développant le premier membre de l'inégalité (.36), on note qu'il est

égal à 1. Quant au deuxième membre de cette inégalité, il donne:

$$\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{G}(n)\boldsymbol{\nu}_n = \frac{\sum_{k=1}^{K}\boldsymbol{\lambda}_k(n)\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_n}{\sum_{l=1}^{K}\left[\boldsymbol{\lambda}_l(n)\right]^2} \tag{.38}$$

Les matrices $\boldsymbol{A}_k = \boldsymbol{Y}^{\top}\boldsymbol{P}_k\boldsymbol{Y}$ étant semi-définies positives, nous avons d'après le théorème de Cauchy-Schwarz:

$$\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_n \leq \sqrt{\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_{n+1}} \times \sqrt{\boldsymbol{\nu}_n^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_n} \tag{.39}$$

Ainsi, nous avons:

$$\mathbf{1} \leq \boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{G}(n)\boldsymbol{\nu}_n \leq \frac{\sum_{k=1}^{K}\boldsymbol{\lambda}_k(n)\sqrt{\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_{n+1}} \times \sqrt{\boldsymbol{\nu}_n^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_n}}{\sum_{l=1}^{K}\left[\boldsymbol{\lambda}_l(n)\right]^2} \tag{.40}$$

En utilisant à nouveau l'inégalité de Cauchy-Schwarz, il vient que le dernier membre de l'inégalité (.40) est plus petit que:

$$\frac{\sqrt{\sum_{k=1}^{K}(\boldsymbol{\lambda}_k(n))^2} \times \sqrt{\sum_{k=1}^{K}\left(\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_{n+1}\right)\left(\boldsymbol{\nu}_n^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_n\right)}}{\sum_{l=1}^{K}\left[\boldsymbol{\lambda}_l(n)\right]^2} =$$

$$\frac{\sqrt{\sum_{k=1}^{K}\boldsymbol{\lambda}_k(n)\boldsymbol{\nu}_{n+1}^{\top}\boldsymbol{A}_k\boldsymbol{\nu}_{n+1}}}{\sqrt{\sum_{l=1}^{K}\left[\boldsymbol{\lambda}_l(n)\right]^2}} \tag{.41}$$

En utilisant pour une dernière fois l'inégalité de Cauchy-Schwarz, il s'ensuit que:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k(n)} \boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1} \leq \sqrt{\sum_{k=1}^{K} \left[\boldsymbol{\lambda_k(n)}\right]^2} \times \sqrt{\sum_{k=1}^{K} (\boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1})^2} \tag{.42}$$

En combinant les inégalités (.40), (.41) et (.42), nous avons:

$$\boldsymbol{1} \leq \frac{\left\{\sum_{k=1}^{K} \left[\boldsymbol{\lambda_k(n)}\right]^2\right\}^{1/4} \left[\sum_{k=1}^{K} (\boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1})^2\right]^{1/4}}{\sqrt{\sum_{l=1}^{K} (\boldsymbol{\lambda_l(n)})^2}} =$$

$$\frac{\left[\sum_{k=1}^{K} (\boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1})^2\right]^{1/4}}{\left\{\sum_{l=1}^{K} \left[\boldsymbol{\lambda_l(n)}\right]^2\right\}^{1/4}} \tag{.43}$$

En remarquant que $\boldsymbol{\lambda_k(n)} = \boldsymbol{\nu}_n^{\top} \boldsymbol{A_k} \boldsymbol{\nu_n}$, il s'ensuit que:

$$\left[\sum_{l=1}^{K} (\boldsymbol{\nu}_n^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_n)^2\right]^{1/4} \leq \left[\sum_{k=1}^{K} (\boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1})^2\right]^{1/4} \tag{.44}$$

Ce qui équivaut à:

$$\sum_{k=1}^{K} (\boldsymbol{\nu}_n^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_n)^2 \leq \sum_{k=1}^{K} (\boldsymbol{\nu}_{n+1}^{\top} \boldsymbol{A_k} \boldsymbol{\nu}_{n+1})^2. \tag{.45}$$

D'où $\boldsymbol{C(n)} \leq \boldsymbol{C(n+1)}$.

5

Développements autour de la méthode ComDim

5.1 Introduction

Aussi bien pour les méthodes multiblocs non supervisées que pour les méthodes multiblocs supervisées, nous avons identifié, en particulier, deux familles de méthodes (figure 5.1):

- une famille de méthodes s'apparentant à l'analyse canonique pour les méthodes non supervisées et à multiblock redundancy analysis pour les méthodes supervisées;
- une famille de méthodes s'apparentant à l'analyse en composantes principales multiblocs pour les méthodes non supervisées et à la régression PLS multiblocs, pour les méthodes supervisées.

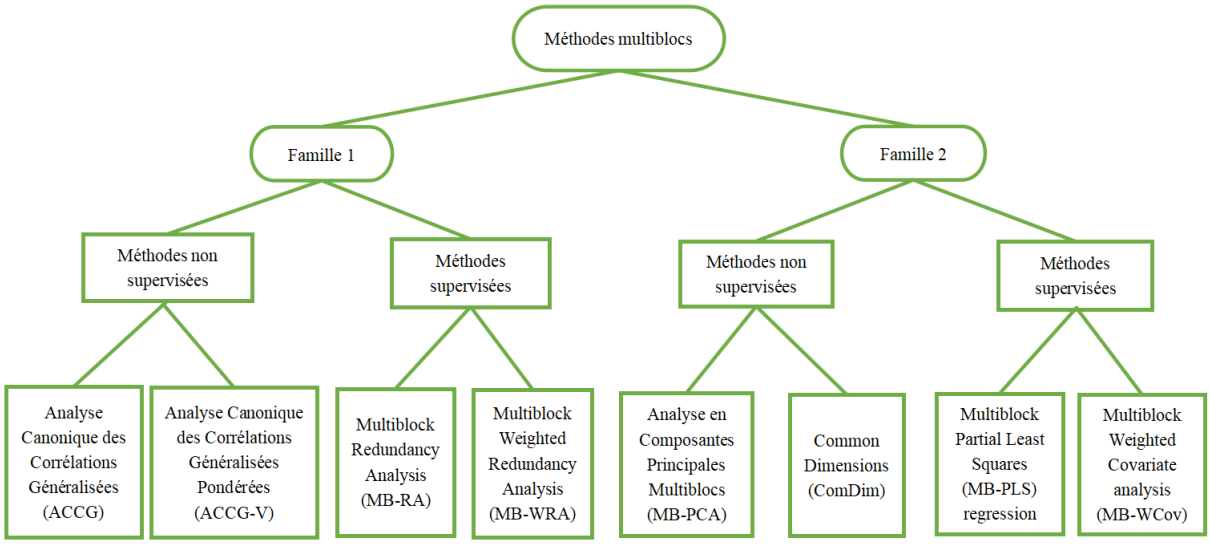


Fig. 5.1: Familles de méthodes multiblocs.

La première famille regroupe, pour ce qui concerne les méthodes non supervisées, l'analyse canonique des corrélations généralisées (ACCG) et l'analyse canonique des corrélations généralisées pondérées (ACCG-V) et, pour ce qui concerne les méthodes supervisées, nous trouvons les méthodes multiblock redundancy analysis (MB-RA) et multiblock weighted redundancy analysis (MB-WRA).

La deuxième famille, quant-à elle, regroupe l'analyse en composantes principales multiblocs (MB-PCA) et ComDim pour les méthodes non supervisées et les méthodes régression PLS multiblocs (MB-PLS) et multiblock weighted covariate analysis (MB-WCov) pour les méthodes supervisées.

Nous avons bien noté qu'un trait caractéristique de ces approches d'analyse est que les méthodes appartenant à la première famille sont vulnérables à la présence de multicolinéarité entre les variables des différents blocs; ce qui n'est pas le cas pour les méthodes de la deuxième famille.

Dans les deux chapitres qui suivent, nous focalisons sur les méthodes de

la deuxième famille. Il est entendu que les développements qui sont faits ici peuvent assez facilement être adaptés aux méthodes de la première famille. De manière plus précise, nous focalisons sur les méthodes ComDim, d'un côté, et MB-WCov, d'un autre côté. Le trait commun de ces deux méthodes est qu'elles exhibent de manière explicite des poids associés aux différents tableaux, indiquant leur contribution dans l'analyse.

Un point important de ce chapitre et du chapitre suivant est la proposition de versions "sparse" pour ComDim et MB-WCov. Ceci consiste à modifier les critères d'optimisation de ces deux méthodes en rajoutant des contraintes de "sparsité" de manière que, pour chaque composante, n'interviennent que les tableaux qui ont une contribution significative à la détermination de cette composante.

Un autre aspect qui nous intéresse dans ces deux chapitres est l'application de ComDim et MB-WCov à des cas particuliers où, par exemple, chaque tableau est constitué d'une seule variable.

Le présent chapitre est consacré à la méthode ComDim alors que le chapitre suivant sera consacré à la méthode MB-WCov.

Toutes les démarches proposées seront illustrées sur la base de données simulées ou d'études de cas réels.

5.2 Méthodes

5.2.1 ComDim: rappels et compléments

Soit $\boldsymbol{K}$ tableaux de données $\boldsymbol{X_1, X_2, ..., X_K}$ portant sur les mêmes $\boldsymbol{n}$ individus. Chaque tableau est supposé être centré et réduit (si nécessaire).

De plus, chaque tableau est normé de sorte à avoir $||\boldsymbol{X_k}|| = \mathbf{1}$. Comme indiqué dans le chapitre 3, la méthode ComDim procède par étapes en ce sens que les composantes globales et les composantes par bloc sont déterminées de manière séquentielle en procédant à des déflations à chaque fois.

A la première étape, la composante globale, $\boldsymbol{t}$ ($||\boldsymbol{t}|| = \mathbf{1}$), ainsi que les poids spécifiques $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ sont déterminées de manière à minimiser le critère [4]:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} ||\boldsymbol{X_k X_k^\top} - \boldsymbol{\lambda_k t t^\top}||^{\mathbf{2}} \tag{5.1}$$

Comme nous l'avons vu, plusieurs algorithmes peuvent être déroulés pour résoudre ce problème d'optimisation. Nous rappelons deux d'entre eux.

Algorithme 1:

0. Initialisation: $\boldsymbol{\lambda_k = 1}$ pour $\boldsymbol{k = 1, 2, ..., K}$;

1. $\boldsymbol{t}$ est le vecteur propre de $\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k X_k X_k^\top}$ associé à la plus grande valeur propre;

2. $\boldsymbol{t_k = X_k X_k^\top t}$: composante du bloc $\boldsymbol{X_k}$;

3. $\boldsymbol{\lambda_k = cov(t, t_k)}$;

4. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Algorithme 2:

0. Initialisation de $\boldsymbol{t}$ avec $||\boldsymbol{t}|| = \mathbf{1}$;

1. $\boldsymbol{t_k = X_k X_k^\top t}$;

2. $\boldsymbol{\lambda_k = cov(t, t_k)}$;

3. $\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$: composante globale;

4. $\boldsymbol{t = t/||t||}$;

5. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Une propriété notable que nous avons exposé dans le chapitre 3 est que la détermination de la composante globale $\boldsymbol{t}$ réalise le maximum du critère:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(t, t_k)} \tag{5.2}$$

où $\boldsymbol{t_k = X_k X_k^\top t}$ est la composante par bloc associée à $\boldsymbol{X_k}$. Ce critère permet de distinguer ComDim par rapport à MB-PCA. En effet, nous avions vu que, pour cette dernière méthode, la composante globale $\boldsymbol{t}$ est obtenue en maximisant le critère:

$$\sum_{k=1}^{K} \boldsymbol{cov(t, t_k)} \tag{5.3}$$

5.2.2 Analyse d'un tableau de données: ComDim-PCA

Lorsque nous ne disposons que d'un seul tableau $\boldsymbol{X}$, l'application de la méthode ComDim peut s'effectuer de deux manières différentes. La première manière consiste à chercher la composante $\boldsymbol{t}$ (de longueur 1) de manière à minimiser le critère:

$$||\boldsymbol{X X^\top - \lambda t t^\top}|| \tag{5.4}$$

qui est le critère de ComDim appliqué à un seul tableau. Naturellement, nous sommes conduits à l'ACP. La deuxième manière est de considérer que chacune des variables de $\boldsymbol{X}$ forme un tableau par elle-même et, par la suite, appliquer la méthode ComDim. Ce cas de figure a été étudié par des chercheurs [48, 49] qui ont souligné que cette démarche présente une alternative intéressante à l'analyse en composantes principales. Ces auteurs ont désigné cette méthode par l'acronyme CCA (Common Components Analysis) mais nous préférons utiliser l'acronyme ComDim-PCA afin d'éviter la confusion avec la méthode Canonical Correlation Analysis (CCA).

En considérant que chaque variable, $\boldsymbol{x_k}$, forme un tableau, l'application de ComDim conduit à déterminer une composante globale $\boldsymbol{t}$ de manière à maximiser:

$$\sum_{k=1}^{K} \boldsymbol{cov^4(t, x_k)} \tag{5.5}$$

En effet, nous avons:

$$\sum_{k=1}^{K} \boldsymbol{cov^2(t, t_k)} = \sum_{k=1}^{K} \boldsymbol{cov^2(t, x_k x_k^\top t)} = \boldsymbol{n^2} \sum_{k=1}^{K} \boldsymbol{cov^4(t, x_k)} \tag{5.6}$$

Par comparaison, l'application de la méthode MB-PCA à ce cas de figure aurait conduit à maximiser:

$$\sum_{k=1}^{K} \boldsymbol{cov(t, t_k)} = \sum_{k=1}^{K} \boldsymbol{cov(t, x_k x_k^\top t)} = \boldsymbol{n} \sum_{k=1}^{K} \boldsymbol{cov^2(t, x_k)} \tag{5.7}$$

Nous savons que la maximisation de ce dernier critère conduit à choisir $\boldsymbol{t}$ comme la première composante principale du tableau $\boldsymbol{X}$ [19]. Ainsi, il

apparaît qu'en remplaçant $\boldsymbol{cov^2(t, x_k)}$ par $\boldsymbol{cov^4(t, x_k)}$, la composante $\boldsymbol{t}$ serait liée de manière privilégiée aux variables qui auraient des covariances relativement élevées avec elle. Ce constat sera ultérieurement illustré sur la base d'un exemple.

5.2.3 Sparse ComDim

L'intérêt des poids $\boldsymbol{\lambda_k}$ exhibés par ComDim a été souligné dans Jouan-Rimbaud Bouveresse et al. [50]. En particulier, ces auteurs ont défini une stratégie d'analyse appelée ANOVA-ComDim qui utilise ces poids pour évaluer l'impact de facteurs opérant sur un tableau de données. Nous présentons une extension de ComDim qui va souligner un autre intérêt des poids $\boldsymbol{\lambda_k}$. Elle consiste à déterminer des composantes globales "sparses" en ce sens que les tableaux qui n'ont pas une contribution significative à la détermination d'une composante donnée sont écartés de cette composante. En d'autres termes, il s'agit de mettre à 0 les valeurs $\boldsymbol{\lambda_k}$ qui ne reflètent pas une contribution significative.

Les méthodes dites "sparses" ont connu un essor considérable depuis leur introduction dans le cadre de la régression linéaire en tant que méthodes de régularisation pour contourner le problème de colinéarité. Le principe de base est d'imposer une contrainte de type $\boldsymbol{L^1}$ sur le vecteur des coefficients de la régression linéaire. De ce fait, ceci se traduit par la mise à 0 des coefficients insignifiants. Par la suite, l'approche a été adoptée dans le cadre des analyses multivariées telles que l'ACP [31, 32], la régression PLS [51], l'analyse discriminante PLS [52], etc. Dans la pratique, il s'est avéré que ces méthodes conduisent à des modèles parcimonieux, dont les

résultats sont plus faciles à interpréter que ceux des modèles standards. Pour autant, la performance de ces modèles en termes de prédiction, par exemple, n'est pas affectée de manière significative.

Dans le cadre de ComDim, la sparsité porte sur les poids $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$. Cela signifie que les poids associés aux tableaux qui ont une contribution non significative à la définition de la composante globale seront mis à zéro.

Nous reformulons la recherche d'un modèle sparse pour ComDim sous forme d'un problème de maximisation basé sur une idée très intuitive. Pour cela, nous nous inspirons de la stratégie basée sur le principe dit "soft-thresholding" (seuillage doux).

Dans un premier temps, nous proposons un nouveau critère de détermination des composantes globales et par blocs pour ComDim. Par suite, ce critère sera adapté pour définir la méthode "Sparse ComDim".

Nous proposons de chercher des paramètres $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ vérifiant la contrainte $\sum_{k=1}^{K} \boldsymbol{\lambda_k^2} = \mathbf{1}$, une composante globale $\boldsymbol{t}$ $\boldsymbol{(||t|| = 1)}$ et des composantes par blocs $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$ de manière à maximiser la quantité:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k cov(t_k, t)} \tag{5.8}$$

Pour $\boldsymbol{t}$ fixé, nous avons $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$. D'après le théorème de Cauchy-Schwarz, les paramètres $\boldsymbol{\lambda_k}$ qui réalisent le maximum du critère ci-dessus sont donnés par $\boldsymbol{\lambda_k} = \boldsymbol{\alpha cov(t_k, t)}$ où $\boldsymbol{\alpha}$ est un scalaire que nous fixons en considérant la contrainte imposée à $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$. Il s'ensuit que: $\boldsymbol{\alpha} = \frac{1}{\sqrt{\sum_{l=1}^{K} \boldsymbol{cov^2(t_l, t)}}}$. En remplaçant $\boldsymbol{\lambda_k}$ par sa valeur dans le critère

à maximiser, nous obtenons la quantité $\sqrt{\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)}}$.
La maximisation de cette quantité par rapport à $\boldsymbol{t}$ et $\boldsymbol{t_k}$ est équivalente à la maximisation de la quantité $\sum_{k=1}^{K} \boldsymbol{cov^2(t_k, t)}$. Nous retrouvons, par conséquent, le même critère qui permet de définir ComDim. Ainsi, l'algorithme qui permet de résoudre le problème de maximisation ci-dessus est le suivant:

0. Initialisation de $\boldsymbol{t}$ avec $||\boldsymbol{t}|| = \mathbf{1}$;

1. $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$;

2. $\boldsymbol{\lambda_k} = \boldsymbol{cov(t_k, t)}$;

3. $\boldsymbol{\lambda_k} = \boldsymbol{\lambda_k} / \sqrt{\sum_{l=1}^{K} \boldsymbol{\lambda_l^2}}$, avec $\sum_{l=1}^{K} \boldsymbol{\lambda_l^2} \neq \mathbf{0}$;

4. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}$: composante globale;

5. $\boldsymbol{t} = \boldsymbol{t} / ||\boldsymbol{t}||$;

6. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Il est clair que la standardisation des coefficients $\boldsymbol{\lambda_k}$ (étape 3) n'a aucune incidence sur le calcul de $\boldsymbol{t}$ du fait de la standardisation opérée à l'étape 5. En d'autres termes, l'algorithme basé sur le nouveau critère conduit exactement à la même solution que l'algorithme de base de ComDim.

Comme indiqué ci-dessus, ce critère peut être adapté assez facilement pour établir une procédure Sparse ComDim. Considérons le critère de maximisation suivant:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k sup\left[cov(t, t_k) - \tau, 0\right]} \tag{5.9}$$

où $\boldsymbol{\tau}$ est un paramètre de sparsité fixé. Nous imposons les mêmes contraintes que pour le critère de ComDim, à savoir $\sum_{k=1}^{K} \boldsymbol{\lambda}_k^2 = \mathbf{1}$, $||\boldsymbol{t}|| = \mathbf{1}$ et $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$. La motivation derrière ce critère est très claire. Le paramètre $\boldsymbol{\tau}$ définit un seuil en deçà duquel la covariance entre $\boldsymbol{t}$ et $\boldsymbol{t_k}$ est considérée comme pouvant être insignifiante et, par conséquent, remplacée par zéro. Pour $\boldsymbol{t}$ fixée et d'après le théorème de Cauchy-Schwarz, nous avons:

$$\boldsymbol{\lambda_k = \alpha sup\left[cov(t, t_k) - \tau, 0\right]} \tag{5.10}$$

D'après la contrainte d'orthogonalité, il vient $\boldsymbol{\alpha} = \frac{\mathbf{1}}{\sqrt{\sum_{l=1}^{K}\{\boldsymbol{sup}[\boldsymbol{cov}(\boldsymbol{t},\ \boldsymbol{t_l}) - \boldsymbol{\tau},\ \mathbf{0}]\}^2}}$. Si maintenant, nous supposons $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ fixés, nous pouvons remarquer que:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k sup\left[cov(t,\ t_k) - \tau, 0\right]} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k \left[cov(t,\ t_k) - \tau\right]} \tag{5.11}$$

En effet, il est facile de vérifier que cette relation est vraie dans le cas où $\boldsymbol{sup\left[cov(t,\ t_k) - \tau, 0\right] = cov(t,\ t_k) - \tau}$ ou, alternativement, dans le cas où $\boldsymbol{sup\left[cov(t,\ t_k) - \tau, 0\right] = 0}$. Il s'ensuit que nous sommes conduits à maximiser:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k cov(t,\ t_k)} - \boldsymbol{\tau} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k} = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{t^\top} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k t_k} - \boldsymbol{\tau} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\lambda_k} \tag{5.12}$$

Par conséquent, toujours d'après le théorème de Cauchy-Schwarz, ce maximum est réalisé par $\boldsymbol{t}$ proportionnel à $\sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}$. En résumé, l'algorithme proposé pour Sparse ComDim est le suivant:

0. Initialisation de $\boldsymbol{t}$ avec $||\boldsymbol{t}|| = \mathbf{1}$;

1. $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top t}$: composante par bloc;

2. $\boldsymbol{\lambda_k} = \boldsymbol{sup}\,[\boldsymbol{cov}(\boldsymbol{t},\ \boldsymbol{t_k}) - \boldsymbol{\tau},\ \mathbf{0}]$ pour $\boldsymbol{k} = \mathbf{1},\ \mathbf{2},\ \boldsymbol{...},\ \boldsymbol{K}$;

3. $\boldsymbol{\lambda_k} = \boldsymbol{\lambda_k} / \sqrt{\sum_{l=1}^{K} \boldsymbol{\lambda_l^2}}$, avec $\sum_{l=1}^{K} \boldsymbol{\lambda_l^2} \neq \mathbf{0}$;

4. $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}$: composante globale;

5. $\boldsymbol{t} = \boldsymbol{t} / ||\boldsymbol{t}||$;

6. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

Étant donné que:

$$\boldsymbol{cov}(\boldsymbol{t},\ \boldsymbol{t_k}) = \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{t}^\top \boldsymbol{X_k X_k^\top t} \leq \frac{\mathbf{1}}{\boldsymbol{n}} ||\boldsymbol{t}||^{\mathbf{2}} \times ||\boldsymbol{X_k X_k^\top}|| = \frac{\mathbf{1}}{\boldsymbol{n}} \tag{5.13}$$

une manière simple et empirique pour sélectionner le paramètre $\boldsymbol{\tau}$ consiste à parcourir l'intervalle $[\mathbf{0}, \frac{\mathbf{1}}{\boldsymbol{n}}]$ avec un pas d'incrémentation de 0.02, par exemple. Pour chaque valeur de $\boldsymbol{\tau}$, on compte le nombre, $\boldsymbol{n_\tau}$, de tableaux pour lesquels $\boldsymbol{\lambda_k} = \mathbf{0}$ et on calcule l'inertie totale restituée par la composante globale. Le choix de $\boldsymbol{\tau}$ devrait être fait en considérant un compromis entre $\boldsymbol{n_\tau}$ qu'on souhaite aussi grand que possible et une inertie totale qu'on souhaite préserver au maximum. En pratique, il n'est pas nécessaire de parcourir tout l'intervalle $[\mathbf{0}, \frac{\mathbf{1}}{\boldsymbol{n}}]$ car à partir d'une certaine valeur, tous les poids sont mis à 0, indiquant que le seuil fixé est trop élevé. Tous ces aspects seront illustrés sur la base d'exemples.

5.2.4 Sparse ComDim-PCA

Comme indiqué ci-dessus, ComDim-PCA consiste à analyser un tableau de données $\boldsymbol{X}$ en considérant que chaque variable forme un tableau et, par la suite, appliquer ComDim aux tableaux ainsi formés. La proposition d'une version "sparse" pour ComDim-PCA est facile. Soit $\boldsymbol{t}$ la composante globale. La composante $\boldsymbol{t_k}$ associée à $\boldsymbol{\{x_k\}}$ est donnée par:

$$\boldsymbol{t_k = x_k x_k^\top t = n \times cov(x_k, t) x_k}. \tag{5.14}$$

La covariance entre $\boldsymbol{t}$ et $\boldsymbol{t_k}$ est donnée par:

$$\boldsymbol{cov(t,\ t_k) = n \times cov^2(x_k, t)}. \tag{5.15}$$

En appliquant l'algorithme général de Sparse ComDim à ce cas particulier, nous sommes conduits à l'algorithme suivant.

0. Initialisation de $\boldsymbol{t}$ avec $\boldsymbol{||t|| = 1}$;

1. $\boldsymbol{t_k = x_k x_k^\top t}$;

2. $\boldsymbol{\lambda_k = sup\left[n \times cov^2(x_k, t) - \tau, 0\right]}$;

3. $\boldsymbol{\lambda_k = \lambda_k / \sqrt{\sum_{l=1}^{K} \lambda_l^2}}$, avec $\boldsymbol{\sum_{l=1}^{K} \lambda_l^2 \neq 0}$;

4. $\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$;

5. $\boldsymbol{t = t / ||t||}$;

6. Réitération du processus à partir de l'étape 1, jusqu'à convergence.

5.2.5 *ComDim-Quali: ComDim appliquée à des variables qualitatives*

Nous proposons d'adapter ComDim à l'analyse d'un ensemble de variables qualitatives. Pour cela, chaque variable est exprimée sous forme d'un tableau disjonctif complet. Un pré-traitement (centrage et standardisation) est appliqué à chacun de ces tableaux. Par la suite, nous appliquons la méthode ComDim.

Soit $\boldsymbol{Z}$ une variable qualitative ayant $\boldsymbol{q}$ modalités. Nous indiquons également par $\boldsymbol{Z}$ le tableau disjonctif complet dont les colonnes sont les $\boldsymbol{q}$ indicatrices des modalités de $\boldsymbol{Z}$. S'agissant d'une variable binaire, la moyenne d'une colonne est égale à $\boldsymbol{p}$, qui est la proportion d'individus qui ont pris la modalité associée à la colonne en considération (c'est-à-dire, la proportion de 1 dans la colonne). Le pré-traitement que nous opérons sur le tableau $\boldsymbol{Z}$ consiste à centrer chaque colonne et la diviser par la racine carrée de cette moyenne (c'est-à-dire, proportion de 1 dans la colonne). De manière formelle, à la colonne $\boldsymbol{z_j}$, indicatrice de la $\boldsymbol{j^{ème}}$ modalité, nous associons la colonne $\boldsymbol{x_j} = \frac{\boldsymbol{z_j - p_j}}{\sqrt{\boldsymbol{p_j}}}$ où $\boldsymbol{p_j}$ est la proportion de 1 dans $\boldsymbol{z_j}$. Ce pré-traitement est très courant pour les méthodes liées à l'analyse des correspondances multiples (ACM). En effet, il arrive souvent que certaines modalités soient davantage observées que d'autres, ce qui pourrait conduire, en l'absence d'une standardisation, à ce que les modalités sur-représentées jouent un rôle prépondérant au détriment des modalités sous-représentées [19].

En présence d'un ensemble de variables qualitatives, la méthode ComDim-Quali consiste à appliquer ComDim aux tableaux $\boldsymbol{X_k}$ obtenus par codage

disjonctifs complets suivi d'une standardisation selon la procédure indiquée ci-dessus. De la même manière, nous pouvons appliquer la procédure de sparsité conduisant ainsi à une méthode Sparse ComDim-Quali.

5.3 Illustrations

5.3.1 Simulation

Nous reprenons le même schéma de simulation que dans le chapitre 3. Nous considérons deux variables orthogonales $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$, et nous formons quatre tableaux $\boldsymbol{X_1}$ à $\boldsymbol{X_4}$ portant sur cinquante individus et qui sont définis comme suit: $\boldsymbol{X_1} = [\boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}, \boldsymbol{d_1}]$, $\boldsymbol{X_2} = \boldsymbol{X_3} = \boldsymbol{X_4} = [\boldsymbol{d_2}, \boldsymbol{randn(4)}]$ où randn(4) désigne quatre variables aléatoires normalement distribuées. Dans chaque tableau, nous rajoutons 20% de bruit aux variables $\boldsymbol{d_1}$ et $\boldsymbol{d_2}$. Nous pré-traitons ensuite ces tableaux puis, nous les soumettons à la méthode ComDim et Sparse ComDim avec différentes valeurs du paramètre $\boldsymbol{\tau}$ ($\boldsymbol{\tau = 0.02}$, $\boldsymbol{\tau = 0.04}$ et $\boldsymbol{\tau = 0.06}$).

Le tableau 5.1 donne, pour les deux premières dimensions, les corrélations entre la composante globale et les composantes par bloc associées à la méthode ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim (pour $\boldsymbol{\tau = 0.02}$). Lorsque $\boldsymbol{\tau = 0}$ ou $\boldsymbol{\tau = 0.02}$, il y a une parfaite corrélation entre la première composante globale et la composante associée au tableau $\boldsymbol{X_1}$. Quant à la deuxième composante, elle est très fortement corrélée avec les composantes associées aux tableaux $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$.

Les pourcentages d'inerties restituées par les deux premières composantes globales en fonction du paramètre $\boldsymbol{\tau}$, ainsi que le nombre, $\boldsymbol{n_\tau}$, de tableaux

dont les contributions à la détermination des deux premières dimensions sont mises à zéro sont présentés sur la figure 5.2. Pour $\boldsymbol{\tau = 0}$ (c'est-à-dire, ComDim), aucun tableau n'est mis à zéro pour les deux premières dimensions. Cependant, nous notons une très faible contribution des tableaux $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$ pour la détermination de la première dimension et du tableau $\boldsymbol{X_1}$ pour la deuxième dimension. En passant à $\boldsymbol{\tau = 0.02}$, les contributions des trois tableaux $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ et $\boldsymbol{X_4}$ pour la détermination de la première dimension sont toutes mises à zéro, de même que la contribution du tableau $\boldsymbol{X_1}$ pour la détermination de la deuxième dimension. Il est important de souligner que cette simplification s'opère sans perte d'inertie restituée (25.17%) pour la première composante globale. De même, pour la deuxième composante globale, l'inertie restituée reste constante à 16.65%.

Tab. 5.1: Données simulées: corrélations entre la composante globale et les composantes par bloc pour les méthodes ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.02}$).

	Dimension 1				Dimension 2			
	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$
$\boldsymbol{\tau = 0}$	1	0.28	0.21	0.18	0.37	0.95	0.95	0.95
$\boldsymbol{\tau = 0.02}$	1	0.28	0.21	0.18	0.37	0.95	0.95	0.95

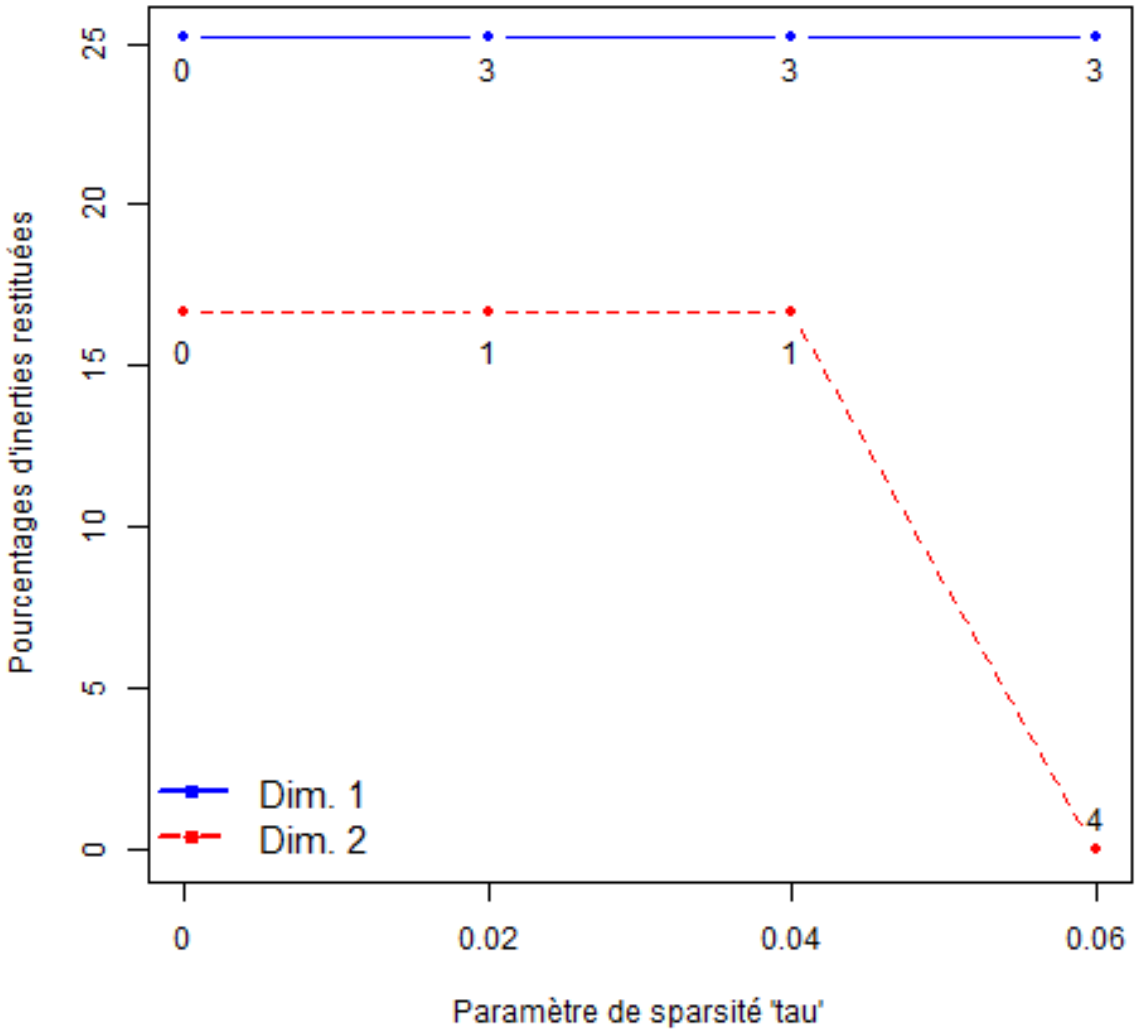


Fig. 5.2: Pourcentages d'inerties restituées par les deux premières dimensions de ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.02}$, $\boldsymbol{\tau = 0.04}$, $\boldsymbol{\tau = 0.06}$) et le nombre, $\boldsymbol{n_\tau}$, de tableaux dont les contributions sont mises à zéro pour les deux premières dimensions.

5.3.2 Etude de cas: Données de "pommes de terre"

Nous reprenons les données de pommes de terre présentées par Thybo et al. [30]. Vingt variétés de pommes de terre ont été analysées après un mois de stockage et six autres variétés ont été analysées après huit mois de stockage. Des mesures expérimentales ont été faites sur ces variétés, con-

duisant à l'obtention de huit tableaux. Le premier tableau, $\boldsymbol{X_1}$, contient les mesures chimiques (nommé "chemical"), le deuxième tableau, $\boldsymbol{X_2}$, contient les mesures de la compression uniaxiale ("compression") de ces variétés de pommes de terre. Les six tableaux, $\boldsymbol{X_3}$ à $\boldsymbol{X_8}$, portent sur les courbes de relaxation Time Domain (TD)-NMR (CMPG et FID) et des mesures faites à partir de la spectroscopie par proche infrarouge (NIR), avant et après cuisson. Les tableaux $\boldsymbol{X_3}$ à $\boldsymbol{X_8}$ sont respectivement nommés "cpmg", "nir", "cpmg.cooked", "nir.cooked", "fid" et "fid.cooked". Ces différents tableaux ont été pré-traités puis soumis à la méthode ComDim et Sparse ComDim, avec différentes valeurs du paramètre de sparsité $\boldsymbol{\tau}$.

Le tableau 5.2 montre les corrélations entre la composante globale et les composantes par bloc pour les deux premières dimensions de la méthode ComDim et Sparse ComDim. De ce tableau, il est clair que la première dimension des méthodes ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.05}$, $\boldsymbol{\tau = 0.1}$ et $\boldsymbol{\tau = 0.15}$) est bien corrélée avec les composantes par bloc associées à tous les tableaux à l'exception de "cpmg" et "cpmg.cooked". De même, il existe une bonne corrélation entre la deuxième composante globale de ComDim et Sparse ComDim avec les composantes par bloc associées à tous les tableaux, sauf pour les tableaux "compression" et "fid.cooked".

La figure 5.3 présente les pourcentages d'inerties restituées par les deux premières dimensions de ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.05}$, $\boldsymbol{\tau = 0.1}$, $\boldsymbol{\tau = 0.15}$) et le nombre, $\boldsymbol{n_\tau}$, de tableaux dont les contributions sont mises à zéro pour la détermination des deux premières dimensions. A partir de cette figure, nous notons que plus nous incrémentons la

valeur du paramètre de sparsité, $\boldsymbol{\tau}$, plus le nombre, $\boldsymbol{n_\tau}$, de tableaux dont la contribution pour la détermination des deux premières dimensions est mise à zéro augmente, avec une légère diminution de l'inertie restituée par les deux premières dimensions. De manière très précise, lorsque $\boldsymbol{\tau = 0}$ (c'est-à-dire, pour la méthode ComDim), aucun des tableaux n'a une contribution nulle pour la détermination des deux premières dimensions. Toutefois, il est à noter une très faible contribution des tableaux "cpmg" et "cpmg.cooked" pour la détermination de la première dimension et "compression", "fid" et "fid.cooked" pour la détermination de la deuxième dimension (résultats non reproduits). Lorsque nous passons à $\boldsymbol{\tau = 0.05}$, les contributions de ces tableaux marginaux tombent à zéro. De plus, nous notons une très faible contribution du tableau "chemical" qui s'annule lorsque nous passons à $\boldsymbol{\tau = 0.1}$. A $\boldsymbol{\tau = 0.15}$, trois tableaux sont mis à zéro sur la première dimension et cinq sur la deuxième dimension. Avec ce degré de sparsité, il résulte que seuls les tableaux "fid" et "fid.cooked" déterminent la première dimension alors que la deuxième dimension est déterminée par les tableaux "cpmg" et "cpmg.cooked". Pour ce qui est de l'inertie restituée par les deux premières dimensions, elle décroît légèrement de 32.36% ($\boldsymbol{\tau = 0}$) à 31.09% ($\boldsymbol{\tau = 0.15}$) pour la première dimension et de 25.32% ($\boldsymbol{\tau = 0}$) à 24.62% ($\boldsymbol{\tau = 0.15}$) pour la deuxième dimension.

Tab. 5.2: Données de pommes de terre: corrélations entre la composante globale et les composantes par bloc pour les deux premières dimensions de la méthode ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.05,\ \tau = 0.1,\ \tau = 0.15}$).

	Dimension 1				Dimension 2			
	$\boldsymbol{\tau = 0}$	$\boldsymbol{\tau = 0.05}$	$\boldsymbol{\tau = 0.1}$	$\boldsymbol{\tau = 0.15}$	$\boldsymbol{\tau = 0}$	$\boldsymbol{\tau = 0.05}$	$\boldsymbol{\tau = 0.1}$	$\boldsymbol{\tau = 0.15}$
chemical	**0.81**	**0.80**	**0.79**	**0.78**	**0.67**	**0.67**	**0.66**	**0.65**
compression	**0.51**	**0.50**	0.48	0.46	0.31	0.31	0.30	0.31
cpmg	0.14	0.13	0.12	0.10	**0.85**	**0.87**	**0.90**	**0.93**
nir	**0.71**	**0.70**	**0.68**	**0.66**	**0.66**	**0.65**	**0.62**	**0.58**
cpmg.cooked	0.38	0.42	0.45	0.49	**0.93**	**0.92**	**0.89**	**0.85**
nir.cooked	**0.80**	**0.79**	**0.78**	**0.76**	**0.75**	**0.74**	**0.72**	**0.68**
fid	**0.97**	**0.97**	**0.98**	**0.98**	**0.60**	**0.59**	**0.58**	**0.56**
fid.cooked	**0.96**	**0.96**	**0.96**	**0.96**	0.20	0.20	0.21	0.24

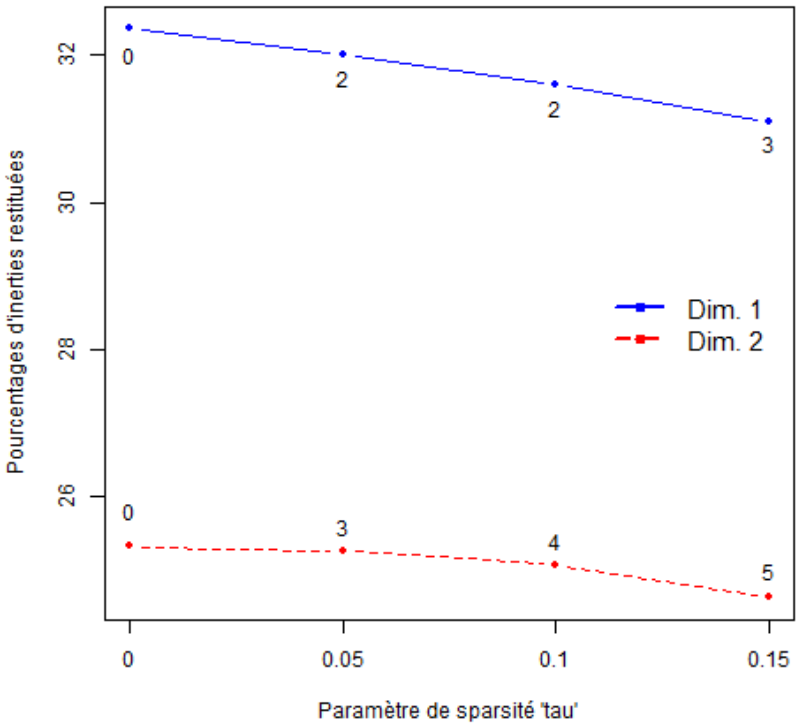


Fig. 5.3: Pourcentages d'inerties restituées par les deux premières dimensions de ComDim ($\boldsymbol{\tau = 0}$) et Sparse ComDim ($\boldsymbol{\tau = 0.05,\ \tau = 0.1,\ \tau = 0.15}$) et le nombre, $\boldsymbol{n_\tau}$, de tableaux dont les contributions sont mises à zéro pour la détermination de ces deux dimensions.

5.3.3 *ComDim-PCA et Sparse ComDim-PCA: données sensorielles*

A travers cette étude de cas, l'objectif est d'illustrer sur la base des données réelles deux aspects importants:

- l'application de ComDim-PCA à un tableau de données et la comparaison des résultats de cette méthode avec ceux de l'analyse en composantes principales (ACP);
- l'application de Sparse ComDim-PCA et la comparaison des résultats obtenus avec ceux de la méthode Sparse PCA.

Les données concernent l'évaluation sensorielle de 18 mousses de poisson selon 17 variables sensorielles (tableau 5.3). L'évaluation consiste à attribuer à chaque mousse de poisson, une note allant de 0 à 9. Cette note exprime l'intensité perçue pour chaque variable sensorielle. Par exemple, une note de 0 signifie que la variable n'est pas perçue et une note de 9, signifie qu'elle est extrêmement perçue. Les notes moyennes attribuées par les juges sont consignées dans un tableau de dimension 18×17 sur lequel sont effectuées les analyses ci-après.

Tab. 5.3: Données sensorielles: description des 17 variables sensorielles.

Variables	Libellé court	Variables	Libellé court
Lisse au toucher	TLISS	Humidité en bouche	BHUMI
Humidité au toucher	THUMI	Huile en bouche	BHUIL
Collant au toucher	TCOLL	Gras en bouche	BGRAS
Fermeté au toucher	TFERM	Mousse en bouche	BMOUS
Déformable au toucher	TDEFO	Fermeté en bouche	BFERM
Cassant au toucher	TCASS	Morceaux en bouche	BMORC
Gras au toucher	TGRAS	Râpeux en bouche	BRAPE
Fermeté au couteau	CFERM	Collant en bouche	BCOLL
Collant au couteau	CCOLL		

Nous avons appliqué ComDim-PCA à ce tableau de données. A titre de comparaison, nous avons aussi appliqué une ACP standardisée.

Le tableau 5.4 présente les pourcentages d'inerties restituées par les cinq premières composantes de ComDim-PCA et de l'ACP. Les pourcentages d'inerties restituées par les deux premières composantes de ComDim-PCA sont respectivement égaux à **53.37**% et **26.76**%. Ces pourcentages sont très légèrement inférieurs à ceux associés aux deux premières composantes principales de l'ACP qui sont respectivement **53.73**% et **26.89**%. Cette différence est très insignifiante et tout à fait prévisible car l'ACP vise à optimiser ce critère (c'est-à-dire, la restitution maximale de l'inertie pour la première composante principale).

Tab. 5.4: Données sensorielles: pourcentages d'inerties restituées par les cinq premières dimensions de ComDim-PCA et de l'ACP.

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
ComDim-PCA	53.37	26.76	5.64	4.51	2.67
ACP	53.73	26.89	6.25	4.38	2.65

Le tableau 5.5 donne les corrélations des variables avec les deux premières composantes de ComDim-PCA et de l'ACP. Comme nous pouvons le remarquer, les résultats des deux méthodes se recoupent dans une large mesure. Sur la première dimension, nous pouvons voir que les mousses de poisson qui sont fermes au toucher, le sont aussi en bouche et au couteau et se morcellent facilement en bouche. Par contre, ces mousses de poisson ne sont pas lisses, collants, déformables et gras au toucher. Ils ne sont non plus collants au couteau et en bouche et huileux en bouche. Sur la deuxième dimension, il apparaît que les mousses de poisson qui sont gras au toucher le sont aussi en bouche et sont huileux en bouche. Ils s'opposent à ceux qui sont humides au toucher puis humides, mousseux et râpeux en bouche.

Tab. 5.5: Données de sensorielles: corrélations des variables avec les deux premières composantes de ComDim-PCA et les deux premières composantes principales de l'ACP.

Variables	Dimension 1		Dimension 2	
	ComDim-PCA	ACP	ComDim-PCA	ACP
TLISSE	**0.68**	**0.67**	0.20	0.28
THUMIDE	0.05	-0.04	**0.98**	**0.97**
TCOLLANT	**0.84**	**0.87**	**−0.50**	-0.43
TFERME	**−0.94**	**−0.91**	-0.23	-0.32
TDEFORMABL	**0.94**	**0.90**	0.28	0.37
TCASSANT	-0.42	-0.47	0.48	0.45
TGRAS	**0.74**	**0.79**	**−0.57**	**−0.52**
CFERME	**−0.89**	**−0.86**	-0.36	-0.44
CCOLLANT	**0.79**	**0.82**	-0.40	-0.36
BHUMIDE	0.33	0.28	**0.85**	**0.85**
BHUILEUX	**0.62**	**0.69**	**−0.53**	**−0.50**
BGRAS	**0.75**	**0.80**	**−0.58**	**−0.53**
BMOUSSEUX	**0.70**	**0.66**	**0.54**	**0.60**
BFERME	**−0.94**	**−0.91**	-0.28	-0.38
BMORCEAUX	**−0.95**	**−0.92**	-0.18	-0.28
BRAPEUX	0.04	-0.02	**0.61**	**0.61**
BCOLLANT	**0.81**	**0.85**	-0.43	-0.37

La figure 5.4*a* représente le cercle de corrélations des variables sensorielles avec les deux premières composantes de ComDim-PCA. La fig-

ure 5.4*b* donne la représentation des mousses de poisson sur la base des deux premières composantes de ComDim-PCA. Les figures 5.4*c* et 5.4*d* présentent respectivement le cercle de corrélations et la configuration des mousses de poisson correspondant aux deux premières composantes principales de l'ACP.

A première vue, les résultats obtenus à l'aide des deux méthodes d'analyses semblent très concordants. Ceci est corroboré par le fait que le coefficient RV entre les deux configurations des mousses de poisson sur les figures 5.4*b* et 5.4*d* est égal à **0.96**.

Pour illustrer la méthode Sparse ComDim-PCA, nous l'avons appliquée aux données sensorielles à partir de différentes valeurs du paramètre de sparsité $\boldsymbol{\tau}$ $\boldsymbol{(\tau = 0.1,\ \tau = 0.2)}$. A titre de comparaison, nous avons aussi appliqué la méthode Sparse PCA avec différentes valeurs du paramètre 'para' (para=0.1, para=0.2). La figure 5.5 présente les pourcentages d'inerties restituées par les deux premières dimensions et le nombre de tableaux dont les contributions pour la détermination des deux premières dimensions sont mises à zéro. De la figure 5.5a, il vient que toute incrémentation du paramètre de sparsité $\boldsymbol{\tau}$ entraîne une augmentation du nombre de tableaux mis à zéro pour les deux premières dimensions. Quant à l'inertie restituée, elle décroît légèrement de 53.37% ($\boldsymbol{\tau = 0}$) à 51.97% ($\boldsymbol{\tau = 0.2}$) pour la première dimension. Pour la deuxième dimension, elle croît légèrement de 26.76% ($\boldsymbol{\tau = 0}$) à 27.94% ($\boldsymbol{\tau = 0.2}$). Sur la figure 5.5b, nous notons une diminution de l'inertie restituée pour les deux premières dimensions. Elle passe de 53.72% (para=0, c'est-à-dire ACP) à 42.55% (para=0.2) pour la première dimension et de 26.89% à 19.44% (para=0.2) pour la deuxième di-

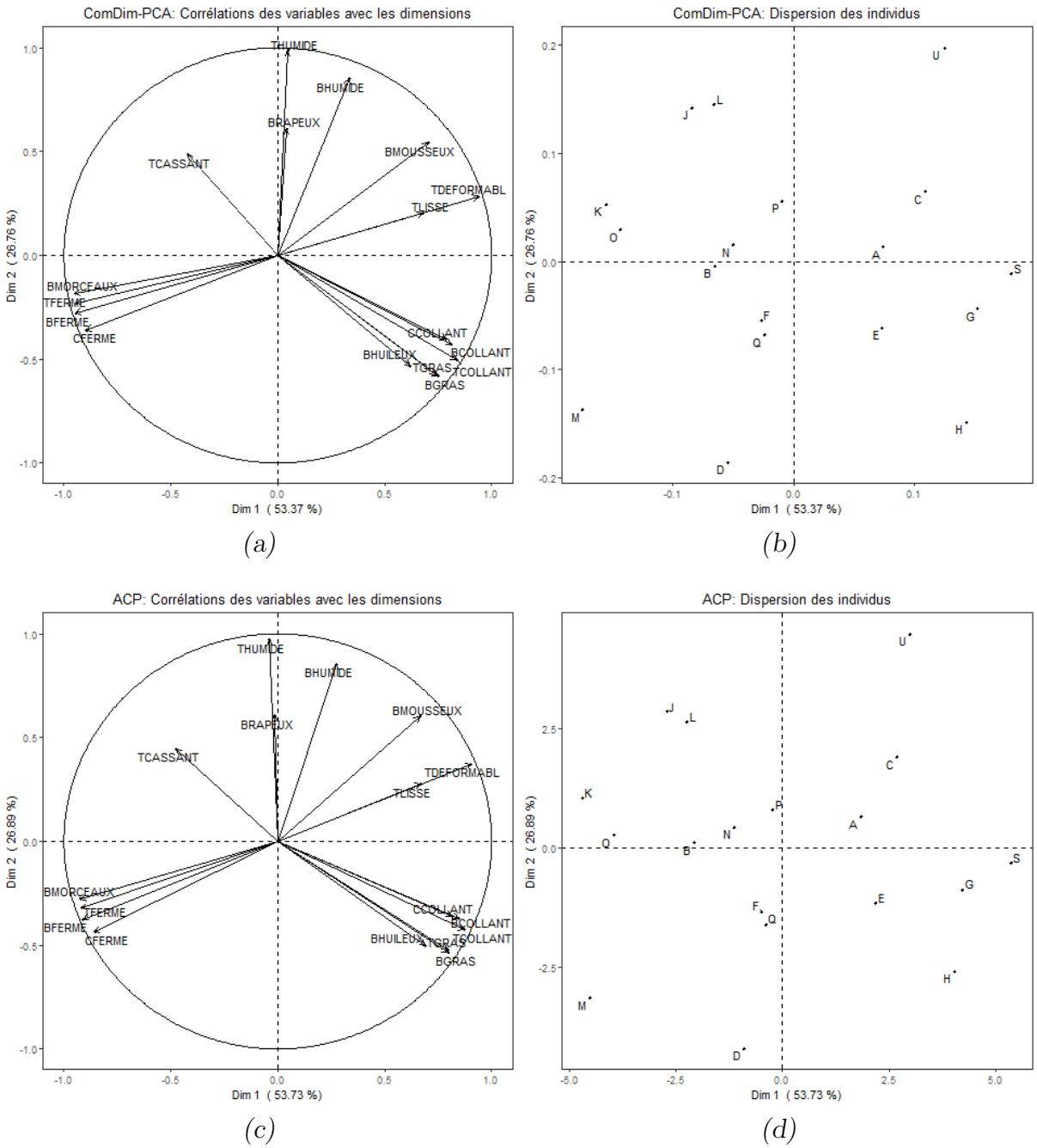
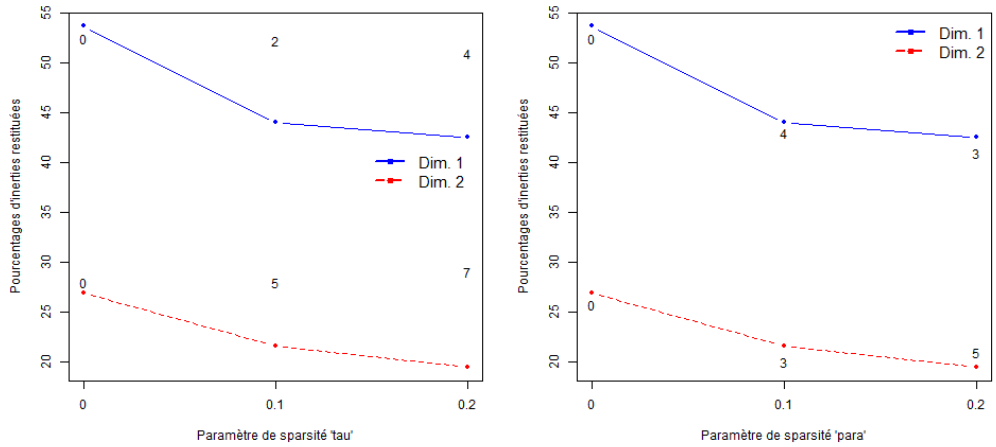


Fig. 5.4: Données sensorielles: cercles de corrélations et représentation des mousses de poisson sur la base des deux premières composantes de ComDim-PCA et de l'ACP.

mension. Cependant, nous remarquons que nous obtenons quatre zéros sur la première dimension (para=0.1) puis trois zéros par la suite (para=0.2). Ceci suggère de retenir comme valeur optimale du paramètre 'para', la valeur 0.1.

Du fait que les méthodes Sparse ComDim-PCA et Sparse PCA visent la même finalité, mais avec des principes et paramètres différents et afin de mieux les comparer, nous proposons de nous appuyer sur le compromis entre le nombre de zéros produit et les pourcentages d'inerties restituées. De ce fait, nous pouvons remarquer sur la première dimension que, pour produire quatre zéros, la méthode Sparse ComDim-PCA restitue une inertie de 51.97%, alors que Sparse PCA ne restitue que 44.06%. Sur la deuxième dimension, pour avoir cinq zéros, Sparse PCA restitue 19.44% d'inertie alors que Sparse ComDim-PCA restitue 26.97% (et produit six zéros).



(a) ComDim-PCA & Sparse ComDim-PCA

(b) PCA & Sparse PCA

Fig. 5.5: Données sensorielles: pourcentages d'inerties restituées par les deux premières dimensions et le nombre de tableaux dont les contributions pour la détermination des deux premières dimensions sont mises à zéro: (a) ComDim-PCA ($\boldsymbol{\tau = 0}$) et Sparse ComDim-PCA ($\boldsymbol{\tau = 0.1,\ \tau = 0.2}$); (b) PCA ($\boldsymbol{para = 0}$) et Sparse PCA ($\boldsymbol{para = 0.1,\ para = 0.2}$).

5.3.4 ComDim-Quali

Les données ayant servi à l'illustration de la méthode ComDim-Quali sont décrites dans Saporta & Niang [53] et dans le package "DiscriMiner" [54]. Ces données sont relatives à l'assurance des voitures en Belgique en 1992. Elles portent sur 1106 individus et 10 variables. Le tableau 5.6 décrit les variables qualitatives ainsi que leurs modalités.

Tab. 5.6: Données d'assurance de véhicules: description des 10 variables et leurs modalités.

Blocs	Variables	Description	Modalités
$\boldsymbol{X_1}$	Claims	Group variable	bad, good
$\boldsymbol{X_2}$	Use	Type d'usage	private, professional
$\boldsymbol{X_3}$	Type	Type d'assurance	companies, female, male
$\boldsymbol{X_4}$	Language	Langage	flemish, french
$\boldsymbol{X_5}$	BirthCohort	Birth Cohort	BD_1890_1949, BD_1950_1973, BD_unknown
$\boldsymbol{X_6}$	Region	Région géographique	Brussels, Other_regions
$\boldsymbol{X_7}$	BonusMalus	Level of bonus-malus	BM_minus, BM_plus
$\boldsymbol{X_8}$	YearSuscrip	Année de souscription	YS<86, YS>=86
$\boldsymbol{X_9}$	Horsepower	Horsepower	HP<=39, HP>=40
$\boldsymbol{X_{10}}$	YearConstruc	Année de construction du véhicule	YC_33_89, YC_90_91

Nous avons pré-traité le tableau de données qualitatives comme décrit dans la section ComDim-Quali (c'est-à-dire, centrage et standardisation des indicatrices des modalités des différentes variables). Par la suite, nous avons appliqué la méthode ComDim-Quali et l'Analyse des Correspondances Multiples (ACM) sur les tableaux obtenus. Les pourcentages d'inerties restituées par les cinq premières dimensions des deux méthodes sont présentées dans le tableau 5.7. Comme nous pouvons le voir, les résultats de ces deux méthodes se recoupent dans une large mesure. De

plus, une conclusion similaire peut être tirée sur la base de la projection des modalités et des variables dans le plan formé par les deux premières dimensions des deux méthodes (figure 5.6).

Tab. 5.7: Pourcentages d'inerties restituées par les cinq premières dimensions de la méthode ComDim-Quali et ACM.

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
ComDim-Quali	21.76	12.30	8.56	8.69	8.38
ACM	23.27	14.21	10.96	9.02	8.25

Le tableau 5.8 présente les corrélations entre la composante globale et les composantes par bloc pour les deux premières dimensions de la méthode ComDim-Quali. A partir de ce tableau, nous notons une forte corrélation entre la première composante globale et les composantes associées aux tableaux $\boldsymbol{X_1}$ (Claims), $\boldsymbol{X_5}$ (BirthCohort), $\boldsymbol{X_7}$ (BonusMalus) et $\boldsymbol{X_8}$ (Year-Suscrip). Quant à la deuxième dimension, elle est fortement corrélée avec les composantes associées aux tableaux $\boldsymbol{X_2}$ (Use) et $\boldsymbol{X_3}$ (Type). Ce tableau confirme bien la projection des variables dans le plan formé par les deux premières dimensions de la méthode ComDim-Quali (figure 5.6c).

Tab. 5.8: Corrélations entre la composante globale et les composantes par bloc pour les deux premières dimensions de la méthode ComDim-Quali.

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$	$\boldsymbol{X_6}$	$\boldsymbol{X_7}$	$\boldsymbol{X_8}$	$\boldsymbol{X_9}$	$\boldsymbol{X_{10}}$
Dim. 1	**0.90**	0.24	0.18	0.18	**0.57**	0.32	**0.90**	**0.60**	0.11	0.28
Dim. 2	0.07	**0.61**	**0.98**	0.10	0.43	0.04	0.13	0.07	0.14	0.03

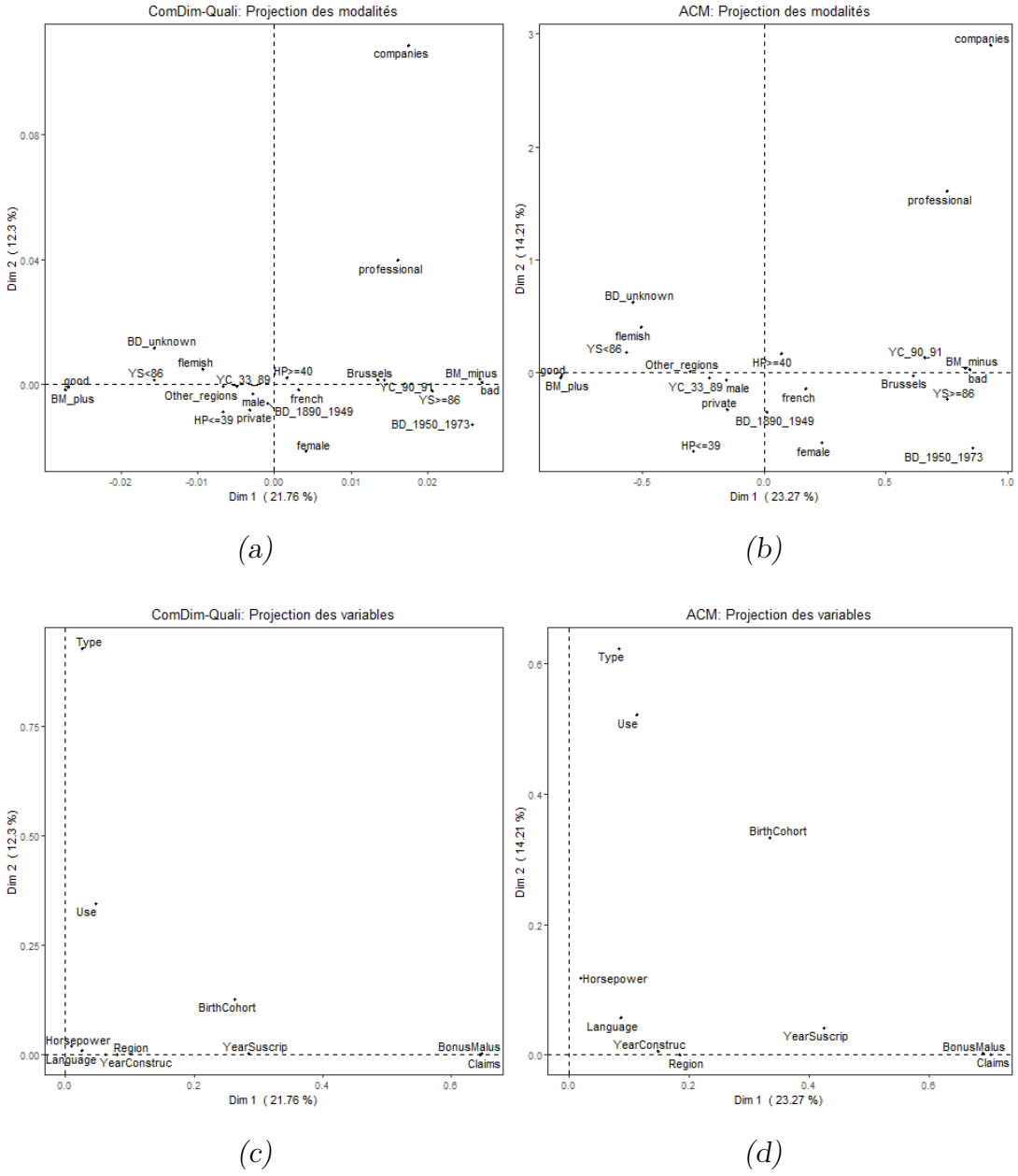


Fig. 5.6: Projection des modalités et des variables dans le plan formé par les deux premières composantes de ComDim-Quali et de l'ACM.

5.3.5 Sparse ComDim-Quali

L'objectif de cette étude de cas est de se focaliser sur la stratégie de sparsité de la méthode ComDim-Quali. Nous avons considéré les données d'une étude qui a été menée auprès de 300 personnes pour étudier leurs usages et habitudes concernant la consommation du thé. Ainsi, 18 ques-

tions leurs ont été posées. Les données de cette étude sont consignées dans un tableau $\boldsymbol{X}$ de dimension $\mathbf{300 \times 18}$ et disponibles dans le package "FactoMineR" [55]. Après avoir pré-traité ce tableau, nous l'avons appliqué la méthode Sparse ComDim-Quali. La figure 5.7 présente les pourcentages d'inerties restituées par les deux premières dimensions de ComDim-Quali ($\boldsymbol{\tau = 0}$) et Sparse ComDim-Quali ($\boldsymbol{\tau = 0.1,\ \tau = 0.2}$) et le nombre de tableaux, $\boldsymbol{n_\tau}$, dont les contributions sont mises à zéro pour ces deux premières dimensions. Comme pour les autres versions sparses abordées plus haut (Sparse ComDim et Sparse ComDim-PCA), au fur et à mesure qu'on incrémente la valeur du paramètre de sparsité $\boldsymbol{\tau}$, le nombre de tableaux qui se mettent à zéro croît. Par exemple, lorsqu'on applique la méthode ComDim-Quali, aucun tableau n'est mis à zéro pour les deux premières dimensions. Mais lorsque $\boldsymbol{\tau = 0.2}$, les contributions de cinq tableaux sont mises à zéro pour la détermination de la première dimension. De même, les contributions de deux tableaux sont mises à zéro pour la détermination de la deuxième dimension. Pour ce qui est de l'inertie restituée, elle est invariante sur les deux premières dimensions.

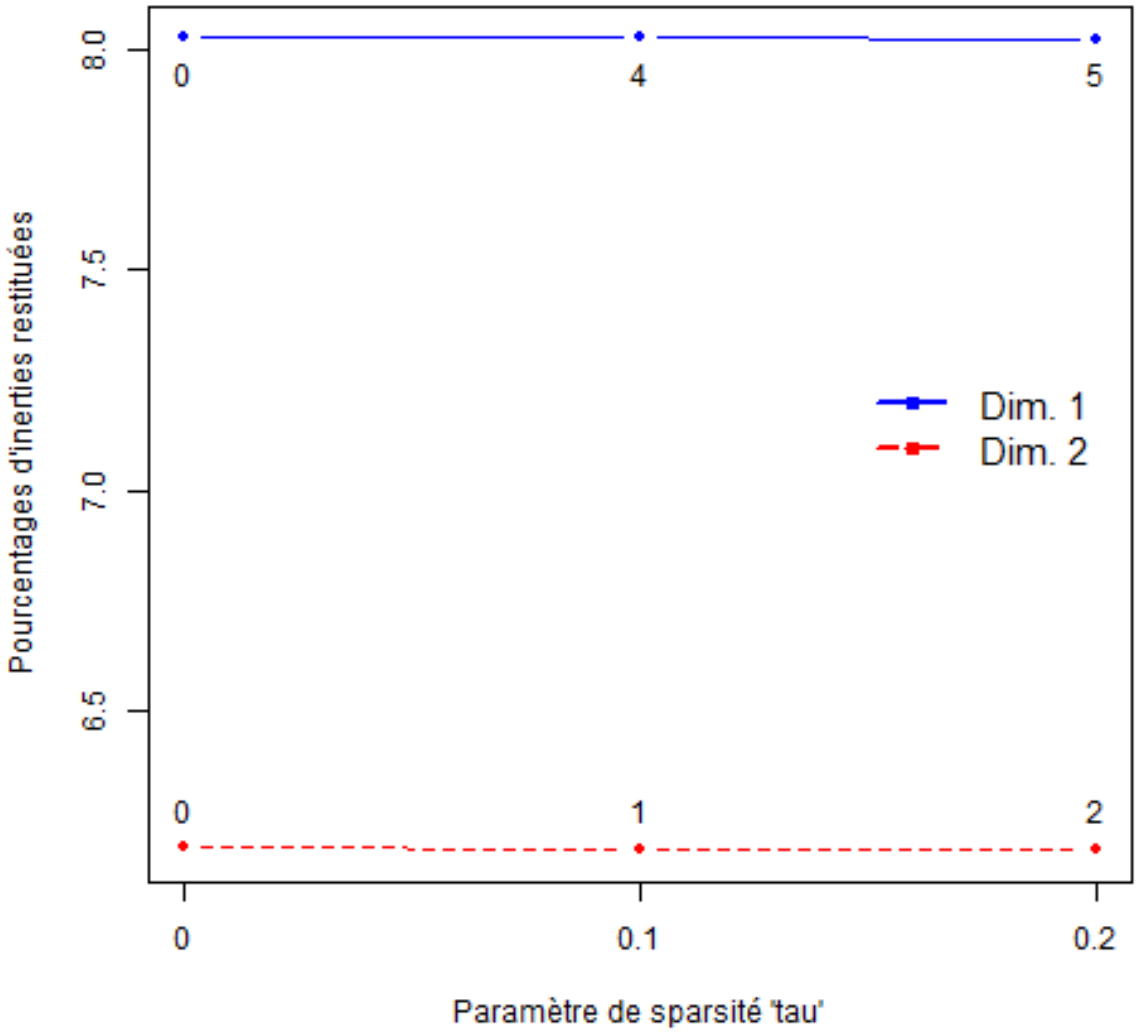


Fig. 5.7: Pourcentages d'inerties restituées par les deux premières dimensions de ComDim-Quali ($\boldsymbol{\tau = 0}$) et Sparse ComDim-Quali ($\boldsymbol{\tau = 0.1,\ \tau = 0.2}$) et le nombre de tableaux, $\boldsymbol{n_\tau}$, dont les contributions sont mises à zéro.

5.4 Discussion et conclusion

La méthode ComDim est initialement conçue pour être appliquée à des données multiblocs quantitatives. Dans ce chapitre, nous avons élargi ce champ d'application en proposant plusieurs variantes.

Dans un premier temps, nous avons étudié le cas particulier où chaque

tableau de données est réduit à une seule variable. En appliquant ComDim, nous avons défini une méthode appelée ComDim-PCA. En pratique, cette méthode a conduit à des résultats qui se recoupent dans une large mesure avec ceux de l'ACP.

Nous avons remarqué que l'ACP multiblocs vise à maximiser la somme des covariances entre la composante globale et les composantes par bloc alors que la méthode ComDim maximise la somme des covariances au carré entre la composante globale et les composantes par bloc. Nous pourrions penser à une extension consistant à considérer une puissance plus élevée. Formellement, nous serons amenés à maximiser $\sum_{k=1}^{K} \boldsymbol{cov}^{\alpha}(\boldsymbol{t},\ \boldsymbol{t_k})$, où $\boldsymbol{t}$ (resp., $\boldsymbol{t_k}$) est la composante globale (resp., par bloc) et $\boldsymbol{\alpha}$ est un entier supérieur ou égal à 1. Pour $\boldsymbol{\alpha = 1}$, nous sommes conduits à MB-PCA et pour $\boldsymbol{\alpha = 2}$, nous retrouvons ComDim. De même, nous savons que l'ACP d'un tableau $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p}]$ consiste à trouver une composante, $\boldsymbol{t}$, de manière à maximiser $\sum_{j=1}^{p} \boldsymbol{cov^2}(\boldsymbol{x_j},\ \boldsymbol{t})$. Nous avons aussi vu que ComDim-PCA vise à maximiser $\sum_{j=1}^{p} \boldsymbol{cov^4}(\boldsymbol{x_j},\ \boldsymbol{t})$. Là aussi, nous sommes en mesure de nous interroger sur l'intérêt (et la résolution) du problème consistant à maximiser $\sum_{j=1}^{p} \boldsymbol{cov^{2\alpha}}(\boldsymbol{x_j},\ \boldsymbol{t})$, où $\boldsymbol{\alpha}$ est un entier naturel supérieur ou égal à 1.

Dans un deuxième temps, nous avons proposé une version "sparse" pour les méthodes ComDim, ComDim-PCA et ComDim-Quali, que nous avons respectivement désigné par Sparse ComDim, Sparse ComDim-PCA et Sparse ComDim-Quali. L'objectif de cette stratégie est d'obtenir des modèles parcimonieux pour faciliter l'interprétation des résultats. De manière spécifique, les tableaux de données qui ne contribuent pas significativement

à la détermination d'une dimension donnée sont ignorés pour cette dimension. L'application de la méthode Sparse ComDim-PCA sur des données réelles a donné des résultats plus satisfaisants que ceux de la méthode Sparse PCA. Toutefois, il est à noter que contrairement à la méthode Sparse PCA pour laquelle la contrainte de sparsité est imposée sur le vecteur de loadings, pour les méthodes Sparse ComDim, Sparse ComDim-PCA et Sparse ComDim-Quali, la contrainte de sparsité est imposée sur les poids spécifiques. Il serait donc intéressant d'explorer ultérieurement, la double sparsité au niveau de ces trois dernières méthodes. De manière plus claire, cette double sparsité se traduira par le fait d'imposer la contrainte de sparsité non seulement sur les poids spécifiques, mais également sur les vecteurs de loadings. Ce qui permettrait d'avoir des modèles beaucoup plus parcimonieux que ceux des méthodes Sparse ComDim, Sparse ComDim-PCA et Sparse ComDim-Quali.

Dans un troisième temps, nous avons proposé une variante de la méthode ComDim qui s'applique à un tableau de données qualitatives. Nous avons appelé cette variante, ComDim-Quali. Sur un jeu de données réelles, les résultats de cette méthode semblent se recouper avec ceux de l'ACM. A l'avenir, il serait intéressant d'étendre la méthode ComDim à d'autres situations courantes. Cette extension concernera l'analyse de données mixtes (c'est-à-dire, un mélange de variables quantitatives et qualitatives).

6

Développements autour de la méthode MB-WCov

6.1 Introduction

Dans le chapitre 4, nous avons introduit la méthode MB-WCov comme alternative à la régression PLS multiblocs (MB-PLS). Nous présentons dans ce chapitre de nouvelles propriétés de cette méthode. Nous en présentons aussi une version "sparse" et nous l'appliquons à des cas particuliers où, par exemple, les tableaux $\boldsymbol{X_k}$ et/ou le tableau $\boldsymbol{Y}$ sont formés d'une seule variable.

En fin de chapitre, les différentes stratégies d'analyse sont illustrées sur la base de données réelles.

6.2 Méthodes

6.2.1 Relation entre deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$

El Ghaziri et Qannari [45] ont étudié les propriétés de plusieurs indices d'association entre deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$. Parmi ces indices, nous nous intéressons en particulier à deux d'entre eux: l'indice de corrélation

multivariée et le coefficient RV.

Soit deux tableaux $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ et $\boldsymbol{Y}$ $(\boldsymbol{n} \times \boldsymbol{q})$ portant sur les mêmes $\boldsymbol{n}$ individus et supposés être centrés. Nous désignons par:

$\boldsymbol{V_{XY}} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{X}^\top\boldsymbol{Y}$, la matrice de covariance entre $\boldsymbol{X}$ et $\boldsymbol{Y}$,

$\boldsymbol{V_{YX}} = \boldsymbol{V_{XY}^\top} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{Y}^\top\boldsymbol{X}$, la matrice de covariance entre $\boldsymbol{Y}$ et $\boldsymbol{X}$,

$\boldsymbol{V_X} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{X}^\top\boldsymbol{X}$, la matrice de variance-covariance de $\boldsymbol{X}$ et

$\boldsymbol{V_Y} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{Y}^\top\boldsymbol{Y}$, la matrice de variance-covariance de $\boldsymbol{Y}$.

Dans un premier temps, nous supposons que les deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$, en plus de porter sur les mêmes $\boldsymbol{n}$ individus, portent également sur les mêmes $\boldsymbol{p}$ variables. L'indice de covariance multivariée entre $\boldsymbol{X}$ et $\boldsymbol{Y}$ est défini par:

$$\boldsymbol{C(X,Y) = trace(V_{XY}) = \sum_{j=1}^{p} cov(x_j, y_j)} \tag{6.1}$$

où $\boldsymbol{x_j}$ et $\boldsymbol{y_j}$ sont respectivement la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{X}$ et de $\boldsymbol{Y}$.

La version standardisée de cet indice nous permet de définir l'indice de corrélation multivariée:

$$\boldsymbol{\rho(X,Y) = \frac{trace(V_{XY})}{\sqrt{trace(V_X)}\sqrt{trace(V_Y)}} = \frac{trace(X^\top Y)}{\sqrt{||X||}\sqrt{||Y||}}} \tag{6.2}$$

L'indice de corrélation multivariée varie entre -1 et 1. Il est égal à -1 ou 1 si la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{X}$ est colinéaire à la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{Y}$ $\boldsymbol{(j = 1, 2, ..., p)}$. Il vaut 0 si la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{X}$ est orthogonale à la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{Y}$ $\boldsymbol{(j = 1, 2, ..., p)}$.

Un autre indice permettant de mesurer le degré de liaison entre les

variables de deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$ est l'indice noté $\boldsymbol{CovV}$ [39]. Contrairement à l'indice de corrélation multivariée, cet indice ne requiert pas que les deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$ portent sur les mêmes variables. Il est défini par:

$$\boldsymbol{CovV(Y,X) = trace(V_{YX}V_{XY}) = \frac{1}{n^2} trace(Y^\top XX^\top Y)}. \quad (6.3)$$

$\boldsymbol{CovV(Y,X)}$ exprime la covariation totale entre les tableaux $\boldsymbol{Y}$ et $\boldsymbol{X}$. En remarquant que $\boldsymbol{trace(Y^\top XX^\top Y) = trace(XX^\top YY^\top)}$ et en utilisant les relations $\boldsymbol{XX^\top = \sum_{i=1}^{p} x_i x_i^\top}$ et $\boldsymbol{YY^\top = \sum_{j=1}^{q} y_j y_j^\top}$, avec $\boldsymbol{x_i}$ et $\boldsymbol{y_j}$, la $\boldsymbol{i^{ème}}$ variable de $\boldsymbol{X}$ et la $\boldsymbol{j^{ème}}$ variable de $\boldsymbol{Y}$ respectivement, nous pouvons déduire la propriété suivante:

$$\boldsymbol{CovV(Y,X) = \sum_{i=1}^{p}\sum_{j=1}^{q} cov^2(x_i, y_j)}. \quad (6.4)$$

Ainsi, il apparaît que $\boldsymbol{CovV(Y,X)}$ reflète la force de la liaison des variables de $\boldsymbol{Y}$ avec celles de $\boldsymbol{X}$; la force de la liaison étant mesurée par la covariance. Dans la suite, nous référons à $\boldsymbol{CovV(Y,X)}$ comme l'indice de covariation entre $\boldsymbol{X}$ et $\boldsymbol{Y}$.

En particulier, nous avons:

$$\boldsymbol{CovV(X,X) = \sum_{i=1}^{p}\sum_{l=1}^{p} cov^2(x_i, x_l)}. \quad (6.5)$$

Nous notons $\boldsymbol{CovV(X,X)}$ par $\boldsymbol{VarV(X)}$. Cet indice reflète la force de la liaison entre les variables de $\boldsymbol{X}$.

Comme $\boldsymbol{CovV(Y,X) = trace(V_{YX}V_{XY})}$, nous avons:

$\boldsymbol{CovV(Y,X)} = \sum_l \boldsymbol{\mu_l}$, où $\boldsymbol{\mu_l}$ sont les valeurs propres de la matrice $\boldsymbol{V_{YX}V_{XY}}$. De même, nous avons $\boldsymbol{VarV(X)} = \sum_i \boldsymbol{\lambda_i^2}$, où $\boldsymbol{\lambda_i}$ sont les valeurs propres de la matrice $\boldsymbol{V_X}$. Ainsi, il apparaît que $\boldsymbol{VarV}$ est intimement lié à l'inertie du tableau $\boldsymbol{X}$ qui est, rappelons-le, égale à $\sum_i \boldsymbol{\lambda_i}$.

La version standardisée de $\boldsymbol{CovV}$ est donnée par le coefficient $\boldsymbol{RV}$:

$$\begin{aligned} \boldsymbol{RV(X,Y)} &= \frac{\boldsymbol{CoV(X,Y)}}{\sqrt{\boldsymbol{VarV(X)}}\sqrt{\boldsymbol{VarV(Y)}}} \\ &= \frac{\boldsymbol{trace(Y^\top XX^\top Y)}}{\sqrt{\boldsymbol{trace((X^\top X)^2)trace((Y^\top Y)^2)}}}. \end{aligned} \tag{6.6}$$

Ce coefficient varie de 0 à 1. Il est égal à 0 si toutes les variables de $\boldsymbol{X}$ sont orthogonales à toutes les variables de $\boldsymbol{Y}$. Il est égal à 1 si les variables de $\boldsymbol{X}$ et celles de $\boldsymbol{Y}$ peuvent être ajustées par une rotation et multiplication par un scalaire [45].

L'étude de l'indice $\boldsymbol{CovV(X,Y)}$ ainsi que sa version standardisée, $\boldsymbol{RV(X,Y)}$, est particulièrement importante pour MB-WCov et la régression MB-PLS car, comme nous l'avons vu, ces méthodes visent à restituer de proche en proche la somme des indices de covariation entre, d'un côté, les tableaux prédictifs et, de l'autre côté, le tableau à prédire. Nous allons, par la suite, exhiber de nouvelles propriétés de $\boldsymbol{CovV}$ qui sont davantage en rapport avec le problème de prédiction.

Considérons une variable $\boldsymbol{y}$ centrée qui peut être manifeste (c'est-à-dire, observable) ou latente. Nous avons:

$$\boldsymbol{CovV(y,X)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{cov^2(y,x_i)}. \tag{6.7}$$

Nous savons aussi que:

$$\begin{aligned} \boldsymbol{CovV(y,X)} &= \frac{\mathbf{1}}{\boldsymbol{n^2}}\boldsymbol{trace}(\boldsymbol{y^\top XX^\top y}) \\ &= \frac{\mathbf{1}}{\boldsymbol{n^2}}\boldsymbol{y^\top XX^\top y} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{y^\top t} = \boldsymbol{cov(y,t)} \end{aligned} \tag{6.8}$$

où $\boldsymbol{t} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{XX^\top y}$. De là, nous pouvons faire deux remarques importantes. La première remarque est que $\boldsymbol{t}$ est proportionnelle à la première composante PLS de $\boldsymbol{y}$ sur $\boldsymbol{X}$. En effet,

$$\boldsymbol{t} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{XX^\top y} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{x_i x_i^\top y} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{cov(x_i, y) x_i}. \tag{6.9}$$

La deuxième remarque est que, d'après le théorème de Cauchy-Schwarz, la quantité $\boldsymbol{CovV(y,X) = cov(y,t)}$ atteint son maximum lorsque $\boldsymbol{y}$ et $\boldsymbol{t}$ sont colinéaires. Cela signifie, en particulier, que $\boldsymbol{y}$ est complètement prédite par sa première composante PLS.
Par ailleurs, le fait que $\boldsymbol{y}$ et $\boldsymbol{t}$ soient colinéaires se traduit par:

$$\boldsymbol{t} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{XX^\top y} = \boldsymbol{\alpha y} \tag{6.10}$$

où $\boldsymbol{\alpha}$ est un scalaire. Cela signifie que $\boldsymbol{y}$ est vecteur propre de $\frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{XX^\top}$ ou, en d'autres termes, $\boldsymbol{y}$ est une composante principale de $\boldsymbol{X}$ (en tant que vecteur propre de $\frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{XX^\top}$). De manière plus précise, la quantité $\boldsymbol{cov(y,t)}$ atteint son maximum, lorsque $\boldsymbol{y}$ est la première composante principale de $\boldsymbol{X}$.

En résumé, il apparaît que $\boldsymbol{CovV(y,X)}$ reflète la capacité de la variable $\boldsymbol{y}$ à être prédite par la première composante PLS de $\boldsymbol{y}$ sur $\boldsymbol{X}$. $\boldsymbol{CovV(y,X)}$

reflète également la proximité de la première composante PLS de $\boldsymbol{y}$ sur $\boldsymbol{X}$ avec la première composante principale de $\boldsymbol{X}$. Nous retrouvons là, le double objectif de PLS consistant à prédire $\boldsymbol{y}$, d'un côté, et à restituer la variabilité de $\boldsymbol{X}$, d'un autre côté.

Considérons maintenant deux tableaux $\boldsymbol{X}$ et $\boldsymbol{Y}$ supposés être centrés. Nous avons:

$$\boldsymbol{CovV(X,Y)} = \frac{\mathbf{1}}{\boldsymbol{n^2}}\boldsymbol{trace(Y^\top X X^\top Y)} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{trace(Y^\top U_{XY})} \quad (6.11)$$

où $\boldsymbol{U_{XY}} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{X X^\top Y}$.

Ainsi:

$$\boldsymbol{CovV(X,Y)} = \boldsymbol{C(X,U_{XY})} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{cov}(\boldsymbol{y_j}, \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{X X^\top y_j}). \quad (6.12)$$

De là, nous déduisons que $\boldsymbol{CovV(X,Y)}$ mesure la capacité des variables $\boldsymbol{y_j}$ à être prédites par leurs premières composantes PLS respectives sur $\boldsymbol{X}$.

6.2.2 Données multiblocs "K+1"

Considérons $\boldsymbol{K}$ tableaux $\boldsymbol{X_1, X_2, ..., X_K}$, chacun de dimension $\boldsymbol{n \times p_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$ et un tableau $\boldsymbol{Y}$ $\boldsymbol{(n \times q)}$. Tous ces tableaux portent sur les mêmes $\boldsymbol{n}$ individus et sont supposés être centrés et réduits (si nécessaire). De plus, les tableaux $\boldsymbol{X_k}$ sont pré-traités en divisant chaque tableau par sa norme. Dans la suite, nous désignons par $\boldsymbol{V_{kY}}$ la matrice de covariance entre $\boldsymbol{X_k}$ et $\boldsymbol{Y}$, et par $\boldsymbol{V_{Yk}} = \boldsymbol{V_{kY}^\top}$ la matrice de covariance entre $\boldsymbol{Y}$ et $\boldsymbol{X_k}$.

Régression MB-PLS

Dans l'objectif de restituer la quantité:

$$\sum_{k=1}^{K} \boldsymbol{CovV(Y, X_k)} = \sum_{k=1}^{K} \boldsymbol{trace(V_{Yk} V_{kY})} \tag{6.13}$$

nous considérons le problème d'optimisation consistant à déterminer un vecteur $\boldsymbol{\nu}$ $(||\boldsymbol{\nu}|| = \mathbf{1})$ et un coefficient $\boldsymbol{\lambda}$, de manière à minimiser le critère:

$$\sum_{k=1}^{K} ||\boldsymbol{V_{Yk} V_{kY}} - \boldsymbol{\lambda \nu \nu^\top}||^2 = \sum_{k=1}^{K} ||\boldsymbol{V_{Yk} V_{kY}}||^2 - \mathbf{2}\boldsymbol{\lambda} \sum_{k=1}^{K} \boldsymbol{\nu^\top V_{Yk} V_{kY} \nu} + \boldsymbol{K\lambda^2}. \tag{6.14}$$

En dérivant le second membre de l'équation (6.14) par rapport à $\boldsymbol{\lambda}$ et en annulant cette dérivée, nous obtenons: $\boldsymbol{\lambda} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\nu^\top V_{Yk} V_{kY} \nu}$.
En remplaçant $\boldsymbol{\lambda}$ par sa valeur dans l'équation (6.14), nous obtenons:

$$\sum_{k=1}^{K} ||\boldsymbol{V_{Yk} V_{kY}} - \boldsymbol{\lambda \nu \nu^\top}||^2 = \sum_{k=1}^{K} ||\boldsymbol{V_{Yk} V_{kY}}||^2 - \frac{\mathbf{1}}{\boldsymbol{K}} \left(\sum_{k=1}^{K} \boldsymbol{\nu^\top V_{Yk} V_{kY} \nu} \right)^{\mathbf{2}} \tag{6.15}$$

Il vient alors que, minimiser cette dernière équation par rapport au vecteur $\boldsymbol{\nu}$ revient à maximiser $\frac{1}{K} \left(\sum_{k=1}^{K} \boldsymbol{\nu^\top V_{Yk} V_{kY} \nu} \right)^2$, qui est équivalent à la maximisation de $\boldsymbol{\nu^\top} \left(\sum_{k=1}^{K} \boldsymbol{V_{Yk} V_{kY}} \right) \boldsymbol{\nu}$. S'agissant de la maximisation d'une forme quadratique, la solution est donnée en considérant $\boldsymbol{\nu}$, comme étant le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{V_{Yk} V_{kY}} = \sum_{k=1}^{K} \boldsymbol{Y^\top X_k X_k^\top Y}$ associé à la plus grande valeur propre. Ainsi, nous sommes conduits à la régression MB-PLS.

Nous pouvons remarquer que:

$$\boldsymbol{\nu}^{\top}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{Y}^{\top}\boldsymbol{X_k}\boldsymbol{X_k}^{\top}\boldsymbol{Y}\boldsymbol{\nu}=\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{u}^{\top}\boldsymbol{t_k}=\boldsymbol{u}^{\top}\boldsymbol{t} \tag{6.16}$$

où $\boldsymbol{u}=\boldsymbol{Y}\boldsymbol{\nu}$, $\boldsymbol{t_k}=\boldsymbol{X_k}\boldsymbol{X_k}^{\top}\boldsymbol{u}$ et $\boldsymbol{t}=\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{t_k}$. Nous retrouvons, là, les critères qui nous avaient permis d'introduire la régression MB-PLS dans le chapitre 4 dévolu aux méthodes supervisées.

Nous pouvons aussi remarquer que:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{u}^{\top}\boldsymbol{t_k}=\boldsymbol{n}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{CovV}(\boldsymbol{u},\boldsymbol{X_k}). \tag{6.17}$$

A partir de là, nous avions proposé l'indice:

$$\frac{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{CovV}(\boldsymbol{u},\boldsymbol{X_k})}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{CovV}(\boldsymbol{X_k},\boldsymbol{Y})}=\frac{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{u}^{\top}\boldsymbol{t_k}}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{CovV}(\boldsymbol{X_k},\boldsymbol{Y})}=\frac{\boldsymbol{u}^{\top}\boldsymbol{t}}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\boldsymbol{CovV}(\boldsymbol{X_k},\boldsymbol{Y})} \tag{6.18}$$

pour mesurer l'importance de la composante $\boldsymbol{t}$ ($\boldsymbol{u}$ et $\boldsymbol{t_k}$) à restituer la covariation entre $\boldsymbol{X_k}$ $(\boldsymbol{k=1,\ 2,\ ...,\ K})$ et $\boldsymbol{Y}$.

MB-WCov

Dans le critère (6.14) de la section précédente, nous avons considéré que le coefficient $\boldsymbol{\lambda}$ était constant. Nous proposons de relaxer cette hypothèse en supposant que le coefficient $\boldsymbol{\lambda}$ varie d'un tableau à un autre. Par conséquent, nous considérons le problème de minimisation suivant:

$$\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}||\boldsymbol{V_{Yk}}\boldsymbol{V_{kY}}-\boldsymbol{\lambda_k}\boldsymbol{\nu}\boldsymbol{\nu}^{\top}||^{\boldsymbol{2}} \tag{6.19}$$

sous la contrainte que $||\boldsymbol{\nu}|| = \mathbf{1}$.

Nous avons:

$$\sum_{k=1}^{K} ||\boldsymbol{V_{Yk}V_{kY}} - \boldsymbol{\lambda_k \nu\nu^\top}||^2 = \sum_{k=1}^{K} ||\boldsymbol{V_{Yk}V_{kY}}||^2 - \mathbf{2}\sum_{k=1}^{K} \boldsymbol{\lambda_k \nu^\top V_{Yk}V_{kY}\nu} + \sum_{k=1}^{K} \boldsymbol{\lambda_k^2}. \tag{6.20}$$

La dérivée du second membre de cette équation par rapport à $\boldsymbol{\lambda_k}$ s'annule pour $\boldsymbol{\lambda_k} = \boldsymbol{\nu^\top V_{Yk}V_{kY}\nu}$. Si nous remplaçons $\boldsymbol{\lambda_k}$ par sa valeur dans l'équation (6.20), nous obtenons:

$$\sum_{k=1}^{K} ||\boldsymbol{V_{Yk}V_{kY}} - \boldsymbol{\lambda_k \nu\nu^\top}||^2 = \sum_{k=1}^{K} ||\boldsymbol{V_{Yk}V_{kY}}||^2 - \boldsymbol{\nu^\top} \left(\sum_{k=1}^{K} \boldsymbol{\lambda_k V_{Yk}V_{kY}} \right) \boldsymbol{\nu}. \tag{6.21}$$

Il est clair que, minimiser le critère (6.19) revient à maximiser: $\boldsymbol{\nu^\top} \left(\sum_{k=1}^{K} \boldsymbol{\lambda_k V_{Yk}V_{kY}} \right) \boldsymbol{\nu}$. Pour des valeurs fixées de $\boldsymbol{\lambda_k}$, le vecteur optimal, $\boldsymbol{\nu}$, est donné par le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k V_{Yk}V_{kY}}$ associé à la plus grande valeur propre. Pour $\boldsymbol{\nu}$ fixé, $\boldsymbol{\lambda_k = \nu^\top V_{Yk}V_{kY}\nu}$.

Ainsi, la minimisation du critère (6.19) par rapport à $\boldsymbol{\nu}$ s'effectue à partir d'un algorithme des moindres carrés alternés:

0. $\boldsymbol{\lambda_k = 1}$ pour $\boldsymbol{(k = 1, 2, ..., K)}$;

1. $\boldsymbol{\nu}$ est le vecteur propre de $\sum_{k=1}^{K} \boldsymbol{\lambda_k V_{Yk}V_{kY}}$ associé à la plus grande valeur propre;

2. $\boldsymbol{\lambda_k = \nu^\top V_{Yk}V_{kY}\nu}$;

3. Réitérer le même processus à partir de l'étape 1, jusqu'à convergence.

Nous pouvons remarquer que le critère à maximiser peut également

s'écrire sous la forme:

$$\boldsymbol{\nu}^\top \sum_{k=1}^{K} \boldsymbol{\lambda}_k \boldsymbol{Y}^\top \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{Y} \boldsymbol{\nu} = \boldsymbol{n} \sum_{k=1}^{K} \boldsymbol{\lambda}_k \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t}_k) = \boldsymbol{n} \times \boldsymbol{cov}(\boldsymbol{u}, \boldsymbol{t}) \quad (6.22)$$

où $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{\nu}$, $\boldsymbol{t}_k = \boldsymbol{X}_k \boldsymbol{X}_k^\top \boldsymbol{u}$, $\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda}_k \boldsymbol{t}_k$.

Nous retrouvons les relations entre $\boldsymbol{u}$, $\boldsymbol{t}_k$ et $\boldsymbol{t}$ qui nous ont permis de définir MB-WCov dans le chapitre 4.

6.2.3 Cas particuliers de la méthode MB-WCov

Pour introduire la méthode MB-WCov, nous avons considéré un tableau $\boldsymbol{Y}$ que nous cherchons à prédire à l'aide $\boldsymbol{K}$ blocs de variables. Nous considérons à présent deux cas particuliers. Le premier cas est celui où chaque tableau est réduit à une seule variable (c'est-à-dire, $\boldsymbol{Y} = [\boldsymbol{y}]$ et $\boldsymbol{X}_k = [\boldsymbol{x}_k],\ \boldsymbol{k} = \boldsymbol{1},\ \boldsymbol{2},\ \boldsymbol{...},\ \boldsymbol{K}$).

La variable $\boldsymbol{u} = \boldsymbol{y}\boldsymbol{\nu}$ se ramène à $\boldsymbol{u} = \boldsymbol{y}$. Les composantes par bloc, $\boldsymbol{t}_k = \boldsymbol{x}_k \boldsymbol{x}_k^\top \boldsymbol{y} = \boldsymbol{n} \times \boldsymbol{cov}(\boldsymbol{y}, \boldsymbol{x}_k)\boldsymbol{x}_k$ $(\boldsymbol{k} = \boldsymbol{1}, \boldsymbol{2}, \boldsymbol{...}, \boldsymbol{K})$. En appliquant la régression PLS multiblocs, nous obtenons:

$$\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{t}_k = \boldsymbol{n} \sum_{k=1}^{K} \boldsymbol{cov}(\boldsymbol{y}, \boldsymbol{x}_k)\boldsymbol{x}_k \quad (6.23)$$

Il est clair que nous retrouvons la régression PLS1. De même, en appliquant la méthode MB-WCov, nous obtenons:

$$\boldsymbol{t} = \sum_{k=1}^{K} \boldsymbol{\lambda}_k \boldsymbol{t}_k = \boldsymbol{n}^2 \sum_{k=1}^{K} \boldsymbol{cov}^3(\boldsymbol{y}, \boldsymbol{x}_k)\boldsymbol{x}_k \quad (6.24)$$

6.2.4 Sparse MB-WCov

Un des intérêts de la méthode MB-WCov est qu'elle exhibe explicitement des poids spécifiques, $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1, 2, ..., K)}$, qui indiquent l'importance de chaque tableau prédictif pour la détermination des composantes globales. Cependant, il arrive très souvent d'avoir des tableaux prédictifs avec une contribution insignifiante. Dans le soucis de faciliter l'interprétation des résultats, nous proposons de ramener ces contributions insignifiantes à la valeur nulle. Cela suggère une nouvelle stratégie d'analyse que nous désignons par Sparse MB-WCov et dont le principe est le suivant. Nous considérons un tableau $\boldsymbol{Y}$ $\boldsymbol{(n \times q)}$ à prédire à partir de $\boldsymbol{K}$ tableaux $\boldsymbol{X_1, X_2, ..., X_K}$, chacun de dimension $\boldsymbol{n \times p_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$. La variable latente associée à $\boldsymbol{Y}$ est $\boldsymbol{u = Y\nu}$ $\boldsymbol{(||\nu|| = 1)}$. Les composantes par bloc associées à $\boldsymbol{X_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$ sont données par $\boldsymbol{t_k = X_k X_k^\top u}$. Sparse MB-WCov consiste à trouver un vecteur normé $\boldsymbol{\nu}$ (et, par conséquent, $\boldsymbol{u}$ et $\boldsymbol{t_k}$) et des coefficients $\boldsymbol{\lambda_k}$ $\left(\boldsymbol{\sum_{k=1}^{K} \lambda_k^2 = 1}\right)$ qui maximisent le critère:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k sup\left[cov(u,\ t_k) - \tau,\ 0\right]}. \tag{6.25}$$

Dans ce critère, $\boldsymbol{\tau}$ est un seuil fixé par l'utilisateur. Il représente le paramètre de sparsité. Pour la résolution de ce critère de maximisation, nous pouvons remarquer que pour $\boldsymbol{u}$ fixé, nous avons $\boldsymbol{t_k = X_k X_k^\top u}$. D'après l'inégalité de Cauchy-Schwarz les valeurs $\boldsymbol{\lambda_k}$ qui maximisent le critère considéré sont données par:

$$\boldsymbol{\lambda_k = \alpha sup\left[cov(u,\ t_k) - \tau,\ 0\right]} \tag{6.26}$$

où $\boldsymbol{\alpha}$ est une constante que nous pouvons calculer à partir de la contrainte imposée sur $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$. Il vient: $\boldsymbol{\alpha} = \frac{1}{\sqrt{\sum_{l=1}^{K}\{sup[cov(u,\ t_l)-\tau,\ 0]\}^2}}$.
Pour $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$ fixés, le critère (6.25) peut s'écrire:

$$\sum_{k=1}^{K} \boldsymbol{\lambda_k} \left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}\right]. \tag{6.27}$$

En effet:

Si $\boldsymbol{sup}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau},\ \boldsymbol{0}\right] = \boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}$, alors $\boldsymbol{\lambda_k} = \boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}$ et donc $\boldsymbol{\lambda_k sup}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau},\ \boldsymbol{0}\right] = \boldsymbol{\lambda_k}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}\right]$.

Si, par contre, $\boldsymbol{sup}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau},\ \boldsymbol{0}\right] = \boldsymbol{0}$, alors $\boldsymbol{\lambda_k} = \boldsymbol{0}$ et nous obtenons $\boldsymbol{\lambda_k sup}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau},\ \boldsymbol{0}\right] = \boldsymbol{\lambda_k}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}\right]\ \ \boldsymbol{(= 0)}$.

Par conséquent, pour $\boldsymbol{\lambda_k}$ $\boldsymbol{(k = 1,\ 2,\ ...,\ K)}$ fixés, nous sommes conduits à maximiser le critère:

$$\begin{aligned}
\sum_{k=1}^{K} \boldsymbol{\lambda_k}\left[\boldsymbol{cov}(\boldsymbol{u},\ \boldsymbol{t_k}) - \boldsymbol{\tau}\right] &= \boldsymbol{cov}\left(\boldsymbol{u}, \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}\right) - \boldsymbol{\tau}\sum_{k=1}^{K} \boldsymbol{\lambda_k} \\
&= \frac{\boldsymbol{1}}{\boldsymbol{n}} \boldsymbol{u}^\top \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k} - \boldsymbol{\tau}\sum_{k=1}^{K} \boldsymbol{\lambda_k} \\
&= \frac{\boldsymbol{1}}{\boldsymbol{n}} \boldsymbol{\nu}^\top \boldsymbol{Y}^\top \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k} - \boldsymbol{\tau}\sum_{k=1}^{K} \boldsymbol{\lambda_k}.
\end{aligned} \tag{6.28}$$

Par application de l'inégalité de Cauchy-Schwarz, le vecteur $\boldsymbol{\nu}$ qui maximise cette quantité est donné par $\boldsymbol{\nu} = \frac{\boldsymbol{Y}^\top \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}}{||\boldsymbol{Y}^\top \sum_{k=1}^{K} \boldsymbol{\lambda_k t_k}||}$.
En résumé, l'algorithme permettant d'exécuter la méthode Sparse MB-WCov est ainsi donné par:

0. Choix de manière aléatoire de $\boldsymbol{\nu}$ $\boldsymbol{(||\nu|| = 1)}$;

1. $\boldsymbol{u} = \boldsymbol{Y\nu}$ et $\boldsymbol{t_k} = \boldsymbol{X_k X_k^\top u}$;

2. $\boldsymbol{\lambda_k = sup\,[cov(u,\ t_k) - \tau,\ 0]}$;

3. $\boldsymbol{\lambda_k = \lambda_k / \sqrt{\sum_{l=1}^{K} \lambda_l^2}}$, avec $\boldsymbol{\sum_{l=1}^{K} \lambda_l^2 \neq 0}$;

4. $\boldsymbol{t = \sum_{k=1}^{K} \lambda_k t_k}$: composante globale;

5. $\boldsymbol{\nu = Y^\top t / ||Y^\top t||}$;

6. Réitération de la procédure à partir de l'étape 1, jusqu'à convergence.

Pour le choix du paramètre de sparsité, nous proposons de commencer par $\boldsymbol{\tau = 0}$ (c'est-à-dire, MB-WCov) et incrémenter $\boldsymbol{\tau}$ par pas fixé (exemple, pas=0.01). Pour chaque valeur de $\boldsymbol{\tau}$, nous évaluons d'un côté, l'inertie de $\boldsymbol{Y}$ restituée par la composante globale $\boldsymbol{t}$ que nous souhaitons préserver aussi grande que possible. D'un autre côté, nous calculons le nombre de paramètres $\boldsymbol{\lambda_k}$ mis à 0. Nous souhaitons que ce nombre soit le plus grand possible. Ainsi le choix de $\boldsymbol{\tau}$ doit réaliser un compromis entre la perte d'inertie de $\boldsymbol{Y}$ et le nombre de paramètres $\boldsymbol{\lambda_k}$ mis à 0. Par ailleurs, comme

$$\boldsymbol{cov(u,t) = \frac{1}{n} u^\top t = \frac{1}{n} \nu^\top Y^\top t \leq \frac{1}{n} ||Y||} \tag{6.29}$$

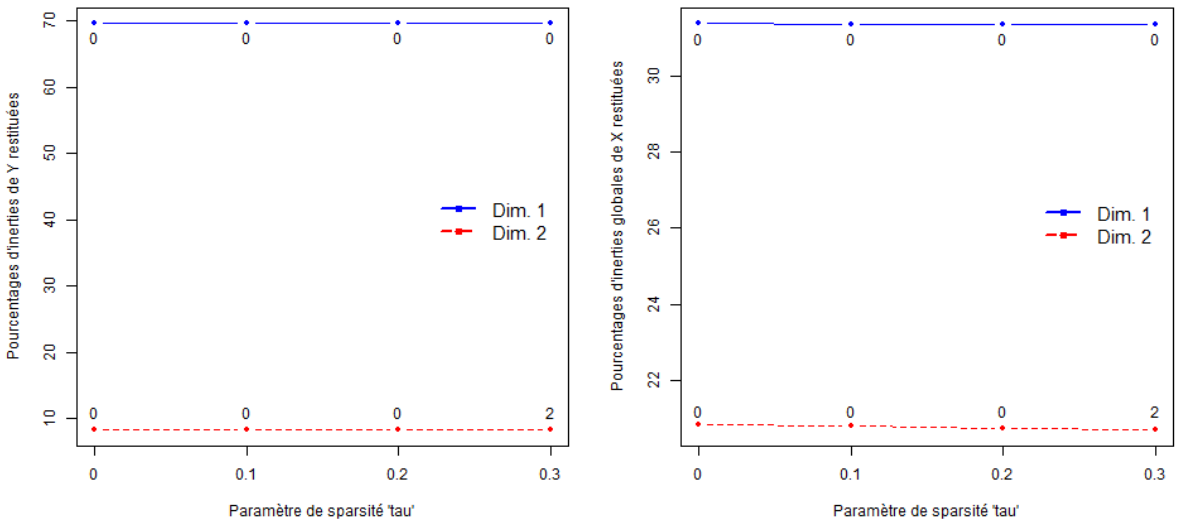
nous en déduisons qu'il suffit de faire varier $\boldsymbol{\tau}$ entre 0 et $\frac{1}{n}||\boldsymbol{Y}||$.

6.3 Illustrations

6.3.1 Illustration de Sparse MB-WCov

Pour illustrer la méthode Sparse MB-WCov, nous utilisons les données de pommes de terre décrites dans le chapitre précédent [30]. Nous disposons ainsi d'un tableau de données sensorielles (tableau à prédire) et

de huit tableaux prédictifs: "chemical", "compression", "cpmg", "nir", cpmg.cooked", "nir.cooked", "fid" et "fid.cooked". Nous pré-traitons ces différents tableaux puis nous appliquons la méthode Sparse MB-WCov. La figure 6.1 présente les pourcentages d'inerties de $\boldsymbol{Y}$ et les pourcentages d'inerties globales des tableaux prédictifs restituées par les deux premières dimensions de la méthode MB-WCov ($\boldsymbol{\tau = 0}$) et Sparse MB-WCov ($\boldsymbol{\tau = 0.1}$, $\boldsymbol{\tau = 0.2}$ et $\boldsymbol{\tau = 0.3}$). De plus, dans ces figures nous indiquons le nombre de tableaux dont les contributions sont jugées insignifiantes pour chaque valeur du paramètre $\boldsymbol{\tau}$. Comme nous pouvons le constater, il semble qu'aucun des tableaux prédictifs n'a une contribution insignifiante pour la détermination de la première dimension. Quant à la deuxième dimension, les contributions de deux tableaux ("fid" et "fid.cooked") sont mises à zéro lorsque le paramètre de sparsité est égal à $\boldsymbol{\tau = 0.3}$. De ces figures, nous pouvons aussi noter que les inerties de $\boldsymbol{Y}$ restituées et les inerties globales des tableaux prédictifs restituées par les deux premières dimensions restent pratiquement invariantes lorsqu'on incrémente le paramètre de sparsité entre 0 et 0.3.



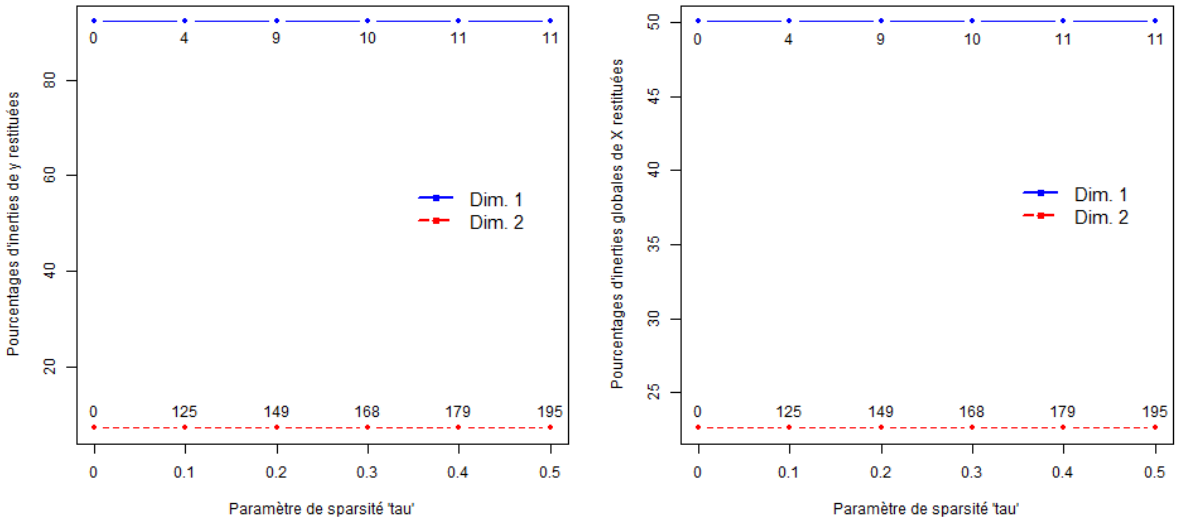
(a) *(b)*

Fig. 6.1: Pourcentages d'inerties de $\boldsymbol{Y}$ et pourcentages d'inerties globales de $\boldsymbol{X}$ restituées par les deux premières dimensions de la méthode Sparse MB-WCov pour différentes valeurs de $\boldsymbol{\tau}$ et nombre de tableaux dont les contributions sont mises à zéro pour chaque dimension et chaque valeur de $\boldsymbol{\tau}$.

6.3.2 Illustration des cas particuliers de la méthode Sparse MB-WCov

La première illustration concerne le cas où tous les tableaux sont univariés. Pour cela, nous utilisons les données "yarn" disponibles dans le package "pls" [56]. L'objectif est de prédire les densités des fils de poly(ethylene terephthalate) (PET) à partir des données spectrales (spectroscopie par proche infrarouge "NIR"). Au total, nous avons 268 tableaux prédictifs. Après avoir pré-traité ces tableaux, nous avons appliqué la méthode Sparse MB-WCov. Les pourcentages d'inerties restituées par le tableau à prédire, ainsi que les pourcentages d'inerties globales restituées par les tableaux prédictifs sont donnés sur la figure 6.2. Dans cette figure, nous indiquons également le nombre de tableaux dont la contribution est mise à zéro pour chaque dimension et pour chaque valeur du paramètre de sparsité

$\boldsymbol{\tau}$. L'inertie restituée par les deux premières dimensions reste constante lorsqu'on incrémente le paramètre de sparsité. De plus, nous pouvons noter que le nombre de tableaux dont la contribution est mise à zéro augmente avec l'incrémentation du paramètre de sparsité.

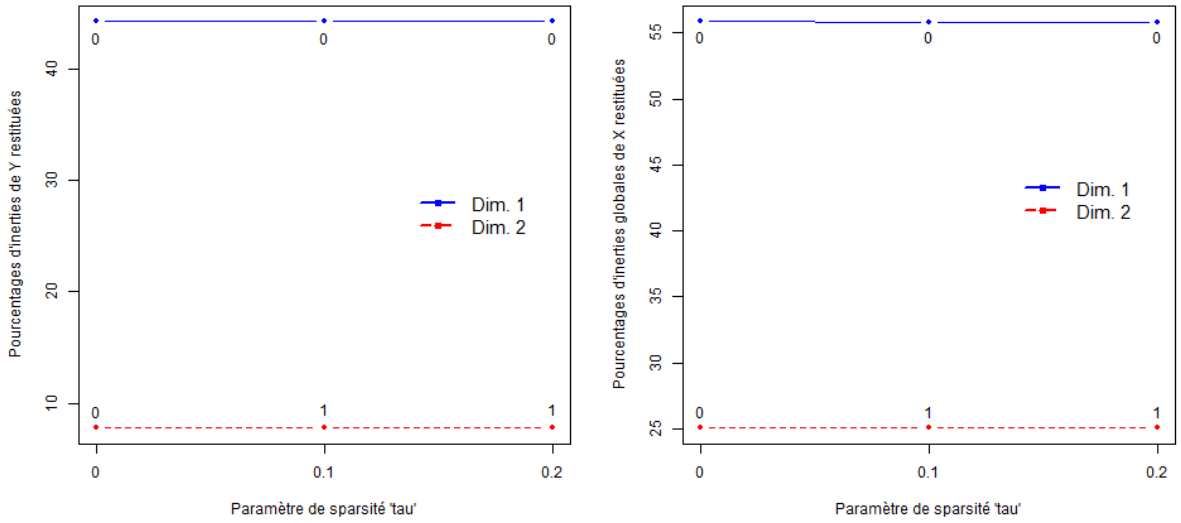


(a) *(b)*

Fig. 6.2: Pourcentages d'inerties de $\boldsymbol{Y}$ et pourcentages d'inerties globales de $\boldsymbol{X}$ restituées par les deux premières dimensions du cas particulier de la méthode Sparse MB-WCov pour différentes valeurs de $\boldsymbol{\tau}$ et le nombre de tableaux dont les contributions sont mises à zéro pour chaque dimension et chaque valeur de $\boldsymbol{\tau}$.

Le deuxième cas particulier concerne la situation où le tableau à prédire est multivarié et les tableaux prédictifs sont univariés. Les données "oliveoil" du package "pls" [56] ont servi à illustrer cette situation. Ces données sont relatives à l'évaluation de six attributs sensoriels de 16 variétés d'huile d'olive à partir de cinq variables physico-chimiques. La figure 6.3 présente les pourcentages d'inerties de $\boldsymbol{Y}$ et les pourcentages d'inerties globales de $\boldsymbol{X}$ restituées par les deux premières dimensions du cas particulier de la méthode Sparse MB-WCov pour différentes valeurs de $\boldsymbol{\tau}$ et le nombre de

tableaux dont les contributions sont mises à zéro pour chaque dimension et chaque valeur de $\boldsymbol{\tau}$. Tout comme pour le cas particulier précédent, les pourcentages d'inerties restituées restent invariables pour chaque dimension. De plus, un seul tableau est mis à zéro sur la deuxième dimension lorsque $\boldsymbol{\tau} = \mathbf{0.1}$ et $\boldsymbol{\tau} = \mathbf{0.2}$.



(a) *(b)*

Fig. 6.3: Pourcentages d'inerties de $\boldsymbol{Y}$ et les pourcentages d'inerties globales de $\boldsymbol{X}$ restituées par les deux premières dimensions du cas particulier de la méthode Sparse MB-WCov pour différentes valeurs de $\boldsymbol{\tau}$ et le nombre de tableaux dont les contributions sont mises à zéro pour chaque dimension et chaque valeur de $\boldsymbol{\tau}$.

6.4 Conclusion

Dans un premier temps, nous avons exhibé de nouvelles propriétés de la méthode MB-WCov en rapport avec la prédiction puis nous avons développé cette méthode à partir d'une nouvelle formulation.

Dans un deuxième temps, nous avons proposé une version sparse pour la méthode MB-WCov et ses cas particuliers. L'application de ces versions

sparses sur des données réelles a montré que lorsque nous incrémentons le paramètre de sparsité, certains tableaux sont mis à zéro, indiquant que leur contribution est négligeable pour la dimension considérée.

7

Conclusion et perspectives

L'analyse des données structurées en plusieurs tableaux (données multi-blocs) a suscité un grand intérêt ces deux ou trois dernières décennies. Ceci se traduit par une offre considérable de méthodes d'analyse. Certaines de ces méthodes, proposées notamment par des chimiométriciens, sont basées sur des idées intuitives et n'ont pas toujours un fondement mathématique. De ce fait, il existe des méthodes qui posent des problèmes de convergence. En tout état de cause, il peut s'avérer difficile pour l'utilisateur de faire un choix éclairé d'une méthode qui serait plus appropriée à ses objectifs.

Nous avons proposé des démarches unificatrices aussi bien pour les méthodes non supervisées que pour les méthodes supervisées. Ceci nous a permis de structurer plusieurs méthodes en familles en soulignant leurs traits distinctifs. De plus, comme les méthodes abordées sont basées sur des constructions mathématiques et des critères d'optimisation clairs, il est possible de mieux les comparer entre elles, montrer la convergence des algorithmes, anticiper des propriétés caractéristiques telles que la vulnérabilité à la quasi-colinéarité. Parmi les méthodes exploratoires que nous avons

étudiées, nous pouvons citer l'ACCG, ComDim/H-PCA, l'ACP multiblocs et GCCA-V, la version pondérée de l'ACCG. De plus, en nous inspirant du principe de la régression par analyse des valeurs latentes, nous avons adapté la méthode non supervisée MB-PCA pour définir une méthode supervisée que nous avons appelé LR-MBPCA.

Pour ce qui concerne les méthodes supervisées, nous avons comparé les méthodes MB-RA et MB-PLS, puis nous avons proposé leurs versions pondérées que nous avons respectivement désigné par MB-WRA et MB-WCov. Nous avons également fourni des indices pour aider à mieux interpréter les résultats de ces différentes méthodes.

La méthode ComDim a été introduite en tant que méthode exploratoire il y a une vingtaine d'années (Qannari et al. [4]). Hanafi et al. [27] ont montré son lien étroit avec la méthode H-PCA. Elle a connu une certaine popularité parmi les utilisateurs [48–50]. Nous avons montré comment cette méthode s'intègre parfaitement dans le cadre de notre démarche unificatrice. Ceci a permis d'exhiber de nouvelles propriétés et d'étudier des cas particuliers comme ComDim-PCA ou ComDim-Quali.

La méthode MB-WCov est une nouvelle méthode supervisée qui a été développée dans le cadre de cette thèse. Nous croyons que c'est une version améliorée et corrigée de la méthode Multiblock Hierarchical PLS (MB-HPLS). Cette dernière méthode, bien que citée dans plusieurs articles, n'est pas vraiment opérationnelle car elle a de sérieux problèmes de convergence.

Le point commun de ComDim, d'un côté, et de MB-WCov, d'un autre côté, est d'exhiber de manière explicite des poids spécifiques associés aux différents tableaux. Nous avons mis à profit cette propriété pour définir des

versions "sparses" des deux méthodes citées. L'application de ces méthodes à des données réelles ou simulées a montré leur intérêt.

Plusieurs perspectives au travail accompli dans cette thèse semblent se dessiner. Dans un premier temps, il nous semble que la méthode MB-WCov mérite davantage de développements et mérite d'être davantage diffusée. La conception du package R nommé "MBAnalysis" va sûrement contribuer à mieux la faire connaître.

Un autre point important que nous n'avons pas eu le temps d'aborder concerne les analyses de tableaux reliés par des liens de causalité. Cela veut dire qu'il s'agit d'étudier un ensemble de tableaux en tenant compte des relations (ou liens) qui existent entre certains d'entre eux. Ce cadre est appelé "Path-modeling" [13, 57–59]. Là aussi, plusieurs méthodes ont été proposées et nous projetons d'explorer comment notre démarche générale pourrait être adaptée à ce contexte.

Le principe général de notre démarche tient en deux étapes clefs. Partant d'une variable latente, nous la projetons, dans une première étape, sur chacun des blocs de variables et, dans une deuxième étape, nous réalisons une synthèse de ces projections. En particulier, nous avons considéré deux manières de réaliser une projection sur les blocs de variables, $\boldsymbol{X}$. La première manière est de considérer l'opérateur $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. La deuxième manière est de considérer $\boldsymbol{W_X} = \boldsymbol{X}\boldsymbol{X}^\top$. $\boldsymbol{W_X}$ n'est pas à proprement parler un projecteur mais nous avons montré dans le chapitre 3 son lien étroit avec la première composante de la régression PLS. Nous avons souligné qu'il est possible d'établir un pont entre ces deux manières de procéder en considérant l'opérateur $\boldsymbol{Q_X} = \boldsymbol{X}\left[\gamma\boldsymbol{I} + (1-\gamma)\boldsymbol{X}^\top\boldsymbol{X}\right]^{-1}\boldsymbol{X}^\top$

où γ est un scalaire compris entre 0 et 1 (paramètre de régularisation) et $\boldsymbol{I}$ est la matrice identité. En considérant les matrices du type $\boldsymbol{Q_X}$, nous réalisons un continuum entre les familles de méthodes. De manière plus importante, il serait intéressant de considérer d'autres types d'opérateurs pour explorer d'autres types de liens entre tableaux de données. L'idée de base est que nos démarches aussi bien dans le contexte exploratoire que prédictif sont valables dès lors que nous considérons des opérateurs de "projection" qui soient semi-définis positifs. Nous pensons qu'avec des choix appropriés, nous pourrions appréhender des problèmes liés à la discrimination, investiguer des relations non linéaires, etc.

Annexe: Package "MBAnalysis"

L'objectif de cet annexe est de présenter les fonctionnalités du package R que nous appelons par "MBAnalysis" et que nous proposons pour la mise en application des méthodes de la deuxième famille (c'est-à-dire, les méthodes non vulnérables au problème de quasi-colinéarité) dans le logiciel R [61]. Dans un premier temps, nous implémentons les méthodes ComDim, ACP multiblocs (MB-PCA), Multiblock Weighted Covariate analysis (MB-WCov) et la régression PLS multiblocs (MB-PLS). Dans un deuxième temps, nous intégrerons les versions sparses de ces différentes méthodes ainsi que des variantes pour l'analyse des variables qualitatives. La figure .1 présente l'arborescence de ce package. Sur les pages suivantes, nous décrivons les données "ham" que nous utilisons pour des illustrations et présentons les méthodes ComDim, MB-PCA, MB-WCov et la régression MB-PLS à travers les fonctions ComDim, MBPCA, MBWCov et MBPLS respectivement. Pour chacune de ces méthodes, nous présentons également la fonction qui permet de faire des représentations graphiques (plot), d'afficher les résultats des différentes analyses (print) et de faire la synthèse des résultats (summary). De plus, pour les méthodes supervisées (régression MB-PLS et MB-WCov), nous présentons la fonction predict qui sert à prédire les valeurs d'un tableau $\boldsymbol{Y}$ à partir de nouvelles observations des blocs de variables explicatifs.

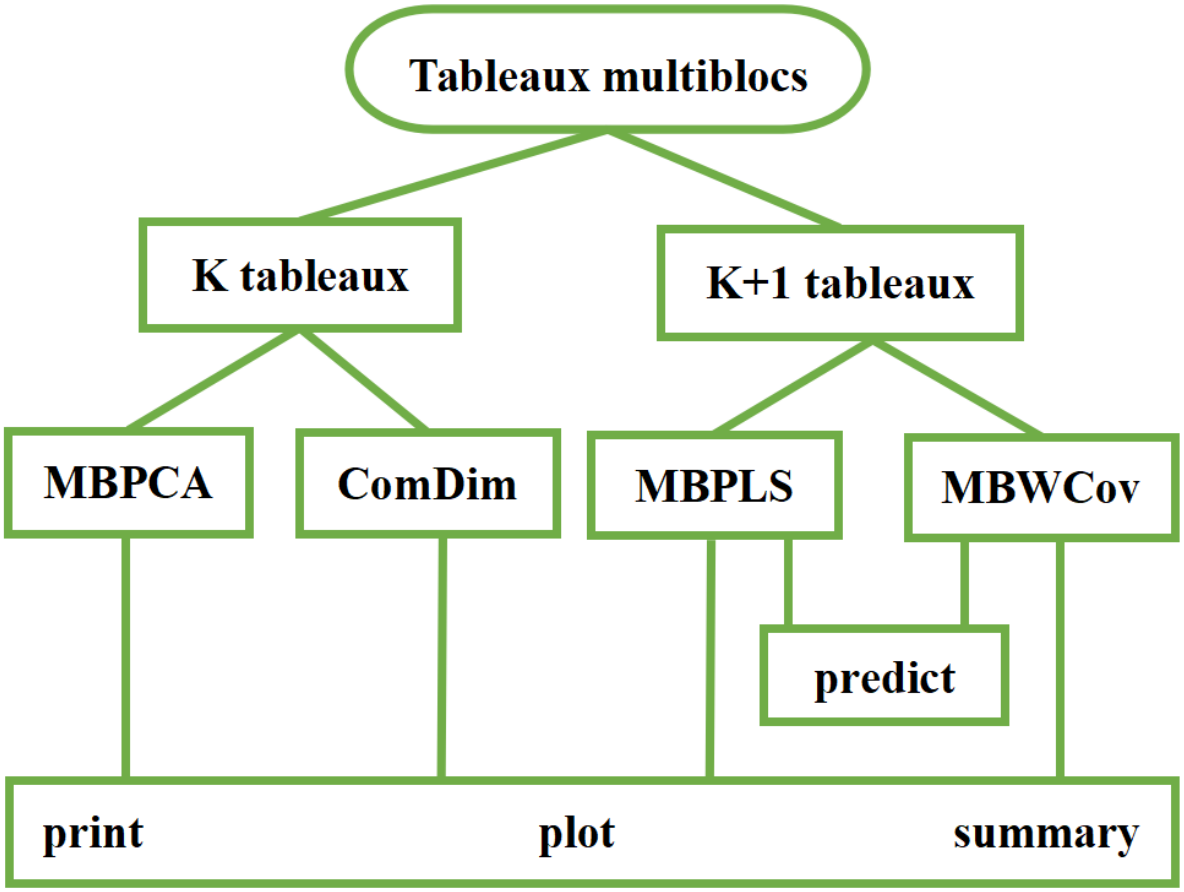


Fig. .1: Arborescence du package "MBAnalysis".

Title: Multiblock exploratory and predictive data analysis.

Version: 0.1.0.

Authors: Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

Maintainer: Essomanda Tchandao Mangamana <tchanesso@yahoo.fr>.

Imports: ggplot2, ggrepel.

Description: Exploratory and predictive methods for the analysis of several blocks of variables measured on the same individuals. The described methods are: Multiblock Principal Components Analysis (MB-PCA), ComDim, Multiblock Partial Least Squares (MB-PLS) regression and Multiblock Weighted Covariate analysis (MB-WCov).

E. Tchandao Mangamana, V. Cariou, E. Vigneau, R. Glèlè Kakaï, E.M. Qannari (2019) <doi:10.1016/j.chemolab.2019.103856>

E. Tchandao Mangamana, R. Glèlè Kakaï, E.M. Qannari (2021) <doi:10.1016/j.chemolab.2021.104388>.

Licence: GPL-3.

Encoding: UTF-8.

LazyData: true.

RoxygenNote: 7.1.1.

Repository: CRAN

Date/Publication: 2021-09-20 09:30:06 UTC

Ham data (ham)

Description

Case study pertaining to the sensory evaluation of eight American dry-cured ham products, performed by a panel of trained assessors.

Usage

data(ham)

Format

An object of class "list" with 8 products, 3 blocks of X variables (Flavor, Aroma, Texture) and 1 block of Y variables corresponding to hedonic measures:

X	dataframe of 8 products and 25 variables structured into 3 blocks: Flavor (11 variables), Aroma (8 variables) and Texture (6 variables).
Y	dataframe of 8 products and 6 vectors of hedonic values corresponding to consumers' segmentation.
group	vector indicating the number of variables per block.

References

M.D. Guardia, A.P. Aguiar, A. Claret, J. Arnau & L. Guerrero (2010). Sensory characterization of dry-cured ham using free-choice profiling. Food Quality and Preference, 21(1), 148-155. doi: 10.1016/j.foodqual.2009.08.014.

Common Dimensions analysis (ComDim)

Multiblock Principal Components Analysis (MB-PCA)

Description

ComDim / MB-PCA applied to a set of quantitative blocks of variables.

Usage

ComDim(X, group, algo = "eigen", ncompprint = NULL, scale = "none", option = "uniform", nstart = 10, threshold = 1e-08, plotgraph = TRUE, axes = c(1, 2))

MBPCA(X, group, algo = "eigen", ncompprint = NULL, scale = "none", option = "uniform", nstart = 10, threshold = 1e-08, plotgraph = TRUE, axes = c(1, 2))

Arguments

X	Block obtained by horizontally merging all the blocks of variables.
group	Vector indicating the number of variables per block.
algo	Type of algorithm to use. Either "eigen" (default) or "nipals".
ncompprint	Number of global components to print. By default (NULL), all the global components of the analysis are printed.

scale Type of standardization applied to the variables. Either "none" (default) or "sd". If scale="sd", each variable is divided by its standard deviation.

option Type of normalization applied to each block of variables (either "none" or "uniform"). If option="uniform" (default), each block of variables is divided by its Frobenius norm.

nstart Number of random initializations of the global component in case of nipals algorithm (by default 10).

threshold Value used to break the iterative loop (by default 1e-8).

plotgraph Boolean (TRUE/FALSE). If TRUE (default), graphs depicting saliences, scores of individuals, correlations of variables with the global components and contributions of blocks of variables to the determination of global components are displayed.

axes Vector of length two which specifies the global components to plot (by default the first two).

Value

Returns a list of the following elements:

components Numeric vector of length two that indicates the number of global components of the analysis and the number of global components to print.

optimalcrit Numeric vector that gives the optimal value of the criterion to be maximized for each dimension.

cumexplained	Two columns matrix of percentages of total inertia of the blocks of variables explained by the successive global components and their cumulative values.
explained.X	Matrix of percentages of inertia explained for each Xb block.
saliences	Matrix containing the specific weights of different blocks of variables on global components (returned by ComDim).
contrib	Matrix of contribution of each Xb block to the determination of global components.
T	Matrix of global components (scores of individuals).
C	Compromise matrix (unnormed global components).
globalcor	Matrix of correlation coefficients between the original variables and the global components.
cor.g.b	Array that gives the correlation of the global components with their respective block components.
T.b	Array that contains the matrices of block components.
blockcor	List of matrices of correlation coefficients between the original variables of each block and the block components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

References

E. Qannari, I. Wakeling, Ph. Courcoux, J.M. MacFie, Defining the underlying sensory dimensions, Food Qual. Prefer. (2000); 11 : 151-154.

E. Tchandao Mangamana, V. Cariou, E. Vigneau, R. Glèlè Kakaï, E.M. Qannari, Unsupervised multiblock data analysis: A unified approach and extensions, Chemometrics and Intelligent Laboratory Systems 194 (2019) 103856.

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

See also

print.ComDim plot.ComDim summary.ComDim

print.MBPCA plot.MBPCA summary.MBPCA

Example

```
data(ham)
X=ham$X
group=ham$group
res.comdim <- ComDim(X, group)
res.comdim
res.mbpca <- MBPCA(X, group)
res.mbpca
```

Multiblock Weighted Covariate analysis (MB-WCov)
Multiblock Partial Least Squares (MB-PLS) regression

Description

MB-WCov / MB-PLS regression applied to multiblock quantitative variables.

Usage

MBWCov(X, Y, group, algo = "eigen", ncompprint = NULL, scale = "none", scaleY = "none", option = "uniform", optionY = "uniform", nstart = 10, threshold = 1e-08, plotgraph = TRUE, axes = c(1, 2))

MBPLS(X, Y, group, algo = "eigen", ncompprint = NULL, scale = "none", scaleY = "none", option = "uniform", optionY = "uniform", nstart = 10, threshold = 1e-08, plotgraph = TRUE, axes = c(1, 2))

Arguments

X	Block obtained by horizontally merging all the explanatory blocks of variables.
Y	Response block of variables.
group	Vector indicating the number of variables in each explanatory block.

algo Type of algorithm to use. Either "eigen" (default) or "nipals".

ncompprint Number of global components to print. By default (NULL), all the global components of the analysis are printed.

scale Type of standardization applied to the variables in the explanatory blocks. Either "none" (default) or "sd". If scale="sd", each variable in the explanatory blocks is divided by its standard deviation.

scaleY Type of standardization applied to the variables in the response block. Either "none" (default) or "sd". If scaleY="sd", each variable of the response block is divided by its standard deviation.

option Type of normalization applied to each explanatory block of variables (either "none" or "uniform"). If option="uniform" (default), each explanatory block of variables is divided by its Frobenius norm.

optionY Type of normalization applied to the response block of variables (either "none" or "uniform"). If optionY="uniform" (default), the response block of variables is divided by its Frobenius norm.

nstart Number of random initializations of the vector of Y loadings in case of nipals algorithm (by default 10).

plotgraph Boolean (TRUE/FALSE). If TRUE (default), graphs depicting saliences, scores of individuals, correlations of variables with the global components and contributions of blocks of variables to the determination of global components are displayed.

axes Vector that indicates the plane in which graphs should be depicted (by default the plane formed by the first two global components).

Value

Returns a list of the following elements:

components Numeric vector of length two that gives the number of global components of the analysis and the number of global components to print.

optimalcrit Numeric vector that gives the optimal value of the criterion to be maximized for each dimension.

cumexplained Four columns matrix of percentages of total inertia of the explanatory blocks, percentages of inertia of the response block explained by the successive global components and their cumulative values.

explained.X Matrix of percentages of inertia explained for each Xb block.

explained.Y Matrix of percentages of inertia explained for each Y variable.

saliences	Matrix containing the specific weights of each explanatory block of variables on global components (returned by MB-WCov).
contrib	Matrix of contribution of each Xb block to the determination of global components.
T	Matrix of global components (scores of individuals).
C	Compromise matrix (unnormed global components).
U	Matrix of components associated with the response block of variables.
globalcor	Matrix of correlation coefficients between the original variables and the global components.
cor.g.b	Array that gives the correlation of the global components with their respective block components.
betaY	Array of regression coefficients.
T.b	Array that contains the matrices of block components.
blockcor	List of matrices of correlation coefficients between the original variables of each block of variables and the block components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

References

S. Wold (1984). Three PLS algorithms according to SW. In: Symposium

MULDAST (Multivariate Analysis in Science and Technology), Umea University, Sweden. pp. 26–30.

E. Tchandao Mangamana, R. Glèlè Kakaï, E.M. Qannari (2021). A general strategy for setting up supervised methods of multiblock data analysis. Chemometrics and Intelligent Laboratory Systems, 217, 104388.

See also

print.MBWCov plot.MBWCov summary.MBWCov

print.MBPLS plot.MBPLS summary.MBPLS

Example

```
data(ham)
X=ham$X
group=ham$group
Y=ham$Y
res.mbwcov <- MBWCov(X, Y, group)
res.mbwcov
res.mbpls <- MBPLS(X, Y, group)
res.mbpls
```

Predict

Description

Predict values of Y, knowing new values of each predictive block of variables.

Usage

predict(Xnew, res)

Arguments

Xnew	Block obtained by horizontally merging all new values of predictive blocks of variables.
res	Results from calibration model.

Value

predY	Predicted values.

Main Graphs for Common Dimensions analysis (ComDim) and Multiblock Principal Components Analysis (MB-PCA)

Description

Plot main graphs for ComDim and MB-PCA.

Usage

S3 method for class 'ComDim' or 'MBPCA'

plot(x, axes = c(1, 2), graphtype = c("saliences", "globalscores", "blockscores", "globalcor", "blockcor", "expl", "cumexpl", "crit", "contrib"), select = NULL, max.overlaps = 20, xlim = NULL, ylim = NULL, title = NULL, color = NULL, ...)

Arguments

x	An object of class ComDim or MBPCA.
axes	A vector of length two which specifies the global components to plot (by default the first two).

graphtype	Type of graph to plot. Either "saliences" (for ComDim), "globalscores", "blockscores", "globalcor", "blockcor", "expl", "cumexpl", "crit" or "contrib". Refer to the details section.
select	Selection of elements to plot (by default, select=NULL). Refer to the details section.
max.overlaps	Exclude text labels that overlap too many things (by default, 20).
xlim	Range for the plotted 'x' values.
ylim	Range for the plotted 'y' values.
title	Title of the graph to draw.
color	Color for the plot.
...	Further arguments.

Details

The arguments graphtype and select are used as follow.
If graphtype="saliences", the relationships between blocks of variables are shown.

If select=NULL, all the blocks are shown, otherwise, only the selected ones are shown.

If graphtype="globalscores", individuals are projected on the space formed by the global components.

In this case, if for example, select=NULL, all the individuals are plotted. However, if select=5, only the first five individuals are plotted.

If graphtype="blockscores", individuals are projected on the space formed

by the block components.

If select=NULL, individuals of each block are plotted on separate figures.

If select=c(1, 3), individuals of blocks 1 and 3 are plotted on separate figures.

If graphtype="globalcor", correlations of original variables with the global components are depicted.

If select=NULL, correlations of the variables of all the blocks are plotted on the same figure.

If select=c(1, 3), correlations of the variables of blocks 1 and 3 are plotted.

If graphtype="blockcor", correlations of original variables with the block components are depicted.

If select=NULL, correlations of the variables of each block are plotted on seperate figures.

If select=c(1, 3), correlations of the variables of blocks 1 and 3 are plotted.

If graphtype="expl", percentages of inertia of all the blocks explained by the global components are drawn.

If graphtype="cumexpl", cumulative percentages of inertia of all the blocks explained by the global components are drawn.

graphtype="crit" plots the values of the maximization criterion.

graphtype="contrib" depicts the contribution of each block of variables to the determination of the global components.

For graphtype="expl", "cumexpl", "crit" and "contrib", if select=NULL, all the dimensions are plotted. But if for example, select=5, only the first five dimensions are plotted.

Value

Returns graphs showing the relationships between blocks of variables (for ComDim), projection of individuals in both global and block components, the correlations of variables with the global and block components, the percentages of inertia explained by the global components and their cumulative values, the values of the maximization criterion and the contributions of the blocks to the determination of global components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

ComDim MBPCA

Examples

```
data(ham)
X=ham$X
group=ham$group
res.comdim <- ComDim(X, group, plotgraph=FALSE)
plot(res.comdim, graphtype="saliences")
plot(res.comdim, graphtype="globalcor")
res.mbpca <- MBPCA(X, group, plotgraph=FALSE)
plot(res.mbpca, graphtype="globalscores")
```

Main Graphs for Multiblock Weighted Covariate analysis (MB-WCov) and Multiblock Partial Least Squares (MB-PLS) regression

Description

Plot main graphs for MB-WCov and MB-PLS regression.

Usage

S3 method for class 'MBWCov' or 'MBPLS'

plot(x, axes = c(1, 2), graphtype = c("saliences", "globalscores", "blockscores", "globalcor", "blockcor", "explY", "cumexplY", "explX", "cumexplX", "crit", "contrib"), select = NULL, max.overlaps = 20, xlim = NULL, ylim = NULL, title = NULL, color = NULL, ...).

Arguments

x	An object of class 'MBWCov' or 'MBPLS'.
axes	A vector of length two which specifies the global components to plot (by default the first two).
graphtype	Type of graph to plot. Either "saliences" (for MB-WCov), "globalscores", "blockscores","globalcor","blockcor", "explY", "cumexplY", "explX", "cumexplX", "crit" or "contrib". Refer to the details section.

select	Selection of elements to plot (by default, select=NULL). Refer to the details section.
max.overlaps	Exclude text labels that overlap too many things (by default, 20).
xlim	Range for the plotted 'x' values.
ylim	Range for the plotted 'y' values.
title	Title of the graph to draw.
color	Color for the plot.
...	Further arguments.

Details

The arguments graphtype and select are used as follow.

If graphtype="saliences", the relationships between blocks of variables are shown.

If select=NULL, all the blocks are shown, otherwise, only the selected ones are shown.

If graphtype="globalscores", individuals are projected on the space formed by the global components.

In this case, if for example, select=NULL, all the individuals are plotted. However, if select=5, only the first five individuals are plotted.

If graphtype="blockscores", individuals are projected on the space formed by the block components.

If select=NULL, individuals of each block are plotted on separate figures.

If select=c(1, 3), individuals of blocks 1 and 3 are plotted on separate fig-

ures.

If graphtype="globalcor", correlations of original variables with the global components are depicted.

If select=NULL, correlations of the variables of all the blocks are plotted on the same figure.

If select=c(1, 3), correlations of the variables of blocks 1 and 3 are plotted.

If graphtype="blockcor", correlations of original variables with the block components are depicted.

If select=NULL, correlations of the variables of each block are plotted on seperate figures.

If select=c(1, 3), correlations of the variables of blocks 1 and 3 are plotted.

If graphtype="explY", percentages of inertia of Y block explained by the global components are drawn.

If graphtype="cumexplY", cumulative percentages of inertia of Y block explained by the global components are drawn.

If graphtype="explX", percentages of inertia of X blocks explained by the global components are drawn.

If graphtype="cumexplX", cumulative percentages of inertia of X blocks explained by the global components are drawn.

graphtype="crit" plots the values of the maximization criterion.

graphtype="contrib" depicts the contribution of each block of variables to the determination of the global components.

For graphtype="explY", "cumexplY", "explX", "cumexplX", "crit" and "contrib", if select=NULL, all the dimensions are plotted.

But if for example, select=5, only the first five dimensions are plotted.

Value

Returns graphs showing relationships between the explanatory blocks of variables (MB-WCov), the projection of individuals in both global and block components, the correlations of variables with the global and block components, the percentages of inertia of both Y block and X blocks explained by the global components and their cumulative values, the values of the maximization criterion and the contributions of the blocks to the determination of global components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

MBWCov MBPLS

Examples

```
data(ham)
X=ham$X; Y=ham$Y; group=ham$group
res.mbwcov <- MBWCov(X, Y, group, plotgraph=FALSE)
plot(res.mbwcov, graphtype="saliences")
plot(res.mbwcov, graphtype="globalscores")
res.mbpls <- MBPLS(X, Y, group, plotgraph=FALSE)
plot(res.mbpls, graphtype="globalscores")
plot(res.mbpls, graphtype="globalcor")
```

Main Results for Common Dimensions analysis (ComDim) and Multiblock Principal Components Analysis (MB-PCA)

Description

Print main results for ComDim and MB-PCA.

Usage

S3 method for class 'ComDim' or 'MB-PCA'

print(x, ...)

Arguments

x	An object of class 'ComDim' or 'MBPCA'.
...	Further arguments passed to or from other methods.

Value

Returns the same results as for the function ComDim or MBPCA.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

ComDim MBPCA

Examples

data(ham); X=ham$X; group=ham$group

res.comdim <- ComDim(X, group, plotgraph=FALSE); print(res.comdim)

res.mbpca <- MBPCA(X, group, plotgraph=FALSE); print(res.mbpca)

Main Results for Multiblock Weighted Covariate analysis (MB-WCov) and Multiblock Partial Least Squares (MB-PLS) regression

Description

Print main results for MB-WCov and MB-PLS regression.

Usage

S3 method for class 'MBWCov' or 'MBPLS'

print(x, ...)

Arguments

x	An object of class 'MBWCov' or 'MBPLS'.
...	Further arguments passed to or from other methods.

Value

Returns the same results as for the function MBWCov or MBPLS.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

MBWCov MBPLS

Examples

data(ham); X=ham$X; Y=ham$Y; group=ham$group

res.mbwcov <- MBWCov(X, Y, group); print(res.mbwcov)

res.mbpls <- MBPLS(X, Y, group); print(res.mbpls)

Summary Results for Common Dimensions analysis (ComDim) and Multiblock Principal Components Analysis (MB-PCA)

Description

Give key results for ComDim and MB-PCA.

Usage

S3 method for class 'ComDim' or 'MBPCA'

summary(object, nvar = NULL, ncompprint = NULL, digits = 2, ...)

Arguments

object	An object of class 'ComDim' or 'MBPCA'.
nvar	Number of variables to print. By default (NULL), all the variables are printed.
ncompprint	Number of global components to print. By default (NULL), the number of global components printed for the main function ComDim or MBPCA.
digits	Number of decimal points (by default 2).
...	Further arguments.

Value

Returns the percentages of inertia explained by successive global components, their cumulative values, the saliences (for ComDim) and the correlations of the original variables with the global components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

ComDim MBPCA

Examples

```
data(ham)
X=ham$X
group=ham$group
res.comdim <- ComDim(X, group, plotgraph=FALSE)
summary(res.comdim)
res.mbpca <- MBPCA(X, group, plotgraph=FALSE)
summary(res.mbpca)
```

Summary Results for Multiblock Weighted Covariate analysis (MB-WCov) and Multiblock Partial Least Squares (MB-PLS) regression

Description

Give key results for MB-WCov and MB-PLS.

Usage

S3 method for class 'MBWCov' or 'MBPLS'

summary(object, nvar = NULL, ncompprint = NULL, digits = 2, ...)

Arguments

object	An object of class 'MBWCov' or 'MBPLS'.
nvar	Number of variables to print. By default (NULL), all the variables are printed.
ncompprint	Number of global components to print. By default (NULL), the number of global components printed for the main function MBWCov or MBPLS.
digits	Number of decimal points (by default 2).
...	Further arguments.

Value

Returns the percentages of inertia explained by successive global components (for both X and Y), their cumulative values, the saliences (for MBWCov) and the correlations of the variables with the global components.

Authors

Essomanda Tchandao Mangamana, Véronique Cariou, Evelyne Vigneau.

See also

MBWCov MBPLS

Examples

```
data(ham)
X=ham$X
Y=ham$Y
group=ham$group
res.mbwcov <- MBWCov(X, Y, group, plotgraph=FALSE)
summary(res.mbwcov)
res.mbpls <- MBPLS(X, Y, group, plotgraph=FALSE)
summary(res.mbpls)
```

Références bibliographiques

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

[2] 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.

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

[4] 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.

[5] 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.

[6] L.E. Wangen, B.R. Kowalski, A multibloc partial least squares algorithm for investigating complex chemical systems, J. Chemom. 3 (1) (1989) 3-20.

[7] S. Bougeard, E.M. Qannari, C. Lupo, M. Hanafi, From Multiblock

Partial Least Squares to Multiblock Redundancy Analysis. A continuum approach, Informatica 22 (2011) 11-26.

[8] S. Bougeard, E.M. Qannari, C. Lupo, C. Chauvin, Multiblock redundancy analysis from a user's perspective. Application in epidemiology, Electron. J. App. Stat. Anal. 4 (2) (2011) 203-214.

[9] S. Bougeard, E.M. Qannari, N. Rose, Multiblock redundancy analysis: interpretation tools and application in epidemiology, J. Chemom. 25 (2011) 467-475.

[10] A. El Ghaziri, V. Cariou, D. Rutledge, E. Qannari, Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K + 1) datasets, J. Chemom. 30 (2016) 420-429. DOI:10.1002/cem.2810.

[11] H. Wold, Soft modelling: the basic design and some extensions. Systems under indirect 435 observation, Part II, (1982) (pp. 1–54).

[12] A. Tenenhaus and V. Guillemot, RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data, R package version 2.1.2, (2017). https://CRAN.R-project.org/package=RGCCA

[13] V. Cariou, E.M. Qannari, D.N. Rutledge, E. Vigneau, ComDim: From multiblock data analysis to path modeling, Food Qual. Pref. 67 (2018) 27–34. doi:10.1016/j.foodqual.2017.02.012 .

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

[15] 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.

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

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

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

[19] I. Jolliffe, Principal Component Analysis. Springer-Verlag, New York, 2nd edition, 2002.

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

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

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

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

[24] 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

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] Zou H., Hastie T. and Tibshirani R., Sparse principal component analysis, Journal of Computational and Graphical Statistics 15 (2) (2006) 265–286.

[32] Shen H. and Huang J. Z., Sparse principal component analysis via regularized low rank matrix approximation, Journal of Multivariate Analysis 99 (6) (2008) 1015–1034.

[33] H. Wold, Estimation of principal components and related models by iterative least squares. In: Krishnaiah, (Ed.), Multivariate Analysis, Academic press, New York (1966).

[34] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228.

[35] M. Vivien, Approches PLS linéaires et non linéaires pour la modélisation de multi-tableaux. Théorie et applications, Phd dissertation, Université de Montpelier, France, (2002).

[36] C.R. Rao, The use and interpretation of principal component analysis in applied research, Sankhya A. 26 (1964) 329–358.

[37] R. Sabatier, Analyse factorielle de données structurées et métriques, Statistique et Analyse des Données 12 (1987) 75-96.

[38] Y. Escoufier, Le traitement des variables vectorielles, Biometrics 29 (1973) 751–761.

[39] P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods: The RV-coefficient, Applied Statistics 25 (1976) 257–265.

[40] P. Schlich. (1996), Defining and validating assessor compromises about product distances and attribute correlations. In T. Ns & E. Risvik (Eds.), Data Handling in Science and Technology (vol. 16, pp. 259–306). Elsevier. Sibson, R. (1978). Studies in the robustness of multidimension.

[41] P. Faye, D. Brémaud, M. Durand Daubin, P. Courcoux, A. Giboreau, H. Nicod, Perceptive free sorting and verbalization tasks with naive subjects: An alternative to descriptive mappings, Food Qual. Pref. 15 (2004) 781–791.

[42] P. Faye, D. Brémaud, E. Teillet, P. Courcoux, A. Giboreau, H. Nicod, An alternative to external preference mapping based on consumer perceptive mapping, Food Qual. Pref. 17 (2006) 604–614.

[43] T. Worch, Prefmfa, a solution taking the best of both internal and external preference mapping techniques, Food Qual. Pref. 30 (2013) 180–191.

[44] H. Abdi, D. Valentin, S. Chollet, C. Chrea, Analyzing assessors and products in sorting tasks: Distatis, theory and applications, Food Qual. Prefer. 18 (2007) 627–640.

[45] A. El Ghaziri, E.M. Qannari, Measures of association between two datasets; Application to sensory data, Food Qual. Pref. 40 (2015) 116–124. http://dx.doi.org/10.1016/j.foodqual.2014.09.010.

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

[47] M. Stone and R. J. Brooks, Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression, J. R. Statist. Soc. B 52 (2) (1990) 237-269.

[48] J. Bouhlel, D. Jouan-Rimbaud Bouveresse, S. Abouelkaram, E. Baéza, C. Jondreville, A. Travel, J. Ratel, E. Engel, D. N. Rutledge, Comparison of Common Components Analysis with Principal Components Analysis and Independent Components Analysis: Application to SPME-GC-MS Volatolomic Signatures, Talanta 178 (2018) 854–863. DOI: 10.1016/j.talanta.2017.10.025.

[49] D. N. Rutledge, Comparison of Principal Components Analysis, Independent Components Analysis and Common Components Analysis, J. Analysis and Testing 2 (3) (2018) 235–248. https://doi.org/10.1007/s41664-018-0065-5.

[50] D. Jouan-Rimbaud Bouveresse, R. Climaco Pinto, L. M. Schmidtke, N. Locquet, D. N. Rutledge, Identification of significant factors by an extension of ANOVA-PCA based on multiblock anal-

ysis, Chemometr. Intell. Lab. Syst. 106(2)(2011) 173-182. DOI: 10.1016/j.chemolab.2010.05.005

[51] K. A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, Sparse PLS: Variable Selection when Integrating Omics data, Statistical Application and Molecular Biology 7(1) (2008) 37.

[52] K. A. Lê Cao, S. Boitard, P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, BMC Bioinformatics 12 (2011) 253. DOI: 10.1186/1471-2105-12-253.

[53] G. Saporta, N. Niang (2006). Correspondence Analysis and Classification. In Multiple Correspondence Analysis and Related Methods, M. Greenacre and J. Blasius, Eds., pp 371-392. Chapman & Hall/CRC, Boca Raton, Florida, USA.

[54] G. Sanchez (2013). DiscriMiner: Tools of the Trade for Discriminant Analysis. R package version 0.1-29. https://CRAN.R-project.org/package=DiscriMiner

[55] S. Le, J. Josse, F. Husson, FactoMineR: An R Package for Multivariate Analysis, Journal of Statistical Software 25 (1) (2008) 1-18. 10.18637/jss.v025.i01

[56] B.-H. Mevik, R. Wehrens, K. H. Liland, P. Hiemstra (2019). pls: Partial Least Squares and Principal Component Regression. R package version 2.7-2. https://CRAN.R-project.org/package=pls.

[57] H.-J. Lee, & Z.-S.Yun, Consumers' perceptions of organic food attributes and cognitive and affective attitudes as determinants of their purchase intentions toward organic food, Food Qual. and Prefer. 39 (2015) 259–267. doi:10.1016/j.foodqual.2014.06.002.

[58] E. Menichelli, T. Almoy, O. Tomic, N.O. Veflen & T. Naes, SO-PLS as an exploratory tool for path modelling, Food Qual. and Prefer. 36 (2014) 122-134.

[59] M. Tenenhaus, V. Esposito, Y.-M. Chatelin, C. Lauro, PLSpath modeling, Computational Statistics & Data Analysis, 48 (2005) 159-205.

[60] E. Tchandao Mangamana, V. Cariou, E. Vigneau, R. Glèlè Kakaï, E.M. Qannari, Unsupervised multiblock data analysis: A unified approach and extensions, Chemometrics and Intelligent Laboratory Systems 194 (2019) 103856.

[61] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

Oniris
École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation — Nantes Atlantique
Santé et alimentation au cœur de la vie

Titre : Analyse des données multiblocs: approche unifiée et développement de nouvelles méthodes

Mots clés : ComDim, régression PLS, analyse canonique des corrélations généralisées, redundancy analysis, analyse en composantes principales, Sparse PCA.

Résumé : L'analyse des données structurées en plusieurs tableaux (données multiblocs) a connu ces deux dernières décennies un développement important. Ceci se traduit par la proposition d'une multitude de méthodes statistiques qu'il n'est pas toujours facile de situer les unes par rapport aux autres. Dans un objectif de clarification, nous avons introduit une démarche analytique qui présente un cadre unifié de plusieurs méthodes, permet d'en définir de nouvelles et dessine des perspectives pour des extensions qui semblent prometteuses.
Dans un premier temps, les méthodes statistiques à caractères exploratoires sont considérées. La démarche unifiée permet d'identifier deux grandes familles de méthodes. La première famille s'apparente à l'analyse canonique généralisée et la deuxième famille s'apparente à l'analyse en composantes principales multiblocs. Dans cette deuxième famille, nous retrouvons en particulier la méthode ComDim.
Dans le cadre des méthodes statistiques pour lesquelles il s'agit d'explorer les relations entre, d'un côté, un tableau de données et, d'un autre côté, un ensemble de tableaux explicatifs, nous identifions également deux familles de méthodes. La première famille de méthodes s'apparente à l'analyse dite «redundancy analysis» et la deuxième famille s'apparente à la régression PLS multiblocs. En particulier, cette deuxième famille inclut une nouvelle méthode que nous désignons par multiblock weighted covariate analysis (MB-WCov).
La spécificité des méthodes ComDim, pour l'analyse exploratoire, et MB-WCov, pour l'analyse prédictive, est qu'elles exhibent explicitement des «poids» associés aux différents tableaux indiquant leurs contributions dans la détermination de chacune des composantes définies par ces méthodes. Nous avons tiré profit de ces poids spécifiques pour définir des analyses dites «sparses» en ce sens que les poids des tableaux qui ne présentent pas une contribution significative à la détermination d'une composante donnée sont systématiquement mis à zéro. Ceci conduit à des modèles parcimonieux, plus faciles à interpréter et plus stables.
Les différentes analyses proposées s'appuient sur des critères d'optimisation clairs et intuitifs. Ceci permet, entre autre, de clarifier davantage les différentes analyses, vérifier la convergence des algorithmes itératifs et suggérer des indices statistiques de nature à aider l'utilisateur dans l'interprétation des résultats.
Les différentes approches sont illustrées sur la base de données simulées et / ou réelles.

Title : Multiblock data analysis: unified approach and development of new methods

Keywords : ComDim, PLS regression, generalized canonical correlation analysis, redundancy analysis, principal components analysis, Sparse PCA.

Abstract : The analysis of data structured into several blocks of variables (multiblock data) has known an important development these last two decades. As a result, several statistical methods have been proposed and it is not easy to situate one method of analysis with respect to the others. For a clarification purpose, we have set up a framework that allows us comparing several existing methods, proposing new other methods and sketching interesting extensions as perspective.
Firstly, exploratory multiblock methods have been considered. The unified framework allowed us identifying two families of methods. The first family is related to generalized canonical correlation analysis and the second family is related to multiblock principal components analysis. In particular, the method ComDim pertains to the second family.
In the frame of statistical methods for which the aim is to explore relationships between a block of variables, on the one hand, and, a set of explanatory blocks of variables, on the other hand, we have also identified two families of methods. The first family is related to redundancy analysis and the second family to multiblock partial least squares regression. In Particular, the second family includes a new method of analysis that we refer to as multiblock weighted covariate analysis (MB-WCov). The specificity of the methods ComDim, for exploratory data analysis, and MB-WCov for predictive data analysis, is that they explicitly exhibit «weights» associated with each block of variables. These weights highlight the contribution of the various blocks to the determination of each of the components defined by these methods. These weights were used to define new sparse methods whereby, blocks of variables that do not have a significant contribution to the determination of a component are discarded (i.e., their weights are set to 0). This leads to parsimonious models that are easily interpretable and more stable.
The different methods of analysis proposed are based on clear and intuitive optimization criteria. This allows us better clarifying the different methods, proving the convergence of iterative algorithms and suggesting new indices to help practitioners interpret the results of the analyses.
The differents approaches are illustrated on the basis of simulated and / or real data.