UNIVERSITE BRETAGNE LOIRE

Oniris ÉCOLE NATIONALE VÉTÉRINAIRE, AGROALIMENTAIRE ET DE L'ALIMENTATION SANTÉ ET ALIMENTATION AU CŒUR DE LA VIE

Thèse de Doctorat

Angélina EL GHAZIRI

Mémoire présenté en vue de l'obtention du
grade de Docteur d'Oniris - École Nationale Vétérinaire Agroalimentaire et de l'Alimentation Nantes-Atlantique
sous le sceau de l'Université Bretagne Loire

École doctorale : Végétal, Environnement, Nutrition, Agroalimentaire, Mer (VENAM)

Discipline : *section 26 : Mathématiques appliquées et applications des mathématiques*
Spécialité : *statistique*
Unité de recherche : *unité de statistique, sensométrie et chimiométrie (StatSC). ONIRIS, Nantes- France*

Soutenue le 20 octobre 2016

Relating datasets: exploration and prediction

Relation entre tableaux de données: exploration et prédiction

JURY

Rapporteurs :	**Lise BELLANGER**, Maître de conférences-HDR, Université de Nantes **Beata WALCZAK**, Professeur, University of Silesia, Institute of Chemistry- Pologne
Examinateurs :	**Véronique CARIOU,** Maître de conférences, ONIRIS, Nantes **Joachim KUNERT**, Professeur, Université Dortmund- Allemagne **Douglas RUTLEDGE**, Professeur, AgroParisTech- Paris
Directeur de Thèse :	**El Mostafa QANNARI,** Professeur, ONIRIS-Nantes

Contents

Acknowledgement

The research work carried out during my thesis was held at ONIRIS, at the Statistic Sensometric and Chemometric unit.

First of all, I would like to thank my advisor Professor El Mostafa Qannari, for accepting to supervise my thesis all throughout the last three years. Thank you for your precious guidance, for giving me your trust and for believing in me. Thank you for your kindness and friendship, for giving me a generous amount of your time, for your positivity, for your patience and for your continuous support. I feel very lucky that I had you as my supervisor. Thank you for all that I have learned in terms of statistical knowledge and on a personal level.

I am truly grateful to Lise Bellanger and Beata Walczak, my two thesis referees. Thank you for your interest in my research work, for your support and for all your constructive remarks.

I am truly grateful to Véronique Cariou, Joachim Kunert and Douglas Rutledge for accepting to be in my theses jury. Thank you for your interest in my research work and for all the new ideas you proposed.

Thank you also Douglas and thank you Achim Kohler for being in my thesis committee for the last three years, for all your advices, your discussions and your questions that helped me progress and reflect on my research work.

Thank you also Véronique and Douglas for the research we worked on together. It was truly a great pleasure working with you. I feel blessed for having shared some time with you and grateful for all that you taught me.

Thank you Marie-Cécile Alexandre-Gouabau and Thomas Moyon for your collaboration and for sharing your dataset with us. It was a pleasure working with you and a rewarding experience for me. I am grateful for your kindness and for the exchange we had.

Thank you Mohamed Hanafi for the research we worked on together. Thank you for the time you gave us when we sought your help, and for your mathematical support that helped to prove what, at times, seemed to be impossible to prove.

I thank all my friends at ONIRIS, Anaïs, Sibylle, Isabelle, Sonia, Saadia, Caroline, Ramona, Huong, Han, Vinh, Pasquale and both Stephanes for the lovely atmosphere in our flour, for the lunch time breaks we took together, for your encouragement and

support during these three years. Thank you for your friendship, and for all the warm moments we shared. I would like also to thank Gislaine and Aida for all the lovely time we spent, for your continuous support and for believing in me.

I would like to thank all the members of the Statistic Sensometric and Chemometric unit at ONIRIS for welcoming me at ONIRIS. Thank you El Mostafa Qannari, Evelyne Vigneau, Philippe Courcoux, Michel Semenou, Véronique Cariou, Mohamed Hanafi, Pasquale Dolce and Benoît Jaillais for your kindness, your friendship and for the fun atmosphere you set up. Thank you for all your constructive remarks during my thesis at each presentation. Thank you for the fun lunches we shared and the warm feelings at the conferences, particularly agrostat 2014 conference at Rabat. I am sincerely grateful for all the kindness and support you showed me.

I would like to thank Jean-Marc Ferrandi, Ivan Dufeu, Pasquale Barillot and Samira Rousselliere from the Marketing unit in the same floor, for their friendly discussions and pleasant smiles during these last three years. Thank you also Bridgite for your support and your kindness.

I would like to thank my friends, Maïté, Nathalie, Noel, Doïna, Sari and Jacqueline for all the time and the lovely moments we shared. Thank you for your continuous support and for believing in me. I'm also thankful for all your experiences that you shared with me, they were for my benefit.

I would like also to thank Adnan Naja for all the advices you shared with me and for your continuous support throughout these three years. I am very grateful for having you by my side.

Last but not the least, I would like to thank my family: my brothers Joseph and Elie, and my sister Maria for believing in me, for supporting me, and for all the fun moments we always share each time we get together. I thank my mother Johane and my father Georges for the continuous presence, for always believing in me, and for your infinite love and support. I can never thank you enough for all the sacrifices you make for us to be where we are today. I would like to thank also my two aunts Nawal and Souraya for all their support during my university studies and my thesis. Thank you for all the phone calls and for welcoming me in your home countless times. I would like to thank Sonia, Georges, Nisrine, Carine, Diala and Daniel for your continuous support, and for welcoming me warmly into your family. Lastly, thank you Charbel, for always believing in me, for pushing me to give the maximum of myself, for your infinite support and for always being by my side.

List of Tables

List of Figures

Chapitre 1

Résumé en français

1.1 Contexte

L'investigation de la relation entre tableaux de données revêt un intérêt primordial en statistique dû au fait de la prolifération des données dans tous les secteurs d'activité, les réseaux sociaux... Cette tendance est accélérée par la conception d'outils de mesure, permettant l'obtention de (grands) tableaux de données, très souvent de manière automatique et très rapide. D'autre part, le souci grandissant de mieux caractériser les produits conduit de plus en plus à évaluer différentes facettes et aspects de ces produits (analyses sensorielles, analyses physico-chimiques, analyse instrumentales...). Par exemple, dans le cadre d'une analyse sensorielle de plusieurs produits, l'évaluation de K juges est effectuée selon des descripteurs. De ce fait, nous sommes en présence de K tableaux, un tableau pour chaque juge. La recherche d'une représentation commune des produits ou de l'étude de l'accord entre les juges sont des objectifs qui nécessitent l'analyse de plusieurs tableaux. De plus, nous pourrions également disposer d'un tableau additionnel contenant, par exemple, les préférences des consommateurs pour les mêmes produits. L'intérêt, dans ce cas, serait de prédire les préférences des consommateurs à partir des évaluations des K juges.

En chimiométrie, dans le cadre d'une analyse de la composition d'un produit, nous pourrions disposer, d'une part, d'un tableau de données spectrales (*e.g.* Résonance Magnétique Nucléaire, ^{1}H NMR) et, d'une autre part, d'un tableau réponse correspondant aux mesures chimiques (*e.g.* acide lactique, potentiel d'hydrogène, acide sorbique ...). Parmi les objectives de l'analyse statistique de ces données, nous pouvons citer l'exploration des données spectrales et la prédiction de la composition chimique des produits. Également, nous pourrions disposer de plusieurs tableaux spectrales correspondant à plusieurs séquences d'impulsions du signal, par exemple. Ainsi, nous avons une situation de K tableaux, un tableau pour chaque séquence d'impulsions et un tableau réponse contenant les mesures chimiques sur les mêmes échantillons. Parmi

les avantages de l'analyse simultanée de K tableaux, nous pourrions citer l'utilisation de plusieurs types de données (les K tableaux) pour la caractérisation du produit à analyser et la possibilité de déterminer le ou les tableaux les plus informatifs pour la description et la prédiction du tableau réponse.

Par ailleurs, dans le cadre d'une collaboration avec l'unité de Physiologie des adaptations nutritionnelles (Phan) du CHU de Nantes, nous nous sommes intéressés à l'étude de l'effet de la sous-nutrition des ratons durant les deux phases de gestation et d'allaitement. De ce fait, nous sommes en présence de données métabolomiques multivariées dépendant de deux facteurs (gestation et lactation).

L'exploration, la synthèse des informations communes et de la relation entre tableaux, la prédiction de nouvelles données, sont parmi les préoccupations les plus courantes des utilisateurs.

1.2 Problèmes et objectifs

Dans un premier temps, pour étudier la similarité entre deux tableaux de données, des investigations relativement approfondies de trois indices d'associations entre deux tableaux ont été entreprises. De manière précise, les indices considérés sont, le coefficient de corrélation multivarié, l'indice de similarité de Procrustes et le coefficient RV.

Parmi les limitations soulignées, il y a particulièrement la sensibilité des coefficients Procrustes et RV au faible nombre d'individus ou au grand nombre de variables. En effet, ces deux coefficients ont tendance à prendre de grandes valeurs dans ces cas de figures même si les deux tableaux de données sont indépendants. Une correction du coefficient RV a été proposée permettant de corriger l'accord dû au hasard entre les tableaux.

Dans un deuxième temps, avant toute étude statistique, l'utilisateur est souvent confronté à un choix de prétraitement des données. En particulier, se pose la question s'il faut standardiser ou non standardiser ces données. Une solution intermédiaire est donnée par la standardisation dite de Pareto qui est utilisée dans des domaines d'applications tels que la métabolomique, la spectroscopie, etc. Elle consiste à diviser chaque variable par la racine carré de son écart type. Les effets du choix de la standardisation sur les données et sur l'étude statistique en termes d'exploration et de prédiction sont des questions importantes qui conditionnent la qualité des résultats. Une approche continuum de standardisation a été proposée dans le cadre de ce travail de recherche pour aider l'utilisateur à choisir la standardisation la plus adaptée à son objectif d'analyse exploratoire ou prédictive.

Dans le cadre d'une régression multivariée, la quasi-colinéarité entre les variables conduit à des modèles de régression instables. Plusieurs remèdes ont été proposés et

utilisés jusqu'à présent, parmi lesquels nous pouvons citer les stratégies consistant à réduire la dimensionnalité des données initiales. C'est le cas, par exemple, de la régression sur composantes principales et la régression PLS. Une autre stratégie consiste en une procédure de régularisation de la matrice de variance-covariance comme c'est le cas par exemple des régressions Ridge et Lasso. Dans le cadre de ce travail, une approche comparable à la régression Ridge est proposée, basée sur une décorrélation graduelle des variables prédictives.

Pour analyser plusieurs tableaux de données portant sur les mêmes individus, dans une perspective exploratoire, une méthode appelée ComDim a été développée à ONIRIS, au sein de l'unité Statistique, Sensométrie et Chimiométrie (StatSC). Elle a suscité un intérêt certain auprès des utilisateurs. Une nouvelle extension de cette méthode pour la prédiction d'un nouveau tableau de donnée est développée. Cette nouvelle approche est appelée P-ComDim (pour prédictive ComDim).

Par ailleurs, pour analyser un tableau de données multivariées dépendant de plusieurs facteurs, plusieurs méthodes statistiques sont proposées, parmi lesquelles nous pouvons citer l'analyse de la variance multivariée (MANOVA), la méthode ANOVA-PCA et l'ANOVA-simultaneous component analysis (ASCA). Une nouvelle méthode pour l'analyse de ce type données est proposée dans le cadre de ce travail, basée également sur la décomposition de la matrice initiale en sous matrices représentatives de chaque facteur, suivie d'une régression PLS.

1.3 Mesures d'association entre deux tableaux

Dans plusieurs situations, nous pouvons être amenés à étudier la similarité entre deux tableaux de données mesurées sur les mêmes individus. Par exemple, en sensométrie, nous pouvons être intéressés par l'étude de l'accord entre l'évaluation sensorielle de deux juges. En métabolomique, nous pouvons être amenés à comparer l'expression d'un gène après un intervalle de temps avant et après un traitement sur le même échantillon d'individus. Plusieurs mesures d'association entre deux tableaux sont proposées (Ramsay *et al.*, 1984). Dans le cadre de ce travail, trois coefficients ont été étudiés :

- Le coefficient de corrélation multivarié (Ramsay *et al.*, 1984) est une généralisation directe du coefficient de corrélation de Pearson entre deux variables au cas de deux tableaux.
- Le coefficient RV (Escoufier, 1973; Robert & Escoufier, 1976) est notamment utilisé dans la méthode STATIS (Lavit, 1988; Lavit *et al.*, 1994).
- Le coefficient de Procrustes (Sibson, 1978) est étroitement lié à la méthode GPA (Generalized Procrustes Analysis, Gower (1975)) qui comme son nom le suggère est une généralisation de l'analyse de Procrustes entre deux tableaux

au cas où il s'agit d'analyser plusieurs tableaux.

Les propriétés et les principales caractéristiques de ces coefficients sont mises en relief et des illustrations sont données sur la base de simulations et de données sensorielles. De plus, nous avons mis en évidence le mauvais usage de ces coefficients par certains utilisateurs. En outre, comme nous avons déjà mentionné ci-dessus, les coefficients RV et Procrustes ont tendance à prendre de grandes valeurs lorsqu'il y a un faible nombre d'individus ou lorsqu'il y a un grand nombre de variables. Ceci a déjà été souligné pour le coefficient RV par Smilde *et al.* (2009). Une correction de ce coefficient a été proposée par ces auteurs. Nous avons proposé une autre correction du coefficient RV qui corrige l'accord dû au hasard. Ceci a conduit à définir un nouveau coefficient nommé Adjusted RV. Une comparaison de ces deux corrections a été effectuée sur la base de données simulées. Le coefficient Adjusted RV a conduit à des meilleurs résultats que ceux du coefficient RV proposé par Smilde *et al.* (2009).

Plus de détails concernant ces coefficients sont donnés au chapitre 3 et une publication avec tous les résultats obtenus se trouve dans l'annexe A.1. Par ailleurs, ces coefficients seront utilisés dans la suite pour évaluer l'effet de la standardisation continuum des variables et de la décorrélation graduelle entre les variables d'un tableau de données.

1.4 Standardisation continuum des variables

Dans le cadre de l'analyse en composantes principales ou la régression PLS, le choix de standardiser ou non-standardiser les données influence la qualité des résultats. Une solution intermédiaire est proposée par la standardisation de Pareto (Robert *et al.*, 2006) consistant à diviser chaque variable par la racine carrée de son écart type. D'où l'idée de proposer une standardisation continuum dépendant d'un paramètre $\boldsymbol{\alpha}$, qui varie entre 0 et 1. Chaque variable est divisée par son écart type à la puissance $\boldsymbol{\alpha}$. Pour les trois valeurs 0, 0.5 et 1 de $\boldsymbol{\alpha}$, nous retrouvons respectivement les trois cas particuliers, données non-standardisées, Pareto et données standardisées. Formellement, la standardisation continuum des variables d'une matrice $\boldsymbol{X}$ est donnée par :

$$\boldsymbol{X}^{(\boldsymbol{\alpha})} = \boldsymbol{X}\boldsymbol{S}^{-\boldsymbol{\alpha}}$$

où $\boldsymbol{S}$ est une matrice diagonale contenant les écarts type des variables de $\boldsymbol{X}$. Ainsi la matrice $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ pourrait être utilisée au lieu de la matrice $\boldsymbol{X}$ dans le cadre de plusieurs méthodes statistiques. L'étude de l'effet du paramètre $\boldsymbol{\alpha}$ sur la matrice de données est investiguée en utilisant le coefficient de variation associé aux variances des variables dans $\boldsymbol{X}^{(\boldsymbol{\alpha})}$. De plus, la similarité entre $\boldsymbol{X}$ et $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ en fonction de $\boldsymbol{\alpha}$ est étudiée à l'aide de deux parmi les trois coefficients d'associations entre deux tableaux introduits au

paragraphe précédent. Nous avons montré que tous ces coefficients décroissent avec $\boldsymbol{\alpha}$. Des illustrations ont été également discutées sur la base de données réelles appartenant à plusieurs domaines d'applications (nutrigénomique, chimiométrique). L'allure de ces coefficients peut donner une idée sur le choix du paramètre $\boldsymbol{\alpha}$ à prendre. De plus, les propriétés relatives à l'application de la standardisation continuum des variables dans le cadre d'une ACP et d'une régression PLS sont démontrées et illustrées. En particulier, en ACP, nous avons montré que la contribution des variables de faible variance à la construction des composantes principales croît avec $\boldsymbol{\alpha}$, contrairement aux variables qui ont des variances relativement plus grandes.
En régression PLS, chaque $\boldsymbol{\alpha}$ conduit à un modèle de régression PLS. Comme en ACP, des propriétés relatives à la contribution des variables dans la construction des variables latentes ont été démontrées. De plus, en termes de prédiction, le paramètre $\boldsymbol{\alpha}$ approprié à chaque situation peut être choisi en utilisant des méthodes de validation comme la technique de validation croisée. A travers des illustrations, nous avons remarqué que le meilleur $\boldsymbol{\alpha}$ en termes de minimum d'erreur de prédiction dépend de la matrice des données predictives utilisée et également des variables à prédire.

Cette méthode de standardisation continuum des variables est introduite au chapitre 4, avec une illustration sur des données génomiques, pour souligner l'effet du paramètre $\boldsymbol{\alpha}$ en termes d'exploration et de prédiction. Une publication sur ce sujet est donnée en annexe A.2 avec plus de détails et d'autres illustrations sur la base de données chimiométriques et spectrales.

1.5 Décorrélation continuum d'un tableau de données

Pour contrer le problème de quasi-colinéarité entre les variables, une nouvelle idée de décorrélation continuum des variables a été développée dans le cadre de ce travail. Cette idée est inspirée de la distance de Mahalanobis. En effet, la nouvelle matrice $\boldsymbol{X_\alpha}$ obtenue par la décorrélation de la matrice $\boldsymbol{X}$, pour $\boldsymbol{\alpha}$ variant entre 0 et 1, est donnée par :

$$\boldsymbol{X_\alpha = X(X^T X)^{-\alpha/2}}$$

Le cas $\alpha = 0$ correspond à la matrice initiale. Pour $\alpha = 1$, $\boldsymbol{X_1 = X(X^T X)^{-1/2}}$ correspond à la transformation de Mahalanobis de la matrice $\boldsymbol{X}$ qui est caractérisée par $\boldsymbol{X_1^T X_1 = I}$, avec $\boldsymbol{I}$ la matrice identité. Lorsque $\boldsymbol{\alpha}$ varie entre 0 et 1, on passe graduellement du cas où les variables sont a priori corrélées (matrice initiale) au cas où les variables sont de covariances égales à 0.

L'impact de cette décorrélation continuum des variables a été étudié à l'aide du coefficient de variation associé aux valeurs propres de la matrice de variance covariance

de $\boldsymbol{X_\alpha}$, ainsi que la similarité entre $\boldsymbol{X}$ et $\boldsymbol{X_\alpha}$ mesurée à l'aide des indices d'associations étudiés dans la première partie. De plus, en termes de prédiction, les propriétés liées à la régression PLS d'une matrice $\boldsymbol{Y}$ sur $\boldsymbol{X_\alpha}$ ont été développées. Particulièrement, quand $\boldsymbol{\alpha}$ varie entre 0 et 1, nous passons d'un modèle de régression PLS à un modèle d'analyse des Redondances (ou ACP sur variables instrumentales). Le choix de $\boldsymbol{\alpha}$ peut être réalisé à l'aide de techniques de validation, comme la validation croisée, pour trouver le meilleur modèle en termes de prédiction. Par ailleurs, le lien entre cette méthode présentée et une méthode appelée continnum power PLS regression, introduite par Wise & Ricker (1993) et discutée par de Jong *et al.* (2001), a été établi.

Cette approche est développée dans le chapitre 5 et en annexe A.3. Deux illustrations sur la base des mêmes données nutrigénomiques et spectrales, utilisées pour la méthode 'standardisation continuum' des variables, sont presentées.

1.6 Nouvelle régression biaisée

Dans le cadre de la régression linéaire multiple, la nouvelle approche de décorrélation continuum des variables a inspiré un modèle de régression biaisée, qui sera appelé Biased Power regression (ou encore BP-regression) dans la suite. En effet, la matrice de variance-covariance $\boldsymbol{X^T X}$, est remplacée par la matrice $\boldsymbol{(X^T X)^{1-\alpha}}$. Cette dernière correspond à la matrice de variance-covariance de $\boldsymbol{X_\alpha}$ introduite dans le paragraphe précédent. Plus précisément, soit le modèle de régression linéaire d'une variable $\boldsymbol{y}$ sur $\boldsymbol{X}$:

$$\boldsymbol{y = X\beta + \epsilon}$$

où $\boldsymbol{\beta}$ est le vecteur des coefficients et $\boldsymbol{\epsilon}$ est le terme correspondant à l'erreur de la régression.
L'estimateur des moindres carrés de $\boldsymbol{\beta}$ est donnée par :

$$\boldsymbol{b_0 = (X^T X)^{-1} X^T y}$$

L'estimateur de la régression Ridge est donnée par :

$$\boldsymbol{b_\kappa = (X^T X + \kappa I)^{-1} X^T y}$$

où $\boldsymbol{\kappa}$ est un scalaire positif et $\boldsymbol{I}$ la matrice identité. L'estimateur de BP-regression est donné par :

$$\boldsymbol{b_\alpha = (X^T X)^{\alpha-1} X^T y}$$

Cette méthode est développée dans le chapitre 6 et en annexe A.4. Des propriétés liées à cette nouvelle régression sont montrées. En particulier la propriété liée au troc de l'augmentation du biais contre la diminution de la variance des estimateurs est démontrée. Une comparaison avec la régression Ridge en termes de prédiction a été effectuée sur la base de trois jeux de données.

1.7 P-ComDim : une nouvelle extension de ComDim

Dans le cadre d'analyse simultanée de plusieurs tableaux portant sur les mêmes individus, une méthode exploratoire appelée Analyse en Composantes Communes et Poids Spécifiques (ACCPS) a été développée au sein de l'unité statistique, sensométrie et chimiométrie à ONIRIS (Qannari *et al.*, 1995).

Cette méthode a suscité beaucoup d'intérêt auprès des utilisateurs en chimiometrie et en sensométrie (Mazerolles *et al.*, 2002, 2006; Hanafi *et al.*, 2006). Elle a été renommée 'ComDim' (Nielsen *et al.*, 2001).

L'idée de cette méthode d'analyse est de determiner des composantes communes pour tous les tableaux $\boldsymbol{X_k}$ (k=1,...,K) avec un poids associé à chaque tableau pour chacune des composantes. Le poids d'un tableau pour une dimension donnée, reflète la contribution de ce tableau à cette composante commune.

Des extensions de la méthode ComDim adaptées aux différents types et problèmes statistiques ont été proposées (Courcoux *et al.*, 2002; Jouan-Rimbaud Bouveresse *et al.*, 2011), ainsi que le lien entre cette méthode et d'autres méthodes d'analyse de plusieurs tableaux (Hanafi *et al.*, 2010).

Dans ce travail de recherche, nous avons proposé une nouvelle extension de cette méthode adaptée à la situation où un nouveau tableau $\boldsymbol{Y}$ est mesuré sur les mêmes individus que les K tableaux $\boldsymbol{X_k}$ (k=1,...,K). D'une manière similaire à ComDim, des composantes communes aux tableaux $\boldsymbol{X_k}$, des poids spécifiques relatifs aux tableaux $\boldsymbol{X_k}$ pour chaque composante, ainsi que les composantes associées au tableau $\boldsymbol{Y}$ sont déterminés d'une manière séquentielle.

A la première étape, la première composante commune $\boldsymbol{t_1}$ ($||\boldsymbol{t_1}|| = 1$) associée aux tableaux $\boldsymbol{X_k}$ et la première composante $\boldsymbol{u_1}$ ($||\boldsymbol{u_1}|| = 1$) associée au tableau $\boldsymbol{Y}$, ainsi que les poids spécifiques $\boldsymbol{\lambda_1^{(k)}}$, sont déterminées de telle sorte à minimiser le critère :

$$C(\lambda_1, \boldsymbol{t}_1, \boldsymbol{u}_1) = \sum_{k=1}^{K} ||\boldsymbol{S}_k - \lambda_1^{(k)} \boldsymbol{t}_1 \boldsymbol{u}_1^T||^2$$

avec $\boldsymbol{S_k} = \boldsymbol{X_k X_k^T Y Y^T}$. Plusieurs algorithmes ont été proposés pour déterminer $\lambda_1^{(k)}$, $\boldsymbol{t_1}$ et $\boldsymbol{u_1}$. Les composantes suivantes sont déterminées de la même manière que dans la première étape, en faisant la déflation des matrices $\boldsymbol{X_k}$ et $\boldsymbol{Y}$ et la mise à jour des matrices $\boldsymbol{S_k}$ (k=1,...,K). Plus précisément, $\boldsymbol{X_k}$ sera remplacé par $\boldsymbol{(I - t_1 t_1^T) X_k}$ et $\boldsymbol{Y}$, par $\boldsymbol{(I - t_1 t_1^T) Y}$.

Un rappel de la méthode ComDim ainsi que les propriétés relatives à P-ComDim sont développés dans le chapitre 7 et l'Annexe A.5. Deux illustrations sont données dans le domaine de l'analyse sensorielle et la chimiometrie. De plus, une comparaison avec la méthode MultiBlock PLS est établie et illustrée.

1.8 AoV-PLS : une nouvelle méthode d'analyse de données multivariées dépendant de plusieurs facteurs

Dans le cadre d'une collaboration avec l'unité Physiologie des Adaptations Nutritionnelles (Phan) du CHU de Nantes, une nouvelle méthode pour l'analyse d'un tableau de données multivariées dépendant de plusieurs facteurs a été développée.

Plusieurs méthodes statistiques sont proposées pour analyser ce type de données, parmi lesquelles nous pouvons citer :

- l'analyse de la variance multivariée (MANOVA). Cette méthode est une extension de l'analyse de la variance au cas multivarié. De manière similaire à l'analyse de la variance, elle est utilisée pour évaluer la significativité des facteurs. Parmi les inconvénients de la méthode MANOVA, signalons l'absence d'un cadre pour une analyse exploratoire des données. De plus, le modèle établi par cette approche peut être instable dans le cas de quasi-colinéarité des données comme c'est souvent le cas en chimiométrie (Smilde *et al.*, 2005).
- la méthode ANOVA-PCA. Cette méthode est introduite par Harrington *et al.* (2005, 2006) dans le cadre d'analyse de données protéomiques. La décomposition d'une variable dépendant de plusieurs facteurs est généralisée au cas de plusieurs variables. Ainsi le tableau de données peut être décomposé en somme des matrices liées aux facteurs et la matrice d'erreur. Une analyse en composante principale est ensuite appliquée sur chacune des matrices des facteurs plus la matrice d'erreur. Un facteur est significatif s'il émerge sur les premières composantes principales. Mais parfois l'erreur admet une grande variance et domine les premières composantes principales. Pour pallier ce problème, une stratégie pour réduire graduellement l'impact du bruit ainsi que des tests de permutation pour évaluer la significativité des facteurs sont proposés (Climaco-Pinto *et al.*, 2009).
- l'ANOVA-simultaneous component analysis (ASCA). Cette méthode est introduite par Smilde *et al.* (2005) dans le cadre de données métabolomiques. Elle est basée également sur la décomposition de la matrice de données en somme de matrices associées aux facteurs plus le résidu. Une analyse en composante principale est ensuite appliquée sur chacune des matrices de cette décomposition. Des tests de permutation ont été proposés ainsi que la projection de la matrice des résidus sur les composantes principales pour évaluer la significativité d'un facteur (Vis *et al.*, 2007; Zwanenburg *et al.*, 2011).

Une nouvelle méthode nommée AoV-PLS pour l'analyse de ce type de données est proposée. Dans le cas où nous disposons d'un seul facteur, le lien entre l'analyse discriminante PLS et l'AoV-PLS a été montré. Ainsi l'AoV-PLS peut être vue comme

une généralisation de l'analyse discriminante PLS au cas de plusieurs facteurs.
Une comparaison entre AoV-PLS, ANOVA-PCA et ASCA est discutée sur la base de données métabolomiques (annexe A.6). Ces résultats sont présentés et illustrés dans le chapitre 8 sur la base de données nutrigénomiques utilisées également dans les chapitres 4 et 5 .

1.9 Conclusion et perspectives

Conclusion

Dans le cadre de ce travail de recherche, plusieurs outils et méthodes statistiques ont été développés.

Dans un premier temps, trois indices d'associations entre deux tableaux de données ont été considérés : le coefficient de corrélation multivarié, l'indice de similarité de Procrustes et le coefficient RV. Les propriétés sont illustrées sur la base de données sensorielles et de données simulées. De plus, la sensibilité des coefficients RV et Procrustes au faible nombre d'individus ou au grand nombre de variables a été soulignée. Une correction du coefficient RV a été proposée. Une comparaison de ce nouveau coefficient (Adjusted RV) avec le coefficient RV modifié de Smilde *et al.* (2009) sur la base de données simulées, a montré que cette nouvelle correction est plus pertinente.

Dans un deuxième temps, une standardisation continuum des variables a été proposée, permettant d'aider dans le choix de la standardisation. Chaque variable est divisée par son écart type à la puissance $\boldsymbol{\alpha}$, avec $\boldsymbol{\alpha}$ un réel entre 0 et 1. En variant $\boldsymbol{\alpha}$, nous passons de la situation où les données sont non-standardisées à la situation ou les données sont standardisées. L'intérêt de cette démarche est de choisir un $\boldsymbol{\alpha}$ qui réalise un compromis entre ces deux extrêmes selon les données utilisées et selon l'objectif de l'analyse : exploration ou prédiction.

Ensuite, un deuxième type de transformation de la matrice de données est introduit. Il s'agit d'une décorrélation continuum des variables en plus de la réduction de leurs variances. L'intérêt d'une telle stratégie est de pallier le problème de quasi-colinéarité de la matrice de données dans le cadre de la régression. Nous avons montré que cette transformation établit un pont entre la méthode d'analyse des redondances (appelée aussi, ACP sur variables instrumentales) et la régression PLS. De plus, le lien entre cette stratégie et la méthode power PLS régression proposée par Wise & Ricker (1993) est établi. Les propriétés liées à cette décorrélation continuum des variables en termes de similarité avec la matrice initiale et en termes de prédiction sont montrées et illustrées. Par la suite, dans le cadre d'une régression linéaire multiple, la décorrélation continuum des variables a inspiré une régression de type régression biaisée. Cette régression est comparée à la régression Ridge et la régression Ridge généralisée. Les propriétés de cette nouvelle régression biaisée sont démontrées, particulièrement

le troc entre biais et réduction de la variance des coefficient estimés. Plusieurs illustrations montrent que cette nouvelle régression est comparable à la régression Ridge en termes de prédiction.

Dans le cadre de l'analyse de plusieurs tableaux mesurés sur les mêmes individus, une extension de la méthode ComDim est proposée et adaptée à la situation où il s'agit de prédir un nouveau tableau de données. Cette extension, nommée P-ComDim, a montré des résultats comparables à la méthode MultiBlock-PLS en termes de prédiction. L'avantage de la méthode P-ComDim est la détermination des poids permettant d'établir l'importance des tableaux à la détermination des composantes communes.

Enfin, dans le cas de données multivariées dépendant de plusieurs facteurs, la méthode AoV-PLS apparaît comme un compromis entre deux méthodes existantes pour l'analyse de ce type de données, à savoir l'ANOVA-PCA et ASCA. Une comparaison à l'aide de données métabolomiques a montré de meilleurs résultats pour l'AoV-PLS.

Perspectives

Ce travail de recherche ouvre de nouvelles perspectives. Tout d'abord, au niveau de la correction du coefficient RV (le coefficient adjusted RV), une comparaison avec le coefficient RV modifié pourrait être également établi dans le contexte de données chimiométriques. En outre, une autre correction du coefficient RV a été proposée par Mayer *et al.* (2011). Cette correction est inspirée par l'expression du R2 ajusté. Il serait intéressant de comparer les propriétés et les résultats de toutes ces corrections du coefficient RV sur des données simulées et réelles.

Dans le cadre de la standardisation continuum des variables, une stratégie directe pour sélectionner le paramètre $\boldsymbol{\alpha}$ approprié dans le cadre de l'ACP serait utile. Ceci pourrait être réalisé à l'aide d'une procédure de validation croisée comme celle utilisée dans le cadre de la régression PLS.

En ce qui concerne la méthode de décorrélation continuum des variables (αM-standardisation), il serait intéressant d'appliquer cette transformation dans un contexte d'analyse discriminante linéaire et quadratique. En effet, l'utilisation de la nouvelle matrice décorrélée, au lieu de la matrice initiale, entraînera une régularisation de la distance de Mahalanobis entre les individus. Par ailleurs, la démarche liée à la décorrélation progressive des variables (αM-standardisation) pourrait être étendue dans le cadre de tableaux multiblocs).

Une extension de BP-régression qui mérite d'être étudiée est liée à la classification et la discrimination. Cela fournirait une alternative à l'analyse discriminate régularisée (Friedman, 1989). La régularisation proposée dans cette dernière méthode est inspirée par celle de la régression Ridge et, par conséquent, devrait être facilement adaptée à une régularisation semblable à BP-régression.

Pour l'analyse de plusieurs tableaux de données, une nouvelle extension de P-

ComDim dans le contexte dit path modeling est en cours de développement au sein de l'unité StatSC.

Enfin, davantage d'investigations sont nécessaires pour comparer la stratégie AoV-PLS aux méthodes utilisées dans le contexte de données multivariées dépendant de plusieurs facteurs (*e.g.* ANOVA-PCA, ASCA, ANOVA-ComDim, Jouan-Rimbaud Bouveresse *et al.* (2011) et ANOVA-Multiblock Orthogonal PLS, Boccard & Rudaz (2016)). De plus, des investigations sont nécessaires pour étudier l'effet de plans déséquilibrés sur les résultats de ces méthodes d'analyse.

Chapitre 2

Introduction

The constant progress in technology in all the fields of daily life, virtual, medical, spatial, agriculture technologies lead to several and, sometimes, large datasets. Extracting the useful information is of primary interest. Moreover, the rising need to better characterize a product, lead to several datasets covering different aspects of this product (sensory, physico-chemical, instrumental ...). Therefore, the investigation of the relation between these datasets and exploring the common information are among the points of interest in this case. The aim of this research work is to contribute to achieve such aims. New strategies and methods of analysis are proposed and illustrated on the basis of several datasets pertaining to different fields of application :

- A "nutrimouse" dataset related to a nutrigenomic study on mice under low-fat nutrition (Martin *et al.*, 2007).
- A metabolimic data pertaining to a nutritional experiment on rat pups. These data were provided as part of a collaboration with the unit Phan (Physiologie des Adaptations Nutritionnelles) at CHU of Nantes.
- A Time Domain-Nuclear Magnetic Resonance data (Rutledge *et al.*, 1999).
- Sensory data on American dry-cured ham products available in Pham *et al.* (2008).
- A conventional sensory profiling data on 10 varieties of cider by four trained assessors.
- An economic dataset on Boston house prices in 1970 (Harrison & Rubinfeld, 1978).
- An economic dataset on total national research and development expenditures between 1972 and 1986 (Gruber, 1998).
- A chemometric study on orange juice authentication by ^{1}H NMR spectroscopy (Vigneau & Thomas, 2012).
- A chemometric study of ^{1}H NMR spectra of table wines (Larsen *et al.*, 2006).
- A chemical study of six contaminants in 18 varieties of fish at the French coast

(IFREMER, 1994).

Firstly, a study of three measures of association between two datasets was undertaken. These three measures are : the multivariate correlation coefficient, the Procrustes similarity index and the RV coefficient. The objective is to investigate the properties related to each of these coefficients, their characteristics and differences. Moreover, to cop with the misuse of these coefficients by some practitioners in terms of interpretation of the obtained values, a permutation test was proposed for each coefficient to asses the significance of the results. Furthermore, as pointed out by several authors, RV coefficient tend to take high values in the case of small number of observations and large number of variables even for the case of completely independent datasets. This finding is also true for the Procrustes similarity index. An adjusted RV coefficient was proposed and compared to the modified RV coefficient introduced by Smilde *et al.* (2009). The three measures of association are illustrated by means of simulated and sensory datasets. A more expanded elaboration of these coefficients can be found in chapter 3 and a publication of the obtained results is given in the appendix A.1. These coefficients are used all along this research work, to assess the similarity between two datasets.

Secondly, before any statistical analysis, among the questions often asked by the practitioners is to whether standardize or not the dataset. A mid way solution is proposed by Pareto standardization where each variable is divided by the square root of its standard deviation (Robert *et al.*, 2006). A new strategy for a continuum standardization of the variables is proposed. Each variable is divided by its standard deviation to the power $\boldsymbol{\alpha}$, with $\boldsymbol{\alpha}$, a parameter ranging between 0 and 1. Therefore, the cases $\alpha = 0$, $\alpha = 0.5$ and $\alpha = 1$ correspond to non-standardization, Pareto standardization and standardization to unit variance, respectively. Particularly, we have investigated the effect of the parameter $\boldsymbol{\alpha}$ when applied to two widely used statistical methods : Principal Component Analysis (PCA) and PLS regression. An elaborated introduction of this strategy is given in the chapter 4 with an illustration using the nutrimouse dataset. A publication of all the results is presented in appendix A.2 with illustrations on the basis of the three chemometric datasets introduced above.

Thirdly, in the context of regression models, several statistical methods have been used to cop with the quasi-collinearity of the $\boldsymbol{X}$-variables. We proposed yet another strategy allowing to reduce gradually the variance and collinearity of the $\boldsymbol{X}$-variables, inspired by the Mahalanobis distance. A transformation of the matrix $\boldsymbol{X}$ is applied leading to $\boldsymbol{X_\alpha = X(X^TX)^{-\alpha/2}}$ with $\boldsymbol{\alpha}$ a parameter ranging between 0 and 1. The cases $\alpha = 0$ and $\alpha = 1$ correspond to $\boldsymbol{X_\alpha = X}$ and $\boldsymbol{X_\alpha = X(X^TX)^{-1/2}}$ which are respectively the initial matrix and the Mahalanobis transformation of $\boldsymbol{X}$. This latter case is characterized by the fact that $\boldsymbol{X_1^TX_1 = I}$ (with $\boldsymbol{I}$ the identity matrix). Therefore, for $\boldsymbol{\alpha}$ moving from 0 to 1, we achieve a gradual decorrelation of the variables and a reduction of their variance to unit variance. The effect of this transformation

of $\boldsymbol{X}$, named αM-standardization, is investigated. The interest of this strategy is highlighted by applying a PLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X_\alpha}$. We show that this method encompass PLS regression ($\alpha = 0$) and Redundancy analysis ($\alpha = 1$). Chapter 5 and appendix A.3 show more developments on this continuum decorrelation strategy. Illustrations are given based on nutrimouse and orange datasets. Subsequently, a new biased regression estimator was proposed. Considering the regression of a dependent variable $\boldsymbol{y}$ on a dataset $\boldsymbol{X}$, the new estimator of the vector of coefficients is given by $\boldsymbol{\beta_\alpha} = \boldsymbol{(X^T X)^{\alpha-1} X^T y}$. We note that $\boldsymbol{(X^T X)^{1-\alpha}}$ is the variance-covariance matrix associated to $\boldsymbol{X_\alpha}$, the αM-standardization of $\boldsymbol{X}$. Therefore, the gradual decorrelation of the variables in $\boldsymbol{X_\alpha}$ ensures a better stability of the regression model. Properties related to this new regression, named BP-regression, which stands for Biased Power regression are shown. Particularly, we highlight the trade off between a small increase of the bias against an important reduction of the variances of the estimated coefficients. A comparison with Ridge regression and Generalized Ridge regression is also performed. Chapter 6 give a more detailed introduction with an illustration based on Boston dataset. An elaborated study is given in appendix A.4 shedding more light on this new biased regression with illustrations using the orange juice and the economic total national research and development expenditures datasets.

For the treatment of several datasets measured on the same individuals, a method called "Common Component and Specific Weights analysis" (CCSWA) was developed at ONIRIS within statistics, sensometrics and chemometrics unit (Qannari *et al.*, 1995). The rationale behind this method of analysis is to seek Common Components for several datasets (K, say), and specific weights (called also saliences) associated to each dataset reflecting the contribution of this dataset to the construction of the common component. This method was later known as 'ComDim', standing for "Common Components". In this research work, we extend ComDim to the case ok K+1 datasets. The aim is to predict a dataset from K other datasets. An introduction of this new strategy is given in chapter 7 with an illustration on sensory analysis (American dry-cured ham products). A publication is also given in appendix A.5 presenting an overview of ComDim strategy and developing the properties and algorithms related to P-ComDim. Moreover a comparison with Multiblock PLS is also made and an illustration using a Time Domain-Nuclear Magnetic Resonance data is shown.

Finally, as part of a collaboration with Phan (Physiologie des Adaptations Nutritionnelles) unit at CHU of Nantes, we proposed a new method for the analysis of multivariate data depending on several factors. The provided dataset pertain to an experiment on rat pups during gestation and lactation stages. The objective of the analysis is to assess the impact of maternal nutrition during these two stages on the metabolomic status of the offsprings. This new method, named AoV-PLS, is based on the decomposition of the dataset into several matrices, one for each factor plus the residual matrix. AoV-PLS was compared to several methods used for this type of

datasets, such as ANOVA-PCA (Harrington *et al.*, 2005) and ANOVA Simultaneous Component Analysis (ASCA, Smilde *et al.* (2005)). An elaborated introduction can be found in chapter 8 with an illustration on the basis of the nutrimouse dataset. A publication is given in appendix A.6 developing the properties of this new method with an illustration on the rat pups metabolomics dataset.

Chapitre 3

Association indices between two datasets

Introduction

In different contexts and in many situations, one may be led to compare the similarity between two sets of variables, measured on the same samples. For instance, in sensory field to compare the evaluation of two assessors of the same set of products ; in ecological studies to find the relation between species and environmental factors on the same sites ; in metabolomics data to compare two gene expressions datasets measured at two time points on the same set of experimental animals.

Several association indices between two datasets are proposed in the literature, we studied the properties related to three of these coefficients, namely : the multivariate correlation coefficient, the Procrustes similarity index and the RV coefficient. An emphasis was made on the particularity of each of this coefficients, and in which situations they are used. We will restrict the illustration to the context of sensory analysis. We pinpoint the misuse of these coefficients and the importance of permutation tests. An investigation was made to highlight the influence of the data dimension on these coefficients where high values were observed for RV coefficient and Procrustes similarity index in the case of small number of observations and large number of variables even for the case of completely independent datasets. We proposed a correction for the RV coefficient and compared the outcomes to another correction of RV coefficient proposed by Smilde *et al.* (2009). The illustration was based on a simulation study and a sensory dataset.
In the following we will denote by $\boldsymbol{X}$ and $\boldsymbol{Y}$ two datasets measured on the same samples and assumed to be column centered.

3.1 Three similarity coefficients

3.1.1 Multivariate correlation coefficient

The multivariate correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ also known as the inner product correlation (Ramsay *et al.*, 1984), is given by :

$$MR(\boldsymbol{X},\boldsymbol{Y}) = \frac{trace(\boldsymbol{X^T Y})}{\sqrt{trace(\boldsymbol{X^T X})}\sqrt{trace(\boldsymbol{Y^T Y})}}$$

To use this coefficient the two datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ must have the same number of columns in addition to having the same number of rows (*i.e.* $\boldsymbol{X}$ and $\boldsymbol{Y}$ with the same dimensions).

It is a straightforward generalization of Pearson correlation coefficient between two variables as $trace(\boldsymbol{X^T Y}) = \boldsymbol{vec}(\boldsymbol{X})^{\boldsymbol{T}}\boldsymbol{vec}(\boldsymbol{Y})$,with 'vec' is the vectorization of a matrix obtained by stacking its columns vertically. The multivariate correlation coefficient varies between -1 and 1. It is equal to ± 1 if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are equal up to a scaling factor : $\boldsymbol{X} = \alpha\boldsymbol{Y}$ with α a positive/negative scalar and it is equal to 0 if on average the variables in $\boldsymbol{X}$ and in $\boldsymbol{Y}$ are uncorrelated.

The multivariate correlation coefficient can be used to compare for instance the evaluation on a set of products between two assessors or between the scores of an assessor $\boldsymbol{X}$ with an overall average of all the other assessors scores, on the basis of a conventional sensory profiling. Since in these cases, a fixed vocabulary is proposed for the sensory analysis to all the assessors, the same attributes are used. In these situations, the multivariate correlation coefficient will reflect how the assessor is performing in comparison to another assessor or to the overall average configuration.

More generally, the multivariate correlation coefficient can also be used to compare between a dataset $\boldsymbol{X}$ and the same matrix after a preprocessing step. The aim in this case is to investigate the influence of the preprocessing on the structure of the matrix in term of correlation, and to assess how much the new matrix is still related to the initial one. This is precisely how this coefficient will be used in subsequent chapters of this thesis.

3.1.2 Procrustes similarity index

The Procrustes similarity coefficient between the two datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by Sibson (1978) :

$$\rho(\boldsymbol{X},\boldsymbol{Y}) = \frac{trace(\boldsymbol{X^T Y H})}{\sqrt{trace(\boldsymbol{X^T X})}\sqrt{trace((\boldsymbol{Y H})^{\boldsymbol{T}}\boldsymbol{Y H})}}$$

where $\boldsymbol{H}$ is the optimal rotation that best adjusts $\boldsymbol{X}$ to $\boldsymbol{Y}$. Practically, $\boldsymbol{H}$ is determined through a singular value decomposition of the matrix $\boldsymbol{Y^T X}$ and is given by :

$$\boldsymbol{H = UV^T} \quad \text{where} \quad \boldsymbol{Y^T X = UDV^T.}$$

With $\boldsymbol{U}$ and $\boldsymbol{V}$ orthogonal and $\boldsymbol{D}$ diagonal matrices.

Procrustes similarity index varies between 0 and 1. It is equal to 1 when the two datasets are equal up to a scalar and a rotation : $\boldsymbol{X} = \alpha \boldsymbol{YH}$, with α a scalar. It is equal to zero if all the variables in $\boldsymbol{X}$ are uncorrelated with all the variables in $\boldsymbol{Y}$. This means that in addition to the Multivariate correlation coefficient, Procrustes similarity index is invariant by reflection and by rotation.

With Procrustes similarity index, the two datasets need only to have the same number of samples. This situation is encountered for instance in free choice profiling (Williams & Langron, 1984) and in flash profiling (Dairou & Sieffermann, 2002; Dehlholm *et al.*, 2012) where for both these methods assessors have a free-choice of attributes. Moreover, the Procrustes similarity index is tightly connected to a well known method, the Generalized Procrustes analysis (GPA, Gower (1975)) which is as its name suggests a generalization of Procrustes idea between two datasets to the case of several datasets. All datasets are simultaneously transformed by means of translation, rotation, reflection and scaled to get the best fit between the datasets.

Another way to write Procrustes similarity index is given by :

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \frac{trace((\boldsymbol{X^T Y Y^T X})^{\boldsymbol{1/2}})}{\sqrt{trace(\boldsymbol{X X^T})}\sqrt{trace(\boldsymbol{Y Y^T})}}$$

since $\boldsymbol{HH^T = I}$, $\boldsymbol{I}$ the identity matrix, we have : $trace((\boldsymbol{X^T YH})(\boldsymbol{X^T YH})^{\boldsymbol{T}}) = trace(\boldsymbol{X^T Y Y^T X})$ and $trace(\boldsymbol{YH}(\boldsymbol{YH})^{\boldsymbol{T}}) = trace(\boldsymbol{Y Y^T})$.

3.1.3 RV coefficient

The RV coefficient between two datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by (Escoufier, 1973; Robert & Escoufier, 1976) :

$$RV(\boldsymbol{X}, \boldsymbol{Y}) = \frac{trace(\boldsymbol{X X^T Y Y^T})}{\sqrt{trace(\boldsymbol{X X^T X X^T})}\sqrt{trace(\boldsymbol{Y Y^T Y Y^T})}}$$

RV coefficient ranges between 0 and 1. It is equal to 1, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be matched by mean of a rotation and scaling factor. It is equal to 0, if all the variables in $\boldsymbol{X}$ are uncorrelated with all the variables in $\boldsymbol{Y}$, this means that the two configurations are in complete disagreement.

RV coefficient was first introduced in sensory analysis by Schlich (1996), and since was very used in this field for many situations :

— pairwise comparisons of the evaluations given by various assessors or with respect to a reference configuration (Worch, 2013; Mammasse & Schlich, 2014; Vidal *et al.*, 2014)
— to compare the similarity of the outputs depicted by two statistical analysis or by two methods of sensory evaluations (Faye *et al.*, 2004, 2006; Reinbach *et al.*, 2014)

Moreover, Robert & Escoufier (1976) showed that most of multivariate analysis techniques amount to maximizing RV coefficient under suitable constraints. RV coefficient is tightly linked to STATIS method (Schlich, 1996; Meyners *et al.*, 2000; Meyners, 2003; Abdi *et al.*, 2007; Pizarro *et al.*, 2013).

3.2 Misuse and remedy

Similarly to the correlation coefficient between two variables, there is a need for a hypothesis testing framework. In the context of multivariate datasets, it is advocated to use permutation test.

Technically, the permutation tests operate as follows : first, compute the initial value of these coefficients, then randomly permute all the rows of one dataset (n ! permutations for n rows) and calculate the associated coefficients after permutation between the permuted dataset and the second dataset. As we expect lower values for the similarity coefficients, we compute the proportion of new coefficients that are higher than the initial value. This proportion stands as a p-value. If it is smaller than a threshold $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha = 0.05}$ for instance), we conclude that the similarity between the two datasets is significant.
Practically, random permutation of the rows (1000 times, say) are done instead of the n ! permutations, obviously to save time. The observed p-value is then an approximation of the real one.

3.3 Limitations

RV coefficient can reach high values in the case of small number of rows and large number of variables even in situations where the two datasets at hand are independent.

This behavior was already noted for several similarity indices, like Rand index (Rand, 1971) in the context of measuring the agreement between two partitions. Hubert & Arabie (1985) proposed a correction of this latter index using the following expression :

$$\frac{Index - ExpectedIndex}{MaximumIndex - ExpectedIndex}$$

With Expected Index is the theoretical value expected for the mean when a random permutation of the rows is realized.

Adjusted RV coefficient

In the same vein, we proposed the adjusted RV coefficient given by :

$$ARV(\boldsymbol{X}, \boldsymbol{Y}) = \frac{RV(\boldsymbol{X}, \boldsymbol{Y}) - m_{RV}}{1 - m_{RV}}$$

where m_{RV} is the mean value expected when all the rows of one of the matrices are randomly permuted. Adjusted RV coefficient can be seen as a linear transformation of RV coefficient, it enjoys the same properties in term of invariance by rotation and multiplication by a scalar. It is equal to 1 if and only if $RV = 1$, that is if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are equal up to a scaling factor. It is equal to 0 if and only if $RV = m_{RV}$ which means RV is equal to the expected value when all the rows of one of the matrices are permuted, thus indicating an agreement due to chance only. Negative values of RV could be attained if and only if $RV \leq m_{RV}$ which implies a very poor agreement between the two configurations $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Another correction

The modified RV correction proposed by Smilde *et al.* (2009) is given by :

$$RV_{mod} = \frac{trace(\widetilde{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}}\widetilde{\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}})}{\sqrt{trace(\widetilde{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}}\widetilde{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}})}\sqrt{trace(\widetilde{\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}}\widetilde{\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}})}}$$

where $\widetilde{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} - diag(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})$ (same definition for $\widetilde{\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}}$). This coefficient ranges between -1 and 1. However, it is not very clear to which situations, in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$, these extreme values correspond to.
A comparison between the Adjusted RV coefficient that we proposed and the modified RV coefficient was performed on a simulation study for small number of samples and high number of variables between two configurations randomly generated. This study showed that the modified RV coefficient keeps increasing for fixed number of rows and an increasing number of columns. Whereas the adjusted RV coefficient is centered around zero and for an increasing number of rows the spread around zero gets smaller. This shows that the Adjusted RV coefficient performs better than the modified RV coefficient for correcting the agreement by chance.

Conclusion

The finding of this introductory chapter will be used in subsequent chapters to assess the effect of the standardization of the variables or the continuum decorrelation of a dataset.

Chapter 4

Continuum standardization

Introduction

When using several statistical methods such as PCA or PLS-regression the question regarding whether to standardize the variables or not is of paramount importance. A mid solution is offered by the so-called Pareto standardization whereby each variable is divided by the square root of its standard deviation. This suggests using a continuum strategy of standardization: each variable is divided by its standard deviation to the power $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ ranges between 0 and 1. The particular cases, $\alpha = 0$, $\alpha = 0.5$ and $\alpha = 1$ correspond to non standardization, Pareto standardization and standardization to unit variance, respectively.

We have investigated the effect of this continuum standardization by means of the multivariate coefficient of correlation and RV coefficient. Moreover, we discussed the interest of the continuum standardization in PCA and PLS regression analysis.

4.1 Effect of the continuum standardization

The properties of the continuum strategy of standardization are intuitively appealing. Let us denote by $\boldsymbol{X}$ the dataset at hand and let us suppose that it is column centered. The α-standardized dataset $\boldsymbol{X^{(\alpha)}}$ is given by:

$$\boldsymbol{X^{(\alpha)} = X S^{-\alpha}}$$

where $\boldsymbol{S}$ is the diagonal matrix which contains the standard deviations of the variables in $\boldsymbol{X}$. We have shown that as $\boldsymbol{\alpha}$ increases, the discrepancy among the variances of the variables in $\boldsymbol{X^{(\alpha)}}$ decreases as assessed by the coefficient of variation, $CV(\boldsymbol{\alpha})$, associated with the variances of the variables in $X^{(\alpha)}$. We also show that $MR(\boldsymbol{X}, \boldsymbol{X^{(\alpha)}})$ and $RV(\boldsymbol{X}, \boldsymbol{X^{(\alpha)}})$ decrease. The curves showing the decrease of these functions can be useful since they give insight into the impact of the α-standardization on $\boldsymbol{X}$.

We have also highlighted the impact of the continuum standardization on the outputs of PCA. The continuum α-standardization is performed on $\boldsymbol{X}$, and we applied PCA to $\boldsymbol{X}^{(\boldsymbol{\alpha})}$. We showed that as $\boldsymbol{\alpha}$ increases the contribution of the variables with small variances to the total variance of $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ is enhanced and, contrariwise, the contribution of the variables with large variances is reduced with $\boldsymbol{\alpha}$.

Of particular interest, is the use of the α-standardization in PLS-regression. As a matter of fact, the continuum standardization yields a family of models indexed by $\boldsymbol{\alpha}$. Each model (*i.e.* each $\boldsymbol{\alpha}$) has its specific performance in terms of prediction. A strategy of selection of the optimal parameter $\boldsymbol{\alpha}$ based on cross-validation was discussed. The interest of the global strategy of analysis was demonstrated on the basis of real case studies.

4.2 Illustration

By way of illustration, we consider hereafter the "nutrimouse" dataset pertaining to a nutrigenomic study on mice under low-fat nutrition. This case study is formed of two datasets, the $\boldsymbol{X}$-variables of dimension 40×120 and the $\boldsymbol{Y}$-variables of dimension 40×21. More details on this dataset can be found in appendix **??**.3.

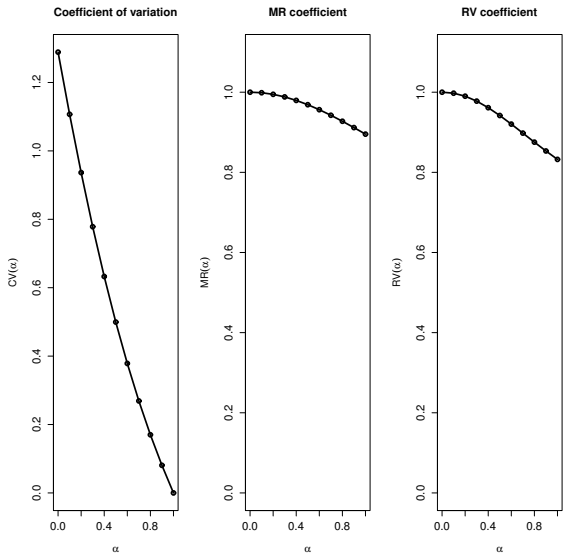

Figure 4.1 – Nutrimouse dataset: (left panel) coefficient of variation associated with the variances of the α-standardized variables. (Middle panel) The multivariate correlation and (Right panel) the RV coefficients between $\boldsymbol{X}$ and $\boldsymbol{X}^{(\boldsymbol{\alpha})}$.

In a first stage, we consider the $\boldsymbol{X}$ dataset. Figure 4.1 shows the variation of the three coefficients in function of $\boldsymbol{\alpha}$. The coefficient of variation, $CV(\boldsymbol{\alpha})$, associated

with the variances of the $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ variables decreases almost linearly with $\boldsymbol{\alpha}$: It starts with relatively a small value, and it is equal to 1 for $\boldsymbol{\alpha}$ around 0.8, indicating an acceptable discrepancy among the variances of $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ variables relatively to their mean value. The MR and RV coefficients between $\boldsymbol{X}$ and $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ show a very similar pattern. They both decrease very slowly with $\boldsymbol{\alpha}$, and reach a minimum value around 0.9 and 0.8 respectively, for $\alpha = 1$. Both coefficients indicate a high similarity between the two datasets $\boldsymbol{X}$ and $\boldsymbol{X}^{(\boldsymbol{\alpha})}$.

We show the results of the application of the $\boldsymbol{\alpha}$-standardization to PLS-regression on the "nutrimouse" dataset of $\boldsymbol{Y}$ dataset on $\boldsymbol{X}$. The Root Mean Square Error associated with a 3fold Cross-Validation (RMSECV) was determined for five values of $\boldsymbol{\alpha}$ (*i.e.* 0, 0.25, 0.5, 0.75 and 1) for both PLS2 regression of $\boldsymbol{Y}$ on $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ and PLS1 regression of each variable $\boldsymbol{y_j}$ $(j = 1, ..., 21)$ in $\boldsymbol{Y}$ on $\boldsymbol{X}^{(\boldsymbol{\alpha})}$.

For PLS2 regression, we computed for each model (i.e. each $\boldsymbol{\alpha}$), the RMSECV as follows:

$$RMSECV = \sqrt{\frac{\sum_{f=1}^{3} ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}_f||^2}{p \times n}}$$

Where $\hat{\boldsymbol{Y}}_{\boldsymbol{f}}$ is the predicted $\boldsymbol{Y}$ obtained by means of the PLS2 regression of $\boldsymbol{Y}$ on $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ corresponding to the f^{th} fold of the three fold Cross-Validation. The scalars p=21 and n=40 are the numbers of columns and rows of $\boldsymbol{Y}$, respectively.

The RMSECV associated with the PLS2 regression of all the $\boldsymbol{Y}$-variables on $\boldsymbol{X}^{(\boldsymbol{\alpha})}$ is presented in figure 4.2 in function of the number of components.

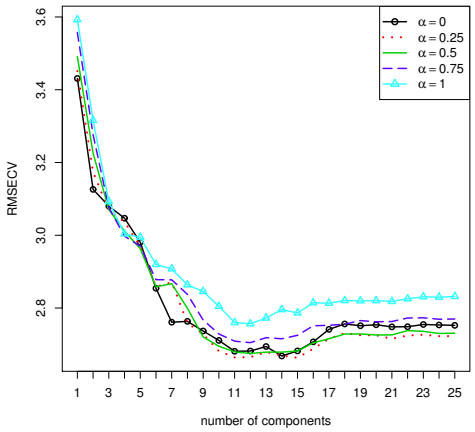

Figure 4.2 – Nutrimouse dataset: An overall RMSE of all the $\boldsymbol{Y}$-variables associated with a 3 fold Cross-Validation obtained by means of a PLS2 regression of $\boldsymbol{Y}$ on $\boldsymbol{X}^{(\boldsymbol{\alpha})}$.

We can see that the minimum error of prediction is attained for small values of $\boldsymbol{\alpha}$

(*i.e.* 0, 0.25 and 0.5). Therefore, a choice of $\boldsymbol{\alpha}$ smaller than 0.5 could be indicated to achieve a good prediction of $\boldsymbol{Y}$.

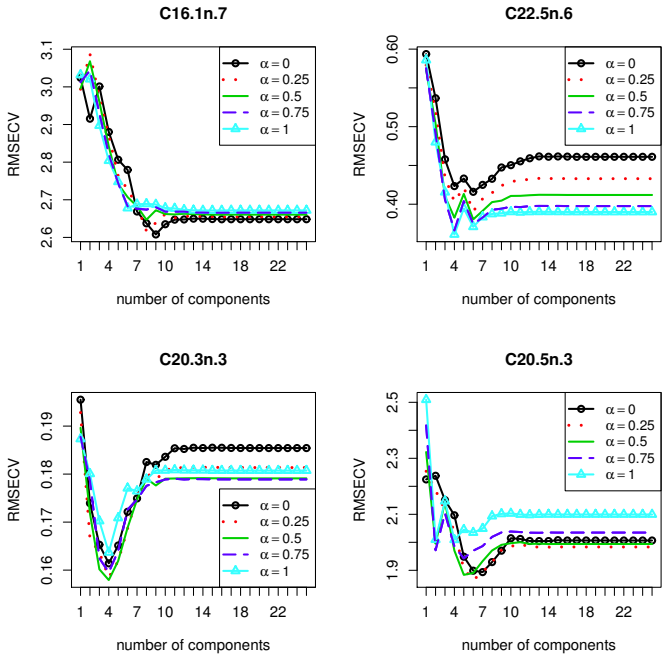

Figure 4.3 – Nutrimouse dataset: RMSECV associated with four $\boldsymbol{Y}$-variables obtained by means of a PLS1 regression of each of these $\boldsymbol{y_j}$ ($j = 5, 16, 18, 19$) variables on $\boldsymbol{X^{(\alpha)}}$.

However, the appropriate choice of $\boldsymbol{\alpha}$ that best fits the model in terms of RMSECV depends on the $\boldsymbol{Y}$-variable to be predicted. For the sake of illustration, four $\boldsymbol{Y}$-variables (*i.e.* 'C16.1n.7', 'C22.5n.6', 'C20.3n.3' and 'C20.5n.3') are selected and PLS1 regression of each of these four variables on $\boldsymbol{X^{(\alpha)}}$ is performed. The RMSECV associated with these four models (*i.e.* each of the four $\boldsymbol{Y}$-variables) is depicted in figure 4.3 as function of the number of components. For instance, the variable 'C16.1n.7' has a minimum error of prediction for $\boldsymbol{\alpha}$ around 0 at nine components. The variable 'C20.5n.3' has also the minimum of the RMSECV for small values of $\boldsymbol{\alpha}$ and a choice of $\boldsymbol{\alpha}$ around 0.25 at six components could be indicated. By contrast, for the variable 'C22.5n.6' high values of $\boldsymbol{\alpha}$ (between 0.75 and 1) perform better in terms of prediction. As for the variable 'C20.3n.3' the minimum of the RMSECV is attained for an intermediate value of $\boldsymbol{\alpha}$ (around 0.5) at four components.

Chapter 5

Continuum decorrelation

Introduction

In regression models, the influence of collinearity of the independent variables in $\boldsymbol{X}$ in the prediction of a dependent $\boldsymbol{Y}$ dataset, has always been a matter of concern for the statisticians. Several methods have been proposed ranging from Principal Component Regression and PLS regression to biased regression models such as Ridge and Lasso regressions. The idea behind these methods is to overcome the collinearity of the $\boldsymbol{X}$-variables problem either by reducing the dimensionality through extraction of uncorrelated latent components, by introducing a bias parameter or by variable selection.

We propose a new continuum strategy that reduces directly and gradually the variance and collinearity of the $\boldsymbol{X}$-variables. This strategy is inspired from the Mahalanobis distance and is based on the transformation of $\boldsymbol{X}$ to $\boldsymbol{X_\alpha = X(X^TX)^{-\alpha/2}}$ where $\boldsymbol{\alpha}$ ranges between 0 and 1. The particular case $\alpha = 0$ amounts to $\boldsymbol{X_\alpha = X}$, no transformation is made, whereas, the case $\alpha = 1$, $\boldsymbol{X_1 = X(X^TX)^{-1/2}}$ is the Mahalanobis transformation of $\boldsymbol{X}$ characterized by the fact that $\boldsymbol{X_1^TX_1 = I}$ (with $\boldsymbol{I}$ the identity matrix). This means that we move gradually from the situation were the variable are correlated to the case where the variables are of unit variance and uncorrelated.

Properties related to this continuum decorrelation are demonstrated in terms of collinearity reduction and prediction. Moreover, the link between this continuum approach, that we will call αM-standardization and another continuum regression method is showed.

5.1 Effect of the continuum decorrelation

Considering a dataset $\boldsymbol{X}$ formed of p variables and n observations assumed to be column centered. A singular value decomposition of $\boldsymbol{X}$ is applied: $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}^{\boldsymbol{1/2}}\boldsymbol{V}^{\boldsymbol{T}}$ with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthogonal matrices and $\boldsymbol{\Lambda}$ a diagonal matrix formed by the eigenvalues $\lambda_1 ... \lambda_p$ of the matrix $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ or $\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$. The αM-standardization of the matrix $\boldsymbol{X}$ is given by:

$$\boldsymbol{X}_{\boldsymbol{\alpha}} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{\alpha/2}}$$

where $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}^{-\boldsymbol{\alpha/2}}\boldsymbol{V}^{\boldsymbol{T}}$. Therefore, the variance-covariance matrix of $\boldsymbol{X}_{\boldsymbol{\alpha}}$ is equal to $\boldsymbol{X}_{\boldsymbol{\alpha}}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{\alpha}} = \boldsymbol{V}\boldsymbol{\Lambda}^{\boldsymbol{1-\alpha}}\boldsymbol{V}^{\boldsymbol{T}}$. As $\boldsymbol{\alpha}$ varies between 0 and 1 the variance covariance matrix of $\boldsymbol{X}_{\boldsymbol{\alpha}}$ varies from the variance-covariance matrix of $\boldsymbol{X}$ to the identity matrix. This means that we gradually reduce the variance of the variables in $\boldsymbol{X}_{\boldsymbol{\alpha}}$ to unit variances and the covariances between these variables to zero.

To illustrate the strategy of continuum αM-standardization, we simulate a dataset $\boldsymbol{X}$ formed of two highly correlated variables. We transform $\boldsymbol{X}$ into $\boldsymbol{X}_{\boldsymbol{\alpha}}$ with three representative values of $\boldsymbol{\alpha}$: 0.3, 0.7 and 1.

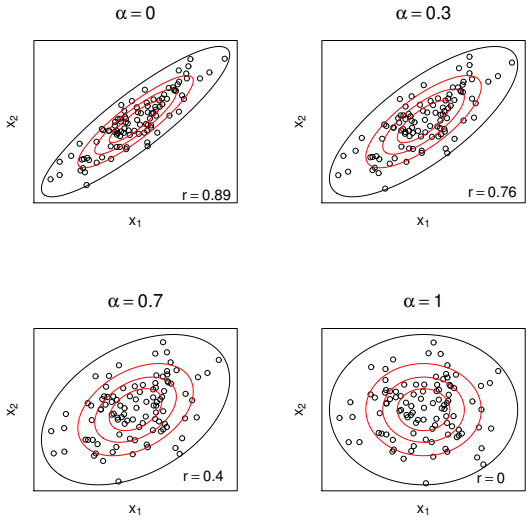

Figure 5.1 – Effect of the αM-Standardization on two correlated variables and ellipses corresponding to various Mahalanobis distances from the origin.

We show in figure 5.1 the scatterplot associated with the two dimensional data ($\alpha = 0$) and the three values of $\boldsymbol{\alpha}$ as well as the ellipses of Mahalanobis at different quantiles (0.975, 0.75, 0.5 and 0.25). The correlation between the two variables corresponding to each $\boldsymbol{\alpha}$ is noted by r on each scatterplot. We can see that the correlation

decreases with $\boldsymbol{\alpha}$ (r=0.89, 0.76, 0.4 and 0) and the ellipses become less elongated gradually taking the shape of circles.

The impact of the continuum decorrelation is investigated by means of the coefficient of variation associated with the eigenvalues of the variance-covariance matrix of $\boldsymbol{X}_{\boldsymbol{\alpha}}$, the multivariate correlation coefficient and the RV coefficient between $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{\alpha}}$. As $\boldsymbol{\alpha}$ increases, these three indices decrease. The evolution of these coefficients give an indication on the influence of αM-standardization on $\boldsymbol{X}$.

To illustrate this finding, we consider the same dataset as in chapter 4, the "nutrimouse" data. Figure 5.2 shows the evolution of these three coefficients in function of $\boldsymbol{\alpha}$.

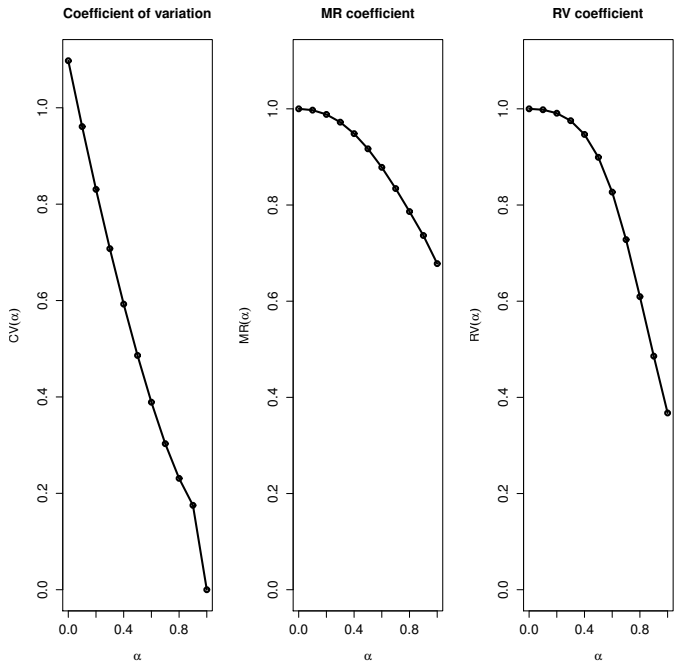

Figure 5.2 – Nutrimouse dataset: evolution of the coefficient of variation (left) associated with the eigenvalues of the variance-covariance matrix of $\boldsymbol{X}_{\boldsymbol{\alpha}}$, the multivariate correlation (center) and the RV (right) coefficients between $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{\alpha}}$.

The coefficient of variation is close to 1 for $\boldsymbol{\alpha}$ around 0.9. The multivariate correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{\alpha}}$ decreases slowly with $\boldsymbol{\alpha}$ and takes smaller values in comparison with the α-standardization. Its minimum value for $\alpha = 1$ is equal to 0.7 which is still relatively high. However, the RV coefficient between $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{\alpha}}$ decreases sharply starting from $\alpha = 0.5$ and reaches a small value around 0.4 for $\alpha = 1$; in contrast with the continuum α-standardization where the minimum value was slightly higher than 0.8. This highlights that the αM-standardization reduces not

only the variance of the variables but also their covariance structure since RV reflects the agreement between the two configurations $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{\alpha}}$.

A PLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_{\boldsymbol{\alpha}}$ is also performed, and some properties related to this regression are shown. Particularly, we showed that this continuum decorrelation amounts to redundancy analysis in the case of $\alpha = 1$. Therefore the αM-standardisation encompasses the PLS regression model ($\alpha = 0$) and the redundancy analysis ($\alpha = 1$). A compromise can be found between focusing on the recovery of the total variance in $\boldsymbol{Y}$ at the risk of defining unstable models due to the collinearity between the $\boldsymbol{X}$ variables or finding stable directions at the risk of not being optimally linked to the $\boldsymbol{Y}$ variables.

To illustrate this particular point, we depict in figure 5.3 the cumulative percentages of explained variance in $\boldsymbol{X}$ (left) and $\boldsymbol{Y}$ (right) by the first four latent components of the PLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_{\boldsymbol{\alpha}}$, with five values of $\boldsymbol{\alpha}$ (0, 0.25, 0.5, 0.75 and 1). As expected, with $\alpha = 1$ we recover a relatively small percentage of variation of $\boldsymbol{X}$, less than 40% with four components whereas we reach almost 100% of explained $\boldsymbol{Y}$ total variance. Contrariwise, for small values of $\boldsymbol{\alpha}$ we have larger explained $\boldsymbol{X}$ total variance but relatively small explained $\boldsymbol{Y}$ total variance.

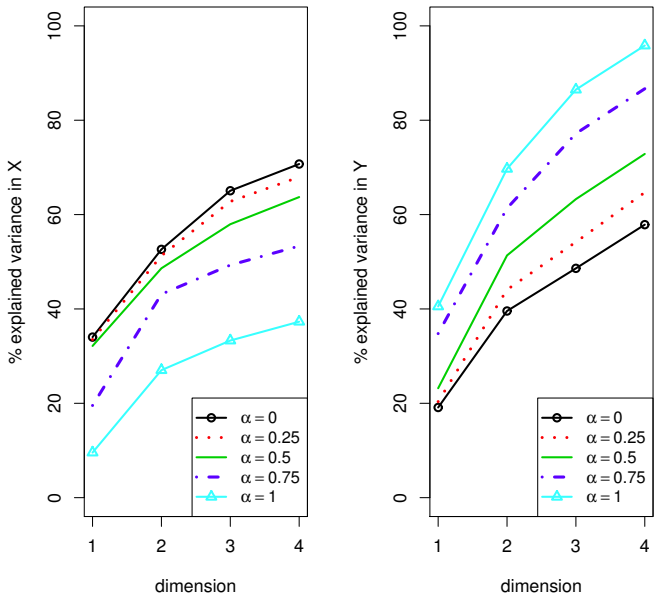

Figure 5.3 – Nutrimouse dataset: percentages of total variance in $\boldsymbol{X}$ and $\boldsymbol{Y}$ explained by the first four latent components obtained by PLS of $\boldsymbol{Y}$ on $\boldsymbol{X}_{\boldsymbol{\alpha}}$

In the case of one variable, $\boldsymbol{y}$, to be predicted, we proved for the first latent component that: its variance decreases with $\boldsymbol{\alpha}$, the covariance with $\boldsymbol{y}$ also decreases with $\boldsymbol{\alpha}$, while $R^2(\boldsymbol{\alpha})$, the percentage of explained variance in $\boldsymbol{y}$ by the first latent component, increases. This confirms, that we move from the situation were a compromise is sought between $\boldsymbol{X}$ and $\boldsymbol{y}$ (the case of PLS regression) to the situation of good prediction of $\boldsymbol{y}$ with $\alpha = 1$ (the case of redundancy analysis).

Chapter 6

Biased Power regression

Introduction

Within the framework of linear regression, we propose a new biased estimator of the vector of coefficients. As a matter of fact, this estimator is indexed by a tuning parameter $\boldsymbol{\alpha}$ which ranges between 0 and 1. There is a connection with this strategy of analysis and the αM-standardization.

The new biased regression estimator, is defined as follows: $\boldsymbol{b_\alpha} = (\boldsymbol{X^T X})^{\alpha-1}\boldsymbol{X^T y}$. Clearly when $\alpha = 0$, we retrieve the ordinary least square estimator. Throughout the paper, the new biased regression, called BP-regression (i.e. Biased Power regression) is compared to Ridge regression (Hoerl & Kennard, 1970). It turns out that similarly to the Ridge estimator, BP-regression shrinks the vector of coefficient of the multiple regression model. It shares the same rationale as Ridge regression since it achieves a trade off of the bias against a substantial reduction of the variance. Moreover, it was shown that BP-regression stands as a mid-way solution between Ridge regression and Generalized Ridge regression.

Several properties related to BP-regression are shown. Among these properties, we single out that similarly to Ridge regression there always exists a value of $\boldsymbol{\alpha}$ such that the Mean Square Error (MSE) associated with BP-regression estimator is less than the MSE of the ordinary least squares estimator. In terms of performance, we found on the basis of several datasets that BP-regression has the same performance compared to Ridge regression.

6.1 Effect of the Biased Power regression

Considering a dataset $\boldsymbol{X}$ and a dependent variable $\boldsymbol{y}$ to predict. We recall that the singular value decomposition of $\boldsymbol{X}$ is given by $\boldsymbol{U\Lambda^{1/2}V^T}$ with $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthog-

onal matrices and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements $\boldsymbol{\lambda_1, ..., \lambda_p}$ are the eigenvalues of the matrix $\boldsymbol{X^T X}$ or $\boldsymbol{X X^T}$. Without loss of generality, a pretreatment is applied on the original dataset $\boldsymbol{X}$ consisting in dividing the $\boldsymbol{X}$-variables by the largest singular value: $\boldsymbol{X}/\sqrt{\boldsymbol{\lambda_1}}$.

The rationale behind BP-regression is the following. Considering a linear regression model of $\boldsymbol{y}$ on $\boldsymbol{X}$:

$$\boldsymbol{y = X\beta + \epsilon}$$

With $\boldsymbol{\epsilon}$ the error term.
The ordinary least squares estimator of $\boldsymbol{\beta}$ is given by:

$$\boldsymbol{b_0 = (X^T X)^{-1} X^T y}$$

The Ridge regression estimator is given by:

$$\boldsymbol{b_\kappa = (X^T X + \kappa I)^{-1} X^T y}$$

Where $\boldsymbol{\kappa}$ a positive scalar and $\boldsymbol{I}$ is the identity matrix.
In the same vein as Ridge regression, we propose to counteract the problem of quasi-collinearity using the αM-standardization method introduced in chapter 5: the matrix $\boldsymbol{X^T X}$ is replaced by the matrix $\boldsymbol{(X^T X)^{1-\alpha}}$ with $\boldsymbol{\alpha}$ a scalar between 0 and 1. This leads to the BP-regression estimator:

$$\boldsymbol{b_\alpha = (X^T X)^{\alpha-1} X^T y}$$

Since $\boldsymbol{(X^T X)^{1-\alpha}}$ is the variance-covariance matrix associated to $\boldsymbol{X_\alpha}$, the αM-standardization of $\boldsymbol{X}$, as $\boldsymbol{\alpha}$ increases, a gradual decorrelation between the variables is achieved. The prediction of $\boldsymbol{y}$ using BP-regression, $\boldsymbol{\hat{y}_\alpha}$, is given by $\boldsymbol{\hat{y}_\alpha = X b_\alpha}$.

On the basis of a real case study, we compare BP-regression and Ridge regression in terms of collinearity reduction using the coefficient of variation associated to the eigenvalues of the matrices $\boldsymbol{(X^T X)^{1-\alpha}}$ and $\boldsymbol{(X^T X + \kappa I)}$, respectively. Moreover, the Root Mean Squared Error associated to a three fold Cross-Validation (RMSECV) was calculated using both these biased regression methods.

6.2 Illustration

In addition to the case studies discussed in appendix A.4, we outline herein another case study.

We consider the dataset of Boston house prices in 1970. This dataset has been used extensively throughout the literature to benchmark algorithms. The dataset is composed of 13 independent variables, 506 observations and one dependent variable,

$\boldsymbol{y}$, related to the house prices. The data was originally published by Harrison & Rubinfeld (1978).

We start by investigating the effect of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ in reducing the collinearity among the $\boldsymbol{X}$-variables using the coefficient of variation. Figure 6.1 (left) shows the variation of the coefficient of variation associated with the eigenvalues of $(\boldsymbol{X^T X})^{\mathbf{1-\alpha}}$ in function of $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha}$ varies between 0 and 1). Figure 6.1 (right) shows the variation of the coefficient of variation associated with the eigenvalues of $\boldsymbol{X^T X + \kappa I}$ ($\boldsymbol{\kappa}$ varies between 0 and 15 000). With BP-regression, the coefficient of variation

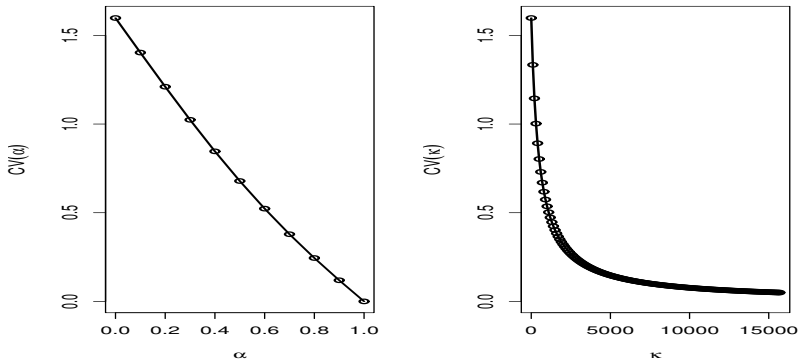

Figure 6.1 – Boston data, evolution of the coefficient of variation associated with the eigenvalues of the matrices $(\boldsymbol{X^T X})^{\mathbf{1-\alpha}}$ *(left) and* $\boldsymbol{X^T X + \kappa I}$ *(right) with* $\boldsymbol{\alpha}$ *between 0 and 1, and* $\boldsymbol{\kappa}$ *between 0 and 15000.*

decreases almost linearly with $\boldsymbol{\alpha}$, it reaches the value 1 for $\boldsymbol{\alpha}$ equals to 0.3. For Ridge regression, the coefficient of variation decreases swiftly with $\boldsymbol{\kappa}$. It is equal to 1 for $\boldsymbol{\kappa}$ equals to 300.

Figure 6.2 shows the decrease of the squared coefficient of correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{\alpha}}$ as a function of $\boldsymbol{\alpha}$. It appears that this curve steadily decreases from 0.75 ($\boldsymbol{\alpha} = \mathbf{0}$) to 0.5 ($\boldsymbol{\alpha} = \mathbf{1}$). This indicates that one should choose a value of $\boldsymbol{\alpha}$ relatively close to 0.

We performed a three fold cross-validation and we computed the RMSECV associated to both BP-regression and Ridge regression. Figure 6.3 depicts these RMSECV as function of $\boldsymbol{\alpha}$ (BP-regression) and $\boldsymbol{\kappa}$ (Ridge regression). These figures show the same pattern and indicate that both $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ should be chosen close to zero.

We run a three fold cross-validation fifty times and at each time, we selected the optimal value $\boldsymbol{\alpha}$ for BP-regression and $\boldsymbol{\kappa}$ for Ridge regression corresponding to the smallest RMSECV. Figure 6.4 shows the boxplots of the 50 optimal RMSECV values from BP-regression and Ridge regression. It turns out that the two biased procedures have the same performance in terms of prediction ability.

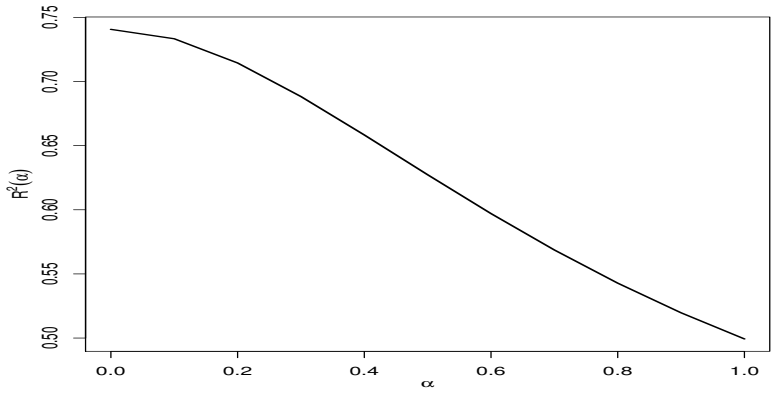

Figure 6.2 – Boston data: Evolution of the squared coefficient of correlation between $\boldsymbol{y}$ *and* $\hat{\boldsymbol{y}}_{\boldsymbol{\alpha}}$ *in function of* $\boldsymbol{\alpha}$

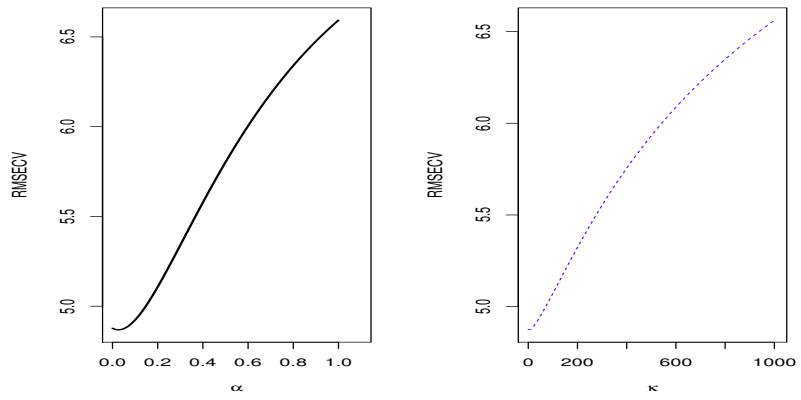

Figure 6.3 – Boston data: comparison of the RMSE associated with a 3fold Cross-Validation applied on BP-regression and Ridge regression

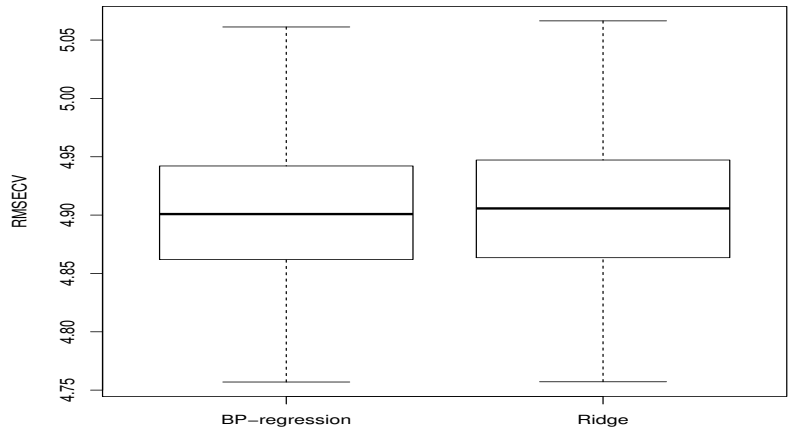

Figure 6.4 – Boston data: Box-plot of fifty minimum RMSE values associated with a three fold cross-validation applied on Biased Power regression (BP-regression) and Ridge regression repeated fifty times

Chapter 7

P-ComDim: ComDim extension to the analysis of (K+1) datasets

Introduction

In the framework of multiblock data analysis, a method called "Common Component and Specific Weights Analysis" (CCSWA) has been developed at ONIRIS within Statistics, sensometrics and chemometrics unit some twenty years ago (Qannari *et al.*, 1995). Since, this method has raised an interest among practitioners in different fields of applications and was soon known as 'ComDim' which stands for "Common Dimensions" (Nielsen *et al.*, 2001).

The rationale behind ComDim is to seek common components for several blocks of variables which refer to the same individuals and, associated with these components, weights or saliences are exhibited highlighting the contribution of the datasets to the determination of the components.

For instance, in a sensory context, consider the evaluation of K assessors with a free choice profiling, on the same set of products. This will result in K datasets, one for each assessor where the rows refer to the products and the columns to the different attributes chosen by each assessor. ComDim method will provide a common space of representation of the products. Moreover the saliences will highlight which assessor contributes most to which dimension. Therefore, ComDim is a powerful tool for the exploration of several datasets at the same time, with not necessarily the same number or type of variables. The common components are orthogonal and sequentially determined which facilitates the interpretation of the outcomes.

Several extensions of ComDim have been proposed to adapt to the type of the data at hand and properties relating ComDim to alternative methods have been proved.

We propose a new extension of ComDim to the analysis of $K+1$ datasets. In this situation, the interest is in the prediction of a new dataset $\boldsymbol{Y}$ using K dataset $\boldsymbol{X_k}$

$(k = 1, ..., K)$. For instance, referring back to the previous example concerning sensory data, we could also have a dataset $\boldsymbol{Y}$ measured on the same products consisting of preference scores given by a panel of consumers. The objective would be to relate and predict the consumers preferences on these products.

Properties related to this new extension, called P-ComDim referring to Predictive ComDim, are studied and compared to MultiBlock PLS.

7.1 Strategy

We will start by recalling ComDim method of analysis. It is first based on computing the cross products matrices: $\boldsymbol{W_k} = \boldsymbol{X_k X_k^T}$. Then in a sequential way, the common components and the associated saliences are determined. At a first stage, the first common component, $\boldsymbol{q_1}$ ($||\boldsymbol{q_1}|| = 1$), and its associated silences $\lambda_1^{(k)}$ (k=1,...,K) are determined such that the following loss function is minimized:

$$L_1 = \sum_{k=1}^{K} ||\boldsymbol{W_k} - \lambda_1^{(k)} \boldsymbol{q_1 q_1^T}||^2 \tag{7.1}$$

It can be shown that, on the one hand, for a fixed values $\lambda_1^{(k)}$ the optimal vector minimizing L_1 is the eigenvector associated to the largest eigenvalue of the matrix $\sum_{k=1}^{K} \lambda_1^{(k)} \boldsymbol{W_k}$ and, on the other hand, for a fixed $\boldsymbol{q_1}$, the optimal saliences associated to this first dimension are given by $\lambda_1^{(k)} = \boldsymbol{q_1^T W_k q_1}$. The first vector $\boldsymbol{q_1}$ and the salience $\lambda_1^{k)}$ are determined iteratively until the criterion L_1 ceases to decrease by more than a specified threshold.This process is repeated sequentially after deflation at each stage of the initial datasets $\boldsymbol{X_k}$ and updating the associated cross products matrices $\boldsymbol{W_k}$. More precisely, for stage r (r=1,...,n), the new dataset is given by: $\boldsymbol{X_k^{(r)}} = \boldsymbol{X_k} - \sum_{\boldsymbol{k=1}}^{\boldsymbol{r}} \boldsymbol{q_j q_j^T X_k}$.

After this brief introduction of ComDim, we will introduce the idea behind P-ComDim. It is based on computing the matrices $\boldsymbol{S_k} = \boldsymbol{X_k X_k^T Y Y^T}$ of dimension $\boldsymbol{n \times n}$. Note that these matrices are not symmetric like $\boldsymbol{W_k}$. Then P-ComDim follows the same strategy as in ComDim with $\boldsymbol{S_k}$ instead of $\boldsymbol{W_k}$. At a first stage, common component $\boldsymbol{t_1}$ ($||\boldsymbol{t_1}|| = 1$) associated to the $\boldsymbol{X_k}$ datasets and a component $\boldsymbol{u_1}$ ($||\boldsymbol{u_1}|| = 1$) associated to $\boldsymbol{Y}$ along with the specific weights $\lambda_1^{(k)}$ are determined such that the following criterion is minimized:

$$C(\lambda_1, \boldsymbol{t_1}, \boldsymbol{u_1}) = \sum_{k=1}^{K} ||\boldsymbol{S_k} - \lambda_1^{(k)} \boldsymbol{t_1 u_1^T}||^2 \tag{7.2}$$

It can be shown, that for fixed values of $\boldsymbol{t_1}$ and $\boldsymbol{u_1}$, $\lambda_1^{(k)} = \boldsymbol{t_1^T S_k u_1}$.

At this stand point, several algorithm have been proposed to determine $\boldsymbol{t_1}$ and $\boldsymbol{u_1}$, including a non iterative, an iterative and a NIPALS algorithm. The iterative algorithm, is defined as follows. For fixed values of $\lambda_1^{(k)}$ and $\boldsymbol{t_1}$, the optimal vector $\boldsymbol{u_1} = \frac{\sum_{k=1}^{K} \lambda_1^{(k)} \boldsymbol{S_k^T t_1}}{\| \sum_{k=1}^{K} \lambda_1^{(k)} \boldsymbol{S_k^T t_1} \|}$. Furthermore, for fixed values of $\lambda_1^{(k)}$ and $\boldsymbol{u_1}$, the optimal vector $\boldsymbol{t_1} = \frac{\sum_{k=1}^{K} \lambda_1^{(k)} \boldsymbol{S_k^T u_1}}{\| \sum_{k=1}^{K} \lambda_1^{(k)} \boldsymbol{S_k^T u_1} \|}$. The algorithm is iterated until the criterion $C(\lambda_1, \boldsymbol{t}_1, \boldsymbol{u}_1)$ ceases to decrease by more than a specified threshold. Subsequent components are determined in the same way as in ComDim after deflating the matrices $\boldsymbol{X_k}$ and $\boldsymbol{Y}$ and updating the associated matrices $\boldsymbol{S_k}$. For instance at the second stage, $\boldsymbol{X_k}$ is replaced by $\boldsymbol{(I - t_1 t_1^T) X_k}$ and $\boldsymbol{Y}$ by $\boldsymbol{(I - t_1 t_1^T) Y}$.

7.2 Illustration

A case study involving Time Domain-Nuclear Magnetic Resonnance data is developed in appendix A.5 to illustrate the P-ComDim strategy of analysis. We outline herein yet another application pertaining to sensory and preference analysis.

The study concerns eight American dry-cured ham products differing in aging times and, accordingly, labeled 1SHORT, 2SHORT, 3SHORT, 4SHORT, 5SHORT, 1LONG, 2LONG and 3LONG. These products were subjected to a sensory evaluation performed by a panel of trained assessors. More precisely, the assessors described the flavor of the hams (11 attributes), the aroma (8 attributes), and the texture (6 attributes). Each assessor was instructed to rate the intensity of each sensory attribute using a 15-point intensity scale, where 0 corresponds to 'not detected' and 15 corresponds to 'extremely strong'. The data were averaged across assessors yielding $\boldsymbol{X_1}$ (8×11) for the flavor, $\boldsymbol{X_2}$ (8×8) for the aroma and $\boldsymbol{X_3}$ (8×6) for the texture, where the rows refer to the ham products and the columns to the sensory attributes.

The consumer acceptability of the eight dry-cured hams was evaluated by a panel of 71 consumers who were regular consumers of this kind of products. The consumers were instructed to evaluate the acceptability of each ham sample using a 9-point hedonic scale. A hierarchical cluster analysis was performed and, eventually, six segments formed of homogeneous consumers within each segment were retained. Hereinafter, the $\boldsymbol{Y}$ data set consists of 8 rows referring to the ham products and 6 columns referring to the average acceptability scores respectively associated to the six segments. More details regarding the data can be found in Pham *et al.* (2008) (with due thanks to the authors for making these data available).

For the investigation of the relationships between $\boldsymbol{X_1}$ (*i.e.* flavor), $\boldsymbol{X_2}$ (*i.e.* aroma), $\boldsymbol{X_3}$ (*i.e.* texture) on the one hand and $\boldsymbol{Y}$ (*i.e.* acceptability) on the other hand, Pham *et al.* performed three separate analyses to relate each block of sensory variables in

turn to $\boldsymbol{Y}$. Each analysis consists in performing a principal components analysis on the block of sensory variables and, thereafter, superimposing the $\boldsymbol{Y}$-variables. We propose to perform P-ComDim on these data.

Figure 7.1 shows the absolute values of the saliences associated with the first four common dimensions. They reflect the importance that $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ attach to the first four couples of components $(\boldsymbol{t_j}, \boldsymbol{u_j})$ $(j = 1, ..., 4)$. It turns out that $\boldsymbol{X_3}$ has the largest salience with respect to $(\boldsymbol{t_1}, \boldsymbol{u_1})$ and a relatively small salience for $(\boldsymbol{t_2}, \boldsymbol{u_2})$. The datasets $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ have more or less even contributions for $(\boldsymbol{t_1}, \boldsymbol{u_1})$ and $(\boldsymbol{t_2}, \boldsymbol{u_2})$. For dimension 3 and 4, all the saliences are very small reflecting the fact that the associated components do not reflect substantial information regarding the relationships between $\boldsymbol{X_k}$ $(k = 1, ..., 3)$ and $\boldsymbol{Y}$.

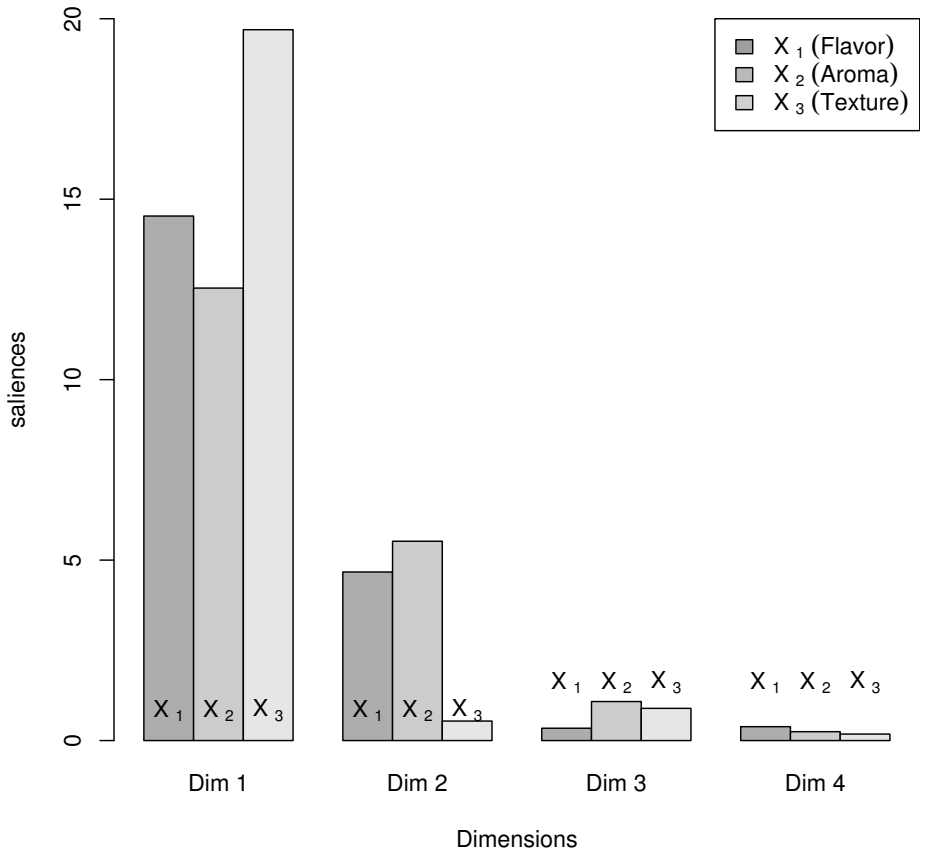

Figure 7.1 – Absolute values of the saliences associated to the three datasets on the first four common components.

For the rest of the discussion, we shall limit ourselves to the interpretation of the outcomes related to $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$. Figure 7.2 shows the representation of the eight products on the basis of $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$. Figure 7.3 shows the correlations of the sensory attributes with $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ and the correlations of the preference scores with $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$. In order to avoid having cumbersome graphical displays only those sensory attributes with a correlation coefficient with either $\boldsymbol{t_1}$ or $\boldsymbol{t_2}$ larger than 0.7 are represented.

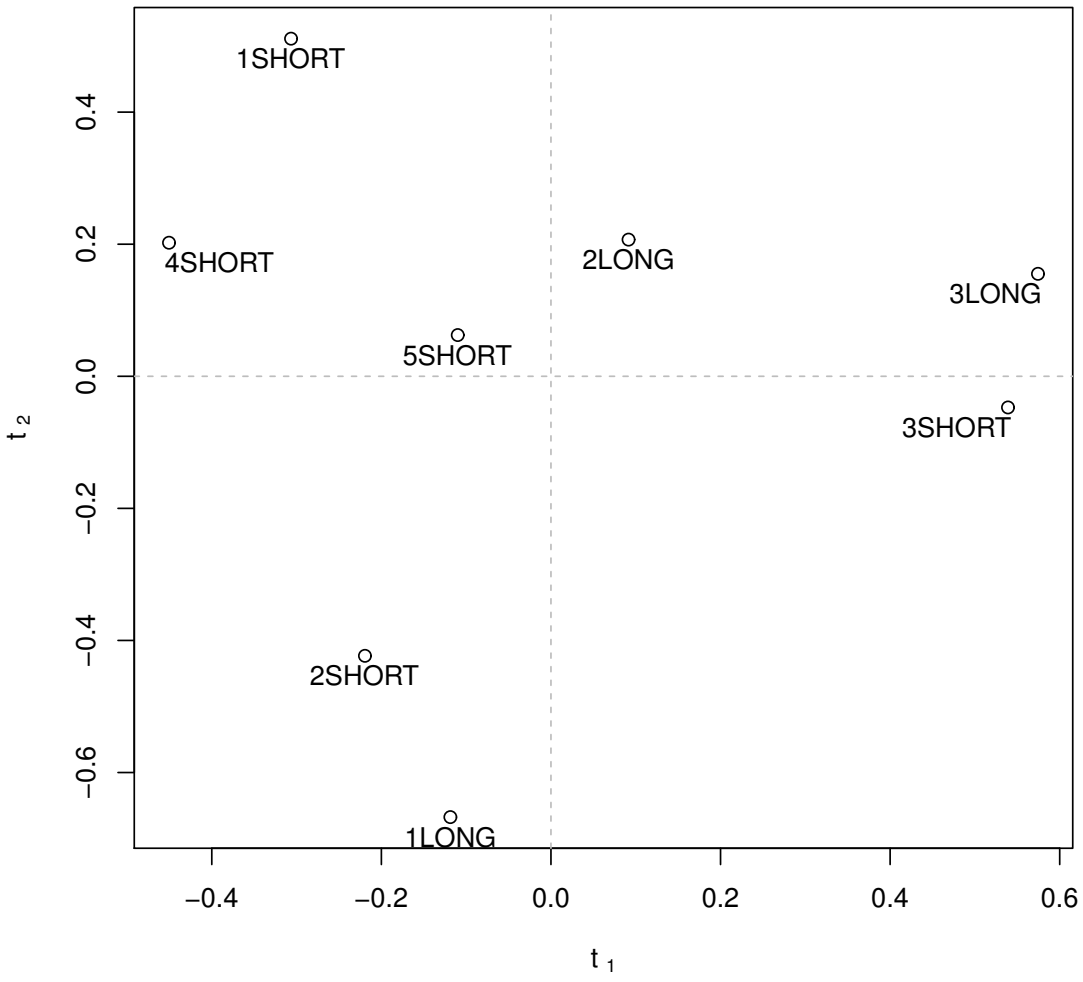

Figure 7.2 – Representation of the eight ham products on the basis of the first two common components $\boldsymbol{t_1}$ *and* $\boldsymbol{t_2}$.

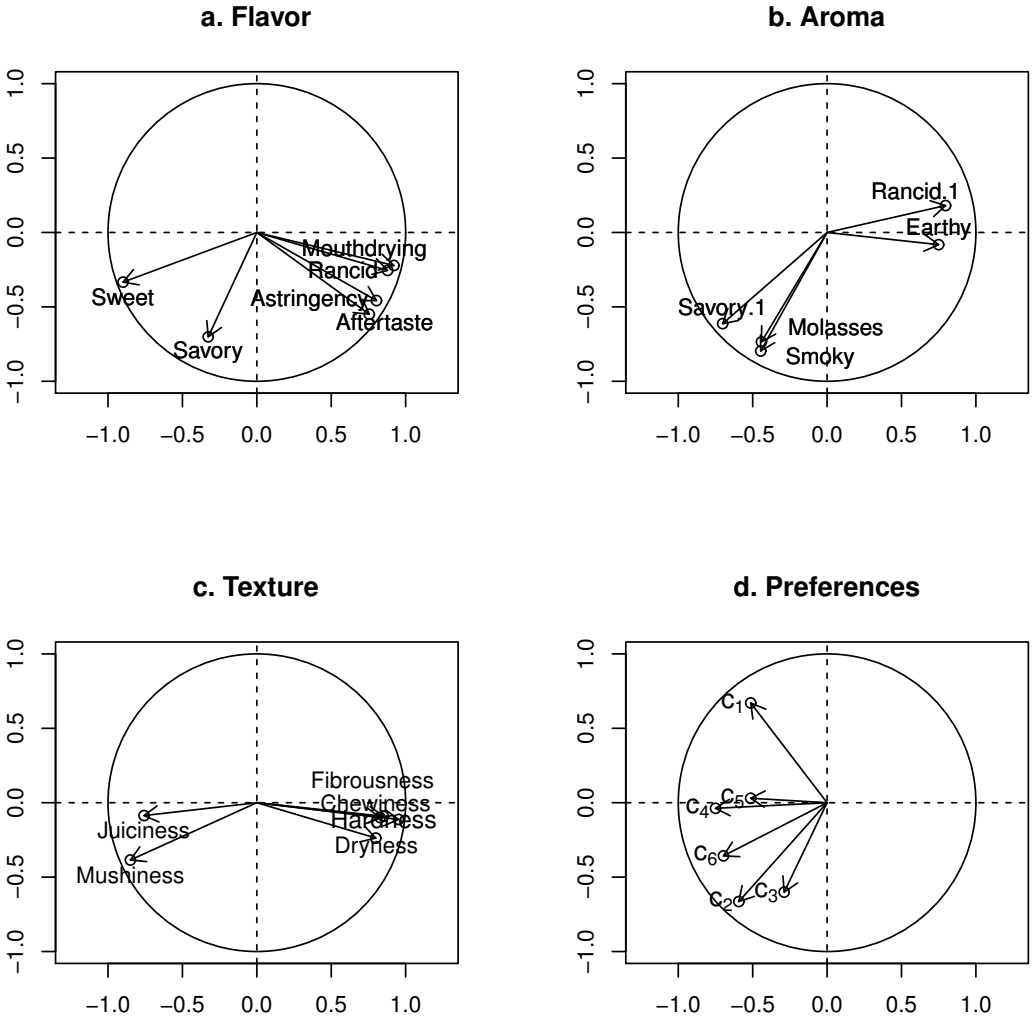

Figure 7.3 – Correlation plots showing the correlations of sensory attributes (figures a, b, c) and preference scores (figure d) with the first two components $\boldsymbol{t_1}$ *and* $\boldsymbol{t_2}$ *of P-ComDim. The labels (*$\boldsymbol{c_1, ..., c_6}$*) refer to the six acceptability variables.*

Considering simultaneously the representation of the eight products (figure 7.2) and the correlations of the preference scores with $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ (figure 7.3, d), it appears that the preferred products by the six groups of consumers (in $\boldsymbol{Y}$) are: 1SHORT, 4SHORT, 5SHORT, 2SHORT and 1LONG. The less liked products are 3LONG and 3SHORT. The correlations of the sensory attributes with $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ indicate that in terms of flavor attributes (Figure 7.3, a), consumers prefer 'sweet' and 'savory'

products and do not like 'mouth drying', 'Rancid', 'astringency' and 'aftertaste' flavor products. In terms of aroma attributes (Figure 7.3, b), consumers prefer mostly 'savory', 'molasses' and 'smoky' products and reject 'rancid' and 'earthy ' products. As for the texture (Figure 7.3, c), the consumers prefer juicy and mushy products and dislike 'fibrousness', 'chewiness', 'hardness' and 'dryness' textures. These findings agree to a large extent with those of Pham *et al.* (2008).

By way of comparing methods, we performed Multiblock PLS on the same data. The configuration of the products on the basis of the first components obtained from Multiblock PLS (not shown herein) was to a very large extent similar to that obtained by means of P-ComDim (figure 7.2). This indicates that the fist two components obtained by means of Multiblock PLS were respectively highly correlated with the corresponding components obtained by means of P-ComDim. This finding is also true for components 3 and 4 (data not shown). Figure 7.4 shows the percentages of total variances in $\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}$ and $\boldsymbol{Y}$ explained by the $\boldsymbol{t}$-components from both Multiblock PLS and P-ComDim. It turns out that the successive components derived from the two strategies of analysis recover very similar percentages of total variance. For the sensory datasets (figure 7.4, a, b, c), in both methods, around 80% of explained variance is recovered by only the first two components. For the acceptability dataset $\boldsymbol{Y}$ (figure 7.4, d) around 55% of the variance is recovered by the first two components alone.

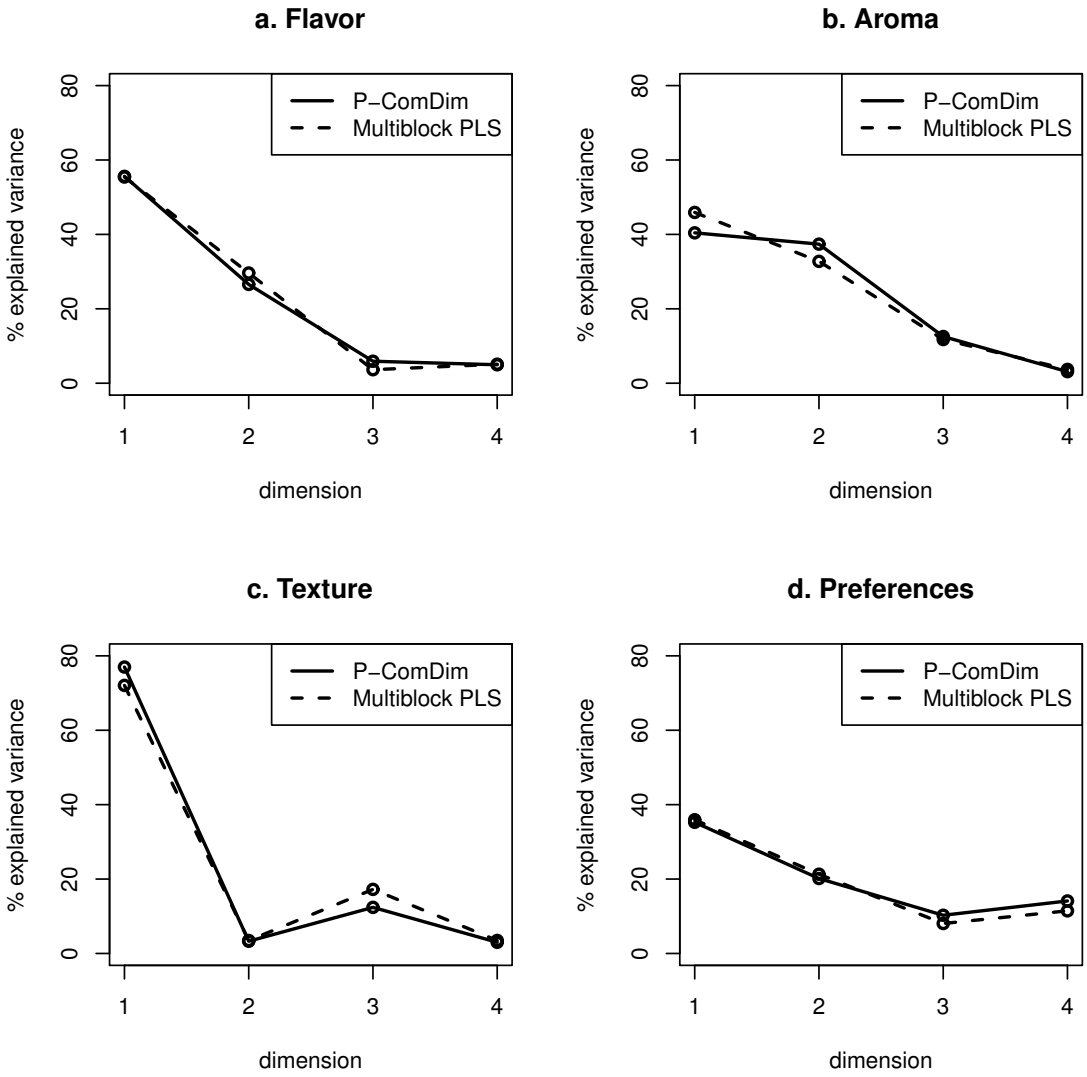

Figure 7.4 – Percentage of total variance explained by the first four components obtained by means of P-ComDim (solid line) and Multiblock PLS (dashed line)

Chapter 8

AoV-PLS: analysis of multivariate data depending on several factors

Introduction

As part of a collaboration with Phan (Physiologie des Adaptations Nutritionnelles) unit at CHU of Nantes, we proposed a new method for the analysis of multivariate dataset depending on several factors. This method called AoV-PLS, which stands for Analysis of Variance-PLS, is based on the decomposition of the original dataset into the main effects, the interaction and the residual matrix similar to that used in Analysis of Variance decomposition in the case of one variable. Then, each effect is considered in turn and assessed using a PLS regression of this effect matrix on the effect matrix plus the residual. The idea behind AoV-PLS, is that if an effect is significant it will emerge on the first latent components of PLS otherwise, it will be merged in the noise.

We compared the outcomes of our approach to several methods used in this context, like ANOVA-Simultaneous Component Analysis (ASCA) and ANOVA-Principal Component Analysis (APCA). Moreover, we showed that in the case of one factor, AoV-PLS amount to PLS-discriminant analysis (PLS-DA). Therefore, this new strategy appears as an extension of PLS-DA to the case of several factors.
Other methods are also proposed in the literature to analyze this kind of data such as ANOVA-ComDim (AComDim, Jouan-Rimbaud Bouveresse *et al.* (2011)) and ANOVA-Multiblock Orthogonal PLS (AMOPLS, Boccard & Rudaz (2016)) but they are not discussed in this document.

8.1 Method

The starting point of all the methods mentioned above is the decomposition of the dataset following a model akin to analysis of variance (AVOVA). We recall that in the case of one variable depending on two factors, a two-way ANOVA decomposition is given by:

$$\boldsymbol{x_{ijk}} = \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\beta_j} + \boldsymbol{\gamma_{ij}} + \boldsymbol{\epsilon_{ijk}} \tag{8.1}$$

where $\boldsymbol{\mu}$ is the overall mean, $\boldsymbol{\alpha_i}$, $\boldsymbol{\beta_j}$ are the effects of the first and the second factors, $\boldsymbol{\gamma_{ij}}$ is the interaction term between these two factors and $\boldsymbol{\epsilon_{ijk}}$ is the model residuals.

An extension to multivariate dataset $\boldsymbol{X}$ with two factors, A and B for instance, is obtained in a direct way by applying the decomposition of one variable (equation (8.1)) to each of the variables in $\boldsymbol{X}$, this leads to:

$$\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{X_A} + \boldsymbol{X_B} + \boldsymbol{X_{AB}} + \boldsymbol{X_\epsilon} \tag{8.2}$$

ASCA method of analysis consist in applying a PCA on each effect, namely $\boldsymbol{X_A}$, $\boldsymbol{X_B}$ and $\boldsymbol{X_{AB}}$. The rationale behind this method of analysis is to maximize the variance among the levels of each factor. However, as pointed out in the literature, with this approach we are unable to assess the significance of a factor since the variation between the factor levels is maximized but the within groups variations are missing. Permutation tests were proposed to overcome this problem.

In APCA, to assess the effect of a factor, B say, a PCA is performed on $\boldsymbol{X_B} + \boldsymbol{X_\epsilon}$. If the factor B is significant, it is likely to emerge on the first principal components. However, the noise might dominate the first components, so even if the factor is significant it will not show up on the first components. To counteract this problem, a strategy to reduce progressively the noise is proposed (Climaco-Pinto *et al.*, 2009) and a permutation test is also advocated to assess the significance of the factor.

In the approach proposed herein, AoV-PLS, to assess the effect of a factor, B say, a PLS regression of $\boldsymbol{X_B}$ on $\boldsymbol{X_B} + \boldsymbol{X_\epsilon}$ is performed. The rationale behind this strategy of analysis is the following. When confronting $\boldsymbol{X_B}$ to $\boldsymbol{X_B} + \boldsymbol{X_\epsilon}$ there are two configurations: (i) factor $\boldsymbol{B}$ is not significant and, in this case, the variations among the levels of this factor (*i.e.* dataset $\boldsymbol{X_B}$) will be completely diluted in the noise (*i.e.* dataset $\boldsymbol{X_\epsilon}$). Therefore, the PLS components will not succeed in distinguishing the signal $\boldsymbol{X_B}$ from the noise $\boldsymbol{X_\epsilon}$; (ii) factor $\boldsymbol{B}$ is significant and, in this case, the PLS components will successfully set apart the levels of this factor.

8.2 Illustration

AoV-PLS is illustrated on the basis of the metablomic dataset provided by Phan unit at CHU of Nantes in appendix A.6. Another illustration is given herein pertaining to the nutrigenomic dataset, already used in the frameworks of the continuum α-standardization (chapter 4) and the continuum αM-standardization (appendix A.3). Only the $\boldsymbol{X}$-dataset will be used in this study. We recall that, the $\boldsymbol{X}$-dataset is composed of 40 individuals (mice) and 120 gene expressions variables. Two factors are studied:

— factor genotype, G, composed of two levels: mice with Peroxisome Proliferator-Activated Receptor-α (PPARα) deficient and mice of wild type (wt). Twenty mice are considered in each level.

— factor diet, D, composed of five levels corresponding to the five diets applied for each genotype level: ref (reference diet), sun (sunflower oil diet), lin (linseed oil diet), coc (hydrogenated coconut oil diet) and fish (fish oil diet). Four mice are submitted to each combination diet$\times$genotype.

More details on the experimental design can be found in appendix A.3 and in Martin *et al.* (2007). Without loss of generality, the dataset is centered and standardized to unit variance. The $\boldsymbol{X}$-dataset is then decomposed into the factors matrices $\boldsymbol{X_G}$, $\boldsymbol{X_D}$, the interaction matrix, $\boldsymbol{X_{GD}}$ and the residual matrix $\boldsymbol{E}$ as described in the method section above.

In the following, we perform AoV-PLS, ANOVA-PCA and ASCA on this dataset. For the sake of saving space, only the results regarding the main effects G and D are discussed. For each factor, we determine the percentages of total variance explained, the significant components in the discrimination of the factor levels and the associated score plots for the relevant components. If more than two components are found to be significant, a Fisher's Linear Discriminant Analysis (LDA) is applied on these components.

Factor Genotype

In order to assess the significance of factor G, we start by the AoV-PLS method. A PLS regression of $\boldsymbol{X_G}$ on $\boldsymbol{X_G} + \boldsymbol{E}$ is performed. Table 8.1 shows the percentages of total variance in $\boldsymbol{X_G}$ and $\boldsymbol{X_G} + \boldsymbol{E}$ explained by the ten first AoV-PLS components. The p-value associated with a one way ANOVA performed on each latent component is also given. The first latent component alone recovers 94.75% of the total variance in $\boldsymbol{X_G}$. It is also the only component which is considered to be significant in discriminating the levels of factor G (for a p-value threshold equal to 0.05)

By way of comparing methods, we also determined the percentages of explained variance in $\boldsymbol{X_G}$ and $\boldsymbol{X_G} + \boldsymbol{E}$ using ANOVA-PCA and ASCA (table 8.2). To assess

Table 8.1 – The percentages of total variance in $\boldsymbol{X_G}$ and $\boldsymbol{X_G + E}$ explained by the ten first AoV-PLS components. P-values associated with one way-ANOVA performed on each component.

	percentage of explained variation		p-values
	$\boldsymbol{X_G}$	$\boldsymbol{X_G + E}$	Factor G
comp 1	94.75	21.31	0
comp 2	1.25	33.90	0.49
comp 3	2.38	11.53	0.34
comp 4	0.96	2.68	0.55
comp 5	0.33	4.12	0.72
comp 6	0.19	2.41	0.79
comp 7	0.09	1.90	0.85
comp 8	0.03	1.37	0.91
comp 9	0.01	1.54	0.94
comp 10	0.00	1.36	0.97

the significance of the components, the p-values associated to a one-way ANOVA on each component are also given in table 8.2.

Table 8.2 – The percentages of total variance in $\boldsymbol{X_G}$ and $\boldsymbol{X_G + E}$ explained by the ten first ANOVA-PCA and ASCA components. P-values associated with one way-ANOVA performed on each component.

	ANOVA-PCA			ASCA		
	$\boldsymbol{X_G}$	$\boldsymbol{X_G + E}$	p-values	$\boldsymbol{X_G}$	$\boldsymbol{X_G + E}$	p-values
comp 1	0.37	42.03	0.71	100	19.94	0
comp 2	92.68	21.16	0.00	0	0.03	1
comp 3	0.10	5.78	0.85	0	1.92	1
comp 4	2.38	4.03	0.34	0	4.83	1
comp 5	1.02	3.02	0.54	0	1.66	1
comp 6	0.07	2.76	0.87	0	2.13	1
comp 7	0.05	2.46	0.89	0	1.31	1
comp 8	0.04	1.80	0.90	0	1.76	1
comp 9	0.09	1.53	0.86	0	5.69	1
comp 10	0.34	1.47	0.72	0	1.96	1

In ANOVA-PCA, the first component does not seem to be significant (p-value equal 0.71) with a very low percentage of total variance in $\boldsymbol{X_G}$ recovered (0.37%) against around 42% of total variance in $\boldsymbol{X_G + E}$ explained. However, the second component is significant (p-value equals 0) with 92.68% of total variance explained

in $\boldsymbol{X_G}$.

Regarding ASCA, not surprisingly the PCA performed on $\boldsymbol{X_G}$ led to a single component that explained 100% of the variation of $\boldsymbol{X_G}$. This is because $\boldsymbol{X_G}$ contains only two different rows, each repeated twenty times, the number of individuals in the two groups of factor G (PPARα or wt). In order to compare methods, the rows of $\boldsymbol{X_G} + \boldsymbol{E}$ were superimposed by projection on PCA components leading to the ASCA components (Zwanenburg *et al.*, 2011). The p-value associated with the one-way ANOVA performed on each ASCA component, shows that only the first component is significant.

Figure 8.1 depicts the box plots associated with the scores on the significant component of the three methods, namely the first AoV-PLS latent component (left), the second ANOVA-PCA component (middle) and the first ASCA component (right).

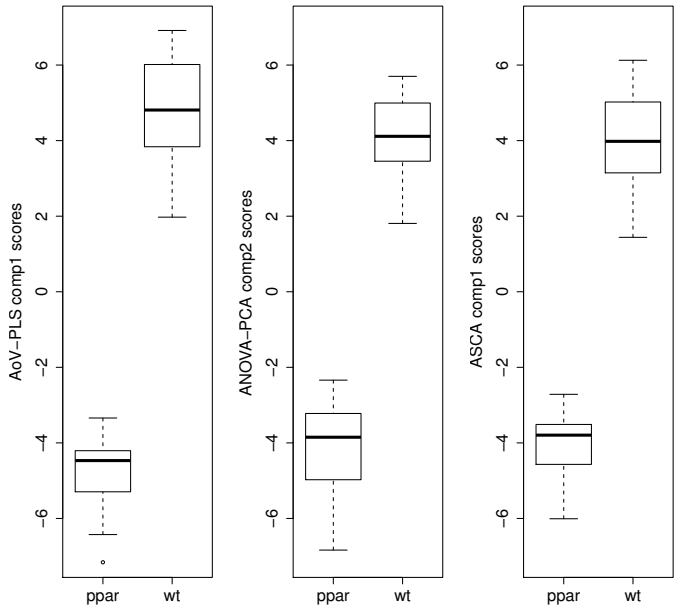

Figure 8.1 – Nutrimouse dataset: factor genotype. Box plot of the first AoV-PLS component, the second ANOVA-PCA component and the first ASCA component for the two groups ppar for PPARα and wt for wild type.

We see a clear separation between the scores of the two levels PPARα and wt of factor G on the selected components for the three used methods.

Factor diet

A PLS-regression of $\boldsymbol{X_D}$ on $\boldsymbol{X_D} + \boldsymbol{E}$ is performed. Table 8.3 shows the resulting percentages of total variance in $\boldsymbol{X_D}$ and $\boldsymbol{X_D} + \boldsymbol{E}$ explained by the ten first AoV-PLS components. The p-values associated with a one way-ANOVA performed on each component considering diet as factor are also given. The first component recovers a

Table 8.3 – The percentages of total variance in $\boldsymbol{X_D}$ and $\boldsymbol{X_D} + \boldsymbol{E}$ explained by the ten first AoV-PLS components. P-values associated with a one way-ANOVA performed on each component.

	percentage of explained variation		p-values
	$\boldsymbol{X_D}$	$\boldsymbol{X_D} + \boldsymbol{E}$	Factor D
comp 1	14.12	44.81	0.01
comp 2	24.26	8.33	0
comp 3	22.66	7.02	0
comp 4	11.66	2.96	0
comp 5	8.09	4.50	0
comp 6	4.64	3.12	0.05
comp 7	3.37	3.19	0.10
comp 8	2.00	2.85	0.25
comp 9	2.80	1.36	0.59
comp 10	2.19	1.50	0.58

relatively small percentage of total variance in $\boldsymbol{X_D}$ (14.12%) in comparison to $\boldsymbol{X_D}+\boldsymbol{E}$ (44.81%). However, it seems to be a significant component in the discrimination of factor diet groups (p-value=0.01). The components 2 and 3 recover higher variance in $\boldsymbol{X_D}$ than the first component.

The one way ANOVA performed on each component indicates to retain the first six components for a significance level of 0.05 (table 8.3). These components were submitted to LDA analysis. Figure 8.2 shows the box plots associated with the four LDA canonical variates derived from this analysis. The first canonical component shows a clear separation between the groups, especially sun and coc. The second component clearly separates the fish group. Moreover, the third canonical variate separates the group sun from the others groups and the fourth canonical variate separates ref and lin groups. The scores plots associated to the first two LDA canonical variates and to the first and fourth LDA canonical variates are shown in figures 8.3 and 8.4, respectively. A clear separation between groups can be seen by the LDA canonical variates obtained by the six AoV-PLS components.

By way of comparing methods, we also performed ANOVA-PCA and ASCA on Diet factor. The percentages of total variance in $\boldsymbol{X_D}$ and $\boldsymbol{X_D} + \boldsymbol{E}$ explained by the

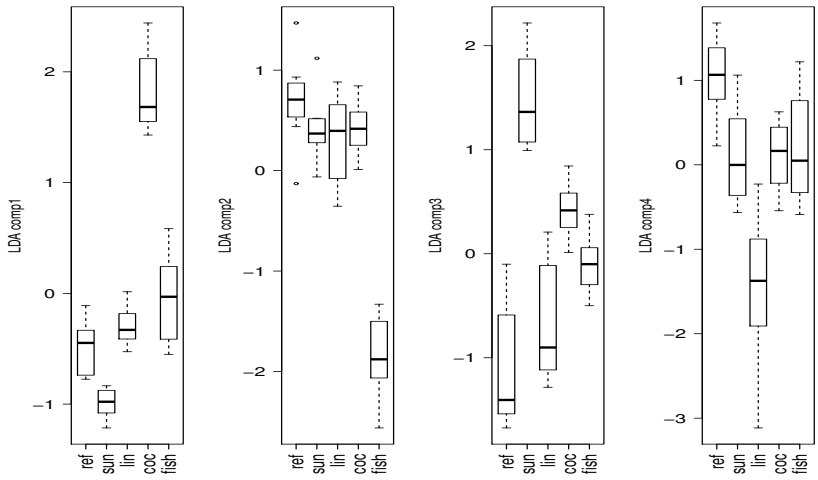

Figure 8.2 – Nutrimouse dataset: Diet factor. Box plots of the four LDA canonical variates associated with the first six AoV-PLS components.

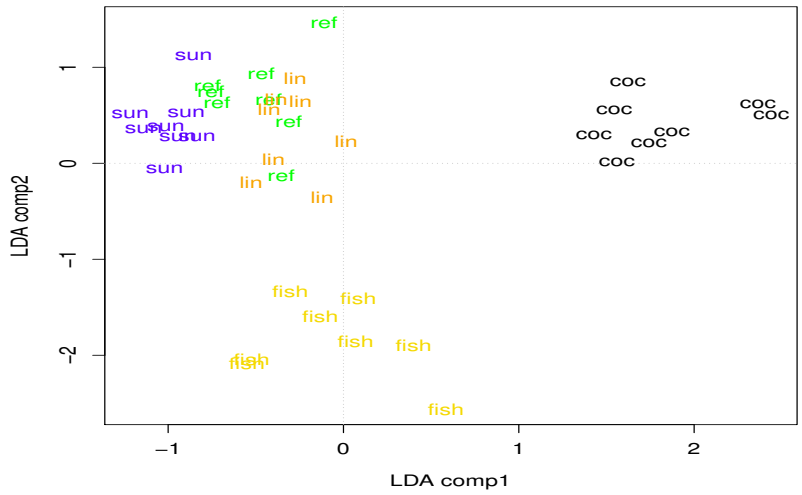

Figure 8.3 – Nutrimouse dataset: diet factor. Representation of the individuals on the first two canonical variates of LDA performed on the first six AoV-PLS components.

ten first ANOVA-PCA and ASCA components as well as the p-values associated with a one way-ANOVA performed on each component are given in table 8.4.

ANOVA-PCA led to six significant components (2, 3, 4, 5, 8 and 9) with the components 2, 3, 4 and 5 explaining up to 50% of the variation in $\boldsymbol{X_D}$. Box plots of

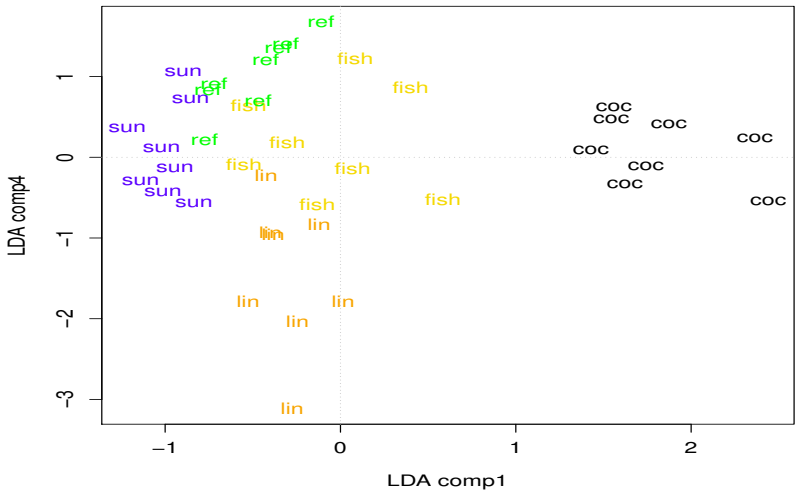

Figure 8.4 – Nutrimouse dataset: diet factor. Representation of the individuals on the first and fourth canonical variates of LDA performed on the first six AoV-PLS components.

Table 8.4 – The percentages of total variance in $\boldsymbol{X_D}$ and $\boldsymbol{X_D + E}$ explained by the ten first ANOVA-PCA and ASCA components. P-values associated with one way-ANOVA performed on each component.

	ANOVA-PCA			ASCA		
	$\boldsymbol{X_D}$	$\boldsymbol{X_D + E}$	p-values	$\boldsymbol{X_D}$	$\boldsymbol{X_D + E}$	p-values
comp 1	5.73	47.66	0.30	47.06	9.29	0.01
comp 2	17.79	7.03	0.00	27.28	5.38	0
comp 3	11.64	6.75	0.01	13.87	2.74	0.01
comp 4	14.85	5.52	0.00	11.79	2.33	0
comp 5	5.28	4.2	0.05	0.05	1.47	1
comp 6	3.52	3.13	0.31	0	1.16	1
comp 7	2.49	2.77	0.11	0	5.24	1
comp 8	8.64	2.4	0.05	0	1.45	1
comp 9	3.28	2.23	0.03	0	1.5	1
comp 10	2.54	1.73	0.13	0	0.95	1

these first four significant components are presented in figure 8.5. Only the second component of ANOVA-PCA shows a clear separation of the groups lin and fish. The fourth component sets apart ref and lin. LDA was also applied to the first four significant components (2,3,4,5). The box plots of the four canonical variates obtained from this analysis are presented in figure 8.6. The first LDA canonical variate separates

the group fish, the other groups are not well discriminated. Likewise the group coc is distinguished in the second canonical variate while the third and the fourth canonical variates do not show a clear discrimination of the groups.

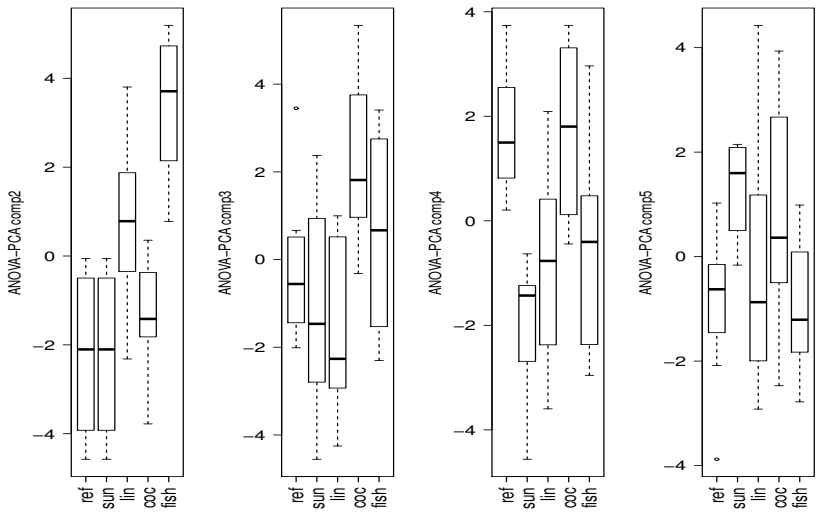

Figure 8.5 – Nutrimouse dataset: diet factor. Box plots of the second, third, fourth and fifth ANOVA-PCA components.

Regarding ASCA, PCA performed on $\boldsymbol{X_D}$ naturally led to four significant components. The data from $\boldsymbol{X_D} + \boldsymbol{E}$ were superimposed on these components and the scores thus obtained were depicted as box plots associated with the five diet groups (figure 8.7). Only the groups fish and ref are well discriminated by the components 2 and 4 of ASCA, respectively. A LDA is performed on the first four significant ASCA components. The Box plots associated to the four LDA canonical variates of this analysis are given in figure 8.8.

We see a clearer separation of the groups on the canonical variates associated to the significant ASCA components compared to the canonical variates associated to ANOVA-PCA. The first canonical variate separates the groups sun and fish, the second canonical variate separates the groups coc and ref and the third canonical variates opposes the groups ref and lin to the other groups. However, the group lin is not separated in any of the canonical variates associated to ASCA. In contrast, we recall that with the LDA applied on the AoV-PLS significant latent components, all the groups of the factor diet were well separated on the canonical variates.

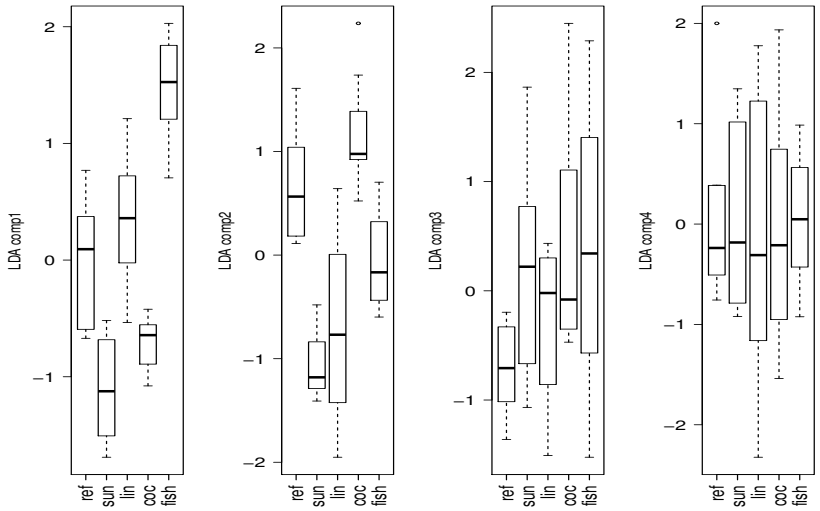

Figure 8.6 – Nutrimouse dataset: diet factor. Box plots of the LDA canonical variate associated with the significant ANOVA-PCA components: 2,3,4,5

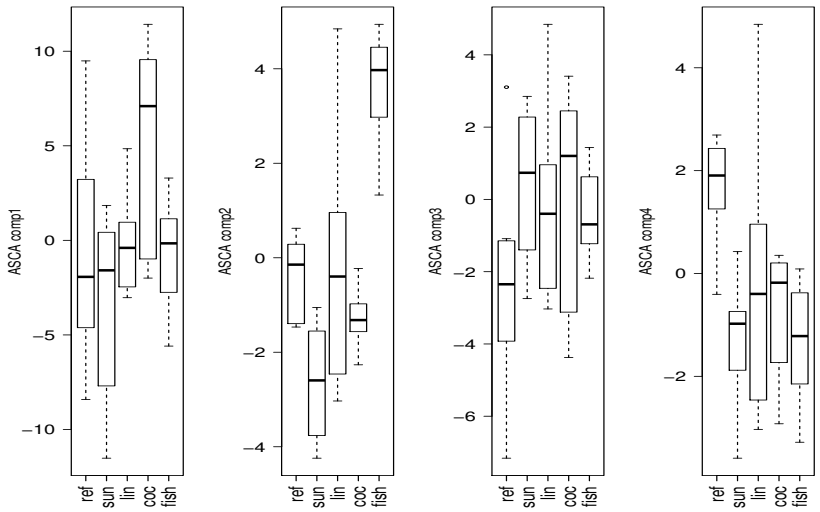

Figure 8.7 – Nutrimouse dataset: diet factor. Box plots of ASCA first four components.

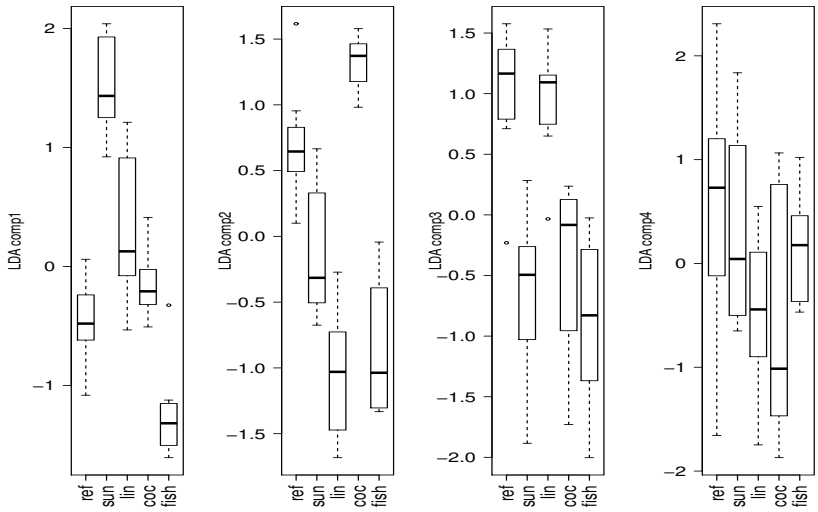

Figure 8.8 – Nutrimouse dataset: diet factor. Box plots of LDA canonical variates associated with the first four ASCA components (using the ACP with supplementary individuals)

Chapter 9

Conclusion and perspectives

Conclusion

The research work carried out herein deals with various aspects of data analysis. New tools, strategies and methods are developed. Starting with the measures of association between two datasets, we investigated three similarity coefficients, namely the multivariate correlation coefficient, the Procrustes index and the RV coefficient. We highlighted the properties, the use and misuse of these coefficients. A new correction of the RV coefficient was proposed, named the adjusted RV coefficient, to cop with the agreement due to chance between two datasets. This correction showed better results compared to a modified RV coefficient proposed by Smilde *et al.* (2009).

To help the practitioners in the choice of the standardization of a dataset before applying a Principal Component Analysis (PCA) or a PLS regression, a new strategy was proposed. It consists of dividing each variable by its standard deviation to the power $\boldsymbol{\alpha}$, with $\boldsymbol{\alpha}$, a parameter between 0 and 1. The particular cases, $\alpha = 0$, $\alpha = 0.5$ and $\alpha = 1$ corresponding to non-standardization, Pareto standardization and standardization to unit variances, respectively. This new continuum called $\boldsymbol{\alpha}$-standardization, is applied to the $\boldsymbol{X}$-dataset before applying PCA or PLS-regression. The effect of the parameter $\boldsymbol{\alpha}$ on the $\boldsymbol{X}$-dataset is assessed. It turns out that one can find an appropriate $\boldsymbol{\alpha}$ according to the data at hand and the objective of the analysis: exploration or prediction.

In the context of regression, the quasi-collinearity of the dataset $\boldsymbol{X}$ often leads to unstable models. A new strategy, named αM-standardization was proposed whereby a gradual decorrelation of the $\boldsymbol{X}$-variables is achieved. This strategy leads to a family of models that encompass PLS-regression and redundancy analysis. The tuning parameter $\boldsymbol{\alpha}$ that realizes a compromise between recovering the total variance of $\boldsymbol{Y}$ and the risk of defining unstable models can be found by means of cross-validation procedure.

The continuum αM-standardization has inspired a new biased regression approach called Biased Power regression (BP-regression). Properties related to this new Biased regression were shown, particularly the trade off between the introduction of a small bias against the reduction of the variation of the regression coefficients. A comparison with Ridge regression and Generalized Ridge regression was also done. BP-regression led to comparable results with Ridge regression in terms of prediction.

For the simultaneous analysis of several datasets (K, say), measured on the same individuals, a method called "Common Components and Specific Weights Analysis" (CCSWA) has gained some popularity among practitioners. It is also often referred as "ComDim", as an abbreviation for Common Dimension. In this research work, an extension of this method to the prediction of a new dataset $\boldsymbol{Y}$ from K datasets is proposed. This new extension is called P-ComDim for Predictive ComDim. Properties related to P-ComDim and several algorithms are developed including non iterative and iterative algorithms. A comparison with MultiBlock PLS is performed. Two illustrations on the basis of sensory and TD-NMR datasets were discussed, showing the advantage of using P-Comdim over MultiBlock-PLS.

Finally, as part of a collaboration with Phan (Physiologie des Adaptations Nutritionnelles) unit at CHU of Nantes, a new method for the analysis of a multivariate dataset depending on several factors is proposed. The new method, called AoV-PLS, can be seen as an extension of PLS-Discriminant Analysis (PLS-DA) to the case of several factors. Moreover, it stands as a compromise between two well known method for the analysis of this kind of data, namely, the ANOVA-Simultaneous Component Analysis, ASCA) and ANOVA-PCA. The application of AoV-PLS and the comparison with the others methods on the basis of a dataset provided by the Phan unit and a nutrigenomic study (Martin *et al.*, 2007) showed promising results.

Perspectives

The research work presented herein brings up new perspectives.

Regarding the new correction of RV coefficient (the adjusted RV coefficient), a comparison with the modified RV coefficient could also be considered in the context of chemometrics, particularly with high dimensional data. Moreover, another correction of the RV coefficient was proposed by Mayer *et al.* (2011) inspired by the adjusted *R-square*. It would be interesting to compare the properties and compare the efficiency of all these corrections of RV coefficient on the basis of simulated data and real case studies.

For the continuum α-standardization, a clear and direct strategy to select the appropriate $\boldsymbol{\alpha}$ in the context of PCA is missing. More work is needed for this respect to investigate procedures of cross-validation for PCA similarly to those proposed for

PLS regression.

For the continuum αM-standardization, it would be interesting to apply this transformation in the contexts of linear and quadratic discriminant analysis. Indeed, using the αM-standardized dataset instead of the original dataset will result in a regularization of the Mahalanobis distance among the individuals.

Another extension of BP-regression which is worth investigating concerns the context of classification and discrimination. This would provide an alternative to the so-called regularized discriminant analysis (Friedman, 1989). The regularization proposed in this latter method draws from Ridge regression and, therefore, should easily be adapted to a regularization akin to BP-regression.

As for the analysis of several datasets simultaneously, a new extension of P-ComDim to the context of path modeling is currently developed within the StatSC-ONIRS unit.

Lastly, more work is required to compare AoV-PLS strategy to other methods used in the context of multivariate data depending on several factors, like ANOVA-ComDim (Jouan-Rimbaud Bouveresse *et al.*, 2011) and ANOVA-Multiblock Orthogonal PLS (Boccard & Rudaz, 2016). Furthermore, investigations are necessary to study the effect of unbalanced experimental designs on these methods of analysis.

Appendix A

Publications

A.1 Association indices between two datasets

Reference:

El Ghaziri, A. and Qannari, E.M. (2015).“Measures of association between two datasets; Application to sensory data”. Food Quality and Preference 40, Part A: 116-124.

A.2 Continuum standardization

Reference:

El Ghaziri, A. and Qannari, E.M. (2015).“A continuum standardization of the variables. Application to Principal Components Analysis and PLS-regression”. Journal of Chemometrics and Intelligent Laboratory Systems 148, 95-105

A.3 Continuum decorrelation

Reference (submitted):
Qannari, E. M., El Ghaziri, A., Hanafi, M. (2016). "A continuum decorrelation of the variables; Application to multivariate regression".

A.4 Biased Power regression

Reference (submitted):
Qannari, E. M. and El Ghaziri, A. (2016). "Biased Power regression: A new biased estimation procedure in linear regression".

A.5 P-ComDim

Reference:

El Ghaziri, A., Qannari, E. M., Cariou, V. and Rutledge, D. N. (2016).“Analysis of multiblock datasets using ComDim. Overview and extension to the analysis of (K+1) datasets”. Journal of Chemometrics, 30: 420-429

A.6 AoV-PLS

Reference:

El Ghaziri, A., Qannari, E.M., Moyon, T., Alexandre-Gouabau, M.-C. (2015). " AoV-PLS : a new method for the analysis of multivariate data depending on several factors ". Electronic Journal of Applied Statistical Analysis 8, (2), 214-235.

Bibliography

Abdi, H., Valentin, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: Distatis, theory and applications. *Food Quality and Preference*, *18*, 627–640.

Boccard, J., & Rudaz, S. (2016). Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica Chimica Acta*, *920*, 18–28.

Climaco-Pinto, R., Barros, A. S., Locquet, N., Schmidtke, L., & Rutledge, D. N. (2009). Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability. *Analytica Chimica Acta*, *653*, 131–142.

Courcoux, P., Devaux, M.-F., & Bouchet, B. (2002). Simultaneous decomposition of multivariate images using three-way data analysis: Application to the comparison of cereal grains by confocal laser scanning microscopy. *Chemometrics and Intelligent Laboratory Systems*, *62*, 103–113.

Dairou, V., & Sieffermann, J.-M. (2002). A comparison of 14 jams characterized by conventional profile and a quick original method, the flash profile. *Journal of Food Science*, *67*, 826–834.

Dehlholm, C., Brockhoff, P. B., Meinert, L., Aaslyng, M. D., & Bredie, W. L. P. (2012). Rapid descriptive sensory methods - comparison of free multiple sorting, partial napping, napping, flash profiling and conventional profiling. *Food Quality and Preference*, *26*, 267–277.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, *29*, 751–761.

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., & Nicod, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. *Food Quality and Preference*, *15*, 781–791.

Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., & Nicod, H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference*, *17*, 604–614.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165–175.

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

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage : the James-Stein and ridge regression estimators*. New York : Marcel Dekker. Includes bibliographical references (pages 591-617) and indexes.

Hanafi, M., Kohler, A., & Qannari, E. M. (2010). Shedding new light on hierarchical principal component analysis. *Journal of Chemometrics*, *24*, 703–709.

Hanafi, M., Mazerolles, G., Dufour, E., & Qannari, E. M. (2006). Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques. *Journal of Chemometrics*, *20*, 172–183.

Harrington, P. d. B., Vieira, N. E., Chen, P., Espinoza, J., Nien, J. K., Romero, R., & Yergey, A. L. (2006). Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry. *Chemometrics and Intelligent Laboratory Systems*, *82*, 283–293.

Harrington, P. d. B., Vieira, N. E., Espinoza, J., Nien, J. K., Romero, R., & Yergey, A. L. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta*, *544*, 118–127.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*, 81–102.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

IFREMER (1994). *spécial environnement littoral*. 47-48. EQUINOXE.

de Jong, S., Wise, B. M., & Ricker, N. L. (2001). Canonical partial least squares and continuum power regression. *Journal of Chemometrics*, *15*, 85–100.

Jouan-Rimbaud Bouveresse, D., Climaco Pinto, R., Schmidtke, L. M., Locquet, N., & Rutledge, D. N. (2011). Identification of significant factors by an extension of ANOVA-PCA based on multi-block analysis. *Chemometrics and Intelligent Laboratory Systems*, *106*, 173–182.

Larsen, F. H., van den Berg, F., & Engelsen, S. B. (2006). An exploratory chemometric study of ^{1}H NMR spectra of table wines. *Journal of Chemometrics*, *20*, 198–208.

Lavit, C. (1988). *ANALYSE CONJOINTE DE TABLEAUX QUANTITATIFS*. Paris: Masson.

Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The act (statis method). *Computational Statistics & Data Analysis*, *18*, 97–119.

Mammasse, N., & Schlich, P. (2014). Adequate number of consumers in a liking test. insights from resampling in seven studies. *Food Quality and Preference*, *31*, 124–128.

Martin, P. G. P., Guillou, H., Lasserre, F., Déjean, S., Lan, A., Pascussi, J.-M., SanCristobal, M., Legrand, P., Besse, P., & Pineau, T. (2007). Novel aspects of PPAR$\boldsymbol{\alpha}$-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*, *45*, 767–777.

Mayer, C., Lorent, J., & Horgan, G. (2011). Exploratory analysis of multiple omics datasets using the adjusted rv coefficient, . *10*. Statistical Applications in Genetics and Molecular Biology.

Mazerolles, G., Devaux, M. F., Dufour, E., Qannari, E. M., & Courcoux, P. (2002). Chemometric methods for the coupling of spectroscopic techniques and for the extraction of the relevant information contained in the spectral data tables. *Chemometrics and Intelligent Laboratory Systems*, *63*, 57–68.

Mazerolles, G., Hanafi, M., Dufour, E., Bertrand, D., & Qannari, E. M. (2006). Common components and specific weights analysis: A chemometric method for dealing with complexity of food products. *Chemometrics and Intelligent Laboratory Systems*, *81*, 41–49.

Meyners, M. (2003). Methods to analyse sensory profiling data – a comparison. *Food Quality and Preference*, *14*, 507–514.

Meyners, M., Kunert, J., & Qannari, E. M. (2000). Comparing generalized procrustes analysis and statis. *Food Quality and Preference*, *11*, 77–83.

Nielsen, J. P., Bertrand, D., Micklander, E., Courcoux, P., & Munck, L. (2001). Study of NIR spectra, particle size distributions and chemical parameters of wheat flours: a multi-way approach. *Journal of Near Infrared Spectroscopy*, *9*, 275–285.

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

Pizarro, C., Esteban-Díez, I., Rodríguez-Tecedor, S., & González-Sáiz, J. M. (2013). A sensory approach for the monitoring of accelerated red wine aging processes using multi-block methods. *Food Quality and Preference*, *28*, 519–530.

Qannari, E. M., Wakeling, I., & MacFie, H. J. H. (1995). Second sensometrics meetinga hierarchy of models for analysing sensory data. *Food Quality and Preference*, *6*, 309–314.

Ramsay, J. O., Berge, J. T., & Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, *49*, 403–423. Psychometrika.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, *66*, 846–850.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L. P., & Bom Frøst, M. (2014). Comparison of three sensory profiling methods based on consumer perception: Cata, cata with intensity and napping. *Food Quality and Preference*, *32, Part B*, 160–166.

Robert, A. v. d. B., Huub, C. H., Johan, A. W., Age, K. S., & Mariët, J. v. d. W. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, *7*, 142–142.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Applied Statistics*, *25*, 257–265.

Rutledge, D. N., Barros, A. S., Vackier, M. C., Baumberger, S., & Lapierre, C. (1999). In *Advances in Magnetic Resonance in Food Science* chapter Analysis of Time Domain NMR and Other Signals. (pp. 203–216). Oxford: McGraw Hill Professional. (Royal Society of Chemistry ed.). doi:10.1533/9781845698133.4.203.

Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T. Næs, & E. Risvik (Eds.), *Data Handling in Science and Technology* (pp. 259–306). Elsevier volume 16.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. Royal Statistical Society, Serie B*, *40*, 234–238.

Smilde, A. K., Jansen, J. J., Hoefsloot, H., Lamers, R.-J., Greef, J., & Timmerman, M. (2005). Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data. *Bioinformatics*, *21*, 3043–3048.

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., & Van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*, *25*, 401–405.

Vidal, L., Cadena, R. S., Antùnez, L., Giménez, A., Varela, P., & Ares, G. (2014). Stability of sample configurations from projective mapping: How many consumers are necessary? *Food Quality and Preference*, *34*, 79–87.

Vigneau, E., & Thomas, F. (2012). Model calibration and feature selection for orange juice authentication by ^{1}H NMR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *117*, 22–30.

Vis, D., Westerhuis, J., Smilde, A. K., & van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, *8*, 322.

Williams, A. A., & Langron, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, *35*, 558–568.

Wise, B. M., & Ricker, N. L. (1993). Identification of finite impulse response models with continuum regression. *Journal of Chemometrics*, *7*, 1–14.

Worch, T. (2013). Prefmfa, a solution taking the best of both internal and external preference mapping techniques. *Food Quality and Preference*, *30*, 180–191.

Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., & Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics*, *25*, 561–567.

UNIVERSITE BRETAGNE LOIRE

Oniris ÉCOLE NATIONALE VÉTÉRINAIRE, AGROALIMENTAIRE ET DE L'ALIMENTATION SANTÉ ET ALIMENTATION AU CŒUR DE LA VIE

Thèse de Doctorat

Angélina EL GHAZIRI

Relation entre tableaux de données: exploration et prédiction

Relating datasets: exploration and prediction

Résumé

La recherche développée dans le cadre de cette thèse aborde différents aspects relevant de l'analyse statistique de données.

Dans un premier temps, une analyse de trois indices d'associations entre deux tableaux de données est développée.

Par la suite, des stratégies d'analyse liées à la standardisation de tableaux de données avec des applications en analyse en composantes principales (ACP) et en régression, notamment la régression PLS sont présentées. La première stratégie consiste à proposer une standardisation continuum des variables. Une standardisation plus générale est aussi abordée consistant à réduire de manière graduelle non seulement les variances des variables mais également les corrélations entre ces variables. De là, une approche continuum de régression a été élaborée regroupant l'analyse des redondances et la régression PLS. Par ailleurs, cette dernière standardisation a inspiré une démarche de régression biaisée dans le cadre de régression linéaire multiple. Les propriétés d'une telle démarche sont étudiées et les résultats sont comparés à ceux de la régression Ridge.

Dans le cadre de l'analyse de plusieurs tableaux de données, une extension de la méthode ComDim pour la situation de K+1 tableaux est développée. Les propriétés de cette méthode, appelée P-ComDim, sont étudiées et comparées à celles de Multiblock PLS.

Enfin, la situation où il s'agit d'évaluer l'effet de plusieurs facteurs sur des données multivariées est considérée et une nouvelle stratégie d'analyse est proposée.

Mots clés
Comparaison de deux tableaux; tableaux multiples; Analyse en Composante Principale; ComDim; Analyse des Redondances; régression biaisée; régression PLS.

Abstract

The research developed in this thesis deals with several statistical aspects for analyzing datasets.

Firstly, investigations of the properties of several association indices commonly used by practitioners are undergone.

Secondly, different strategies related to the standardization of the datasets with application to principal component analysis (PCA) and regression, especially PLS-regression were developed. The first strategy consists of a continuum standardization of the variables. The interest of such standardization in PCA and PLS-regression is emphasized.

A more general standardization is also discussed which consists in reducing gradually not only the variances of the variables but also their correlations. Thereafter, a continuum approach was developed combining Redundancy Analysis and PLS-regression. Moreover, this new standardization inspired a biased regression model in multiple linear regression. Properties related to this approach are studied and the results are compared on the basis of case studies with those of Ridge regression.

In the context of the analysis of several datasets in an exploratory perspective, the method called ComDim, has certainly raised interest among practitioners. An extension of this method for the analysis of K+1 datasets was developed. Properties related to this method, called P-ComDim, are studied and compared to Multiblock PLS.

Finally, for the analysis of datasets depending on several factors, a new approach based on PLS regression is proposed.

Key Words
Comparison between two datasets; mutiblocks datasets; Principal Component Analysis; ComDim, Redundancy Analysis; biased regression; PLS-regression