UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES ET DES TECHNIQUES

ÉCOLE DOCTORALE VENAM :
Végétal, Environnement, Nutrition, Agroalimentaire et Mer.

Année 2014

Classification de variables autour de variables latentes avec filtrage de l'information : application à des données en grande dimension

---

THÈSE DE DOCTORAT

Discipline : Sciences Mathématiques Appliquées
Spécialité : Statistique Appliquée

Présentée

et soutenue publiquement par

Mingkun CHEN

Le 23 octobre 2014, devant le jury ci-dessous

Rapporteurs	Robert SABATIER	Professeur à Université Montpellier 1
	Christian DERQUENNE	Chercheur Senior à EDF R&D, Clamart
Examinateurs	Jean-Benoît HARDOUIN	Maître de Conférences à Université de Nantes
	Jean-Philippe ANTIGNAC	Ingénieur de Recherche à Oniris, Nantes
	Joachim KUNERT	Professeur à Université Technique de Dortmund
	El Mostafa QANNARI	Professeur à Oniris, Nantes
	Evelyne VIGNEAU	Professeur à Oniris, Nantes

Directeur de thèse : Evelyne VIGNEAU

Remerciements

Je voudrais adresser mes sincères remerciements à ma directrice de thèse, Madame Evelyne Vigneau, pour la confiance qu'elle m'a accordée, pour ses conseils avisés, pour son encadrement et pour toutes les heures qu'elle a consacrées à diriger cette recherche et à m'aider à rédiger cette thèse. Merci beaucoup !

Je voudrais remercier mon co-directeur de thèse, Monsieur El Mostafa Qannari, pour m'avoir également accordé sa confiance, pour ses conseils précieux, et le temps consacré à la relecture des documents que je lui ai adressés.

Je tiens également à remercier toutes les personnes du laboratoire qui m'ont accompagnées durant ces trois années. Je me souviendrai toute ma vie de vous et des moments passés ensemble.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Données en grande dimension

Avec les récents développements des méthodes de mesures analytiques (chromatographie, spectroscopie infra-rouge, spectroscopie en résonance magnétique nucléaire, spectrométrie de masse, etc.), on est souvent confronté à la grande dimension des données (nombre important de variables). Par exemple, dans le domaine de la biologie intégrative, les techniques à haut débit produisent une énorme quantité de données, appelées données-omiques (génomiques, transcriptomiques, protéomiques et métabolomiques). Les techniques de spectroscopie conduisent également à considérer un très grand nombre de variables fortement redondantes par nature.

Trois caractéristiques importantes des données en grande dimension remettent en cause la plupart de méthodes classiques en statistique :

1. Fort déséquilibre entre le nombre d'observations n (échantillons, patients ...), de l'ordre $10 \sim 10^2$, et le très grand nombre de variables p (gènes, protéines ...) de l'ordre $10^2 \sim 10^4$,
2. La présence de forte redondance (corrélations simples ou multiples) parmi l'ensemble des variables.
3. La présence de variables sur-numéraires, non-informatives, parmi les variables mesurées.

1.2 Objectif et directions

Dans ce contexte de données en grande dimension, l'objectif statistique est de développer des méthodes permettant de surmonter les difficultés inhérentes à la structure de ce type de données afin d'augmenter la qualité des modèles statistiques en termes de précision, robustesse et d'interprétabilité. Toutes ces conditions sont requises et répondent aux attentes des utilisateurs. Deux niveaux d'analyse de données peuvent être identifiés, l'un surtout exploratoire, l'autre lié à la modélisation :

1. L'exploration de données en utilisant des techniques de réduction du nombre de variables, que ce soit sur la base de variables latentes (analyse en composantes principales, analyse factorielle, ...) ou que ce soit par sélection de variables, est une étape préliminaire, utile, voire indispensable. Cette étape peut être considérée en tant que telle pour l'identification des éléments principaux structurant les données, ou comme base à l'extraction d'une partie de l'information. Elle peut également être directement combinée avec la phase de modélisation. Nous nous intéressons spécifiquement aux méthodes utilisant l'analyse des redondances et corrélations entre les variables. Parmi ces méthodes, l'approche par classification de variables autour de variables latentes est au cœur de nos travaux.
2. La sélection de variables utilisées dans un modèle de régression ou de discrimination est une étape clé non seulement du fait de l'impossibilité d'utiliser des méthodes classiques de régression (méthode des moindres carrés) quand le nombre d'observations est inférieur au nombre de variables explicatives, mais aussi parce que, souvent, la réponse que l'on veut prédire n'est liée qu'à un petit nombre de variables mesurées (tels que les biomarqueurs recherchés en biologie intégrative). Ajuster un modèle n'incluant qu'un petit nombre de variables utiles, dit modèle parcimonieux, permet non seulement d'avoir de meilleures qualités statistiques mais également de faciliter l'interprétation des phénomènes sous-jacents.

Dans les deux sections suivantes, nous allons rapidement balayer le contexte bibliographique dans lequel nous nous plaçons, tant du point de vue de l'analyse exploratoire que de celui de l'analyse prédictive.

1.2.1 Analyse exploratoire

L'analyse en Composantes Principales (ACP) est beaucoup utilisée pour l'exploration et la représentation des données quantitatives. Il s'agit d'une méthode de réduction de la dimensionnalité qui consiste à rechercher des variables latentes, qui sont des combinaisons linéaires des variables d'origine, restituant la plus grande part possible de l'inertie. Ces composantes principales sont orthogonales et permettent de révéler les traits saillants du nuage des points-variables. Cependant, dans le cas des données en grande dimension, une composante principale, calculée par la combinaison linéaire d'un très grand nombre de variables, peut devenir difficile à interpréter. Pour l'analyse du nuage des points-variables, la structure de groupe de variables n'est pas facile à identifier.

Dans le but d'explorer la structure de variables, la classification de variables offre une alternative qui permet d'identifier des groupes de variables, mais aussi de réduire la dimensionalité de l'espace factoriel étudié. La procédure VARCLUS du logiciel SAS (Sarle, 1990) et la procédure de Classification de Variables autour de Variables Latentes (CLV) proposée par Vigneau et Qannari (2003) sont deux méthodes de classification de variables basées sur une approche factorielle. La méthode CLV repose sur la recherche de variables latentes, chacune d'elles représentant un groupe de variables. Ces nouvelles variables synthétiques de groupes ne sont pas forcément orthogonales entre elles mais elles sont plus facilement interprétables que les composantes principales de l'ACP. En effet chaque variable latente de groupe est une combinaison linéaire qui n'implique que les variables du groupe qui lui est associé.

Réduire la dimension de l'espace factoriel en associant au mieux les variables de départ à de nouvelles composantes rejoint la notion de recherche de *simple structure* qui sous-tend les techniques de rotation (Varimax, Quartimax, ...) (Jolliffe, 2002). Dans le domaine de la psychométrie, en relation avec les solutions de l'analyse factorielle, Revelle et Rocklin (1979) introduisent un critère de recherche d'une *Very Simple Structure* (VSS), et Bernaards et Jennrich (2003) évoquent la recherche d'une *perfect simple structure*. Dans ces deux cas, l'idée est que chaque variable n'a finalement qu'un seul coefficient non nul sur l'ensemble des composantes extraites. Ceci

conduit donc à associer aux composantes une partition des variables, à l'instar de ce qui est fait par la méthode de classification de variables CLV.

Les variables latentes de groupe, obtenues par CLV, peuvent également être vues comme une version *sparse* des composantes principales de l'ACP. En effet, pour chaque variable latente CLV, les variables qui n'appartiennent pas à ce groupe ont d'office un coefficient nul dans la combinaison linéaire, et l'ensemble des variables latentes est construit de sorte à restituer au mieux la variabilité totale.

La notion de solutions *sparse* est souvent mise en avant comme un avantage dans le cas de grand ensembles de variables. Elle est considérée comme une approche permettant d'améliorer l'interprétation du modèle. L'idée principale est de chercher une solution en ne considérant qu'une partie de l'information. Par exemple, pour une analyse non supervisée par ACP, chercher une composante principale restituant une grande part de la variance totale mais ne combinant qu'un sous-ensemble de variables peut faciliter à l'interprétation des résultats. Dans cet objectif, plusieurs approches ont été proposées parmi lesquelles les méthodes *Sparse PCA* (Zou *et al.*, 2006), SCoTLASS (Jolliffe *et al.*, 2003), sPCA-rSVD (Shen et Huang, 2008) qui utilisent une pénalité *Lasso* pour trouver les composantes *sparse* et l'approche *gene shaving* (Hastie *et al.*, 2000) qui consiste à supprimer une proportion des gènes ayant les plus petits *loadings* sur les premières composantes principales. Wang et Wu (2012) ont, quant à eux, proposé récemment un algorithme d'élimination itérative pour trouver des composantes *sparse*.

Dans le chapitre 3, nous introduirons l'idée de *sparsité* dans le cadre de la classification de variables pour éliminer les variables "de bruit" qui sont potentiellement présentes lorsque de très nombreuses variables sont mesurées simultanément.

La définition de variables "de bruit" dans un ensemble de variables est en fait complexe et mérite d'être un peu précisée. Dans notre contexte, une variable est assimilée à du bruit si elle n'a pas d'intérêt pour l'objectif visé qui est de mettre en lumière la structure du nuage des points-variables. Donc pour la classification de variables, les variables qui ne rentrent pas dans un groupe avec d'autres variables ou entravent la recherche de la "vraie" structure de groupe vont être considérées comme des variables "de bruit". Les méthodes proposées dans chapitre 3 ont pour objectif

d'améliorer l'interprétation des résultats de classification en éliminant l'influence de cette sorte de variables.

1.2.2 Analyse prédictive

Lorsque l'objectif est de prédire une réponse, les caractéristiques des données en grande dimension posent une question difficile pour les statisticiens : la sélection de variables. Elle consiste à chercher un sous-ensemble de variables explicatives afin de construire un modèle simple et aussi prédictif que celui qui intégrerait toutes les variables. Plusieurs raisons justifient cette démarche de sélection de variables :

1. le nombre de variables pertinentes est souvent beaucoup plus petit que le nombre total de variables. Dans les données sur l'expression des gènes, il est supposé qu'une réponse, comme par exemple le type de cancer, est déterminée seulement par quelques groupes de gènes.
2. ajouter des variables inutiles dans un modèle peut affecter l'estimation des coefficients des variables importantes. Il est alors naturel de penser à éliminer les variables inutiles pour ne retenir que les variables importantes.
3. la présence de forte redondances (colinéarités) entre certains prédicteurs conduit à des estimations peu précises des coefficients du modèle, qui sont également fortement liés entre eux.

La performance prédictive et l'interprétation sont deux aspects complémentaires pour évaluer un modèle de régression. Certaines méthodes telles que les réseaux de neurones privilégient l'aspect prédictif mais comprendre un phénomène est tout aussi important aux yeux de nombreux utilisateurs. Dans le domaine des données-omiques, le biologiste cherche non seulement un modèle prédictif, par exemple pour la prédiction de maladies, mais aussi aimerait pouvoir identifier les biomarqueurs les plus pertinents.

Dans les articles scientifiques liés à la régression linéaire, il est possible d'identifier différents types d'approches visant à obtenir des modèles "parcimonieux" et donc plus simples à interpréter :

1. L'élimination de variables a priori : il s'agit d'une étape préalable pour éliminer un grand nombre de variables qui n'ont pas forcément d'intérêt. L'étape suivante de sélection de variables peut ensuite être améliorée à la fois en termes de vitesse et de précision. Des règles d'élimination comme le *Sure independence screening* de Fan et Lv (2008), la *Safe Feature Elimination* de El Ghaoui *et al.* (2010), et *Strong rule* de Tibshirani *et al.* (2012) ont été proposées.
2. La sélection de sous-ensembles de prédicteurs : le moyen le plus direct est d'essayer tous les sous-ensembles de variable possibles, et de choisir le plus pertinent sur la base des critères de qualité comme le R^2, le R^2 ajusté, le AIC (*Akaike's Information Criterion*), le BIC (*Bayesian information criterion*) , le Cp de Mallows (1973) etc. Mais il n'est pas raisonnable, ni même possible, d'évaluer les 2^p modèles possibles lorsque p , le nombre du variables, est grand. Chercher un bon chemin parmi l'ensemble des solutions est plus réaliste. C'est ce qui est fait avec les méthodes de sélection pas à pas (*forward, backward ou stepwise*) ou la régression *stagewise* (Hastie *et al.*, 2001a).
3. La régularisation : parmi les méthodes de régularisation, la régression *Ridge* et la régression *Lasso* sont les plus connues. La régression *Ridge* (Hoerl et Kennard, 1970) utilise une pénalité L^2 qui permet, au prix de l'introduction d'un biais, de trouver une solution lorsque le nombre de variables, p, est supérieur au nombre d'observations, n, et qui, d'une manière générale, permet d'améliorer la qualité de prédiction du modèle dans les problèmes mal conditionnés. La régression *Lasso* (Tibshirani, 1996), avec une pénalité L^1, produit souvent des valeurs estimées pour les coefficients non nulles ou exactement égales à 0. Ainsi une sélection de variables est faite de façon automatique pendant la régression. Cette propriété a suscité beaucoup d'intérêt et de nombreuses extensions comme *Elastic Net* (Zou et Hastie, 2005), ou *Group lasso* (Yuan et Lin, 2006). Hesterberg *et al.* (2008) proposent une revue très complète des méthodes de régression linéaire avec pénalité L^1.
4. La réduction de la dimension : les procédures de réduction de la dimension comme l'ACP, la Régression sur Composantes Principales (PCR) et la régression PLS (Wold, 1966) sont particulièrement adaptées aux données en grande

dimension ($p \gg n$) principalement lorsqu'il y a de fortes colinéarités entre les variables. Cependant, dans ce type de procédures, aucune sélection de variables n'est faite. La combinaison d'une approche par PCR et de "nettoyage" a été proposée par Bair *et al.* (2006). D'autres approches combinent la régression PLS et la sélection de variables comme l'UVE-PLS (Centner *et al.*, 1996), la GA-PLS (Hasegawa *et al.*, 1997). La revue de Mehmood *et al.* (2012) présente différentes méthodes de sélection de variables dans le cadre de la régression PLS. La combinaison d'une pénalité L^1 avec la régression PLS a aussi été proposée dans des approches récentes. Leur implémentation est similaire à celle des méthodes d'analyse en composantes principales *sparse* mais concerne le modèle de régression. On citera par exemple la méthode *Sparse* PLS de Chun et Keles (2010) qui est similaire aux approches SCoTLASS et ACP *sparse*, et la méthode *Sparse* PLS de Le Cao *et al.* (2008) qui est similaire à sPCA-rSVD.

Cependant, la très grande majorité des méthodes évoquées ci-dessus ne permettent pas d'explorer en même temps la structure des inter-corrélations entre les prédicteurs et d'établir un modèle de prédiction. Dans le cadre des données-omiques, la présence de groupes de variables n'est plus considérée à l'issue d'une phase de sélection. En effet cette étape ne permet de retenir que quelques-unes des variables de départ. L'identification de groupes de variables, chacun d'eux étant associé à une variable latente prédictive, a pour avantage de révéler les structures réelles et les formes intéressantes dans l'espace des variables. Dans le chapitre 4, nous proposons de combiner la classification de variables et la sélection de variables afin de prendre en compte la structure des inter-corrélations entre les variables, tout en mettant en lumière les informations ayant le plus grand pouvoir prédictif.

1.3 Structure de la suite du manuscrit

Dans le chapitre 2, un rappel sera fait au sujet de la méthode de Classification de variables autour de Variables Latentes (CLV), qui est le point de départ de ce travail. Le principe des approches par pénalisation de type L^1 (méthode *Lasso*, en particulier), sur lequel nous nous sommes également appuyé, est détaillé ensuite.

Dans le chapitre 3, nous proposons deux nouvelles approches basées sur la méthode CLV qui s'attachent à traiter le problèmes des variables "de bruit" dans la classification. Les variables qui ne présentent pas vraiment de lien avec la structure principale, qui sont isolées ou qui ne portent qu'une information de type bruit blanc, seront ainsi écartées dans la formation des groupes de variables ou de leurs variables latentes.

Dans le chapitre 4, nous proposons deux algorithmes de sélection de variables basés sur la classification de variables. L'idée principale est d'utiliser les variables latentes des groupes, obtenues par la méthode de classification CLV, comme nouveaux prédicteurs dans le modèle afin de tenir compte de la structure de corrélation entre variables et surmonter les difficultés liées à la sélection de variables. Les groupes de prédicteurs mis en exergue pourraient être intéressants pour mieux comprendre les processus sous-jacents au phénomène étudié.

Dans le chapitre 5, nous présentons le package **ClustVarLV**, fonctionnant sous R, pour la mise en œuvre des méthode proposées. Il comporte (i) la classification de variables CLV de base ; (ii) la classification de variables avec variables externes ; (iii) la classification de variables en présence de variables de *noise* ; (iv) des approches de régression basées sur les variables latentes CLV.

Le chapitre 6 est consacré aux conclusions de ce travail.

Les travaux présentés dans les chapitres 3, 4 et 5 ont fait l'objet de communications lors de conférences et de publications dont la liste figure en Annexe A.

Chapitre 2

Rappel des méthodes

Dans ce chapitre nous allons détailler deux méthodes de base sur lesquelles nous nous sommes appuyés au cours de ce travail. Dans un premier temps, la méthode de Classification de variables autour Variables Latentes (CLV) est décrite, avec en particulier, les critères d'optimisation pris en compte et les algorithmes de classification mis en œuvre. Dans un second temps, nous présentons la méthode *Lasso*, à la base des approches *sparse*, et un algorithme efficace (LARS) pour la solution de *Lasso*. Les notations utilisées tout au long de ce document sont tout d'abord précisées.

2.1 Notations

Considérons un jeu de données contenant n observations (ou individus) sur lesquelles p variables quantitatives ont été mesurées. Notons $\mathbf{X} = [\mathbf{x}_1|\ldots|\mathbf{x}_p]$ la matrice formée par ces p variables. $\mathbf{x}_j = (x_{ij}, \ldots, x_{nj})^T$ représente le vecteur des observations pour la $j^{\text{ième}}$ variable ($j = 1, \ldots, p$). Dans certains cas, la matrice $\mathbf{X}$ représente un ensemble de p prédicteurs potentiels vis-à-vis d'une variable de réponse $\mathbf{y} = (y_1, \ldots, y_n)^T$. Nous supposons que toutes les variables sont centrées : $\bar{y} = 0$, $\bar{x}_j = 0$, pour $j = 1, 2, \ldots, p$. Dans la partie théorique de la régression *Lasso* (§ 2.3), nous supposons que les prédicteurs sont standardisés : $s_j = 1$. Cependant, l'utilisateur peut choisir le type de standardisation qui lui convient dans les applications réelles selon les données.

2.2 La méthode CLV

La méthode de classification de variables (CLV), proposée par Vigneau et Qannari (2003), est une approche qui consiste à regrouper les variables en groupes homogènes, permettant d'identifier ainsi la structure sous-jacente au sein de l'ensemble des variables mesurées. Le principe de la méthode est de définir des groupes de variables autour variables latentes de sorte que la similarité entre une variable observée et la variable latente du groupe dans lequel elle est affectée, soit maximale. Dans CLV, le critère de similarité considéré est soit la covariance, soit la covariance au carré (en fonction du type de groupes recherché). Lorsque les variables sont standardisées, ce sont les corrélations, ou corrélations au carré, qui sont prises en compte dans l'algorithme de classification.

CLV est une approche parmi les diverses méthodes existentes de classification de variables. Les approches les plus connues (implémentées dans de nombreux logiciels de statistique), sont des approches hiérarchiques, fondées sur une matrice de disimilarité (ou similarité) entre les variables, et utilisant pour progresser dans la hiérarchie une stratégie de lien (lien minimum, maximum, moyen, ...). En fonction de la nature des variables (quantitatives ou qualitatives) et en fonction de l'indice de similarité adopté (coefficient de corrélation linéaire, son carré, sa valeur absolue ; coefficient de corrélation de rang de Spearman, de Kendall, indice de Jaccard, ...), l'éventail des possibilités est large. Pour les variables quantitatives, la méthode CLV se distingue des approches hiérarchiques sur matrice de (dis)similarité par le fait qu'elle vise non seulement à définir une partition des variables mais, également, qu'à chaque classe de variables est associée un noyau, c'est-à-dire une variable latente. La procédure VARCLUS du logiciel SAS (Sarle, 1990), est une autre approche, particulièrement connue, qui s'appuie sur l'analyse factorielle. De fait, la procédure VARCLUS et la méthode CLV présentent de nombreux points communs. Une des principales différences est que l'algorithme de la procédure VARCLUS est un algorithme divisif alors que CLV procède selon un mode hiérarchique ascendant ou par partitionnement (voir § 2.2.2). Par ailleurs, CLV est fondée sur des critères d'optimisation précis, ce qui rend la généralisation du problème de classification avec la prise en compte d'informations

externes assez aisée (voir § 2.2.3).

Ces dix dernières années, d'autres approches de classification de variables qui, comme CLV, visent à définir des groupes de variables organisées selon certaines directions principales ont également été proposées. Citons notamment l'algorithme *Diametrical clustering* présenté par Dhillon *et al.* (2003) dans le domaine de la Bioinformatique pour la classification de gènes. Enki *et al.* (2012) ont proposé une procédure dite *Weighted-variance clustering* qui est un algorithme hiérarchique basé sur un critère similaire à celui de CLV. Très récemment, Bühlmann *et al.* (2013) ont adopté une démarche très similaire à ce qui sera développé au chapitre 4. Ils suggèrent de réaliser une classification de variables avant de réaliser une modélisation *sparse* de type Lasso à partir des représentants des groupes de variables. Pour la classification de variables, ils ont envisagé deux algorithmes différents : (i) un algorithme hiérarchique classique à partir de la matrice des valeurs absolues des corrélations entre les variables (critère d'agrégation du lien moyen) ; (ii) un nouvel algorithme hiérarchique ascendant qui, à chaque pas, agrège les deux groupes de variables qui ont la plus grande corrélation canonique. Ce critère est différent de celui qui est considéré dans l'algorithme hiérarchique de CLV (voir (2.5), (2.6)), mais il fait également référence au contexte de l'analyse factorielle de données.

2.2.1 Critères de base de CLV

Etant donné un nombre de groupes K, la méthode CLV consiste à chercher une partition en K groupes, $G_1, \ldots, G_K$, et K variables latentes, $\mathbf{c}_1, \ldots, \mathbf{c}_K$, de sorte à maximiser un critère de cohérence interne des groupes. Deux critères sont considérés permettant de définir deux types de groupes (Figure 2.1) : (i) "groupes directionnels" où les variables corrélées (positivement ou négativement) vont être agrégées ensemble, sans tenir compte du signe de leur coefficient de corrélation ; (ii) "groupes locaux" qui visent à rassembler des variables positivement corrélées entre elles.

Pour les groupes directionnels, le critère à maximiser est noté T. Il est défini comme suit :

$$T = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov}^2 \left(\mathbf{x}_j, \mathbf{c}_k \right), \text{ avec } \|\mathbf{c}_k\| = 1 \tag{2.1}$$

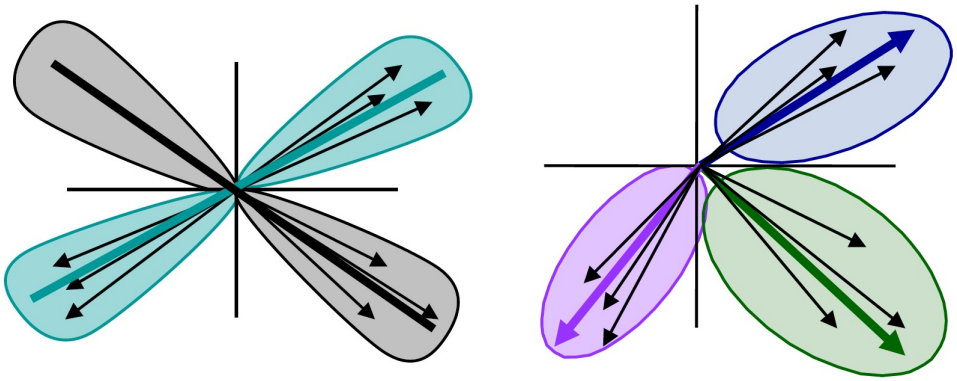

FIGURE *2.1 – Type de groupes dans CLV : "groupes directionnels" (à gauche) et "groupe locaux" (à droite).*

Pour les groupes locaux, le critère à maximiser est noté S, et s'exprime par :

$$S = \sqrt{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}(\mathbf{x}_j, \mathbf{c}_k), \text{ avec } \|\mathbf{c}_k\| = 1 \tag{2.2}$$

Dans les équations (2.1) et (2.2), $\delta_{kj} = 1$ si la variable $\mathbf{x}_j$ appartient au groupe G_k et $\delta_{kj} = 0$ sinon. $\operatorname{cov}(\mathbf{x}_j, \mathbf{c}_k) = \mathbf{x}_j^T \mathbf{c}_k / n$ représente la covariance entre la variable $\mathbf{x}_j$ et la variable latente $\mathbf{c}_k$, calculée sur la base de n observations.

2.2.2 Algorithmes de classification

La solution du problème considéré peut être obtenue, pour un nombre de groupes K, en utilisant un algorithme de partitionnement, de type *k-means*, en alternant deux étapes : l'étape d'affectation des variables aux différents groupes et l'étape d'estimation du noyau ("centre") de chaque groupe. Cet algorithme nécessite de partir d'une partition initiale des variables. La démarche courante est de répéter R fois une initialisation au hasard, et de retenir la meilleure (au sens du critère T ou S) des R partitions obtenues à convergence de l'algorithme. C'est une stratégie efficace. De manière alternative, la méthode CLV permet à l'utilisateur de mettre en œuvre une initialisation de l'algorithme de partitionnement en considérant la partition en K groupes obtenue par la coupure d'un dendrogramme.

En effet, un algorithme de classification ascendante hiérarchique basé sur la maximisation du critère T ou S (selon le choix effectué) est également disponible. Sa construction est simple et consiste à chaque étape à agréger les deux groupes de variables qui conduisent à la plus petite diminution possible du critère de classification. On montre aisément que cette règle d'agrégation conduit à une évolution monotone, décroissante, du critère (Vigneau et Qannari, 2003). Par ailleurs, le suivi des variations du critère de classification (ou critère de qualité de la partition) fournit une aide pour le choix du nombre de groupes. Le niveau auquel on observe une forte diminution du critère de classification pourra être choisi comme niveau de coupure du dendrogramme.

Ces deux algorithmes sont décrits plus en détail ci-après :

Algorithme de partitionnement

1. Initialisation. Initialiser au hasard une partition en K groupes ou utiliser la partition en K groupes obtenue par coupure du dendrogramme de l'algorithme hiérarchique.
2. Étape d'estimation. Pour chaque groupe G_k ($k = 1, \ldots, K$), la variable latente $\mathbf{c}_k$ est définie
   - pour les groupes directionnels, par la première composante principale normée de $\mathbf{X}_k$, la matrice formée des variables appartenant au groupe G_k.
   - pour les groupes locaux, par la variable moyenne (*centroids*) normée du groupe G_k.
3. Étape d'affectation. Chaque variable est ré-affectée dans le groupe dont elle est le plus proche. Plus précisément,
   - pour les groupes directionnels :

$$\delta_{kj} = 1 \quad \text{si} \quad \text{cov}^2\left(\mathbf{x}_j, \mathbf{c}_k\right) = \max_{l=1,\ldots,K} \left\{\text{cov}^2\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\} \tag{2.3}$$

   - pour les groupes locaux :

$$\delta_{kj} = 1 \quad \text{si} \quad \text{cov}\left(\mathbf{x}_j, \mathbf{c}_k\right) = \max_{l=1,\ldots,K} \left\{\text{cov}\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\} \tag{2.4}$$

4. Les étapes 2 et 3 sont répétées jusqu'à stabilisation de la partition des variables. Dans le cas où le nombre de variables est très important, le nombre d'itérations peut être élevé avant stabilisation de la partition. C'est pourquoi la règle d'arrêt intègre également un paramètre sur la convergence du critère de classification, ainsi qu'un nombre maximal d'itérations possibles.

Algorithme hiérarchique

1. À la première étape, chaque variable forme un groupe à elle seule. Dans le cas des groupes directionnels, $T_1 = \sum_{j=1}^{p} \text{var}(\mathbf{x}_j)$. Pour les groupes locaux, $S_1 = \sum_{j=1}^{p} \sigma(\mathbf{x}_j)$, où $\sigma(\mathbf{x}_j)$ représente l'écart-type de $\mathbf{x}_j$.
2. À l'étape $i > 1$, les deux groupes qui conduisent à la plus petite diminution de critère sont fusionnés. Si A^* et B^* sont ces deux groupes, ils vérifient $\arg\min_{A,B} \Delta T_i$ ou $\arg\min_{A,B} \Delta S_i$, avec :
   – pour les groupes directionnels,

$$\Delta T_i = T_i - T_{i+1} = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)} \tag{2.5}$$

   où $\lambda_1^{(G_k)}$ est la première valeur propre de $1/n\mathbf{X}_k\mathbf{X}_k^T$.
   – pour les groupes locaux,

$$\Delta S_i = S_i - S_{i+1} = p_A\sigma(\bar{\mathbf{x}}_A) + p_B\sigma(\bar{\mathbf{x}}_B) - (p_A + p_B)\sigma(\bar{\mathbf{x}}_{A \cup B}) \tag{2.6}$$

   où p_{G_k} est le nombre de variables dans le groupe G_k, et $\bar{\mathbf{x}}_{G_k}$ représente la variable moyenne du groupe G_k.
3. À la dernière étape, toutes les variables sont agrégées dans un seul groupe.

2.2.3 Variables externes

L'un des avantages de la méthode CLV, par rapport à d'autres méthodes (notamment VARCLUS), est qu'elle permet de tenir compte de variables externes. La stratégie adoptée est de contraindre la variable latente dans chaque groupe à être une

combinaison linéaire des variables externes. L'intérêt de cette démarche est de pouvoir expliquer les variables latentes des groupes en fonction d'informations connexes. Par exemple, dans le domaine de l'analyse des préférences sensorielles de consommateurs, il peut être intéressant de former, simultanément, des groupes de consommateurs ayant des préférences similaires (ce sont des segments de consommateurs) et d'expliquer le vecteur de préférence représentatif de chaque segment en fonction des caractéristiques sensorielles des produits (les variables externes). On peut également envisager de relier des variables mesurées sur des échantillons préparés selon un certain plan d'expériences en fonction des paramètres du plan d'expériences.

Notons $\mathbf{Z} = [\mathbf{z}_1 | \dots | \mathbf{z}_q]$ la matrice (centrée) des variables externes. Les q variables de $\mathbf{Z}$ sont mesurées sur les mêmes individus. La démarche de CLV consiste alors à maximiser :

– pour les groupes directionnels :

$$\tilde{T} = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}^2 (\mathbf{x}_j, \mathbf{c}_k), \text{ avec } \mathbf{c}_k = \mathbf{Z}\mathbf{a}_k \text{ et } \|\mathbf{a}_k\| = 1 \tag{2.7}$$

– pour les groupes locaux :

$$\tilde{S} = \sqrt{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov} (\mathbf{x}_j, \mathbf{c}_k), \text{ avec } \mathbf{c}_k = \mathbf{Z}\mathbf{a}_k \text{ et } \|\mathbf{a}_k\| = 1 \tag{2.8}$$

Même si cette démarche répond à des problématiques précises dans différents domaines d'application (Vigneau et Qannari, 2002; Vigneau, 2012; Vigneau *et al.*, 2014), l'expérimentateur n'a pas forcément l'objectif d'expliquer les variables latentes de groupes en fonction de variables externes. Au contraire, son objectif peut plutôt être d'utiliser les variables latentes de groupes comme prédicteurs (et non comme des réponses). Ce point de vue n'avait jamais encore été traité dans le cadre de la méthode CLV. C'est à ce problème que le chapitre 4 est consacré.

2.3 La méthode *Lasso*

Fondement de nombreuses approches *sparse*, pour lesquelles un certain nombre de variables se verront attribuer un poids nul dans l'analyse descriptive ou prédictive des données, la contrainte L^1, de type *Lasso*, est rappelée dans cette partie.

Considérons le modèle de régression linéaire :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2.9}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients à estimer et ϵ est le terme d'erreur. L'estimateur le plus souvent utilisé est un estimateur des moindres carrés ordinaires $\hat{\boldsymbol{\beta}}^{OLS}$ minimisant la somme des carrés résiduels :

$$\hat{\boldsymbol{\beta}}^{OLS} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{2.10}$$

où $\|\cdot\|_2$ représente la norme euclidienne usuelle. Cependant, il est bien connu que, lorsque les prédicteurs ne sont pas linéairement indépendants, l'estimation du vecteur de coefficients β est instable et la régression des Moindres Carrés (OLS) n'a pas de bonnes performances en prédiction. La régression OLS n'a, en outre, pas de solution unique lorsque le nombre de prédicteurs, p, dépasse le nombre d'observations, n. Dans ces cas là, l'estimation est souvent contrôlée sous contrainte de régularisation.

La méthode *Lasso* (*Least Absolute Shrinkage and Selection Operator*) proposée par Tibshirani (1996) est la plus populaire des méthodes de régularisation en régression linéaire. Elle consiste à chercher une solution du problème d'optimisation :

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{sous la contrainte :} \quad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j| < t \tag{2.11}$$

où $t \geqslant 0$ est un paramètre de réglage. Une façon équivalente d'écrire le problème précédent est de définir un critère des moindres carrés pénalisé :

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{2.12}$$

où $\lambda \geqslant 0$ est un paramètre de pénalité. Les deux équations (2.11) et (2.12) sont équivalentes dans le sens où il existe une correspondance unique entre t et λ qui donnent la même solution.

2.3.1 Solutions

Avec le terme de régularisation, ou la contrainte sur la norme L^1 du vecteur de coefficients $\boldsymbol{\beta}$, la solution obtenue par *Lasso* comporte des valeurs de β exactement égales à zéro. Ainsi une sorte de sélection de variables est réalisée dans le modèle. Si $t = 0$, ou λ tend vers l'infini, alors tous les coefficient de β sont nuls. Pour $t > \|\hat{\boldsymbol{\beta}}^{OLS}\|_1$, où $\hat{\boldsymbol{\beta}}^{OLS}$ une solution de problème (2.10), ou quand $\lambda = 0$, nous retrouvons l'estimateur des moindres carrés.

Dans le cas spécifique où $\mathbf{X}$ est une matrice orthonormée ($\mathbf{X}^T\mathbf{X} \propto \mathbf{I}$), la solution de la régression *Lasso* (2.12) peut être explicitement calculée en dérivant la fonction objectif. Elle correspond à un seuillage "doux" de l'estimateur des moindres carrés. Pour tout coefficient β_j ($j = 1, \ldots, p$) :

$$\hat{\beta}_j^{lasso} = \text{Signe}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda/2)_+ = \begin{cases} \hat{\beta}_j^{OLS} - \lambda/2 & \text{si} \quad \hat{\beta}_j^{OLS} > \lambda/2 \\ \hat{\beta}_j^{OLS} + \lambda/2 & \text{si} \quad \hat{\beta}_j^{OLS} < -\lambda/2 \\ 0 & \text{si} \quad |\hat{\beta}_j^{OLS}| \leqslant \lambda/2 \end{cases} \tag{2.13}$$

Dans le cas général, calculer la solution de *Lasso* est un problème d'optimisation quadratique. L'algorithme proposé par Tibshirani (1996) n'est malheureusement pas assez rapide en pratique. Un algorithme assez efficace, LARS (*Least Angle Regression*), a ensuite été proposé par Efron *et al.* (2004). Il permet d'obtenir assez rapidement l'ensemble de toutes les solutions de *Lasso*, pour différentes valeurs de λ. Dans la mesure où l'algorithme LARS sera mis en oeuvre ultérieurement, une description détaillée accompagnée d'une représentation graphique (Figure 2.2) est donnée après.

L'algorithme LARS

Notation : Soit $\mathcal{A}$ un ensemble d'indices, $\mathbf{X}_\mathcal{A} = [\ldots s_j\mathbf{x}_j \ldots]_{j\in\mathcal{A}}$, où $s_j = \text{signe}(\mathbf{x}_j^T\mathbf{y})$, et $\hat{\boldsymbol{\mu}}_\mathcal{A}$ une estimation de $\mathbf{y}$ basée les variables $\mathbf{X}_\mathcal{A}$.

1. Commencer par l'ensemble actif $\mathcal{A} = \emptyset$, alors $\hat{\boldsymbol{\mu}}_{\mathcal{A}} = 0$ et $\mathbf{r} = \mathbf{y}$.

2. Trouver le prédicteur $\mathbf{x}_j$ le plus corrélé avec $\mathbf{y}$: $j = \arg\max_{j=1,\ldots,p} |\mathbf{x}_j^T \mathbf{y}|$. Mise à jour $\mathcal{A} = \{j\}$.

3. Calculer le vecteur équiangle des variables actives : $\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} \omega_{\mathcal{A}}$, où $\omega_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T \mathbf{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2} \mathbf{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$ est une direction équiangle avec $\mathbf{G}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$ et $\mathbf{1}_{\mathcal{A}}$ un vecteur de 1 avec cardinal $|\mathcal{A}|$ éléments.

4. Mise à jour de l'estimation $\hat{\boldsymbol{\mu}}_{\mathcal{A}} = \hat{\boldsymbol{\mu}}_{\mathcal{A}} + \gamma \mathbf{u}_{\mathcal{A}}$ et $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{A}}$ avec un pas γ, jusqu'à trouver un nouveau prédicteur $\mathbf{x}_j$, ($j \in \mathcal{A}^c$, le complémentaire de $\mathcal{A}$), tel que la corrélation de $\mathbf{x}_j$ avec le résidu courant soit la même que celles entre les variables actives $\mathbf{x}_i$, $\forall i \in \mathcal{A}$ et le résidu courant :

$$j = \arg \left\{ \max_{j \in \mathcal{A}^c} |\mathbf{x}_j^T \mathbf{r}| = \mathbf{x}_i^T \mathbf{r}, \forall i \in \mathcal{A} \right\}$$

5. Mise à jour $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{A}}$, $\mathbf{y} = \mathbf{r}$ et $\mathcal{A} = \mathcal{A} \cup \{j\}$.

6. Répéter 3-6 jusqu'à ce que toutes les variables soient entrées dans le modèle. Arrêter lorsque $|\mathbf{x}_j^T \mathbf{r}| = 0, \forall j$.

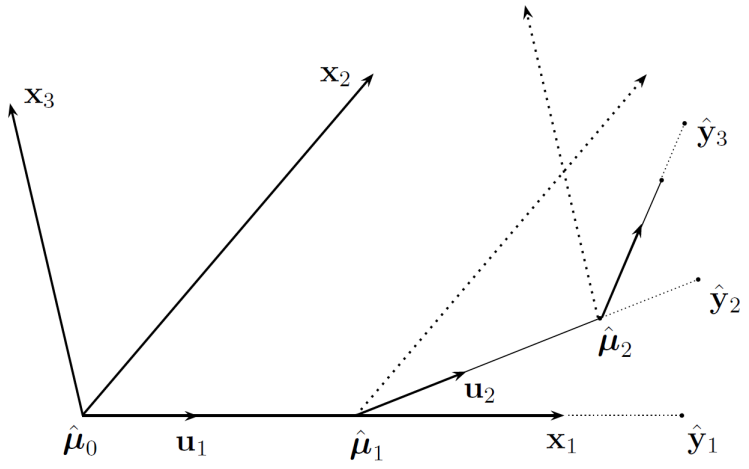

FIGURE *2.2 – L'algorithme LARS dans le cas de 3 prédicteurs. A chaque étape k, $\hat{\boldsymbol{\mu}}_k$ est l'estimation de LARS. $\hat{\mathbf{y}}_k$ est l'estimation de OLS correspondante.*

2.3.2 Propriétés et développement de *Lasso*

Elastic Net Dans le cas de données en grande dimension, Zou et Hastie (2005) soulignent les limites rencontrées avec la régression *Lasso* :

- dans le cas où $p > n$, la méthode *Lasso* ne peut pas sélectionner plus de variables qu'il y a d'observations, autrement dit il y aura au maximum n coefficients de régression non nuls.
- s'il existe des groupes de variables fortement corrélées entre elles, la méthode *Lasso* tend à ne sélectionner qu'une seule variable au hasard parmi un groupe. Cela peut conduire l'utilisateur à passer à côté d'éléments importants pour une compréhension plus complète des phénomènes sous-jacents à son étude.

Zou et Hastie (2005) propose avec *Elastic Net* de combiner la pénalité L^1 et L^2 pour surmonter ces problèmes. La régression *Elastic Net* revient alors à :

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{2.14}$$

Elastic Net bénéficie donc à la fois des propriétés de la régression *Lasso* et de la régression *Ridge* : il produit, d'une part, des solutions *sparse* et il permet, d'autre part, de sélectionner des prédicteurs corrélés entre eux (effet de groupement).

Lorsque $\lambda_2 = 0$, la méthode *Lasso* est un cas particulier de *Elastic Net*. Lorsque $\lambda_2 \to \infty$, la solution peut être explicitement exprimée par :

$$\hat{\beta}_j(\infty) = \text{Signe}(\mathbf{y}^T\mathbf{x}_j)(|\mathbf{y}^T\mathbf{x}_j| - \lambda_1/2)_+ \quad j = 1, 2, \ldots, p \tag{2.15}$$

qui est simplement un seuillage doux (tel qu'il a été défini en (2.13)) sur le coefficient de la régression uni-variée pour le prédicteur $\mathbf{x}_j$.

Néanmoins, cet effet de groupement ne fournit pas explicitement des informations sur les groupes de prédicteurs. Pour mieux explorer les inter-corrélations de prédicteurs entre eux, nous allons discuter dans le chapitre 4 d'une combinaison de la classification et la sélection de variables.

LARS-OLS, Relaxed Lasso et Adaptive Lasso La pénalité L^1 permet de sélectionner les variables en "tirant" les estimations des coefficients vers zéro. En conséquence, la performance du modèle est pénalisée par l'introduction de ce biais. Par exemple, dans une application de données en grande dimension ($p \gg n$) où la plupart de variables sont non pertinentes, une grande valeur de pénalité λ est souvent nécessaire pour identifier les variables importantes. Dans ce cas, les coefficients de ces variables seront fortement pénalisés, ce qui peut conduire à une mauvaise qualité du modèle en prédiction.

Certaines extensions ont été proposées pour dé-biaiser les estimateurs de *Lasso*. L'idée principale est de séparer la sélection et l'estimation en deux étapes. Par exemple LARS-OLS proposé par Efron *et al.* (2004) utilise tout d'abord LARS pour sélectionner un sous-ensemble de prédicteurs suivi d'une estimation OLS sur les prédicteurs sélectionnés. De même, Meinshausen (2007) propose avec *Relaxed Lasso* de calculer, pour chaque valeur de pénalité λ possible, une solution de *Lasso* dans la première étape, et dans la deuxième étape, d'effectuer encore une fois une régression *Lasso* avec les variables retenues mais avec une petite pénalité $\phi\lambda$, où $\phi \in [0, 1]$. Pour $\phi = 0$, *Relaxed Lasso* est équivalente à LARS-OLS. Une autre méthode en deux étapes, appelée *Adaptive Lasso* (Zou, 2006) consiste, dans un premier temps, à considérer les estimateurs des moindres carrés $\hat{\boldsymbol{\beta}}^{OLS} = (\hat{\beta}_1^{OLS}, \ldots, \hat{\beta}_p^{OLS})$, et, dans un second temps, à réaliser une régression de type *Lasso* avec la pénalité pondérée par les valeurs des coefficients $\hat{\beta}_j^{OLS}$:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j| \tag{2.16}$$

où $\hat{w}_j = 1/|\hat{\beta}_j^{OLS}|^\gamma$, $\gamma > 0$. Ainsi les variables qui ont une forte valeur pour l'estimation $\hat{\beta}_j^{OLS}$ (considérées comme variables importantes, les variables étant toutes normées) sont moins pénalisées. Les propriétés théoriques de *Adaptive Lasso* pour les données en grande dimension sont données par Huang *et al.* (2008).

Après avoir rappelé les outils nécessaires pour la compréhension de l'algorithme CLV et la régression *Lasso*, nous abordons dans la suite des extensions de ces algo-

rithmes dans le cadre de la classification de variables en présence de variables atypiques (Chapitre 3) et de la régression basée sur la classification de variables (Chapitre 4).

Chapitre 3

Classification en présence de variables atypiques ou de *noise*

Les méthodes traditionnelles de classification ne tiennent pas compte de l'existence d'information atypique ou de bruit dans les données. Si on se place dans le cadre de la classification de variables, chaque variable sera forcément affectée à un groupe même si elle est isolée ou si elle correspond à une variable de bruit. Dans ce chapitre, nous proposons deux approches basées sur la méthode CLV pour répondre à ce type de problème. L'une est une approche "$K + 1$" qui consiste à introduire un groupe supplémentaire, dit groupe de bruit. L'autre "*Sparse* LV" est fondée sur la définition de variables latentes de groupes *sparse*.

3.1 Introduction

L'objectif des méthodes de classification est de chercher des groupes le plus homogènes possibles avec en même temps une hétérogénéité la plus grande possible entre les groupes. Cependant, les méthodes traditionnelles de classification ne donnent pas forcément des résultats satisfaisants en présence de données aberrantes ou de données bruitées. Par exemple, la présence d'*outliers* peut conduire, pour une partition de taille donnée, à agréger de 'bons" groupes. Pour les données en grande dimension, il est également possible que les "bons" groupes soient augmentés d'une partie

non-informative. Lorsqu'il y a une quantité importante de bruit dans les données, les résultats de classification peuvent être difficilement exploitables parce que chaque objet est forcément affecté dans un groupe, même si cela n'a pas de sens. Une bonne méthode devrait montrer de bonnes performances en termes de qualité de partitionnement, mais aussi en termes d'interprétabilité des groupes. Idéalement un groupe homogène ne devrait pas comporter d'objets inutiles.

Dans le cadre de la classification de variables, la plupart des méthodes, parmi lesquelles CLV ou VARCLUS, sont fondées sur des algorithmes de regroupement "dur" (*hard clustering*) où chaque variable est affectée exactement à un, et un seul, groupe. Les méthodes de regroupement "flou" (*fuzzy clustering*) qui renvoient une matrice d'appartenance avec des valeurs comprises entre 0 et 1, sont souvent proposées pour pallier ce défaut. Une variable ayant peu de lien avec les classes sous-jacentes devrait, dans ce cas, avoir un degré d'appartenance faible pour tous les groupes. Pour le regroupement des gènes, Berget *et al.* (2005) ont proposé de chercher séquentiellement des groupes "intéressants" en utilisant la méthode *fuzzy c-means*. Une fonction de pénalité est ajoutée dans le critère de classification de sorte que les groupes "intéressants" soient moins pénalisés. Dans le domaine de l'analyse sensorielle, Dahl et Næs (2009) ont aussi adapté le *fuzzy clustering* en considérant une partition en deux groupes, un "bon" groupe et un groupe de bruit, afin d'identifier les juges atypiques. Ces deux méthodes visent un objectif proche du nôtre, mais les critères de classification utilisés sont basés sur la distance euclidienne. Dans notre cas, nous considérons spécifiquement la classification de variables en utilisant les critères de similarité de la méthode CLV, fondés sur la covariance ou la corrélation entre variables.

Nous proposons de modifier la méthode CLV pour écarter ou réduire les poids des variables atypiques. Par conséquent, les variables qui ne présentent pas vraiment de lien avec la structure principale, qui sont isolées ou qui ne portent qu'une information de type bruit blanc, ne devraient pas être affectées à un "bon" groupe. Pour plus de simplicité, ces variables sont appelées variables atypiques ou de bruit dans la suite. Ce terme doit s'entendre au sens large par rapport à la "vraie" structure sous-jacente, qui est inconnue. Pour fixer les idées, considérons un petit exemple pour lequel il y a deux groupes de variables et quelques variables isolées. Les variables au sein d'un groupe

sont fortement corrélées entre elles, positivement ou négativement. La structure des corrélations entre les variables de l'exemple est représentée dans la partie gauche de la Figure 3.1. Ce que l'on souhaite, c'est que les variables représentées en gris, dans la partie droite de la Figure 3.1, soient écartées dans le *noise cluster* car ce sont les variables qui n'appartiennent pas à la structure principale en deux groupes.

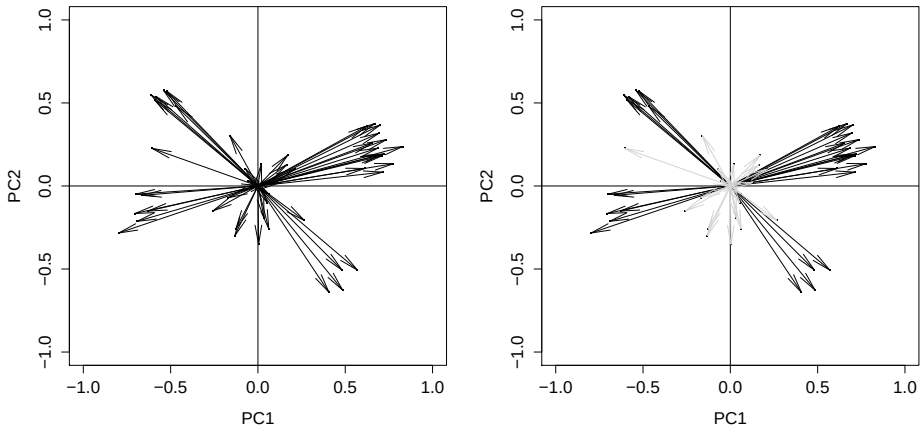

FIGURE *3.1 – Structure de corrélations dans un jeu de données contenant deux groupes de variables et plusieurs variables isolées. L'objectif recherché est d'écarter les variables identifiées en gris sur le schéma de droite.*

Nous proposons deux approches pour atteindre cet objectif. La première, appelée stratégie "$K+1$" , est basée sur une modification des critères utilisés dans l'approche CLV en ajoutant un terme supplémentaire de sorte qu'un groupe additionnel (*noise cluster*) soit introduit. Cette approche s'inspire de la notion de *noise cluster* introduite par Dave (1991). Dans cette démarche, au cours de la classification, tout ce qui n'a pas un degré d'affiliation suffisamment clair à un des groupes en formation sera affecté au groupe additionnel. La seconde stratégie, appelée stratégie "*Sparse* LV", est basée sur le principe de l'Analyse en Composantes Principales (ACP) *sparse* (Zou *et al.*, 2006). Les variables d'un groupe qui ont un poids petit dans la variable latente du groupe se verront attribuées *in fine* un poids nul.

3.2 Contexte méthodologique

3.2.1 *noise cluster*

Pour la classification d'observations, certaines méthodes comme les méthodes de classification robuste (García-Escudero *et al.*, 2010; Dave et Krishnapuram, 1997) ou les méthodes basées sur la détection des données aberrantes (Marghny et Taloba, 2011; Al-Zoubi *et al.*, 2010) ont été proposées. Parmi ces approches, la classification avec identification d'un *noise cluster* est une stratégie très intéressante proposée par Dave (1991). L'algorithme est développé dans l'optique de capturer tous les objets non pertinents dans le *noise cluster* et de partitionner les autres objets dans des groupes représentatifs.

Selon les auteurs, il peut y avoir différentes définitions pour le *noise cluster*, ce qui correspond à différentes notions derrière le terme "objets non pertinents". Par exemple Cuesta-Albertos *et al.* (1997) proposent de chercher un sous-ensemble de taille prédéfinie de sorte à ce que la qualité de la classification des objets qui y sont affectés soit la meilleure possible. Le sous-ensemble non retenu, que l'on peut qualifier de *noise cluster*, comporte donc des objets les moins importants pour le partitionnement. Plus précisément, soit $\mathcal{O} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n\}$ l'ensemble des observations. La méthode consiste à partitionner un sous-ensemble d'observations $\mathcal{S}$, de taille $n(1-a)$, avec $a \in [0, 1]$, dans K groupes $(G_1, \ldots, G_K)$ afin de minimiser la distance entre les observations au sein de chaque groupe :

$$\min_{\mathcal{S}} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{S}, G_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2 \tag{3.1}$$

où $\bar{\mathbf{x}}_k$ est le barycentre des individus dans le groupe G_k.

L'algorithme DBSCAN (*Density-based Spatial Clustering of Applications with Noise*) de Ester *et al.* (1996) est, quant-à-lui, basé sur la densité des points. Les objets qui ne sont pas situés dans une région d'une certaine densité vont être considérés comme objets à mettre de côté. Plus précisément, étant donnés deux paramètres, un rayon ϵ et le nombre minimum de poins *minPts*, l'algorithme catégorise les obser-

vations dans trois familles :

- point noyau : un point qui a plus de *minPts* points à l'intérieur d'un rayon ϵ.
- point frontière : un point non noyau, mais qui se trouve dans le voisinage d'un point noyau pour le rayon ϵ.
- *noise* : un point qui n'est ni point noyau, ni point frontière.

En ce qui nous concerne, nous avons privilégié le raisonnement de Dave (1991) qui introduit la notion de prototype pour le *noise cluster*. Par définition, le prototype du *noise cluster* est une entité qui est située à la même distance δ (à fixer) de tous les objets du jeu de données. En intégrant ce prototype du *noise cluster* et la distance δ dans le critère de classification, les objets qui sont plus proches du prototype du *noise cluster* que des prototypes des autres groupes seront affectés au *noise cluster*. Dave (1991) considère K "bonnes" classes d'observations et une $K+1^{\text{ième}}$ classe qui est le *noise cluster*. Il définit le problème suivant :

$$\min \sum_{i=1}^{n} \sum_{k=1}^{K+1} (u_{ik})^m (d_{ik})^2 \tag{3.2}$$

où (d_{ik}) représente la distance euclidienne d'un objet i au centre du groupe G_k pour $k = 1, \ldots, K$ et $(d_{ik}) = \delta$, pour tout i, si $k = K+1$. Dans l'expression précédente, u_{ik} désigne le degré d'appartenance d'une observation i à un groupe G_k et $m > 1$ est un paramètre contrôlant le degré de "flou" dans un algorithme de type *Fuzzy c-means*. Si on considère un algorithme de classification "dure", non floue, les valeurs u_{ik} sont égales 0 ou 1 et le paramètre m n'a plus d'intérêt (on choisit alors $m = 1$).

Dans ce travail, nous proposons de transposer les idées de Dave (1991) lorsque les objets à classer sont des variables et en considérant une classification "dure" comme c'est le cas actuellement dans l'approche CLV. Comme dans CLV, les prototypes de groupes, c'est-à-dire les variables latentes $\mathbf{c}_k$ dans chaque groupe de variables, sont un élément clé de la procédure, le raisonnement de Dave (1991) s'adapte facilement.

3.2.2 Composantes Principales *sparse*

L'interprétation d'une composante en ACP peut être faite en regardant les coefficients du vecteur propre associé (*loadings*). Cependant, avec plusieurs centaines, voire plusieurs milliers, de variables, de trop nombreux coefficients rendent la composante difficile à interpréter. L'idée de *sparsité* est de chercher des composantes avec un maximum de coefficients nuls pour les variables qui contribuent peu à une composante donnée. De ce point de vue, la classification de variables qui procède au regroupement de variables dans différents groupes disjoints, associés chacun à une variable latente, offre une alternative pour extraire des composantes interprétables. Cependant, en présence de variables atypiques ou de bruit, les groupes de variables doivent encore être affinés. Les approches *sparse* sont alors envisagées dans chaque groupe de variables pour améliorer l'interprétation de la variable latente du groupe.

Des méthodes de l'ACP *sparse* ont été proposées en utilisant la pénalité L^1 ou des fonctions de seuillage. Par exemple, Jolliffe *et al.* (2003) ont proposé l'approche SCoTLASS pour chercher des combinaisons linéaires $\mathbf{X}\boldsymbol{\alpha}_k$ $(k = 1, \ldots, p)$ qui maximisent la variance, comme en ACP, mais avec une contrainte L^1 supplémentaire sur les coefficients de la combinaison linéaire :

$$\text{maximiser} \qquad \boldsymbol{\alpha}_k^T (\mathbf{X}^T \mathbf{X}) \boldsymbol{\alpha}_k \tag{3.3}$$

$$\text{avec} \quad \boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k = 1, \quad k \geqslant 1 \quad \text{et} \quad \boldsymbol{\alpha}_h^T \boldsymbol{\alpha}_k = 0, \quad \text{pour} \quad h < k,\, k \geq 2$$

$$\text{sous la contrainte :} \quad \textstyle\sum_{j=1}^{p} |\alpha_{kj}| < t$$

Cependant, le problème SCoTLASS est un problème d'optimisation non convexe dont la résolution a un coût de calcul élevé. Il est difficile d'essayer toutes les valeurs possibles du paramètre t et de choisir une valeur optimale.

A la différence de SCoTLASS, la *Sparse PCA* proposée par Zou *et al.* (2006) consiste à ne chercher qu'une approximation des *loadings* de l'ACP. Le problème de recherche du vecteur de *loadings* en ACP est transformé en un problème de régression. Ainsi, cela revient, pour une composante donnée, à un problème de régression avec une pénalité L^1 qui peut se résoudre efficacement en utilisant l'algorithme LARS

(voir § 2.3). Soit $\mathbf{A}_{p\times k} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k]$ et $\mathbf{B}_{p\times k} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k]$, le problème *sparse PCA* devient :

$$\arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j}\|\beta_j\|_1 \tag{3.4}$$

$$\text{avec } \mathbf{A}^T\mathbf{A} = I_{k\times k}$$

Un algorithme alterné sur la matrice de *loadings* $\mathbf{A}$ et l'approximation $\mathbf{B}$ est proposé.

Une autre approche d'approximation a été proposée par Shen et Huang (2008) sous le nom *Sparse PCA via regularized SVD* (sPCA-rSVD). Le problème d'ACP est vu comme un problème de régression via une décomposition en valeurs singulières (SVD). Soit $\tilde{\mathbf{u}}$ un n-vecteur normé et $\tilde{\mathbf{v}}$ un p-vecteur, sPCA-rSVD consiste à chercher successivement les composantes et *loadings* en résolvant le problème suivant :

$$\min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^T\|^2 + P_\lambda(\tilde{\mathbf{v}}) \tag{3.5}$$

avec une mise à jour à chaque étape k de $\mathbf{X}$ ($\mathbf{X} = \mathbf{X} - \tilde{\mathbf{u}}_{k-1}\tilde{\mathbf{v}}_{k-1}^T$). P_λ représente différent types de seuillage sur le vecteur de *loadings* ; par exemple un seuillage dur $h_\lambda^{\text{dur}}(\mathbf{v}) = I(|\mathbf{v}| > \lambda)\mathbf{v}$ avec I une fonction indicatrice, ou un seuillage doux $h_\lambda^{\text{doux}}(\mathbf{v}) = \text{signe}(\mathbf{v})(|\mathbf{v}| - \lambda)_+$.

3.3 La méthode CLV modifiée

3.3.1 Stratégie "$K + 1$"

La même logique que celle du *noise cluster* de Dave (1991), Berget *et al.* (2005) et Dahl et Næs (2009) est adoptée pour la classification de variables. Nous considérons ici que le prototype du *noise cluster* a une même corrélation, ρ, avec toutes les variables observées $\mathbf{x}_j$ ($j = 1, \ldots, p$). Un groupe additionnel qui contiendra les variables atypiques ou les variables de bruit est ainsi défini. Sa contribution est ajoutée dans les critères considérés dans la méthode CLV. Selon le type de groupes cherchés, il consiste à maximiser :

– pour les groupes directionnels, un nouveau critère T défini par :

$$T_{new} = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov}^2 \left(\mathbf{x}_j, \mathbf{c}_k\right) + \sum_{j=1}^{p} (1 - \sum_{k=1}^{K} \delta_{kj}) \, \rho^2 \, \mathrm{var} \left(\mathbf{x}_j\right) \tag{3.6}$$

– pour les groupes locaux, un nouveau critère S :

$$S_{new} = \sqrt{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov} \left(\mathbf{x}_j, \mathbf{c}_k\right) + \sum_{j=1}^{p} (1 - \sum_{k=1}^{K} \delta_{kj}) \, \rho \, \sqrt{\mathrm{var} \left(\mathbf{x}_j\right)} \tag{3.7}$$

avec $\|\mathbf{c}_k\| = 1$.

Dans les équations (3.6) et (3.7), le dernier terme ajouté correspond à la contribution du groupe additionnel, ou *noise cluster*. Plus précisément, si une variable $\mathbf{x}_j$ appartient à l'un des groupes principaux G_k $(k = 1, \ldots, K)$, nous avons $\delta_{kj} = 1$ pour un indice k donné. Ainsi $(1 - \sum_k \delta_{kj}) = 0$. Au contraire, si une variable n'appartient à aucun des groupes G_k, on a $(1 - \sum_k \delta_{kj}) = 1$.

La taille du groupe additionnel dépend de la valeur du paramètre ρ, qui sera choisie entre 0 et 1. Si une valeur élevée de ρ est prise, beaucoup de variables seront traitées comme des variables atypiques et ne rentreront dans aucun des groupes principaux. Si ρ est très petit, les variables seront affectées dans un des groupes G_k plutôt qu'au *noise cluster*. $\rho = 0$ et $\rho = 1$ sont deux cas particuliers. Pour $\rho = 0$, les équations (3.6) et (3.7) reviennent aux critères classiques de CLV (équations (2.1) et (2.2)) et le *noise cluster* est vide. Pour $\rho = 1$, les variables seront affectées dans le *noise cluster*. La démonstration du comportement attendu de la démarche est donnée en Annexe B.

Les critères T_{new} et S_{new} peuvent être maximisés en utilisant l'algorithme de partitionnement décrit dans la section (§ 2.2.2), à une modification près dans l'étape d'affectation. La modification apportée est la suivante :

Étape d'affectation. Une variable $\mathbf{x}_j$ est normalement affectée dans le groupe qui permet d'avoir la plus grande covariance (ou covariance au carré) entre la variable $\mathbf{x}_j$ et l'une des variables latentes de groupes $\mathbf{c}_l$, $(l = 1, \ldots, K)$. Mais si cette valeur

est faible par rapport à la valeur du paramètre ρ, compte-tenu de l'écart-type (ou la variance) de la variable $\mathbf{x}_j$, cette variable sera considérée comme étant plus proche de la variable latente du *noise cluster* et sera affectée à ce groupe. Plus formellement, nous avons :

– pour les groupes directionnels :

$$\begin{cases} \delta_{kj} = 0 \ \forall k & \text{si} \max\left\{\max_l\left\{n\,\mathrm{cov}^2\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho^2 \mathrm{var}(\mathbf{x}_j)\right\} = \rho^2 \mathrm{var}(\mathbf{x}_j) \\ \delta_{kj} = 1 & \text{si} \max\left\{\max_l\left\{n\,\mathrm{cov}^2\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho^2 \mathrm{var}(\mathbf{x}_j)\right\} = n\,\mathrm{cov}^2\left(\mathbf{x}_j, \mathbf{c}_k\right) \end{cases}$$

– pour les groupes locaux :

$$\begin{cases} \delta_{kj} = 0 \ \forall k & \text{si} \max\left\{\max_l\left\{\sqrt{n}\,\mathrm{cov}\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho\sqrt{\mathrm{var}(\mathbf{x}_j)}\right\} = \rho\sqrt{\mathrm{var}(\mathbf{x}_j)} \\ \delta_{kj} = 1 & \text{si} \max\left\{\max_l\left\{\sqrt{n}\,\mathrm{cov}\left(\mathbf{x}_j, \mathbf{c}_l\right)\right\}, \rho\sqrt{\mathrm{var}(\mathbf{x}_j)}\right\} = \sqrt{n}\,\mathrm{cov}\left(\mathbf{x}_j, \mathbf{c}_k\right) \end{cases}$$

Dans la construction de critère T_{new} (3.6) et S_{new} (3.7), le paramètre de réglage ρ est analogue à un coefficient de corrélation. Il peut être choisi de manière simple par l'utilisateur qui a préalablement une idée du seuil de corrélation souhaité. Sinon, il peut être choisi comme étant la valeur critique du coefficient de corrélation de *Pearson*, pour un nombre d'observations donné et un niveau de risque α donné. Lorsque l'on veut comparer des solutions différentes ayant un nombre de groupes K différents, nous suggérons d'ajuster le risque α pour tenir compte, au mieux, de la modification du risque global de première espèce lorsque K varie. En effet, le test d'affectation effectué pour chaque variable x_j englobe implicitement une série de comparaisons de la corrélation de la variable avec chacune des variables latentes de groupe.

3.3.2 Stratégie "*Sparse* LV"

Considérons la mise en œuvre de la méthode CLV avec les critères de base pour groupe directionnel (équation (2.1)). Pour une solution en K groupes, nous verrons sur la base de données simulées dans la section (§ 3.4) que les variables atypiques ou de bruit sont affectées au hasard à l'un des groupes principaux, par exemple G_k. Cependant on s'attend à ce que leur contribution à la variable latente $\mathbf{c}_k$ de ce groupe

soit faible, comparée à celle des variables dans G_k fortement corrélées à $\mathbf{c}_k$.

Dans cette stratégie, nous considérons la même logique que celle de *Sparse PCA*. Elle consiste à chercher pour chaque groupe de variables G_k $(k = 1, \ldots, K)$, une variable latente *sparse* $\mathbf{c}_k^*$ pour laquelle les variables qui contribuent peu à la variable latente auront des *loadings* de valeur nulle. L'hypothèse est que ce seront justement les variables atypiques ou les variables de bruit, placées dans le groupe G_k, qui auront ces *loadings* nuls. En nous appuyant sur l'algorithme de partitionnement décrit dans la section (§ 2.2.2), nous proposons une modification pour l'étape d'estimation :

1. Pour un groupe G_k $(k = 1, \ldots, K)$ qui contient p_k variables, on note $\boldsymbol{\alpha}_k$ le vecteur de *loadings* associé à la première composante $\mathbf{c}_k$ de $\mathbf{X}_k$.
2. Calcul d'un vecteur de *loadings sparse* $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{jk}, \ldots, \beta_{p_k k})^T$. Pour chaque variable $\mathbf{x}_j$ appartenant à ce groupe, β_{jk} est défini par :

$$\beta_{jk} = \text{Signe}\left(\text{cov}\left(\mathbf{x}_j, \mathbf{c}_k\right)\right)\left(\left|\text{cov}\left(\mathbf{x}_j, \mathbf{c}_k\right)\right| - \rho\sqrt{\text{var}\left(\mathbf{x}_j\right)/n}\right)_+ \tag{3.8}$$

   qui est un seuillage doux (tel qu'il a été défini en (2.13)) sur la covariance entre la variable $\mathbf{x}_j$ et la variable latente $\mathbf{c}_k$.
3. Mise à jour du vecteur de *loadings* $\boldsymbol{\alpha}_k$ afin de maximiser $\boldsymbol{\alpha}_k^T \mathbf{X}_k^T \mathbf{X}_k \boldsymbol{\beta}_k$, sous la contrainte $\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k = 1$. La solution est $\boldsymbol{\alpha}_k = \mathbf{X}_k^T \mathbf{X}_k \boldsymbol{\beta}_k / \|\mathbf{X}_k^T \mathbf{X}_k \boldsymbol{\beta}_k\|$. Mise à jour de $\mathbf{c}_k = \mathbf{X}_k \boldsymbol{\alpha}_k / \|\mathbf{X}_k \boldsymbol{\alpha}_k\|$.
4. Répétition des étapes 2 et 3 jusqu'à convergence du vecteur de *loadings* $\boldsymbol{\beta}_k$
5. Le vecteur de *loadings sparse* et la variable latente *sparse* du groupe G_k sont respectivement $\mathbf{v}_k = \boldsymbol{\beta}_k / \|\boldsymbol{\beta}_k\|$ et $\mathbf{c}_k^* = \mathbf{X}_k \mathbf{v}_k / \|\mathbf{X}_k \mathbf{v}_k\|$.

Pour les groupes locaux, une stratégie similaire est adoptée mais avec comme variable latente dans un groupe G_k, la variable moyenne du groupe. L'algorithme avec définition de variables latentes *sparse* est construit comme suit :

1. Pour un groupe G_k $(k = 1, \ldots, K)$, on note $\boldsymbol{\alpha}_k = \left(\frac{1}{p_k}, \frac{1}{p_k}, \ldots, \frac{1}{p_k}\right)^T$ le vecteur de *loadings* associée à la variable moyenne normée $\mathbf{c}_k$ de G_k.

2. Calcul d'un vecteur de *loadings sparse* $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{jk}, \ldots, \beta_{p_k k})^T$. Pour chaque variable $\mathbf{x}_j$ appartenant à ce groupe, β_{jk} est défini par :

$$\beta_{jk} = \begin{cases} 1 & \text{si} \quad \text{cov}\left(\mathbf{x}_j, \mathbf{c}_k\right) > \rho\sqrt{\text{var}\left(\mathbf{x}_j\right)/n} \\ 0 & \text{sinon} \end{cases} \tag{3.9}$$

   qui est une fonction indicatrice permettant d'annuler les *loadings* des variables pour lesquelles la covariance avec $\mathbf{c}_k$ est inférieure à un seuil.

3. Pour un vecteur de *loadings* $\boldsymbol{\beta}_k$ fixé, mis à jour de $\boldsymbol{\alpha}_k = \boldsymbol{\beta}_k/(\sum_j \beta_{jk})$ et de $\mathbf{c}_k = \mathbf{X}_k\boldsymbol{\alpha}_k/\|\mathbf{X}_k\boldsymbol{\alpha}_k\|$.

4. Répétition des étapes 2 et 3 jusqu'à convergence de $\boldsymbol{\beta}_k$

5. Le vecteur de *loadings sparse* et la variable latente *sparse* de groupe G_k sont respectivement $\mathbf{v}_k = \boldsymbol{\beta}_k/(\sum_j \beta_{jk})$, et $\mathbf{c}_k^* = \mathbf{X}_k\mathbf{v}_k/\|\mathbf{X}_k\mathbf{v}_k\|$.

Nous pouvons remarquer que le paramètre du seuillage ($\rho\sqrt{\text{var}\left(\mathbf{x}_j\right)/n}$) utilisé dans les expressions (3.8) et (3.9) pour la stratégie "*Sparse* LV" est le même que le seuil d'affectation utilisé dans la stratégie "$K+1$". De fait, le paramètre ρ est un seuil de corrélation qui est à choisir entre 0 et 1.

3.4 Simulations

L'objectif de la simulation est de montrer que les deux approches proposées sont capables d'identifier des variables générées complètement au hasard parmi un ensemble de variables structurées en groupes. Deux modèles de simulation sont considérés : le premier correspond à un cas simple avec deux groupes de variables bien séparés. Dans le deuxième cas, quatre groupes de variables de différentes tailles, avec des corrélations non nulles entre les groupes, sont générés. L'intérêt de ce deuxième modèle est aussi d'étudier le comportement des approches lorsque le nombre de groupes spécifié au départ n'est pas le vrai nombre de groupes.

Modèle 1

Dans ce premier exemple, 100 jeux de données de 90 variables mesurées sur 50 observations sont générées. Chaque jeu de données est formé de deux groupes indépendants, de 30 variables chacun, et 30 variables isolées, générées indépendamment des deux groupes principaux. Pour les deux groupes, deux variables de support, indépendantes, $\mathbf{Z}_1$ et $\mathbf{Z}_2$, ont été créées et dans chacun de ces groupes, les variables ont été générées autour de $\mathbf{Z}_1$ et $\mathbf{Z}_2$ en ajoutant un terme d'erreur aléatoire. Les 30 variables isolées ont été générées, de manière indépendante, selon une distribution gaussienne. On a donc :

$$\begin{aligned}
&\mathbf{x}_j = \omega_j \mathbf{Z}_1 + \boldsymbol{\varepsilon}_j, \quad \mathbf{Z}_1 \sim \mathcal{N}(\mathbf{0}, I), \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \sigma I), \quad \text{pour} \quad j = 1, \ldots, 30, \\
&\mathbf{x}_j = \omega_j \mathbf{Z}_2 + \boldsymbol{\varepsilon}_j, \quad \mathbf{Z}_2 \sim \mathcal{N}(\mathbf{0}, I), \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \sigma I), \quad \text{pour} \quad j = 31, \ldots, 60, \\
&\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, I), \quad \text{pour} \quad j = 61, \ldots, 90
\end{aligned}$$

où $\mathbf{0}$ est un vecteur nul et I la matrice identité. Le facteur $\omega_j \in \{+1, -1\}$ est utilisé pour produire au hasard un coefficient de corrélation positif ou négatif entre chaque variable et sa variable de support. L'écart-type des erreurs σ est fixé à 0.8.

Le Tableau (3.1) résume les résultats de classification, lorsqu'une partition en deux groupes directionnels est envisagée, en utilisant la méthode CLV (décrite dans la section § 2.2), la stratégie "$K + 1$" (décrite dans la section § 3.3.1), ou la stratégie "*Sparse* LV" (décrite dans la section § 3.3.2). Les vrais groupes sont indiqués par G_1 et G_2, les classes trouvées par l'algorithme de classification sont notées C_1 et C_2.

Avec la méthode CLV ordinaire, chacune des deux classes trouvées inclue un des groupes principaux. Le modèle de simulation étant assez simple et tranché, ce résultat n'est pas étonnant. Cependant, les deux classes contiennent aussi, respectivement, 15.15 et 14.84 variables isolées, en moyenne. A l'aide de l'algorithme de la stratégie "$K + 1$" (avec $\rho = 0.5$), les deux classes C_1 et C_2 correspondent aux deux groupes de départ, G_1 et G_2, et la classe additionnelle C_A rassemble presque parfaitement les variables isolées. La stratégie "*Sparse* LV" (avec $\rho = 0.5$) ne produit pas explicitement de groupe additionnel mais les variables peu importantes vis-à-vis de la structure sont repérées avec des *loading* nuls. Nous pouvons observer que la plupart de variables

TABLE *3.1 – Tableaux croisés entre la structure de données simulées (G_1, G_2 et variables isolées) et les classes trouvées par la classification (C_1 et C_2)(moyenne des résultats sur 100 répétitions). Pour la stratégie "$K+1$" ($\rho = 0.5$), C_A représente le noise cluster. Pour la stratégie "Sparse LV" ($\rho = 0.5$), les chiffres correspondent au nombre de variables ayant un loading non nul (nombre de variables ayant un loading nul).*

$K=2$	CLV		"$K+1$"			"*Sparse* LV"	
$\rho = 0.5$	C_1	C_2	C_1	C_2	C_A	C_1	C_2
G_1	30	0	30	0	0	30	0
G_2	0	30	0	30	0	0	30
var.isolées	15,15	14,85	0.01	0	29,99	0.01(15,06)	0(14,93)

isolées (29.99 sur 30 en moyenne) ont eu un *loading* nul pour l'une ou l'autre des deux variables latentes *sparse*.

Concernant l'impact du paramètre ρ, nous avons fait varier la valeur de ρ et avons calculé trois critères pour évaluer la qualité des résultats : (i) θ_1 : la proportion de variables appartenant à l'un des deux groupes principaux correctement affectées dans les classes C_1 et C_2 ; (ii) θ_2 : la proportion de variables isolées correctement affectées au *noise cluster*, ou ayant un *loading* nul, selon la stratégie mise en œuvre. (iii) θ_3 : le pourcentage de variables placées dans le *noise cluster*, C_A, ou ayant reçu un *loading* nul, selon la stratégie mise en œuvre.

La Figure 3.2 montre la performance de deux stratégies pour ρ variant de 0 à 1 avec un pas de 0.05. Les évolutions de θ_1, θ_2 et θ_3 pour deux stratégies sont très similaires. Comme attendu, le pourcentage des variables rejetées à tort, θ_3, augmente avec ρ. Si une grande valeur de ρ est utilisée ($\rho \geqslant 0.9$), toutes les variables seront traitées comme du bruit. Dans cet exemple, prendre une valeur de ρ dans l'intervalle $[0.35, 0.7]$ conduit à une solution très satisfaisante (θ_1 et $\theta_2 > 0.95$ et $\theta_3 \simeq 0.3$).

Afin d'évaluer l'impact de la "compacité" des groupes de variables sur la performance, l'écart-type des erreurs, σ, a été modifié de 0.1 à 2. La Figure 3.3 montre les changements de θ_1 et θ_2 en utilisant la stratégie "$K+1$" avec différentes valeurs de σ et de ρ (les résultats de la stratégie "*Sparse* LV" ne sont pas représentés parce qu'ils présentent exactement le même profil). Dans la partie gauche de la Figure 3.3, quand σ augmente, θ_1 diminue plus rapidement. Ainsi, moins les groupes de variables

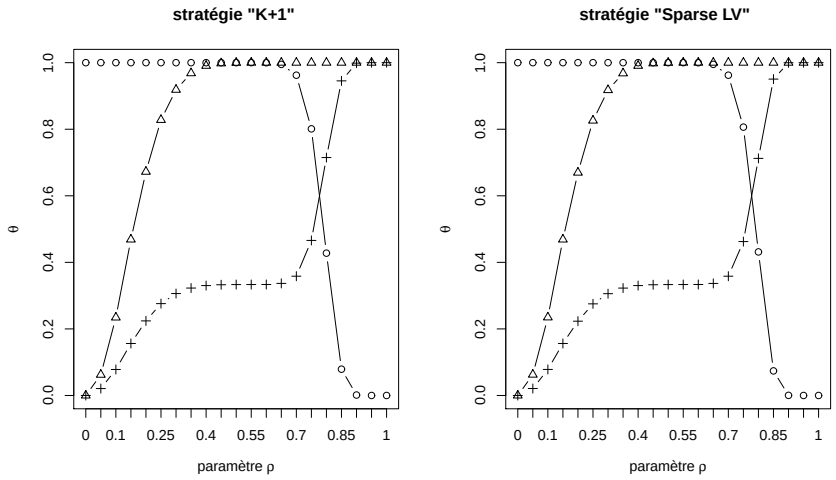

FIGURE *3.2 – Évolutions de $\theta_1(\circ)$, $\theta_2(\triangle)$ et $\theta_3(+)$, en fonction du paramètre ρ. θ_1 : pourcentage de variables de groupes correctement identifiées. θ_2 : pourcentage de variables isolées correctement identifiées. θ_3 : pourcentage de variables considérées comme des variables de bruit.*

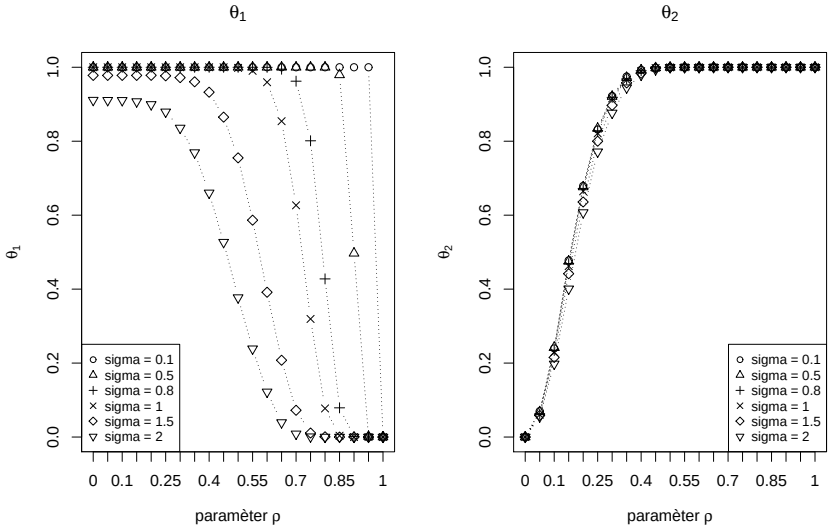

FIGURE *3.3 – Évolution de θ_1 (pourcentage de variables de groupes correctement identifiées) sur le côté gauche, et l'évolution de θ_2 (pourcentage de variables isolées correctement identifiées) sur le côté droit, en utilisant la stratégie "$K+1$", avec différentes valeurs de σ et ρ.*

sont compacts, plus la valeur du paramètre ρ devra être faible pour limiter les erreurs d'affectation des variables des groupes principaux au *noise cluster*. Au contraire dans la partie droite, les proportions de variables isolées correctement identifiées, θ_2, ne paraissent pas être affectées par σ. Cela pourrait s'expliquer par le fait qu'ici les variables isolées ont été standardisées, comme les variables de groupes, et qu'une fois les variables latentes des groupes correctement estimées, la probabilité qu'une variable de bruit soit, ou ne soit pas, affectée dans les groupes principaux ne dépend pas de la variabilité intra-groupe. Dans cette simulation, les deux prototypes de groupes, $\mathbf{Z}_1$ et $\mathbf{Z}_2$ ont été simulés pour ne pas être corrélés, ce qui conduit à des variables latentes estimées, $\mathbf{c}_1$ et $\mathbf{c}_2$, proches des prototypes $\mathbf{Z}_1$ et $\mathbf{Z}_2$.

Modèle 2

Dans cette seconde simulation, 100 jeux de données de 85 variables mesurées sur 50 observations ont été générés. Ils contiennent quatre groupes de variables, de taille 30, 20, 10, et 5 respectivement, et 20 variables isolées. Les quatre variables de support $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}_3$, et $\mathbf{Z}_4$ sont générées suivant une loi Normale multivariée avec un vecteur de moyenne $\mu = (0, 0, 0, 0)$ et une matrice de covariance Σ :

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.3 & 0.2 \\ 0 & 0.3 & 1 & 0.1 \\ 0 & 0.2 & 0.1 & 1 \end{pmatrix} \tag{3.10}$$

Comme pour le modèle 1, les variables sont générées de la manière suivante :

$$\begin{aligned} \mathbf{x}_j &= \omega_j \mathbf{Z}_1 + \boldsymbol{\varepsilon}_j, \quad j = 1, \ldots, 30, \\ \mathbf{x}_j &= \omega_j \mathbf{Z}_2 + \boldsymbol{\varepsilon}_j, \quad j = 31, \ldots, 50, \\ \mathbf{x}_j &= \omega_j \mathbf{Z}_3 + \boldsymbol{\varepsilon}_j, \quad j = 51, \ldots, 60, \\ \mathbf{x}_j &= \omega_j \mathbf{Z}_4 + \boldsymbol{\varepsilon}_j, \quad j = 61, \ldots, 65, \\ \mathbf{x}_j &\sim \mathcal{N}(\mathbf{0}, I), \quad j = 66, \ldots, 85 \end{aligned}$$

TABLE *3.2 – Tableaux croisés entre la structure de données simulées (G_1, G_2, G_3, G_4 et variables isolées) et les classes trouvées par la classification en fonction de la méthode de classification utilisée et du nombre de groupes K choisi (moyenne des résultats sur 100 répétitions). Pour la stratégie "$K+1$" ($\rho = 0.5$), C_A représente le noise cluster. Pour la stratégie "Sparse LV" ($\rho = 0.5$), les chiffres correspondent au nombre de variables ayant loading non nul (nombre de variables ayant un loading nul).*

$K=3$	CLV			"$K+1$"				"*Sparse* LV"		
$\rho = 0.5$	C_1	C_2	C_3	C_1	C_2	C_3	C_A	C_1	C_2	C_3
G_1	30	0	0	30	0	0	0	30	0	0
G_2	0	20	0	0	20	0	0	0	20	0
G_3	0	0	10	0	0	10	0	0	0	10
G_4	0.42	3.41	1.17	0	0	0	5	(0.69)	(3.17)	(1.14)
var.isolées	6.76	6.34	6.90	0.01	0.02	0.03	19.94	0.01(7.05)	0.02(6.29)	0.02(6.61)

$K=4$	CLV				"$K+1$"					"*Sparse* LV"			
$\rho = 0.5$	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_A	C_1	C_2	C_3	C_4
G_1	30	0	0	0	30	0	0	0	0	30	0	0	0
G_2	0	20	0	0	0	20	0	0	0	0	20	0	0
G_3	0	0	10	0	0	0	10	0	0	0	0	10	0
G_4	0	0	0	5	0	0	0	5	0	0	0	0	5
var.isolées	4.77	4.44	4.97	5.82	0.01	0.03	0.03	0.09	19.84	0.01(5.12)	0.03(4.74)	0.02(5.06)	0.06(4.96)

$K=5$	CLV					"$K+1$"						"*Sparse* LV"				
$\rho = 0.5$	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	C_A	C_1	C_2	C_3	C_4	C_5
G_1	30	0	0	0	0	30	0	0	0	0	0	30	0	0	0	0
G_2	0	20	0	0	0	0	20	0	0	0	0	0	20	0	0	0
G_3	0	0	10	0	0	0	0	10	0	0	0	0	0	10	0	0
G_4	0	0	0	5	0	0	0	0	5	0	0	0	0	0	2	0
var.isolées	2.7	2.34	2.7	3.04	9.22	0.01	0.03	0.03	1.2	3.53	16.58	0.01(3.43)	0.03(3.29)	0.02(3.56)	0.06(3.19)	2.27(4.14)

avec $\omega_j \in \{+1, -1\}$ et $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, 0.8I)$.

Dans la suite, la méthode CLV ordinaire et les deux stratégies proposées, avec le paramètre $\rho = 0.5$, ont été comparées dans trois situations différentes : le nombre de groupes choisi $K < 4$, $K = 4$, $K > 4$, 4 étant le vrai nombre de groupes. Comme dans l'exemple précédent, G_k (avec $k = 1, \ldots, 4$) représente un des groupes réels et C_k (avec $k = 1, \ldots, 4$) représente une des classes obtenues. La synthèse des résultats figure dans le Tableau 3.2.

- Si le nombre de groupes est bien choisi, ($K = 4$, milieu du Tableau 3.2), des résultats similaires à la première simulation ont été observés. Les trois approches permettent de retrouver les groupes réels, mais avec CLV ordinaire, les variables de bruit sont aussi distribuées au hasard dans les quatre classes. Avec les deux stratégies proposées, les variables isolées sont généralement bien identifiées. Elles sont affectées dans le *noise cluster*, C_A, avec la stratégie "$K+1$" ou ont eu un *loading* nul avec la stratégie "*Sparse* LV".
- Si une partition en $K = 3 < 4$ groupes est envisagée (partie haute du Tableau 3.2), la méthode CLV ordinaire aboutit à la redistribution du groupe de plus

petite taille (ici G_4) dans les trois autres groupes même si les corrélations avec les variables de support sont modestes (0.3 maximum). Dans ce type de situation, les variables générées autour de $\mathbf{Z}_4$ pourraient être considérées à tort comme des variables de bruit. Par ailleurs, le fait qu'elles soient allouées au hasard dans les trois autres groupes principaux complique l'interprétation des groupes trouvés. Au contraire, la stratégie "$K + 1$" donne des résultats intéressants et plus raisonnables. L'existence du petit groupe, G_4, n'a pas été repérée mais les autres groupes ont été correctement identifiés. Avec la stratégie "*Sparse* LV", les variables issues du groupe G_4 ont obtenu un *loading* nul sur la variable latente du groupe dans lequel elles ont été affectées.

– Lorsque le nombre de groupes K demandé est 5 au lieu de 4 (partie basse du Tableau 3.2), dans la solution de CLV classique, un certain nombre de variables isolées (9.22 sur 20, en moyenne) forment le $5^{\text{ième}}$ groupe. Les vrais groupes, G_1 à G_4 ont parfaitement été retrouvés et le reste des variables isolées a été attribué de façon aléatoire dans les classes C_1 à C_4. Pour les stratégies "$K + 1$" et "*Sparse* LV", la taille de $5^{\text{ième}}$ groupe est relativement faible (3.53 variables ou 2.27 avec un *loading* nul, en moyenne). En fait, l'objectif de ces deux approches est d'identifier des groupes homogènes en écartant les variables non pertinentes. Il se trouve qu'un groupe qui comprend surtout des variables disparates sera débarrassé de la plupart de ces variables, qui vont être déplacées dans le *noise cluster*.

3.5 Données réelles

3.5.1 Données de RMN

Nous considérons un jeu de données contenant les spectres RMN de jus de fruit collectés dans le cadre d'une étude visant l'authentification de jus d'orange (Vigneau et Thomas, 2012). Les observations correspondent à 150 mélanges de jus d'orange avec différentes proportions de jus de clémentine. La zone spectrale s'étend de 0.5 à 2.3 ppm (Figure 3.4), et comprend 180 variables spectrales (déplacements chimiques).

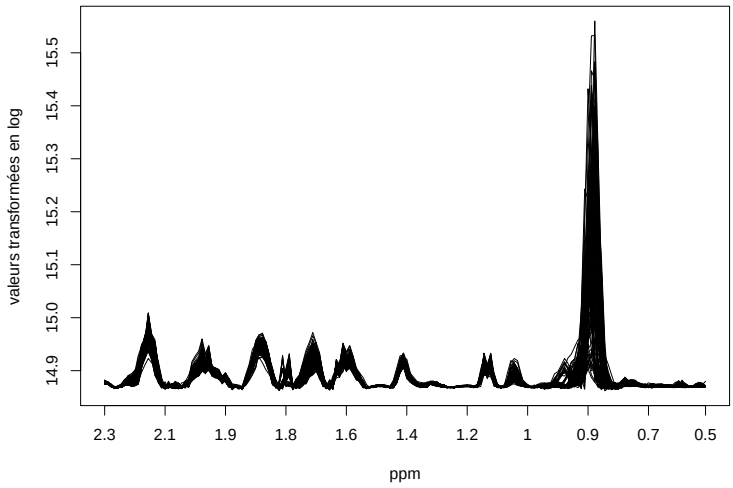

FIGURE *3.4 – Spectres RMN des jus de fruit correspondant à la zone spectrale allant de 0.5 à 2.3 ppm. Les valeurs de signaux ont été transformées en logarithme.*

La méthode CLV est utilisée pour identifier des groupes de variables spectrales afin de réduire la complexité d'interprétation des signaux, et ainsi mieux comprendre quelles sont les composantes latentes contenues dans les spectres RMN. Dans cet exemple, nous avons décidé de rechercher des groupes locaux qui permettent d'agréger des variables positivement corrélées entre elles. L'objectif est, en effet, de mettre dans un même groupe de variables, les variables spectrales qui appartiennent à des bandes spectrales et qui correspondent à la même information chimique.

Les deux stratégies, stratégie "$K+1$" et stratégie "*Sparse LV*", sont utilisées avec une valeur de paramètre $\rho = 0.5$ et une initialisation de l'algorithme de partitionnement par coupure du dendrogramme de l'algorithme hiérarchique (voir section § 2.2.2). La Figure 3.5 montre l'évolution du pourcentage des variables mises dans le *noise cluster*, ou obtenant un *loading* nul, en fonction de nombre de groupes K. Pour les deux stratégies, le nombre de variables écartées diminuent de la même manière en fonction de K. Si l'on cherche un seul groupe de variables, environ 80% de variables sont considérées comme atypiques ou comme de bruit. Avec une seule variable latente, il s'avère que la plupart des variables ne sont pas bien associées à cette composante.

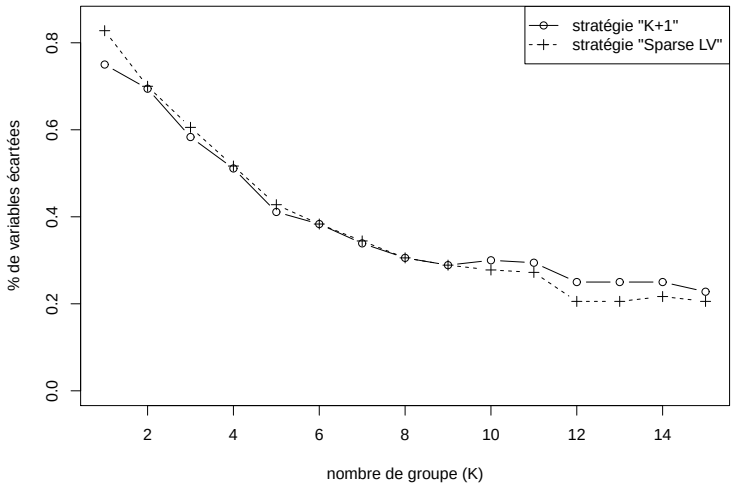

FIGURE *3.5 – Proportions de variables écartées en fonction de nombre de groupes avec la stratégie "$K+1$" et la stratégie "Sparse LV".*

Un partitionnement en cinq groupes a été retenu. Les cinq groupes définis par la procédure CLV classique sont indiqués par des numéros le long du spectre moyen dans la Figure 3.6 . En regardant en détail la corrélation de chaque variable spectrale avec la variable latente du groupe dans lequel elle est affectée, on observe que les coefficients de corrélation les plus importants sont respectivement de 0.76, 0.76, 0.91, 0.82, 0.87 dans les groupes 1 à 5, mais que les coefficients de corrélation les plus faibles ne sont que de 0.10, 0.22, 0.16, 0.18, 0.20, respectivement. De nombreuses variables spectrales s'avèrent être assez modestement corrélées avec leur variable latente. Avec les stratégies "$K+1$" et "*Sparse* LV", il y a respectivement 74 et 77 variables (soit 41% et 43%) qui sont écartées. Leur position est indiquée en dessous du spectre moyen dans la Figure 3.6. On observe que les variables du *noise cluster*, ou de *loading* nul, correspondent souvent à des variables spectrales dont l'intensité est faible (proches de la ligne base). En revanche, la plupart de variables associées aux pics spectraux sont conservées dans les groupes principaux, ou ont un *loading* non nul.

Par la suite, nous avons comparé les résultats obtenus avec seulement 106 variables (les variables qui ne sont pas dans *noise cluster* en utilisant la stratégie "$K+1$") et ceux

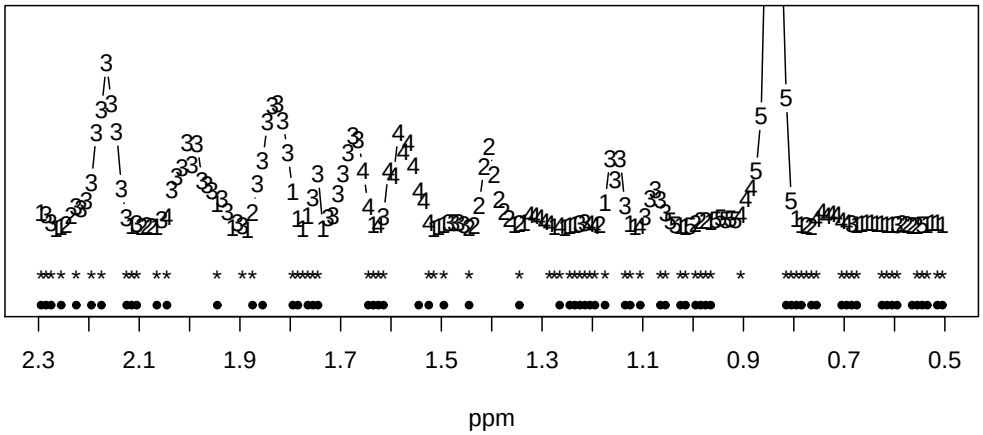

FIGURE *3.6 – Spectre moyen des jus de fruit et identification des groupes obtenus avec CLV. La position des variables exclues est identifiée par ∗ pour la stratégie "$K + 1$" et • pour la stratégie "Sparse LV" (L'échelle en ordonnée est tronquée pour une meilleure lisibilité générale du spectre).*

obtenus avec l'ensemble des 180 variables. Du point de vue de l'analyse exploratoire par ACP, il se trouve que les résultats pour ces deux cas sont très similaires. En écartant les variables du *noise cluster*, on observe une très légère augmentation du pourcentage de variance expliquée par les deux premières composantes principales (haut du Tableau 3.3). En regardant les corrélations entre les paires des cinq premières composantes principales définies dans chacun des deux cas (bas du Tableau 3.3), on observe que la réduction du nombre de variables prises en compte n'a pas modifié les deux premières composantes principales et très peu les composantes trois à cinq.

Du point de vue de l'analyse prédictive, comme dans cette étude de cas, nous nous intéressons également à la prédiction de pourcentage de jus de fruit de clémentine ajouté dans le jus d'orange, nous avons évalué la performance de prédiction du modèle de régression PLS ajusté avec respectivement les 180 variables ou les 106 variables retenues. La racine carrée de l'erreur quadratique moyenne (RMSE) de validation croisée (selon la procédure décrite dans Vigneau et Thomas (2012)) en fonction de nombre de composantes PLS utilisées dans le modèle, sont présentées dans la Fi-

TABLE *3.3 – Pourcentages d'inertie expliquée et matrice de corrélations entre les cinq premières composantes principales obtenues par ACP avec les 180 variables de départ (cas A) et les 106 variables qui n'ont pas été placées dans le noise cluster (cas B) .*

	Pourcentage de variance				
	comp. 1	comp. 2	comp. 3	comp. 4	comp. 5
A(180 var.)	35.0%	17.0%	9.7%	6.0%	3.7%
B(106 var.)	42.6%	19.5%	7.1%	5.6%	3.5%
	matrice de corrélations				
	B comp. 1	B comp. 2	B comp. 3	B comp. 4	B comp. 5
A comp. 1	1.00	0.04	-0.03	-0.03	0.01
A comp. 2	-0.04	0.99	-0.08	-0.03	0.02
A comp. 3	0.05	0.10	0.84	0.38	-0.15
A comp. 4	0.01	-0.01	-0.44	0.88	-0.01
A comp. 5	-0.00	-0.02	0.20	0.10	0.91

gure 3.7. Comme on peut l'observer, écarter les variables du *noise cluster* n'a pas eu d'incidence notoire sur la qualité attendue en prédiction. Il semble que les variables informatives ont donc bien été retenues parmi les 106 variables. De plus, en réduisant le nombre de variables considérées qui sont, ici, principalement localisées au niveau des pics des spectres l'interprétabilité du modèle est potentiellement améliorée.

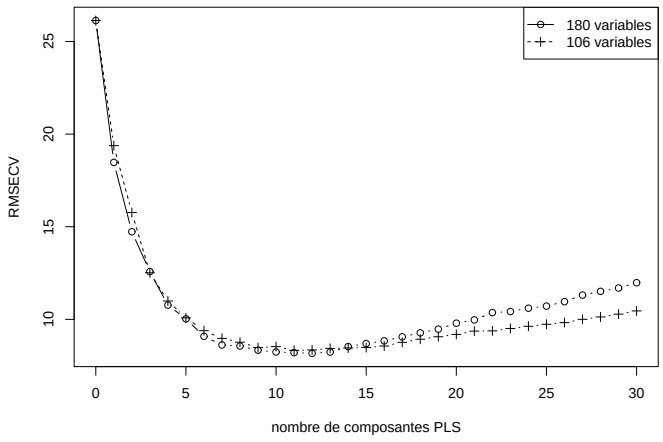

FIGURE *3.7 – La racine carrée de l'erreur quadratique moyenne (RMSE) de validation croisée du modèle PLS ajusté avec respectivement les 180 variables ou les 106 variables retenues avec la stratégie "K+1", en fonction de nombre de composantes PLS.*

3.5.2 Données métabolomiques

Ce jeu de données métabolomiques contient 4515 variables (ions) identifiées sur des échantillons urinaires de 32 veaux. La technique analytique de séparation des ions est une RPLC-HRMS (*Reverse-Phase Liquid Chromatography-High Resolution Mass Sectrometry*) (Jacob *et al.*, 2014). La caractérisation et la comparaison des empreintes métabolomiques urinaires obtenues pour des animaux non traités (contrôles) et traités, a pour objectif de détecter l'utilisation de promoteurs de croissance pour les veaux. Du point de vue d'analyse statistique, les méthodes comme l'ACP, la classification de variables, l'ANOVA, la régression logistique permettent de mieux comprendre la structure sous-jacente des métabolites, d'identifier des biomarqueurs et de construire un modèle prédictif. Cependant, le grand nombre de variables rend les méthodes usuelles difficiles à utiliser et complique l'interprétation des résultats. Dans la suite, nous allons, dans un premier temps, comparer les deux stratégies proposées avec la méthode CLV classique, et dans un second temps, regarder de plus près les solutions obtenues avec mise à l'écart d'une partie des métabolites.

Les échantillons urinaires pour 32 veaux (16 contrôles et 16 traités) ont été collectés 2 jours avant le début du traitement puis 7, 21, 28, 35, 42, 48, 56, 63 et 68 jours après le début du traitement. Dans cette analyse, nous considérons seulement le jeu de données à J42, c'est-à-dire à la $6^{\text{ième}}$ semaine. Une transformation logarithmique ($\log(\mathbf{x}+1)$) et une standardisation de type Pareto ($\mathbf{x}/\sqrt{\sigma(\mathbf{x})}$) sont utilisées comme pré-traitement de données. Le *Volcano Plot* représenté Figure 3.8, donne la position de chaque métabolite en fonction de la significativité (*p-value* du test de *Student*) de la différence entre les deux moyennes observées dans les deux conditions (contrôle et traité) et en fonction du rapport de ces moyennes (*fold change*).

Dans un premier temps, les deux stratégies proposées sont utilisées avec une valeur de paramètre ρ déterminé pour un niveau de risque α approximativement de 0.01, en fonction du nombre de groupes (K). Ici on a, par exemple, pour $K = 1$, $\rho = 0.44$; pour $K = 2$, $\rho = 0.49$; ... ; pour $K = 10$, $\rho = 0.61$; pour $K = 20$, $\rho = 0.65$, etc. La Figure 3.9 donne les pourcentages des variables écartées dans le *noise cluster* avec la mise en œuvre de la stratégie "$K + 1$", ou ayant un *loading* nul suite à la mise en

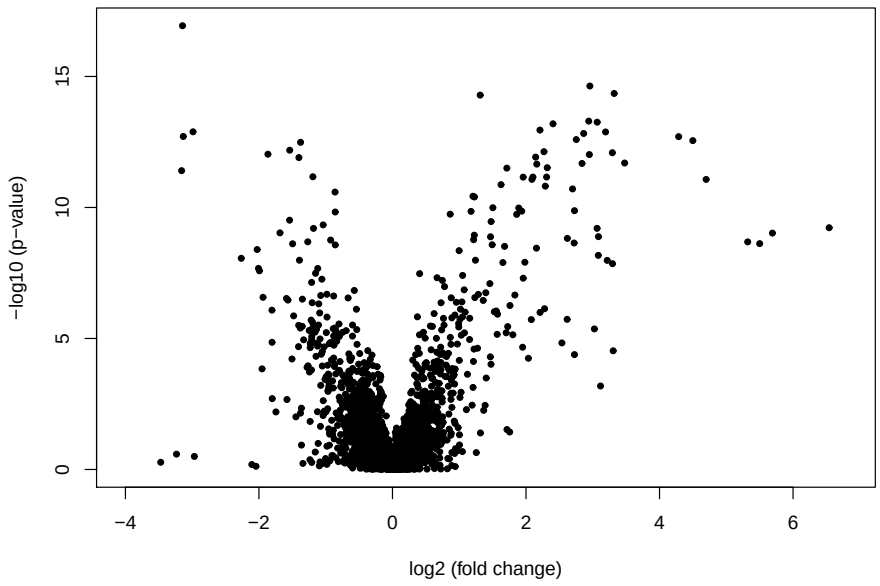

FIGURE *3.8 – Volcano plot : pour chaque variable* $\mathbf{x}_j$ *en abscisse, le fold change*$= \bar{\mathbf{x}}_c/\bar{\mathbf{x}}_t$ *(en log) ; en ordonnée, la p-value pour le test de Student de comparaison des deux moyennes (en log, est inversée de sorte à ce que les variables les plus discriminantes aient les plus fortes valeurs en ordonnée).*

œuvre de la stratégie "*Sparse* LV". Les deux courbes se stabilisent à partir de $K = 6$. Comme nous sommes plus intéressés au repérage des variables atypiques qu'au choix du nombre de groupes adéquats, nous avons choisi $K = 20$. Il y a environ 60% de variables écartées, 2763 pour la stratégie "$K + 1$" et 2706 pour la stratégie "*Sparse* LV". Avec un autre nombre de groupes (> 6) on aurait aboutit à des résultats assez similaires.

Les distributions des *p-value* (lié au pouvoir discriminant) des variables au sein de chaque groupe, sont représentés en Figure 3.10 pour les groupes issus de la méthode CLV classique, pour les groupes trouvés avec la stratégie "$K + 1$" (le groupe 0 correspond au *noise cluster*) et pour les variables ayant des *loadings* non nuls avec la stratégie "*Sparse* LV" (le cas 0 rassemble les variables ayant un *loading* nul).

On observe dans la Figure 3.10, une concentration de ces distributions dans certains groupes en utilisant l'une ou l'autre des stratégies proposées par rapport à la

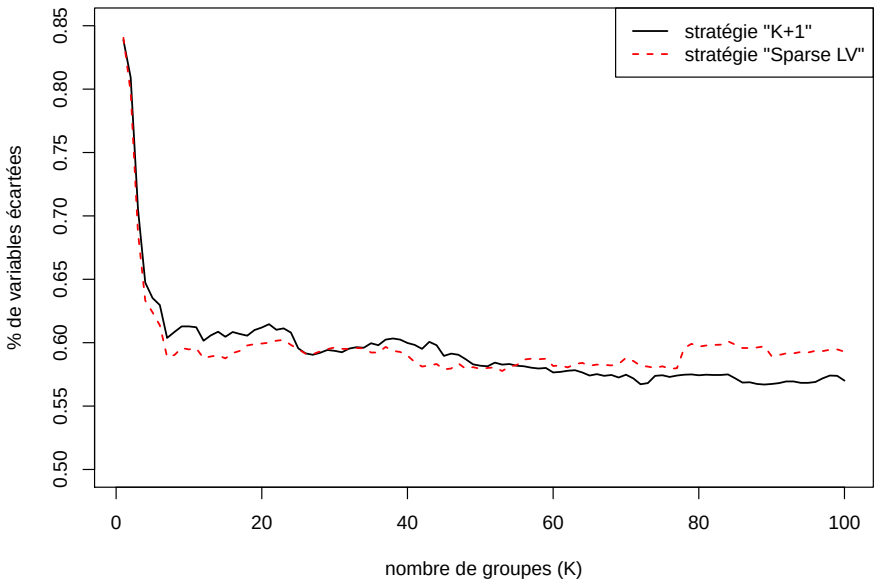

FIGURE *3.9 – Proportions de variables écartées en fonction de nombre de groupes avec la stratégie "$K+1$" et la stratégie "Sparse LV".*

CLV classique. Ceci est net pour les groupes 1, 2, 4, et 18 qui contiennent principalement des variables discriminantes (avec de petites valeurs de *p-value*), ainsi que pour le groupe 12 qui contient des métabolites non discriminants vis-à-vis du traitement administré. On s'intéresse donc spécifiquement aux quatre groupes 1, 2, 4 et 18, chacun d'eux rassemblent des métabolites corrélés à une même variable latente et orientés vers la discrimination des groupes de veaux. Le *Volcano plot* de la Figure 3.8 est reproduit en Figure 3.11 en mettant en évidence les positions des métabolites des groupes repérés (les résultats étant quasiment identiques, seuls ceux correspondant à la stratégie "$K+1$" sont représentés).

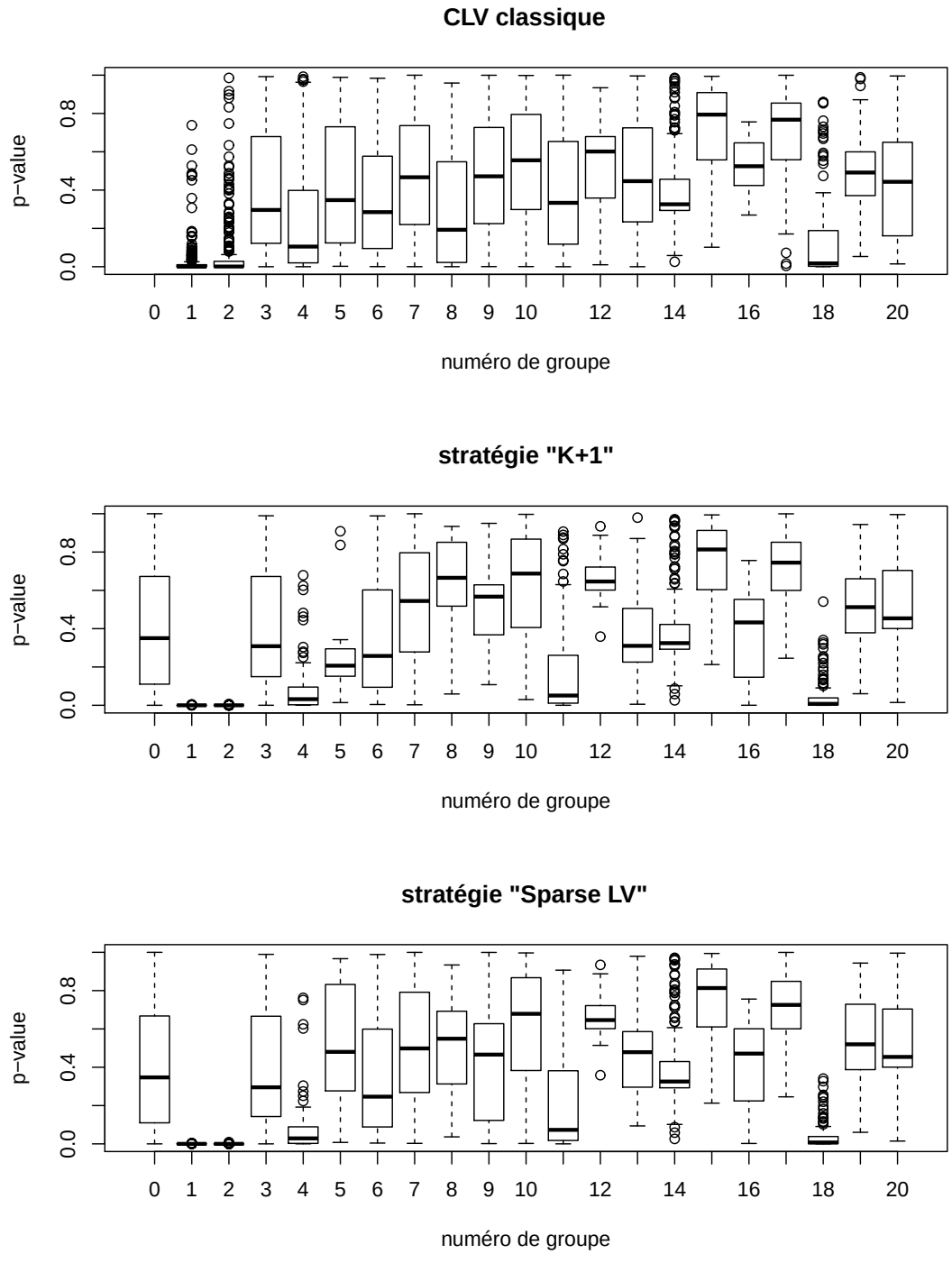

FIGURE *3.10 – Répartition des p-value (test de Student) des variables dans les différents groupes de variables. Le groupe 0 pour la stratégie "$K+1$" et la stratégie "Sparse LV" représente respectivement le noise cluster et les variables qui ont un loading nul.*

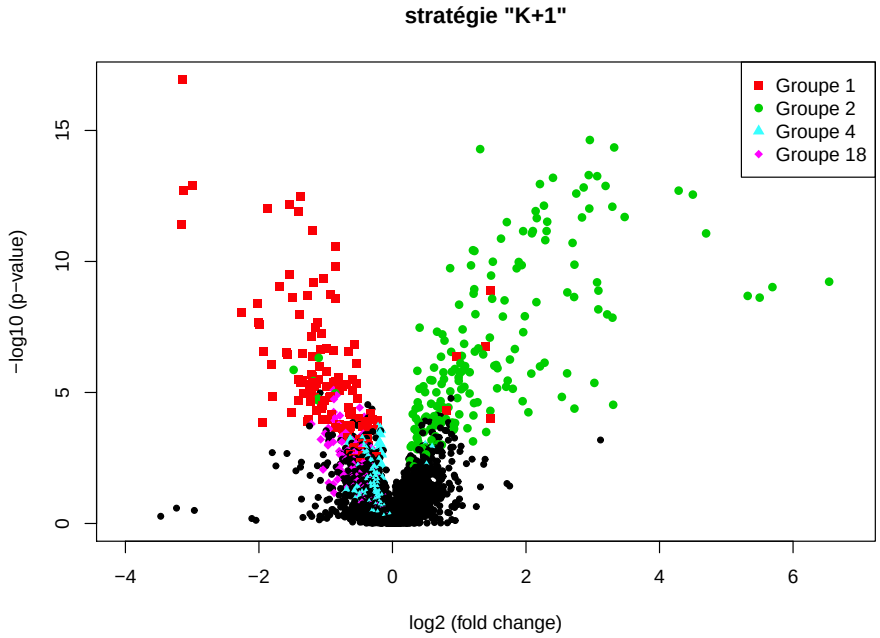

FIGURE *3.11 – Le Volcano plot sur lequel les variables appartenant aux quatre groupes repérés sont identifiées.*

Finalement, nous avons comparé les ACP réalisées en considérant les 4515 métabolites initiaux, les 1752 métabolites qui n'appartiennent pas au *noise cluster* de la stratégie "$K+1$" et les 520 métabolites des groupes 1, 2, 4 et 18. Les premiers plans factoriels, pour chaque analyse, sont représentés dans la Figure 3.12. La similitude des configurations des individus obtenues en utilisant les 4515 métabolites ou seulement 40% (1752) d'entre eux est claire. On en déduit que les variables qui ne sont pas liées aux structures principales du jeu de données ont été écartées. Ainsi les premières composantes principales ne sont pas modifiées. La configuration obtenue avec seulement 520 métabolites (associés aux quatre groupes repérés) révèle une meilleure séparation des individus en fonction du traitement. Ceci montre l'intérêt d'une sélection raisonnée de métabolites, en tenant compte de leur structure de groupes et de leur pouvoir discriminant.

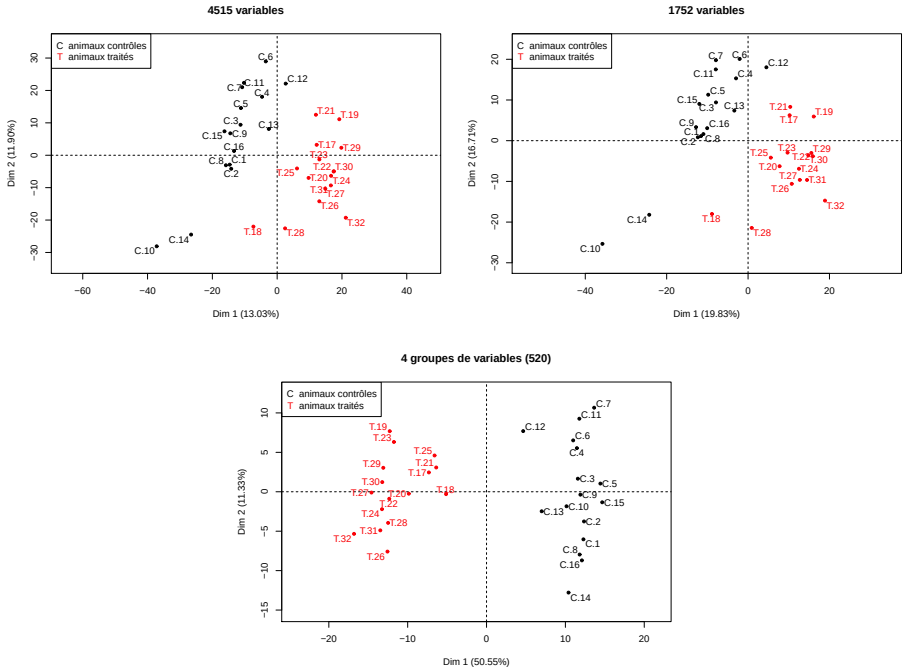

FIGURE *3.12 – Configurations des individus sur les deux premiers axes des ACP réalisées en utilisant les 4715 métabolites initiaux, les 1752 métabolites non écartés par la stratégie "$K+1$" et les 520 métabolites associés aux groupes 1, 2, 4 et 18. Un individu "contrôle" est identifié par un "C", un individu traité est identifié par un "T".*

3.6 Conclusion

Nous avons proposé deux modifications de la méthode CLV, la stratégie "$K+1$" et la stratégie "*Sparse* LV", afin de tenir compte des variables atypiques ou de bruit. Cette notion de variable "atypique" ou "de bruit", dans le contexte de classification, s'applique aux informations de type bruit blanc, mais également aux variables isolées, dispersées ou qui n'entrent pas bien dans la structure principale des données. Avec la stratégie "$K+1$", un groupe additionnel (*noise cluster*) est défini. Il est supposé contenir toutes les variables atypiques ou de bruit. Avec la stratégie "*Sparse* LV", les variables qui ne sont pas bien associées à la variable latente de chaque groupe se voient attribuées un *loading* nul. Du point de vue de l'algorithme du partitionnement de CLV, les deux stratégies correspondent à une modification, respectivement, de l'étape d'affectation ou de l'étape d'estimation des variables latentes. Le paramètre ρ

utilisé représente un seuil de corrélation qui doit être choisi entre 0 et 1.

Les deux stratégies montrent de bonnes performances et une grande similitude de comportement sur la base des données simulées ou de données réelles. Dans la simulation, leur comportement est logique lorsque le nombre de groupes spécifié au départ n'est pas le vrai nombre de groupes. Dans les applications réelles, les deux stratégies offrent un moyen de chercher des groupes de variables plus homogènes et donc plus simples à interpréter. De plus, elles peuvent être considérées comme des approches de pré-nettoyage, suivies de méthodes d'analyse exploratoire ou prédictive selon le contexte.

Bien que les résultats obtenus avec les deux stratégies soient très similaires, elles représentent deux aspects différents de la classification de variables CLV, qui sont, (i) l'identification des groupes de variables aussi homogènes que possible, et (ii) la réduction de la dimension des données en utilisant les variables latentes de groupes. Du point de vue du partitionnement des variables, la stratégie "$K + 1$" extrait un groupe additionnel en plus des groupes principaux. Cette stratégie est plus naturelle et directe si l'objectif au départ est d'identifier des groupes de variables. Cependant, si la méthode CLV est utilisée à l'instar des approches de l'analyse factorielle pour extraire les variables latentes associées à la structure des variables, la stratégie "*Sparse* LV" est recommandée.

Chapitre 4

Régression basée sur la classification des variables

Dans ce chapitre, nous proposons deux algorithmes de sélection de variables (CLVR et CLVlar) basés sur la classification de variables CLV. Les variables latentes de groupes vont être utilisées comme de nouveaux prédicteurs. Ils permettent de tenir compte de la structure de corrélation entre les variables et surmonter les difficultés rencontrées pour la sélection de variables lorsque celles-ci sont fortement corrélées. Avec les algorithmes proposés, une séquence pertinente de groupes, et donc de variables latentes peut être identifiée. Le modèle de régression est construit pas à pas en intégrant ces variables de groupes.

4.1 Contexte méthodologique

La construction d'un modèle de régression en grande dimension (avec de nombreux prédicteurs) est une problématique de plus en plus souvent rencontrée dans divers domaines. Les méthodes traditionnelles se heurtent à deux difficultés qui sont le "fléau de la dimension" ($p \gg n$) et la forte redondance (corrélations) entre les variables.

Deux grandes familles de stratégies ont été adoptées pour ces situations : la sélection de variables qui consiste à chercher un sous-ensemble de variables pertinentes, et la réduction de dimensionalité en utilisant des composantes latentes définies comme

des combinaisons linéaires des variables initiales (§ 1.2.2). La méthode Lasso (Tibshirani, 1996) et la méthode PLS (Wold, 1966) sont deux méthodes parmi les plus utilisées dans chacune de ces deux familles d'approches. Cependant, Zou et Hastie (2005) montrent le manque de robustesse du sous-ensemble de variables conservé par la méthode Lasso. Cette méthode tend, en effet, à ne sélectionner, au hasard, qu'une seule variable parmi un groupe de variables très corrélées. Cependant, avoir une idée des variables fortement corrélées entre elles est intéressant en tant que tel. Par exemple, dans le domaine des données-omiques, ces groupes de variables peuvent représenter des formes biologiques intéressantes (Eisen *et al.*, 1998). En ce qui concerne les méthodes de réduction de la dimensionalité, comme les composantes sont souvent des combinaisons linéaires de toutes les variables d'origine, la solution n'est pas toujours facile à interpréter et l'identification de groupes éventuels de variables n'est pas directe.

Pourtant, les deux familles d'approches évoquées ci-dessus présentent un grand intérêt. La sélection de variables pertinentes et l'exploration de la structure des corrélations entre les variables sont deux aspects complémentaires. Une certain nombre de stratégies ont été proposés pour profiter des avantages de ces deux familles d'approches. Par exemple, des pénalisations de type *Lasso* ont été proposées pour la sélection de variables avec un effet de groupement des variables fortement corrélées. Le point de vue de départ est de faire en sorte que les variables qui sont potentiellement dans un même groupe aient le même coefficient de régression. Nous avons déjà cité *Elastic Net* proposé par Zou et Hastie (2005) qui utilise des pénalités L^1 et L^2. Citons également OSCAR proposé par Bondell et Reich (2008) dont la pénalité permet de mettre à zéro certains coefficients et de regrouper d'autres variables en contraignant leurs coefficients de régression à être égaux. La pénalisation d'*Elastic Net* a aussi été introduite dans le cadre de la régression PLS par Chun et Keles (2010); Le Cao *et al.* (2008). Ces approches de *Sparse* PLS intègrent la sélection de variables dans la régression PLS et tirent ainsi avantage des approches de sélection et des approches factorielles. Cependant, la structure de groupes de variables n'est pas explicitement intégrée dans les pénalisations.

Comme l'objectif principal est de chercher une structure de groupes au sein de l'en-

semble des variables prédictives, un autre point de vue plus direct est de construire une procédure qui combine la classification et la sélection : une classification de variables pour réduire le nombre de variables d'entrée suivie d'une régression. Cela a été proposé par Hastie *et al.* (2001b); Au *et al.* (2005); Ma et Huang (2007); Ma *et al.* (2007); Yousef *et al.* (2007); Park *et al.* (2007). Une comparaison de méthodes combinant classification de variables et modélisation est présentée dans Tolosi et Lengauer (2011). Ces approches consistent à créer un nouveau prédicteur pour chaque groupe de variables, qui est simplement la variable moyenne des variables du groupe. Ces nouvelles variables sont parfois appelées "méta-variables" ou "*super*-variables". Les modèles de régression construits sur les "méta-variables" n'intègrent ainsi qu'un nombre réduit de prédicteurs, dont le degré de redondance est moindre que celui des variables initiales. Il convient de noter que les techniques de classification utilisées dans ces approches ne concernent que la formation de "groupes locaux" (variables positivement corrélées). De plus, comme dans beaucoup de méthodes, la détermination du nombre optimal de groupes de variables à retenir est une question difficile. Hastie *et al.* (2001b) proposent avec l'algorithme "*tree harvesting*" de considérer tous les nœuds d'un arbre hiérarchique et de définir la "méta-variable" associée à chaque nœud. Avec le même objectif, Park *et al.* (2007) proposent d'effectuer une régression *Lasso* à tous les niveaux de la hiérarchie.

Les approches "*tree harvesting*" (Hastie *et al.*, 2001b) et "*averaged gene expression*" (Park *et al.*, 2007) ayant particulièrement retenues notre attention, elles sont décrites plus en détail ci-dessous.

Tree harvesting Hastie *et al.* (2001b) sont parmi les premiers à avoir combiner une classification de variables et une étape de modélisation avec la méthode *tree harvesting*. Tout d'abord, de nombreux groupes de candidats sont générés par classification ascendante hiérarchique (groupes locaux). Tous les *centroids* de groupes à tous les niveaux de l'arbre hiérarchique sont considérés comme des variables prédictives potentielles en plus des variables initiales. Ensuite une régression *stepwise* au cours de laquelle variables et *centroids* entrent petit à petit dans le modèle est réalisée.

Schéma général de l'algorithme

1. Classification hiérarchique des p variables aboutissant à la construction d'un arbre hiérarchique avec $2p-1$ nœuds. Les $2p-1$ *centroids* des groupes formés au niveau des $2p-1$ nœuds sont utilisés comme candidats.
2. Identification du prédicteur (variable ou *centroids*) qui améliore le plus l'ajustement du modèle au sens d'un score, $\mathcal{S}$ (à maximiser).
3. Ajout au modèle du prédicteur associé au groupe de plus grande taille dont le score est d'au moins $0.9 * \mathcal{S}$.
4. Itération jusqu'à ce que le nombre maximum de termes à ajouter soit atteint.

Plutôt que de couper l'arbre hiérarchique à un niveau fixe, *tree harvesting* permet de prendre en considération tous les groupes, à tous les niveaux de la résolution. L'intérêt est de ne pas avoir à choisir a priori le nombre de groupes. Cependant, lorsqu'il y a un grand nombre de variables initiales, le nombre de nœuds est important et la construction d'un modèle prédictif reste délicate.

Averaged gene expression Park *et al.* (2007) proposent une autre méthode en deux étapes utilisant conjointement une classification hiérarchique et une régression de type *Lasso* : les variables (des gènes ici) sont regroupées à l'aide d'une procédure de classification hiérarchique et les *centroids* de groupes (supergènes) sont utilisés pour la formation du modèle linéaire.

Schéma général de l'algorithme

1. Réalisation d'une classification ascendante hiérarchique des p gènes.
2. A chaque niveau de la hiérarchie, création des supergènes en faisant la moyenne des expressions des gènes appartenant à un même groupe. Cela nous donne p partitions différentes de gènes et donc p ensembles de supergènes, chaque ensemble représentant un niveau donné de la hiérarchie.
3. Pour chaque ensemble de supergènes, ajustement d'un modèle de régression *Lasso*, pour un critère de pénalité, λ, fixé.

4. Le degré optimal de la pénalité λ et le niveau optimal de coupure de l'arbre hiérarchique sont déterminés en utilisant une validation croisée.

Les deux paramètres (paramètre λ de *Lasso* et niveau de coupure du dendrogramme) sont optimisés simultanément en utilisant la validation croisée. Comme la recherche de toutes les combinaisons de deux paramètres est très longue, un chemin de recherche dans l'espace bidimensionnel est proposé par les auteurs.

Dans ce chapitre, nous considérons une autre démarche combinant la classification et la sélection de variables dans une démarche de modélisation. Dans la phase de classification, la méthode CLV est mise en œuvre pour la recherche de "groupes directionnels" avec extraction de la première composante principale comme méta-variable pour chaque groupe (ce qui diffère des approches décrites précédemment). Très récemment, Bühlmann *et al.* (2013) ont également proposé un algorithme de classification de variables qui est, là aussi, une première étape avant une phase de modélisation. Leur méthode de classification est fondée sur l'analyse canonique des corrélations entre des paires de groupes de variables dans une procédure hiérarchique. Cela aboutit aussi à l'extraction de groupes de variables directionnels, mais étonnamment, Bühlmann *et al.* (2013) calculent la variable moyenne dans chaque groupe comme méta-variable de groupe.

Pour le problème du choix du nombre de groupes à retenir, nous proposons deux stratégies (appelées par commodité CLVR et CLVlar) qui consiste, non pas à choisir une partition des variables en un nombre fini de groupes, mais à identifier, pas à pas, les groupes potentiellement les plus intéressants vis-à-vis de la variable à prédire. Ainsi les phases de classification et de sélection sont intimement liées dès de départ. Les deux stratégies peuvent être considérées comme un compromis entre l'approche de Hastie *et al.* (2001b) qui considèrent tous les nœuds d'un arbre et l'approche de Park *et al.* (2007) qui considèrent tous les niveaux de la hiérarchie. Plus précisément, nous considérons seulement les nœuds qui sont liés à la variable de réponse et les niveaux qui représentent des groupes homogènes. Une séquence de groupes de variables, et

donc une séquence de variables latentes CLV est identifiée. Concernant la phase de modélisation proprement dite, ces méta-variables sont intégrées, pas à pas, dans un modèle de régression des moindres carrés usuelle (ou régression logistique si la réponse est binaire).

4.2 CLV supervisée

La démarche envisagée dans ce travail, dit CLV supervisée, consiste à chercher, séquentiellement, des groupes de variables homogènes qui soient également liées à la variable de réponse. Cette recherche est réalisée, de manière itérative, dans un arbre de classification hiérarchique.

Deux stratégies sont envisagées. Dans la première stratégie, dite CLVR, la classifications implique les prédicteurs $\mathbf{X}$ et la réponse $\mathbf{y}$. Les groupes de prédicteurs identifiés le long de l'arbre hiérarchique et qui sont pertinents vis-à-vis de la réponse $\mathbf{y}$ sont repérés. Un de ces groupes est extrait sur la base d'un critère d'homogénéité. Puis, de nouveau, une classification hiérarchique ascendante est construite sur la base des prédicteurs $\mathbf{X}$, sauf ceux qui ont été extraits précédemment, et la réponse $\mathbf{y}$. La classification et l'extraction sont réalisées itérativement jusqu'à ce que tous les prédicteurs soient sélectionnés, groupe par groupe. Nous obtenons à la fin une séquence de groupes, ordonnée selon leur importance par rapport à la réponse $\mathbf{y}$.

Avec la seconde stratégie, appelée CLVlar, on cherche une séquence de groupes en utilisant la procédure LARS (§ 2.3). A la différence de la première stratégie, la classification hiérarchique des variables n'est réalisée qu'une seule fois et ne porte que sur les variables $\mathbf{X}$. Le lien entre les groupes de prédicteurs candidats et la réponse est estimé en considérant à chaque étape le résidu "LARS" de $\mathbf{y}$ calculé par rapport aux variables latentes des groupes déjà extraits à l'étape précédente. La séquence obtenue comporte ainsi des groupes de prédicteurs moins corrélés entre eux.

4.2.1 Stratégie CLVR

Considérons la matrice concaténée $\mathbf{U}_{n\times(p+1)} = [\mathbf{x}_1|\ldots|\mathbf{x}_p|\mathbf{y}]$. Comme les prédicteurs $\mathbf{x}_j$, $(j = 1,\ldots,p)$ et la réponse $\mathbf{y}$ n'ont pas nécessairement les mêmes unités, un coefficient d'échelle est appliqué de sorte à ce que $\mathbf{y}$ ait un écart-type égal à la moyenne des écarts-types de tous les prédicteurs. Si les prédicteurs sont standardisés, alors $\mathbf{y}$ aura également un écart-type de 1.

L'objectif en réalisant une classification sur la matrice $\mathbf{U}$ est de savoir quels sont les prédicteurs qui forment un groupe avec $\mathbf{y}$. Nous avons ainsi la possibilité de chercher des groupes de prédicteurs à la fois homogènes (les plus unidimensionnels possible) et prédictifs.

Le schéma général de l'algorithme est décrit comme suit :

1. Soit $\mathbf{U} = [\mathbf{x}_1|\ldots|\mathbf{x}_p|\mathbf{y}]$, $\tilde{\mathbf{U}} = \mathbf{U}$ and $\tilde{p} = p + 1$. Ici la variable de réponse peut être notée $\mathbf{y}$ ou $\mathbf{x}_{\tilde{p}}$.

2. Application de la classification CLV sur $\tilde{\mathbf{U}}$.

   L'objectif est de chercher des partitions emboîtées, constituées de groupes de variables directionnels, en utilisant l'algorithme hiérarchique (§ 2.2.2). Pour une partition en K groupes, nous rappelons que la démarche de CLV vise la maximisation du critère T suivant :

$$T = n\sum_{k=1}^{K}\sum_{j=1}^{\tilde{p}} \delta_{kj}\ cov^2(\mathbf{x}_j, \mathbf{c}_k) \text{ avec } \|\mathbf{c}_k\| = 1 \tag{4.1}$$

3. Pour chaque niveau de la hiérarchie, mise en évidence du groupe auquel $\mathbf{y}$ appartient.

   Par exemple, à un niveau q donné de la hiérarchie, une partition de l'ensemble de variables en $(\tilde{p}-q)$ groupes est obtenue. Nous pouvons noter G_q^y le groupe qui contient la variable de réponse $\mathbf{y}$ et $\mathbf{a}_q = (\ldots, a_{jq}, \ldots)^T$ avec $j = 1,\ldots,(\tilde{p}-1)$, un vecteur binaire indiquant l'appartenance, ou non, de chaque prédicteur à ce groupe : $a_{jq} = 1$ si la variable $\mathbf{x}_j$ appartient au groupe G_q^y et $a_{jq} = 0$ sinon.

En considérant tous les niveaux de la hiérarchie, nous pouvons définir une matrice binaire $\mathbf{A} = [\mathbf{a}_1|\ldots|\mathbf{a}_{\tilde{p}}]$ représentant les groupes de prédicteurs liés à la réponse $\mathbf{y}$. La matrice $\mathbf{A}_{(\tilde{p}-1)\times\tilde{p}}$ peut comporter des colonnes adjacentes identiques parce que les prédicteurs liés à $\mathbf{y}$ ne sont pas nécessairement modifiés à chaque niveau de la hiérarchie. Dans ce cas, une matrice réduite, ne retenant que la dernière colonne pour chaque bloc des colonnes identiques, est considérée. Elle est notée $\tilde{\mathbf{A}}_{(\tilde{p}-1)\times\tilde{q}}$ et code $\tilde{q}$ groupes différents de prédicteurs.

4. Sélection d'un groupe pertinent $G^y_{l^*}$, $l^* \in \{2,\ldots,\tilde{q}\}$.

   Nous considérons les $\tilde{q}$ groupes repérés à l'étape 3. La Figure 4.1 montre un exemple avec $\tilde{q} = 4$ groupes potentiellement intéressants (G^y_1, G^y_2, G^y_3 et G^y_4, de taille croissante). L'objectif est de sélectionner un des groupes qui soit le plus grand possible mais, également, le plus unidimensionnel, possible. Or, dans la hiérarchie, on s'attend à ce que plus un groupe est de taille importante, moins il y a de chance d'être homogène. Pour une partition donnée, le critère T (4.1) est un critère global d'homogénéité calculé sur l'ensemble des groupes de variables. Mais puisque nous nous intéressons spécifiquement aux groupes qui contiennent la réponse $\mathbf{y}$, alors un critère de variation locale $\Gamma_l = (T_{G(l-1)} - T_{G(l)})/T_{G(l)}$ a été privilégié. L'indice l ici représente un niveau de la hiérarchie associé à une colonne de $\tilde{\mathbf{A}}$ ($l \in \{2,...,\tilde{q}\}$). Au niveau $(l-1)$ pour lequel le groupe d'intérêt est $G^y_{(l-1)}$, nous considérons la partie du critère T concernant les prédicteurs de $G^y_{(l-1)}$, mais aussi les prédicteurs qui rejoindront le groupe G^y_l au niveau l suivant. Cette partie du critère T est notée $T_{G(l-1)}$. Elle est définie par :

$$T_{G(l-1)} = cov^2(\mathbf{y}, \mathbf{c}_{(l-1)}) + \sum_{\{j\,|a_{j(l-1)}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{(l-1)}) + \sum_{\{j\,|a_{j(l-1)}=0;\, a_{jl}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{*/j}) \tag{4.2}$$

   Le dernier terme de l'équation (4.2) concerne les variables $\mathbf{x}_j$ qui n'appartiennent pas au groupe $G^y_{(l-1)}$ au niveau d'indice $(l-1)$ mais qui rejoignent G^y_l au niveau l. $\mathbf{c}_{*/j}$ représente la variable latente du groupe auquel chacune de ces variables appartenait avant de rejoindre G^y_l.

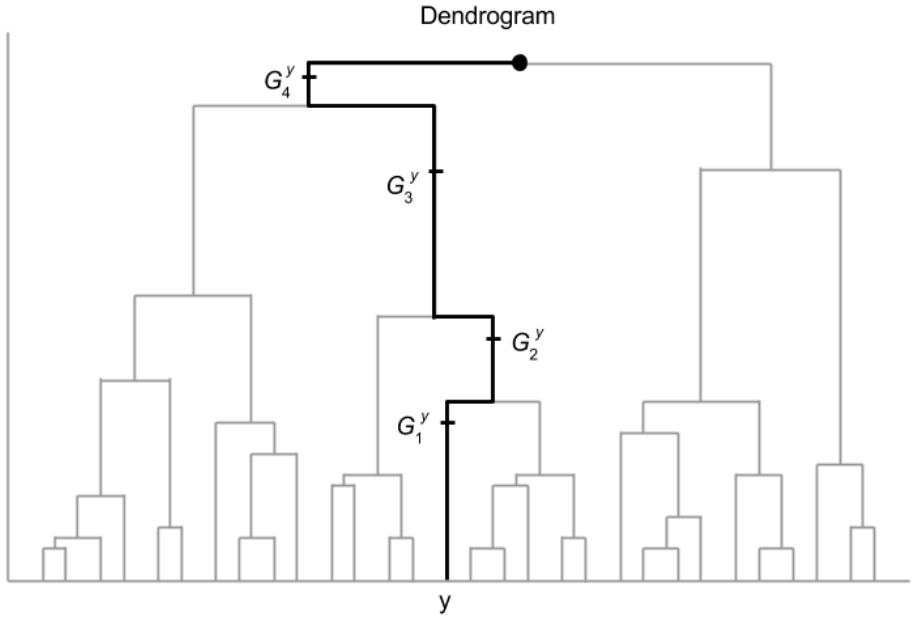

FIGURE *4.1 – Identification des groupes intéressants (dans cet exemple, G_1^y, G_2^y, G_3^y et G_4^y) basée sur l'arbre hiérarchique obtenu par la méthode CLV.*

Au niveau suivant, l, où le groupe intéressant est G_l^y, nous considérons la partie du critère T, notée $T_{G(l)}$, concernant les prédicteurs de G_l^y, mais aussi les prédicteurs qui étaient dans $G_{(l-1)}^y$ et qui ne se retrouvent plus dans G_l^y. Cela peut arriver si une procédure de consolidation est utilisée en combinaison de la classification hiérarchique. À ce niveau, $T_{G(l)}$ est donné par :

$$\begin{aligned} T_{G(l)} = cov^2(\mathbf{y}, \mathbf{c}_l) + & \sum_{\{j\,|a_{jl}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_l) \\ + & \sum_{\{j\,|a_{jl}=0;\, a_{j(l-1)}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{(l-1)}) \end{aligned} \tag{4.3}$$

Le critère de variation locale Γ_l représente, en pourcentage, la perte de l'homogénéité entre deux niveaux intéressants successifs. Il dépend aussi, de manière indirecte, de la taille des groupes des prédicteurs concernés. Un niveau pertinent

l^* est défini comme suit :

$$l^* = \max(l \,|\, \Gamma_l < \gamma) \tag{4.4}$$

Le seuil γ est un paramètre de réglage important de la stratégie CLV supervisée. Un seuil de 5% - 10% donne des résultats satisfaisants dans les exemples que nous avons testés jusqu'à présent.

5. Mise à jour de la matrice $\tilde{\mathbf{U}}$ en éliminant les colonnes associées aux prédicteurs $\mathbf{x}_j$ associés au groupe $G_{l^*}^y$, c'est-à-dire ceux pour lesquels $a_{jl^*} = 1$. Mise à jour $\tilde{p} = \tilde{p} - \sum_j a_{jl^*}$.
6. Répétition des étapes 2 à 5 jusqu'à ce que $\tilde{\mathbf{U}} = \mathbf{y}$

4.2.2 Stratégie CLVlar

La motivation en développant cette nouvelle stratégie a été de limiter le temps de calcul. En effet, la stratégie CLVR suppose, qu'à chaque étape, un nouveau dendrogramme soit re-créé par classification ascendante hiérarchique. Lorsque le nombre de variables est de l'ordre du millier, ceci induit des temps de calcul de plusieurs heures.

Considérons la matrice de prédicteurs $\mathbf{X}_{n\times p} = [\mathbf{x}_1|\ldots|\mathbf{x}_p]$. L'objectif est toujours de trouver des groupes de variables homogènes et de bonne qualité prédictive. Mais à la différence de la stratégie précédente, ici la réponse $\mathbf{y}$ n'est pas intégrée dans la classification. Cependant on s'intéresse, au cours des itérations, à la variable de $\mathbf{X}$ la plus corrélée à la variable de réponse $\mathbf{y}$ ou à son résidu obtenu en utilisant la procédure LARS.

Le schéma général de l'algorithme est décrit comme suit :

1. Classification CLV sur $\mathbf{X}$. On obtient une hiérarchie de p partitions. Au départ, on définit $\mathbf{r} = \mathbf{y}$. On pose $\tilde{p} = p$.
2. On cherche le prédicteur $\mathbf{x}_{max}$ le plus corrélé avec $\mathbf{r}$ ($\mathbf{x}_{max} = \arg\max_{j=1,\ldots,\tilde{p}} |\mathbf{x}_j^T \mathbf{r}|$). Cette variable est utilisée pour représenter la réponse $\mathbf{y}$.
3. Pour chaque niveau, q, de la hiérarchie, mise en évidence du groupe auquel appartient $\mathbf{x}_{max}$. On note G_q^{max} avec $q \in \{2,\ldots,\tilde{p}\}$ chacun de ces groupes.

On définit une matrice binaire $\mathbf{A}_{p\times\tilde{p}} = [\mathbf{a}_1|\ldots|\mathbf{a}_{\tilde{p}}]$ dont le terme générique $a_{jq} = 1$ si la variable $\mathbf{x}_j$ appartient au groupe G_q^{max} et $a_{jq} = 0$ sinon. Comme dans la stratégie précédente une nouvelle matrice $\tilde{\mathbf{A}}$, de dimensions plus réduites, $p \times \tilde{q}$, est définie en supprimant toutes les colonnes identiques de $\mathbf{A}$.

4. Pour choisir un des groupes G_l^{max} associé à un vecteur binaire $\tilde{\mathbf{a}}_l$, avec $l \in \{2,\ldots,\tilde{q}\}$, un critère de variation locale $\Gamma_l = (T_{G(l-1)} - T_{G(l)})/T_{G(l)}$, similaire à celui de la stratégie précédente, est considéré. Cependant, dans cette stratégie, la réponse $\mathbf{y}$ n'entrant pas en compte dans le calcul de T, les quantités $T_{G(l-1)}$ et $T_{G(l)}$ sont re-définies par :

$$T_{G(l-1)} = \sum_{\{j\,|a_{j(l-1)}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{(l-1)}) + \sum_{\{j\,|a_{j(l-1)}=0;\, a_{jl}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{*/j}) \qquad (4.5)$$

$$T_{G(l)} = \sum_{\{j\,|a_{jl}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_l) + \sum_{\{j\,|a_{jl}=0;\, a_{j(l-1)}=1\}} cov^2(\mathbf{x}_j, \mathbf{c}_{(l-1)}) \qquad (4.6)$$

5. Identification du groupe $G_{l^*}^{max}$ telle que $l^* = \max(l\,|\,\Gamma_l < \gamma)$.
   Ce groupe de prédicteurs remplit la condition de pertinence, c'est-à-dire, qu'il est aussi grand que possible et le plus homogène possible. Cependant, la variable latente, $\mathbf{c}_{l^*}^{max}$, de ce groupe n'est pas nécessairement la variable la plus corrélée avec $\mathbf{r}$.
6. On répète donc les étapes 2 à 5 , en remplaçant les variables du groupe $G_{l^*}^{max}$ par leur variable latente, $\mathbf{c}_{l^*}^{max}$, et en mettant à jour $\tilde{p}$ (le nombre de colonnes de la nouvelle matrice $\mathbf{X}$) jusqu'à ce que $\mathbf{x}_{max}$ soit une variable latente de groupe. Notons qu'il est possible que $\mathbf{x}_{max}$ soit une variable latente d'un groupe qui ne contient en fait qu'une seule variable.
7. Extraction de la variable $\mathbf{x}_{max}$ retenue et de son groupe associé. La matrice $\mathbf{X}_{n\times\tilde{p}}$ est mise à jour.
   La mise à jour du résidu $\mathbf{r}$ fait appel à l'algorithme LARS décrit dans § 2.3. Dans ce cas, l'ensemble actif comporte les variables latentes déjà extraites. L'ensemble non actif correspond à la nouvelle matrice $\mathbf{X}$. On calcule le vecteur équiangle

des variables actives (l'étape 3 de l'algorithme LARS), et l'estimation $\tilde{\mu}$ est mise à jour jusqu'à trouver un nouveau prédicteur $\mathbf{x}_j$ de l'ensemble non actif tel que la corrélation de $\mathbf{x}_j$ avec le résidu courant soit la même que celles entre variables actives et le résidu courant (l'étape 4 de LARS). Enfin, mise à jour de $\mathbf{r} = \mathbf{r} - \tilde{\mu}$.

8. Répétition de la procédure des étapes 2 à 7 jusqu'à ce que tous les prédicteurs aient été extraits et remplacés par une variable latente de groupe. La procédure peut également prendre fin au moment où le nombre de groupes extraits est égale au nombre d'observations n.

4.2.3 Construction du modèle

Suite à la démarche itérative d'extraction de groupes de variables, par l'une ou l'autre des stratégies, une séquence de groupes de prédicteurs et de variables latentes associées est obtenue.

Concernant la stratégie CLVR, la première variable latente, $\mathbf{LV}_{G1}$, est définie comme étant une méta-variable qui est la plus corrélée possible avec la réponse $\mathbf{y}$. La deuxième variable latente $\mathbf{LV}_{G2}$ est choisie pour atteindre le même objectif lorsque les prédicteurs du premier groupe sont éliminés. Les dernières méta-variables de la séquence sont moins pertinentes pour expliquer $\mathbf{y}$. Elles peuvent éventuellement être associées avec des variables atypiques ou de bruit comme cela a été discuté dans le chapitre 3.

Concernant la stratégie CLVlar, la séquence de groupes est ordonnée en utilisant le principe de la procédure LARS. Chaque groupe de variables est choisi de sorte à ce que sa variable latente associée soit complémentaire des autres variables latentes extraites aux étapes précédentes pour minimiser la partie de $\mathbf{y}$ non encore expliquée.

Dans tous les cas, la variable latente, $\mathbf{LV}_{Gk}$ associée à G_k est une combinaison linéaire des prédicteurs appartenant au groupe G_k. Ainsi, $\mathbf{LV}_{Gk} = \mathbf{X}_k\boldsymbol{\alpha}_k$ où $\mathbf{X}_k$ la matrice formée des p_k variables appartenant au groupe G_k et $\boldsymbol{\alpha}_k = (\alpha_{1k}, \ldots, \alpha_{jk}, \ldots, \alpha_{p_k k})^T$ le vecteur de *loadings* associé à la première composante principale de $\mathbf{X}_k$.

Un modèle de régression classique est alors construit en introduisant pas à pas (de manière ascendante) les premières variables latentes de la séquence. En considérant

les M premières variables latentes, le modèle de prédiction peut être écrit comme suit :

$$\hat{\mathbf{y}} = \sum_{k=1}^{M} \mathbf{LV}_{Gk} \hat{\beta}_{\mathbf{LV}_{Gk}} \tag{4.7}$$

Ce qui peut également s'exprimer, en fonction des prédicteurs d'origine appartenant aux groupes sélectionnés,

$$\hat{\mathbf{y}} = \sum_{k=1}^{M} \sum_{j \in G_k} \mathbf{x}_j b_j \quad \text{avec} \quad b_j = \alpha_{jk} \hat{\beta}_{\mathbf{LV}_{Gk}} \tag{4.8}$$

Le nombre M de variables latentes à retenir peut être choisi sur la base d'une procédure de validation croisée.

4.3 Simulations

4.3.1 Exemple 1

Dans cette section, les deux stratégies proposées, CLVR et CLVlar, sont illustrées à l'aide du jeu de données simulées pour lequel le nombre de prédicteurs, p, est plus grand que le nombre d'observations n. Ils ont été fixés à $n = 50$ et $p = 80$. De plus, les prédicteurs sont supposées être structurés en groupes. Nous considérons cinq groupes de variables $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_4$ et $\mathcal{G}_5$ avec les cardinalités suivantes : $|\mathcal{G}_1| = 20, |\mathcal{G}_2| = 20, |\mathcal{G}_3| = 10, |\mathcal{G}_4| = 10$ et $|\mathcal{G}_5| = 20$. Ils ont été créés sur la base de cinq variables latentes $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}_3$, $\mathbf{Z}_4$, et $\mathbf{Z}_5$ comme suit :

$$\begin{aligned}
\mathbf{x}_j &= \omega_j \mathbf{Z}_1 + \boldsymbol{\varepsilon}_j, \quad \text{pour} \quad j = 1, \ldots, 20 \\
\mathbf{x}_j &= \omega_j \mathbf{Z}_2 + \boldsymbol{\varepsilon}_j, \quad \text{pour} \quad j = 21, \ldots, 40 \\
\mathbf{x}_j &= \omega_j \mathbf{Z}_3 + \boldsymbol{\varepsilon}_j, \quad \text{pour} \quad j = 41, \ldots, 50 \\
\mathbf{x}_j &= \omega_j \mathbf{Z}_4 + \boldsymbol{\varepsilon}_j, \quad \text{pour} \quad j = 51, \ldots, 60 \\
\mathbf{x}_j &= \omega_j \mathbf{Z}_5 + \boldsymbol{\varepsilon}_j, \quad \text{pour} \quad j = 61, \ldots, 80
\end{aligned} \tag{4.9}$$

où $\omega_j \in \{+1, -1\}$. ω est utilisé pour produire au hasard un coefficient de corrélation positif ou négatif entre chaque variable simulée dans un groupe $\mathcal{G}_k$ et la variable associée $\mathbf{Z}_k$, $(k = 1, \ldots, 5)$. La variable aléatoire $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, 0.4I)$. Les cinq variables latentes sont générées suivant une loi Normale multivariée avec un vecteur de moyenne $\mu = (0, 0, 0, 0, 0)$ et une matrice de covariance Σ :

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0.3 & 0 \\ 0 & 0.5 & 1 & 0.2 & 0 \\ 0 & 0.3 & 0.2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.10}$$

Enfin, une variable de réponse $\mathbf{y}$ est simulée avec le modèle suivant :

$$\mathbf{y} = 6\,\mathbf{Z}_1 + 1.5\,\mathbf{Z}_2 + 2\,\mathbf{Z}_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I) \tag{4.11}$$

La réponse $\mathbf{y}$ est uniquement fonction des variables latentes des groupes $\mathcal{G}_1$, $\mathcal{G}_2$, et $\mathcal{G}_3$. Compte-tenu des valeurs des coefficients du modèle dans (4.11), l'ordre d'importance de ces groupes devrait être $\mathcal{G}_1 > \mathcal{G}_3 > \mathcal{G}_2$. En tenant compte de tous les paramètres du modèle de simulation, on peut montrer que les coefficients de corrélation théoriques entre $\mathbf{y}$ et les cinq variables de support sont, respectivement, de 0.882 avec $\mathbf{Z}_1$, 0.367 avec $\mathbf{Z}_2$, 0.404 avec $\mathbf{Z}_3$, 0.125 avec $\mathbf{Z}_4$ et 0 avec $\mathbf{Z}_5$.

100 jeux de données ont été générées en utilisant ce modèle de simulation. Pour chacun d'entre eux, la méthode CLV supervisée a été réalisée en mettant en oeuvre soit la stratégie CLVR, soit la statégie CLVlar. Différentes valeurs du seuil γ ont été considérées pour le critère de variation locale.

Sur l'ensemble des 100 jeux de données, les fréquences d'extraction des différentes variables, dans l'ordre d'extraction, sont résumées dans le Tableau 4.1. On observe que les deux stratégies sont capables d'identifier très correctement les cinq groupes de prédicteurs avec un seuil $\gamma = 10\%$, avec cependant de petites différences. Avec la stratégie CLVR, et un seuil $\gamma = 10\%$, pour quatre jeux de données simulées seulement, le bloc de variables générées autour de $\mathbf{Z}_2$ a été extrait avant le bloc de variables

générées autour de $\mathbf{Z}_3$. Cela peut s'expliquer par le fait que les variables latentes $\mathbf{Z}_2$ et $\mathbf{Z}_3$ ont un coefficient de corrélation moyen (de 0.5) et leur coefficients du modèle sont assez similaires (1.5 et 2). Avec la stratégie CLVlar, et un seuil $\gamma = 10\%$, la fréquence de ce changement d'ordre est un peu plus forte (18 fois sur 100). De plus, toujours pour un seuil $\gamma = 10\%$, les deux groupes générés autour $\mathbf{Z}_4$ et $\mathbf{Z}_5$ ont été extraits avec quasiment la même fréquence en quatrième ou cinquième position (56 et 44 fois sur 100) avec la stratégie CLVlar alors que l'on a quasiment toujours observé que le groupe $\mathcal{G}_5$ était extrait après $\mathcal{G}_4$ avec la stratégie CLVR. Pour un seuil $\gamma = 5\%$, des résultats similaires ont été observés.

Lorsqu'une grande valeur de seuil, par exemple $\gamma = 30\%$, est utilisée, les deux stratégies ont tendance à produire un grand groupe de variables qui agrége les groupes $\mathcal{G}_2$, $\mathcal{G}_3$ et $\mathcal{G}_4$. Dans ce cas, des erreurs de classification commencent à apparaître. Par exemple, avec la stratégie CLVR, des variables générées autour de $\mathbf{Z}_5$ sont affectées dans $\mathcal{G}_1$ dans certains cas. Avec la stratégie CLVlar, le premier et le troisième groupe extraits contiennent des variables provenant des autres groupes. Il convient de noter que dans la stratégie CLVlar, une variable pourrait être extrait dans plusieurs groupes, parce que l'extraction est toujours réalisée sur le même arbre hiérarchique.

Pour un seuil $\gamma = 40\%$ (résultats non rapportés), toutes les variables ont été agrégées dans un seul groupe. La procédure CLV supervisée s'arrête donc après une seule étape. A l'inverse, avec des seuils plus faibles que 5% (moins de 1% par exemple), la procédure prendra plus que cinq étapes et des vrais groupes de variables seront divisés en plusieurs blocs. Il est donc conseillé d'utiliser un seuil entre 5% et 10%.

Le Tableau 4.2 résume les résultats de modélisation suite à CLVR (en prenant $\gamma = 10\%$). Les résultats obtenus avec CLVlar ne sont pas présentés ici, car ils sont très similaires. En effet avec le seuil γ de 10%, les deux stratégies de classification supervisée conduisent à identifier presque exactement les mêmes groupes de prédicteurs, dans le même ordre. Les erreurs quadratiques moyennes de prédiction (MSEP) ont été évaluées pour chaque jeu de données simulées, en fonction du nombre de variables latentes introduites, par validation croisée. Le minimum observé des MSEP correspond à un modèle basé sur les trois premières variables latentes de groupes. Pour ce modèle, la valeur moyenne des estimations des coefficients (Tableau 4.2) est

TABLE *4.1 – Fréquence des affectations de chacune des variables dans les groupes extraits, en utilisant la stratégie CLVR (gauche) et la stratégie CLVlar (droite), sur les données simulées introduits dans l'exemple 1. Deux valeurs du seuil γ sont considérées : 10% et 30%.*

	CLVR, $\gamma = 10\%$					CLVR, $\gamma = 30\%$			CLVlar, $\gamma = 10\%$						CLVlar, $\gamma = 30\%$		
	1^{er}	2^e	3^e	4^e	5^e	1^{er}	2^e	3^e		1^{er}	2^e	3^e	4^e	5^e	1^{er}	2^e	3^e
x1	100	0	0	0	0	100	0	0	x1	100	0	0	0	0	100	0	0
x2	100	0	0	0	0	100	0	0	x2	100	0	0	0	0	100	1	0
x3	100	0	0	0	0	100	0	0	x3	100	0	0	0	0	100	0	0
x4	100	0	0	0	0	100	0	0	x4	100	0	0	0	0	100	1	0
x5	100	0	0	0	0	100	0	0	x5	100	0	0	0	0	100	0	0
x6	100	0	0	0	0	100	0	0	x6	100	0	0	0	0	100	0	0
x7	100	0	0	0	0	100	0	0	x7	100	0	0	0	1	100	1	2
x8	100	0	0	0	0	100	0	0	x8	100	0	0	0	0	100	0	0
x9	100	0	0	0	0	100	0	0	x9	100	0	0	0	0	100	0	0
x10	100	0	0	0	0	100	0	0	x10	100	0	0	0	0	100	0	0
x11	100	0	0	0	0	100	0	0	x11	100	0	0	0	0	100	1	0
x12	100	0	0	0	0	100	0	0	x12	100	0	0	0	0	100	1	0
x13	100	0	0	0	0	100	0	0	x13	100	0	0	0	0	100	1	0
x14	100	0	0	0	0	100	0	0	x14	100	0	0	0	0	100	0	0
x15	100	0	0	0	0	100	0	0	x15	100	0	0	0	0	100	0	1
x16	100	0	0	0	0	100	0	0	x16	100	0	0	0	0	100	1	0
x17	100	0	0	0	0	100	0	0	x17	100	0	0	0	0	100	1	0
x18	100	0	0	0	0	100	0	0	x18	100	0	0	0	0	100	0	1
x19	100	0	0	0	0	100	0	0	x19	100	0	0	0	0	100	0	0
x20	100	0	0	0	0	100	0	0	x20	100	0	0	0	0	100	1	1
x21	0	4	96	0	0	0	100	0	x21	0	18	82	0	0	0	100	0
x22	0	4	96	0	0	0	100	0	x22	0	18	82	0	0	0	100	0
x23	0	4	96	0	0	0	100	0	x23	0	18	82	0	0	0	100	0
x24	0	4	96	0	0	0	100	0	x24	0	18	82	0	0	0	100	0
x25	0	4	96	0	0	0	100	0	x25	0	18	82	0	0	0	100	0
x26	0	4	96	0	0	0	100	0	x26	0	18	82	0	0	0	100	0
x27	0	4	96	0	0	0	100	0	x27	0	18	82	0	0	0	100	0
x28	0	4	96	0	0	0	100	0	x28	0	18	82	0	0	0	100	0
x29	0	4	96	0	0	0	100	0	x29	0	18	82	0	0	0	100	0
x30	0	4	96	0	0	0	100	0	x30	0	18	82	0	0	0	100	0
x31	0	4	96	0	0	0	100	0	x31	0	18	82	0	0	0	100	0
x32	0	4	96	0	0	0	100	0	x32	0	18	82	0	0	0	100	0
x33	0	4	96	0	0	0	100	0	x33	0	18	82	0	0	0	100	0
x34	0	4	96	0	0	0	100	0	x34	0	18	82	0	0	0	100	0
x35	0	4	96	0	0	0	100	0	x35	0	18	82	0	0	0	100	0
x36	0	4	96	0	0	0	100	0	x36	0	18	82	0	0	0	100	0
x37	0	4	96	0	0	0	100	0	x37	0	18	82	0	0	0	100	0
x38	0	4	96	0	0	0	100	0	x38	0	18	82	0	0	0	100	0
x39	0	4	96	0	0	0	100	0	x39	0	18	82	0	0	0	100	0
x40	0	4	96	0	0	0	100	0	x40	0	18	82	0	0	0	100	0
x41	0	96	4	0	0	0	100	0	x41	0	82	18	0	0	0	100	0
x42	0	96	4	0	0	0	100	0	x42	0	82	18	0	0	0	100	0
x43	0	96	4	0	0	0	100	0	x43	0	82	18	0	0	0	100	0
x44	0	96	4	0	0	0	100	0	x44	0	82	18	0	0	0	100	0
x45	0	96	4	0	0	0	100	0	x45	0	82	18	0	0	0	100	0
x46	0	96	4	0	0	0	100	0	x46	0	82	18	0	0	0	100	0
x47	0	96	4	0	0	0	100	0	x47	0	82	18	0	0	0	100	0
x48	0	96	4	0	0	0	100	0	x48	0	82	18	0	0	0	100	0
x49	0	96	4	0	0	0	100	0	x49	0	82	18	0	0	0	100	0
x50	0	96	4	0	0	0	100	0	x50	0	82	18	0	0	0	100	0
x51	0	0	0	100	0	0	100	0	x51	0	0	0	56	44	0	100	0
x52	0	0	0	100	0	0	100	0	x52	0	0	0	56	44	0	100	0
x53	0	0	0	100	0	0	100	0	x53	0	0	0	56	44	0	100	0
x54	0	0	0	100	0	0	100	0	x54	0	0	0	56	44	0	100	0
x55	0	0	0	100	0	0	100	0	x55	0	0	0	56	44	0	100	0
x56	0	0	0	100	0	0	100	0	x56	0	0	0	56	44	0	100	0
x57	0	0	0	100	0	0	100	0	x57	0	0	0	56	44	0	100	0
x58	0	0	0	100	0	0	100	0	x58	0	0	0	56	44	0	100	0
x59	0	0	0	100	0	0	100	0	x59	0	0	0	56	44	0	100	0
x60	0	0	0	100	0	0	100	0	x60	0	0	0	56	44	0	100	0
x61	1	0	0	0	99	7	0	93	x61	1	0	0	44	56	2	1	100
x62	1	0	0	0	99	8	0	92	x62	0	0	0	44	56	2	1	100
x63	0	0	0	0	100	3	0	97	x63	0	0	0	44	56	0	2	100
x64	0	0	0	0	100	5	0	95	x64	0	0	0	44	56	1	0	100
x65	0	0	0	0	100	1	0	99	x65	0	0	0	44	56	0	1	100
x66	0	0	0	0	100	5	0	95	x66	0	0	0	44	56	1	0	100
x67	0	0	0	0	100	3	0	97	x67	0	0	0	44	56	1	0	100
x68	0	0	0	0	100	7	0	93	x68	0	0	0	44	56	2	1	100
x69	0	0	0	0	100	5	0	95	x69	0	0	0	44	56	2	1	100
x70	0	0	0	0	100	2	0	98	x70	0	0	0	44	56	0	0	100
x71	0	0	0	0	100	4	0	96	x71	0	0	0	44	56	1	1	100
x72	0	0	0	0	100	3	0	97	x72	0	0	0	44	56	0	0	100
x73	0	0	0	0	100	5	0	95	x73	0	0	0	44	56	0	1	100
x74	0	0	0	0	100	6	0	94	x74	0	0	0	44	56	1	1	100
x75	0	0	0	0	100	3	0	97	x75	0	0	0	44	56	0	2	100
x76	0	0	0	0	100	5	0	95	x76	0	0	0	44	56	2	1	100
x77	0	0	0	0	100	1	0	99	x77	0	0	0	44	56	0	0	100
x78	0	0	0	0	100	5	0	95	x78	0	0	0	44	56	0	1	100
x79	0	0	0	0	100	3	0	97	x79	0	0	0	44	56	0	2	100
x80	0	0	0	0	100	5	0	95	x80	0	0	0	44	56	1	0	100

égale, ou quasiment égale, à la valeur théorique (équation (4.11)).

TABLE *4.2 – Coefficients de régression estimés en moyenne (les valeurs minimum et maximum données entre parenthèses) et l'erreur quadratique moyenne de prédiction (MSEP) évalués sur 100 jeux de données, avec un nombre de variables latentes utilisées de 1 à 5.*

nb	Coefficients					MSEP
LV	LV_{G1}	LV_{G2}	LV_{G3}	LV_{G4}	LV_{G5}	
1	6.00 (5.68-6.31)					11.28
2	5.99 (5.59-6.28)		2.74 (2.41-3.14)			3.69
3	6.00 (5.63-6.27)	1.48 (1.09-1.91)	2.00 (1.63-2.30)			1.57
4	6.00 (5.63-6.28)	1.47 (1.08-1.96)	2.01 (1.60-2.29)	0.16 (0-0.49)		1.58
5	6.00 (5.63-6.28)	1.47 (1.08-1.95)	2.01 (1.60-2.30)	0.16 (0-0.48)	0.14 (0-0.70)	1.61

4.3.2 Exemple 2

Dans le modèle de simulation considéré ici, l'équation de régression est le suivant :

$$\mathbf{y} = 6\,\mathbf{Z}_1 + 4\,\mathbf{Z}_3 + 2\,\mathbf{Z}_5 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I) \tag{4.12}$$

Les variables de support ont la même structure de corrélation que précédemment et les variables individuelles sont générées de la même manière (§ 4.3.1).

La différence avec le cas précédent est que $\mathbf{y}$ est liée aux groupes $\mathcal{G}_1$, $\mathcal{G}_3$ et $\mathcal{G}_5$ qui sont indépendants entre eux. Cependant $\mathbf{y}$ ne dépend pas de la variable latente de $\mathcal{G}_2$ qui a pourtant un coefficient de corrélation de 0.5 avec la variable latente de $\mathcal{G}_3$. Comme dans le premier exemple, 100 jeux de données sont générés.

Les résultats des approches CLVR et CLVlar mises en oeuvre avec un paramètre de seuil γ de 10% figurent dans le Tableau 4.3. Nous pouvons observer que les cinq groupes de variables sont parfaitement identifiés quelle que soit la stratégie utilisée. Cependant, l'ordre d'extraction des groupes est différent. Pour la stratégie CLVR, $\mathcal{G}_2$ et $\mathcal{G}_5$ ont été extraits en rang 3 ou 4 avec approximativement la même fréquence. Or les variables du groupe $\mathcal{G}_2$ ne sont pas pertinentes pour la réponse, mais elles ont été sélectionnées parce qu'elles sont assez corrélées avec les variables du groupe $\mathcal{G}_3$ qui est un groupe pertinent dans cet exemple.

En utilisant le résidu de LARS, on observe bien que les variables du $\mathcal{G}_2$ ne sont

extraites qu'en rangs 4 ou 5, comme les variables du groupe non informatif $\mathcal{G}_4$. En effet, une fois que le groupe $\mathcal{G}_3$ est extrait, toutes les variables des groupes corrélés avec $\mathcal{G}_3$ deviendront moins importantes, et ainsi ne seront pas extraites juste après $\mathcal{G}_3$.

Dans les données de grande dimension où il existe éventuellement de nombreux groupes ($K > n$), la stratégie CLVlar permet de sélectionner prioritairement certains groupes avec une combinaison linéaire des variables latentes associées plus proche de $\mathbf{y}$.

4.4 Données réelles

Nous allons considéré dans cette partie trois applications de données réelles prises dans des domaines différents mais pour lesquelles un nombre assez important de variables (entre 100 et 600) a été collecté dans une démarche holistique.

4.4.1 Prédiction de la saveur sucrée de petits pois à partir de spectres Proche-Infrarouges.

Nous considérons un jeu de données contenant 60 échantillons de petits pois analysés avec un spectromètre proche infrarouge (NIR) (Naes et Kowalski, 1989). On dispose des mesures pour 116 longueurs d'ondes (variables spectrales) prises à intervalles réguliers entre 1100 et 2500 nm. Une analyse préliminaire des données révèle que les spectres diffèrent principalement selon leur intensité globale. Afin de réduire cet effet, qui est généralement due à la variation des conditions instrumentales, une correction a été réalisée selon la méthode *Standard Normal Variate* (SNV) (Barnes *et al.*, 1989). Les données spectrales d'origine et corrigées par SNV sont présentées dans la Figure 4.2. Par la suite, on considère les données spectrales corrigées dont les variables seront centrées et réduites.

En plus des mesures Proche-Infrarouge, les échantillons ont été caractérisés par des attributs sensoriels tel que le caractère sucré des petit pois. Dans la mesure où les bandes d'absorption NIR sont principalement liés à l'atome d'hydrogène, dans diverses

TABLE *4.3 – Fréquence des affectations de chacune des variables dans les groupes extraits, en utilisant la stratégie CLVR et la stratégie CLVlar, sur les données simulées introduits dans l'exemple 2.*

CLVR, $\gamma = 10\%$						CLVlar, $\gamma = 10\%$					
	1^{er}	2^{e}	3^{e}	4^{e}	5^{e}		1^{er}	2^{e}	3^{e}	4^{e}	5^{e}
x1	100	0	0	0	0	x1	100	0	0	0	0
x2	100	0	0	0	0	x2	100	0	0	0	0
x3	100	0	0	0	0	x3	100	0	0	0	0
x4	100	0	0	0	0	x4	100	0	0	0	0
x5	100	0	0	0	0	x5	100	0	0	0	0
x6	100	0	0	0	0	x6	100	0	0	0	0
x7	100	0	0	0	0	x7	100	0	1	0	0
x8	100	0	0	0	0	x8	100	0	0	0	0
x9	100	0	0	0	0	x9	100	0	0	0	0
x10	100	0	0	0	0	x10	100	0	0	0	0
x11	100	0	0	0	0	x11	100	0	0	0	0
x12	100	0	0	0	0	x12	100	0	0	0	0
x13	100	0	0	0	0	x13	100	0	0	0	0
x14	100	0	0	0	0	x14	100	0	0	0	0
x15	100	0	0	0	0	x15	100	0	0	0	0
x16	100	0	0	0	0	x16	100	0	0	0	0
x17	100	0	0	0	0	x17	100	0	0	0	0
x18	100	0	0	0	0	x18	100	0	0	0	0
x19	100	0	0	0	0	x19	100	0	0	0	0
x20	100	0	0	0	0	x20	100	0	0	0	0
x21	0	0	48	52	0	x21	0	0	0	51	49
x22	0	0	48	52	0	x22	0	0	0	51	49
x23	0	0	48	52	0	x23	0	0	0	51	49
x24	0	0	48	52	0	x24	0	0	0	51	49
x25	0	0	48	52	0	x25	0	0	0	51	49
x26	0	0	48	52	0	x26	0	0	0	51	49
x27	0	0	48	52	0	x27	0	0	0	51	49
x28	0	0	48	52	0	x28	0	0	0	51	49
x29	0	0	48	52	0	x29	0	0	0	51	49
x30	0	0	48	52	0	x30	0	0	0	51	49
x31	0	0	48	52	0	x31	0	0	0	51	49
x32	0	0	48	52	0	x32	0	0	0	51	49
x33	0	0	48	52	0	x33	0	0	0	51	49
x34	0	0	48	52	0	x34	0	0	0	51	49
x35	0	0	48	52	0	x35	0	0	0	51	49
x36	0	0	48	52	0	x36	0	0	0	51	49
x37	0	0	48	52	0	x37	0	0	0	51	49
x38	0	0	48	52	0	x38	0	0	0	51	49
x39	0	0	48	52	0	x39	0	0	0	51	49
x40	0	0	48	52	0	x40	0	0	0	51	49
x41	0	100	0	0	0	x41	0	100	0	0	0
x42	0	100	0	0	0	x42	0	100	0	0	0
x43	0	100	0	0	0	x43	0	100	0	0	0
x44	0	100	0	0	0	x44	0	100	0	0	0
x45	0	100	0	0	0	x45	0	100	0	0	0
x46	0	100	0	0	0	x46	0	100	0	0	0
x47	0	100	0	0	0	x47	0	100	0	0	0
x48	0	100	0	0	0	x48	0	100	0	0	0
x49	0	100	0	0	0	x49	0	100	0	0	0
x50	0	100	0	0	0	x50	0	100	0	0	0
x51	0	0	0	0	100	x51	0	0	0	49	51
x52	0	0	0	0	100	x52	0	0	0	49	51
x53	0	0	0	0	100	x53	0	0	0	49	51
x54	0	0	0	0	100	x54	0	0	0	49	51
x55	0	0	0	0	100	x55	0	0	0	49	51
x56	0	0	0	0	100	x56	0	0	0	49	51
x57	0	0	0	0	100	x57	0	0	0	49	51
x58	0	0	0	0	100	x58	0	0	0	49	51
x59	0	0	0	0	100	x59	0	0	0	49	51
x60	0	0	0	0	100	x60	0	0	0	49	51
x61	0	0	52	48	0	x61	1	0	100	0	0
x62	0	0	52	48	0	x62	0	0	100	0	0
x63	0	0	52	48	0	x63	0	0	100	0	0
x64	0	0	52	48	0	x64	0	0	100	0	0
x65	0	0	52	48	0	x65	0	0	100	0	0
x66	0	0	52	48	0	x66	0	0	100	0	0
x67	0	0	52	48	0	x67	0	0	100	0	0
x68	1	0	52	47	0	x68	0	0	100	0	0
x69	0	0	52	48	0	x69	0	0	100	0	0
x70	0	0	52	48	0	x70	0	0	100	0	0
x71	0	0	52	48	0	x71	0	0	100	0	0
x72	0	0	52	48	0	x72	0	0	100	0	0
x73	0	0	52	48	0	x73	0	0	100	0	0
x74	0	0	52	48	0	x74	0	0	100	0	0
x75	0	0	52	48	0	x75	0	0	100	0	0
x76	0	0	52	48	0	x76	0	0	100	0	0
x77	0	0	52	48	0	x77	0	0	100	0	0
x78	0	0	52	48	0	x78	0	0	100	0	0
x79	0	0	52	48	0	x79	0	0	100	0	0
x80	0	0	52	48	0	x80	0	0	100	0	0

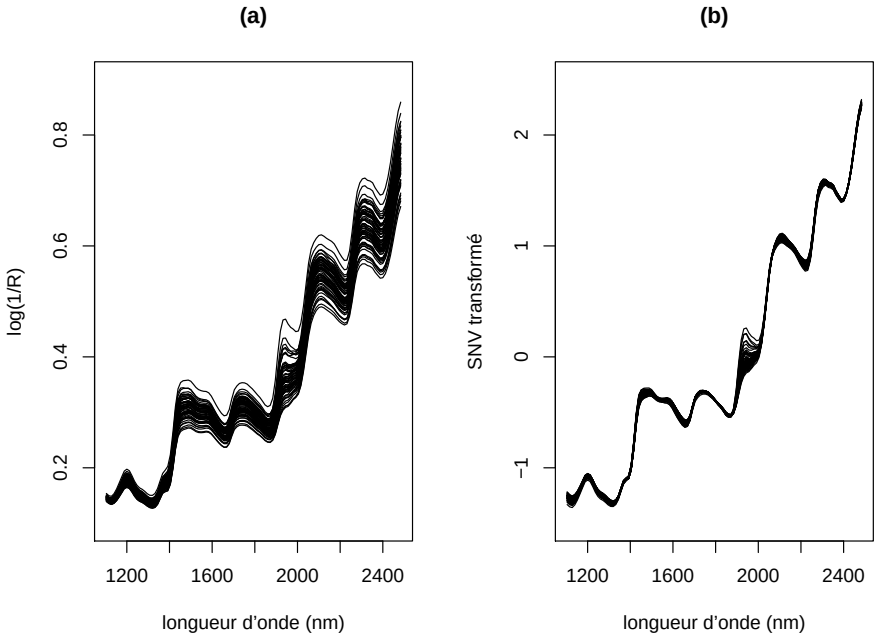

FIGURE *4.2 – (a) Données spectrales d'origine. (b) Données spectrales corrigées avec une transformation SNV.*

configurations moléculaires, les variations de la teneur en amidon des échantillons, et par voie de conséquence leur saveur sucrée devrait pouvoir être appréhendées par le biais des mesures NIR. On cherchera donc à la fois à expliquer les variations de saveur sucrée perçue et à identifier les groupes de variables spectrales associées à ces variations. Comme Vigneau *et al.* (2005) l'ont déjà observé dans un contexte similaire, on s'attend à ce que ces groupes soient organisés en une ou plusieurs bandes spectrales.

Les deux approches CLVR et CLVlar ont été mises en œuvre, en fixant la valeur du paramètre γ à 5%. Afin d'évaluer la stabilité de la procédure d'extraction des groupes et la qualité prédictive des variables latentes associées, une validation croisée, en 10 segments, a été utilisée.

La Figure 4.3 représente la fréquence de sélection des variables qui ont été extraites avec un groupe de variables de la première à la quatrième étapes avec les stratégies CLVR et CLVlar. On peut remarquer, comme attendue, que des longueurs d'onde

voisines ont été retenues, ce qui a conduit à l'identification des bandes spectrales les plus intéressantes.

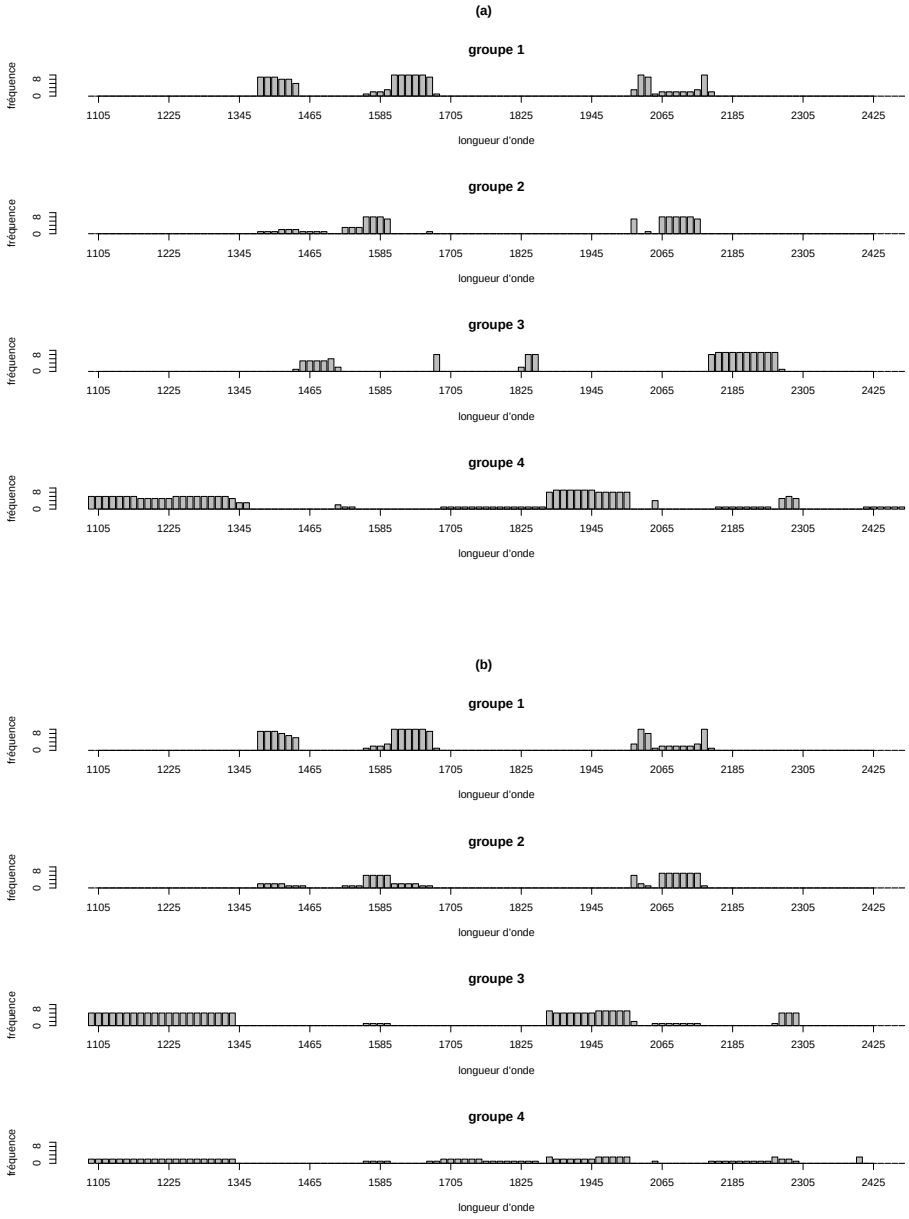

FIGURE *4.3 – Fréquence d'appartenance des longueurs d'onde aux groupes de variables extraits séquentiellement avec la stratégie CLVR (a) et CLVlar(b).*

Les longueurs d'ondes extraites en premier, qui sont a priori les plus corrélées avec

la saveur sucrée, couvrent deux régions principales (1390-1440 nm et 1620-1680 nm) ou sont situées autour de 2050 et 2150 nm. Les longueurs d'ondes du deuxième groupe correspondent à des régions assez complémentaires de celles du premier groupe. Pour un spectroscopiste les régions mises en avant sont interprétables. Les résultats des deux stratégies sont similaires à ce niveau.

Avec la stratégie CLVR, on a constaté que les longueurs d'ondes du troisième groupe sont associées à des creux plutôt qu'à des pics dans le profil spectral, ce qui est plus difficile à interpréter. Le quatrième groupe comporte principalement la région spectrale associée au pic de l'eau (autour 1900 nm) et le début de la zone spectrale de l'infra-rouge. Avec la stratégie CLVlar, ce sont ces longueurs d'onde qui sont extraites au rang 3. Avec la stratégie CLVlar, les zones identifiées au rang 4 sont peu stables d'une répétition à une autre de la procédure de validation croisée. On observe donc des fréquences assez faibles pour chaque variable repérée pour le quatrième groupe.

Les régressions réalisées sur les variables latentes des groupes identifiés par CLVR ou CLVlar sont comparées à la régression PCR (ajustée sur les composantes principales de $\mathbf{X}$), à la régression PLSR (ajustée sur les composantes PLS) et à la méthode *Sparse* PLS (SPLSR). Pour la régression *Sparse* PLS, une méthode de seuillage doux (Zou et Hastie, 2005; Chun et Keles, 2010) est utilisée. Après avoir fait varier le paramètre de *sparsité*, η, entre 0 et 1, une valeur $\eta = 0.8$ a été choisie par validation croisée.

La Figure 4.4 représente l'évolution des erreurs de prédiction (MSEP) pour ces différentes méthodes, en fonction de nombre de composantes introduites. Pour CLVR, CLVlar et SPLSR, une fois la première variable latente introduite, l'erreur de prédiction se stabilise très vite. Ces trois méthodes montrent des performances assez similaires. Au contraire, pour les méthodes PLSR et PCR, un nombre plus important de composantes est nécessaire. Le nombre optimal de composantes, pour chaque méthode, a été déterminé en utilisant une règle dite "1 standard error" sur les MSEP observés dans une démarche de validation croisée (Filzmoser *et al.*, 2009; Vigneau et Thomas, 2012). Le Tableau 4.4 récapitule les valeurs de MSEP (validation croisée avec 10 segments) ainsi que l'écart-type des valeurs en fonction des segments, pour le modèle ayant le plus petit MSEP. Ainsi, 2 composantes ont été choisies pour le

modèle CLVR, 2 pour CLVlar, 5 pour PCR, 3 pour PLSR, et 2 pour SPLSR.

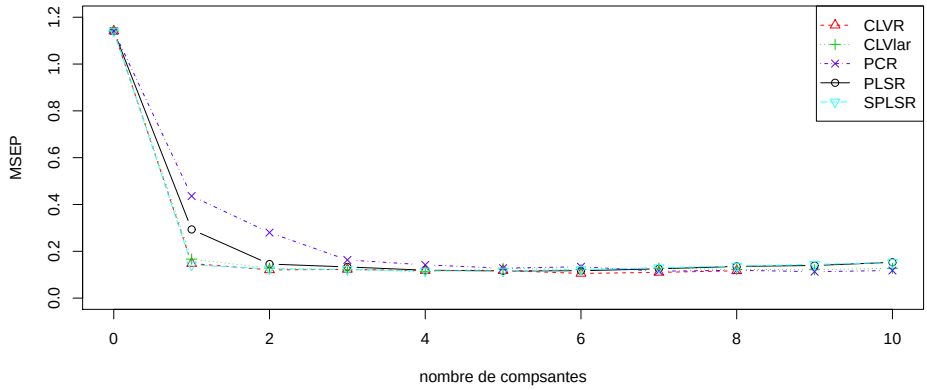

FIGURE *4.4 – Erreur quadratique moyenne de prédiction (MSEP) en fonction de nombre de composantes introduites dans les modèles.*

TABLE *4.4 – Erreur quadratique moyenne de prédiction (MSEP) pour la Régression CLV (CLVR et CLVlar), Régression sur composantes principales (PCR), Régression PLS (PLSR) et Régression Sparse PLS (SPLSR, avec* $\eta = 0.8$*), en fonction de nombre de composantes introduites dans les modèles. Pour la valeur MSEP minimale, l'écart-type évalué sur les 10 segments de la validation croisée est donné entre parenthèses. Le nombre optimal de composantes, déterminé avec la règle "1 standard error" est indiqué en gras.*

nb comp.	0	1	2	3	4	5	6	7	8	9	10
CLVR	1.141	0.148	**0.120**	0.123	0.118	0.118	0.106 (*0.074*)	0.110	0.117		
CLVlar	1.141	0.167	**0.128**	0.120	0.113 (*0.073*)	0.127	0.122	0.128	0.122	0.121	0.127
PCR	1.141	0.436	0.280	0.163	0.142	**0.128**	0.134	0.115	0.119	0.113 (*0.070*)	0.118
PLSR	1.141	0.293	0.145	**0.133**	0.119	0.116 (*0.074*)	0.117	0.125	0.135	0.139	0.153
SPLSR	1.141	0.142	**0.125**	0.123	0.116	0.116 (*0.072*)	0.122	0.130	0.137	0.144	0.152

Les coefficients de régression estimés pour les cinq modèles finalement choisis sont représentés en Figure 4.5. Pour toutes les approches, il y a deux bandes spectrales, de 1400 à 1680 nm et de 2030 à 2150 nm, qui sont repérées comme importantes pour la prédiction de la saveur sucrée. Il s'avère ici que les coefficients des modèles CLVR

et CLVlar sont les mêmes, les méta-variables des rangs 1 et 2 étant quasiment identiques. Le profil de leurs coefficients est assez comparable à celui obtenu avec *Sparse* PLS. Pour ces trois approches, des blocs de coefficients ont des valeurs nulles, ce qui correspond à une sélection de variables associée à la construction du modèle. La différence principale entre ces trois approches tient au fait qu'avec *Sparse* PLS plusieurs longueurs d'onde autour de 1800, 2270, 2420 et 2480 nm ont été sélectionnées, alors qu'elles ne sont pas incluses dans la sélection de CLVR et CLVlar.

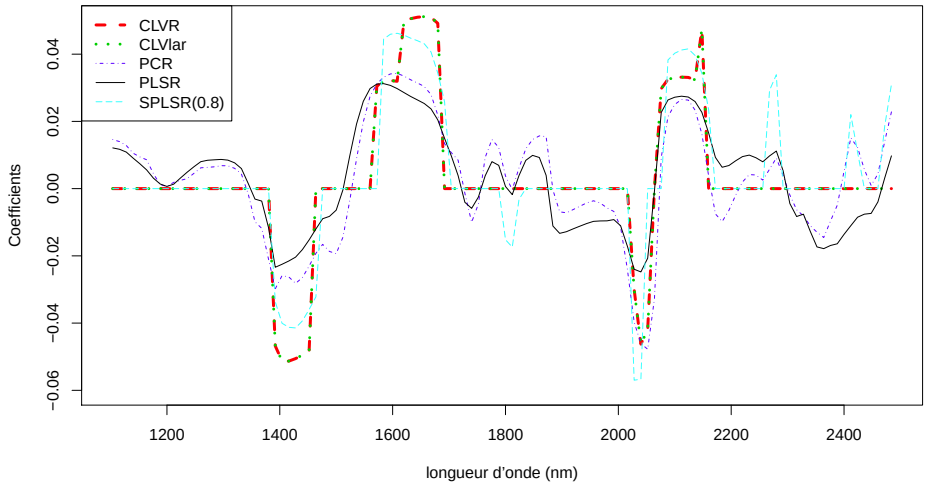

FIGURE *4.5 – Coefficients de régression pour les cinq modèles finaux : CLVR ($K = 2$), CLVlar ($K = 2$), PCR ($K = 5$), PLSR ($K = 3$) et SPLSR ($\eta = 0.8$, $K = 2$).*

Dans cet exemple, la régression *Sparse* PLS et les deux approches proposées sont plus performantes que les méthodes PLSR et PCR. Elles ont abouti à la sélection des mêmes variables, ou presque, et ont ainsi montré des performances en prédiction similaires.

4.4.2 Identification de biomarqueurs pour un dépistage précoce de la maladie d'Alzheimer

Ce jeu de données est décrit dans Craig-Schapiro *et al.* (2011). Il peut être téléchargé à partir du *package* R **AppliedPredictiveModeling** (Kuhn et Johnson, 2014).

L'objectif de l'étude citée est de déterminer si la mesure de certains marqueurs biologiques présents dans le liquide céphalo-rachidien (CSF) pourrait être utilisée pour prédire la survenue de la maladie d'Alzheimer (AD). 333 patients ont été étudiés. Parmi eux, 242 ne présentaient pas de symptôme et 91 avaient des symptômes de la maladie. Ici, la réponse considérée est donc une variable binaire, codée avec 0 pour les patients normaux et 1 pour les patients avec symptômes.

Traditionnellement, trois biomarqueurs CSF sont utilisés comme outils de diagnostic. Ce sont $A\beta 42$, *tau* et *p-tau 181*. Dans l'étude considérée, 124 composés, de natures très diverses (cytokines, facteurs de croissance, hormones, marqueurs métaboliques ...) ont été évaluées en plus. Pour un grand nombre d'entre eux, une transformation logarithmique des données en base 10 a été réalisée par les auteurs.

Ces données vont être utilisées pour tester les stratégies de classification CLV supervisée proposées et pour comparer nos résultats avec ceux présentés dans Craig-Schapiro *et al.* (2011). Les deux stratégies de classification supervisée CLVR et CLVlar ont été mises en oeuvre avec une valeur de paramètre $\gamma = 5\%$. Le Tableau 4.5 représente les deux séquences de groupes de variables extraits.

Il peut sembler surprenant que les deux approches, qui visent à identifier des groupes de variables, produisent des petits groupes contenant seulement un ou deux biomarqueurs. Ces groupes de taille réduite soulignent le fait que les biomarqueurs candidats sont hétérogènes en termes de fonctions et voies biologiques (Craig-Schapiro *et al.*, 2011). Ils ne sont ainsi pas nécessairement fortement corrélés entre eux. A titre de vérification, les coefficients de corrélation entre $A\beta 42$, PP, $CKMB$, $Cortisol$ (qui correspondent aux quatre premiers groupes de la stratégie CLVR) sont faibles. Ils varient entre -0.11 et 0.18. Pour la stratégie CLVlar, les coefficients de corrélation entre les variables des quatre premiers groupes varient entre -0.36 et 0.24.

Les trois biomarqueurs "traditionnels" ont été sélectionnés dans le premier et le cinquième groupes pour la première stratégie, et dans les deux premiers groupes pour la deuxième stratégie. De plus, plusieurs biomarqueurs du Tableau 4.5 se retrouvent dans la liste des quinze prédicteurs publiée dans Craig-Schapiro *et al.* (2011). Pour établir cette liste, les auteurs ont utilisé plusieurs méthodes telles que les forêts aléatoires, les *boosted trees*, la méthode des *Nearest Shrunken Centroids* et la régression

PLS. Les biomarqueurs que l'on retrouve en commun sont mis en évidence en gras dans le Tableau 4.5.

L'AUC ou surface sous la courbe ROC (*Receiver Operating Characteristic*) a été calculée pour évaluer la capacité de discrimination entre patients avec ou sans symptôme. La sensibilité à un niveau de spécificité égale à 80% a également été évaluée. Elle permet d'estimer la capacité à identifier les cas vrais positifs d'un modèle pour lequel le pourcentage de faux négatifs est de 20%. Comme on peut le voir dans le Tableau 4.5, les meilleures performances (AUC de 0.761 et 0.753) ont été obtenues avec les variables latentes des groupes contenant les biomarqueurs "traditionnels".

Des modèles de régression logistique ont été également ajustés en utilisant les premières variables de groupes introduites progressivement. Les résultats obtenus sont résumés dans le Tableau 4.6. Il s'avère que l'AUC, comme la sensibilité pour une spécificité de 80%, augmentent régulièrement. En combinaison, les variables latentes extraites améliorent la précision du diagnostic. Un test statistique portant sur la différence entre les déviances résiduelles de deux modèles logistiques successifs (risque $\alpha < 5\%$) permet de conclure qu'il suffit de conserver le modèle ajusté sur les six premières variables latentes pour la stratégie CLVR et sur les cinq premières variables latentes pour la deuxième stratégie CLVlar.

4.4.3 Données métabolomiques pour l'identification de fraudes éventuelles en élevage

Ce jeu de données métabolomiques est très similaire à celui de § 3.5.2. L'objectif est, là aussi, de détecter l'utilisation de promoteurs de croissance administrés à des veaux à partir d'empreintes métabolomiques urinaires. Les données contiennent 575 ions identifiées par LC-HRMS (chromatographie liquide couplée à la spectrométrie de masse haute résolution) sur 18 échantillons correspondant à 12 veaux non traités (contrôles) et 6 veaux traités (Courant *et al.*, 2009). Les échantillons urinaires ont été aussi collectés six jours avant le début de traitement puis aux jours 1, 3, 6, 13 et 17 de la période de traitement, ainsi qu'aux jours 22, 27, 29, 35, 41 et 48, le traitement étant arrêté au jour 21. Nous considérons ici les jours 17 et 22.

TABLE *4.5 – Groupes de biomarqueurs identifiés en utilisant les deux approches proposées (seuil $\gamma = 5\%$). Les mesures de performance ROC pour la discrimination entre patients normaux et patients avec symptômes ont été évalués pour chaque variable latente de groupe.*

groupes. de var.	biomarqueurs	AUC	Sensibilité (à 80% de Spécificité)
	CLVR		
G1	**Aβ42**	0.761	0.637
G2	**PP**	0.677	0.450
G3	**CKMB**	0.634	0.286
G4	**Cortisol**	0.616	0.253
G5	FABP ; **tau** ; **p-tau 181**	0.723	0.473
G6	MMP-3 ; MMP-7 ; **MMP10**	0.678	0.374
G7	I-309 ; MIP-1β ; **NT-proBNP** ; **TRAIL-R3**	0.669	0.462
G8	MIG ; IP-10 ; **MCP-2**	0.644	0.450
G9	**FAS** ; MIP-1α ; PARC ; TECK	0.659	0.374
G10	**GRO-α** ; IL-8 ; PAI-1 ; PLGF ; PAP ; RANTES ; TIMP-1 ; vWF	0.651	0.407
	CLVlar		
G1	**tau** ; **p-tau 181**	0.753	0.516
G2	**Aβ42**	0.761	0.637
G3	**GRO-α**	0.696	0.418
G4	**PP**	0.677	0.451
G5	**NT-proBNP**	0.673	0.352
G6	**MMP10**	0.717	0.484
G7	**VEGF**	0.628	0.319
G8	**TRAIL-R3**	0.685	0.440
G9	**Cystatin-C**	0.598	0.275
G10	PAI-1	0.677	0.418

Aβ42 : Log Aβ42 ; PP : Log Pancreatic Polypeptide ; CKMB : Log Creatine Kinase ; FABP : Log Fatty Acid Binding Protein ; tau : Log tau ; p-tau 181 : Log p-tau 181 ; MMP-3/-7/-1 :0 Log Matrix Metalloproteinase-3/-7/-10 ; I-309 : Log T Lymphocyte-Secreted Protein I-309 ; MIP-1β : Log Macrophage Inflammatory Protein-1 beta ; NT-proBNP : Log N-terminal pro-brain natriuretic peptide ; TRAIL-R :3 TNF-Related Apoptosis-Inducing Ligand Receptor 3 ;MIG : Log Gamma-Interferon-Induced Monokine ; IP-10 : Log Interferon gamma Induced Protein 10 ; MCP-2 : Monocyte Chemotactic Protein 2 ; FAS : Log FAS ; MIP-1α : Macrophage Inflammatory Protein-1$alpha$; PARC : Log Pulmonary and Activation-Regulated Chemokine ; TECK : Thymus-Expressed Chemokine ; GRO-α : Growth-Regulated alpha protein ; IL-8 : Log Interleukin-8 ; PAI-A : Log Plasminogen Activator Inhibitor 1 ; PLGF : Log Placenta Growth Factor ; PAP : Prostatic Acid Phosphatase ; RANTES : Log Regulated on Activation, Normal T Expressed and Secreted ; TIMP-1 : Tissue Inhibitor of Metalloproteinases 1 ; vWF : Log von Willebrand Factor ; VEGF : Vascular Endothelial Growth Factor ; Cystatin-C : Log Cystatin C

TABLE *4.6 – Les mesures de performance ROC pour la discrimination entre patients avec ou sans symptôme ont été évaluées pour chaque modèle de régression logistique ajusté sur les K premières variables latentes. Pour chaque modèle la déviance résiduelle, ainsi que le degré de liberté résiduel, sont donnés. La dernière colonne du tableau donne les p-value correspondant au test du chi-deux pour la comparaison de deux modèles successifs.*

nb LV	AUC	Sensibilité (à 80% de Spécificité)	Déviance résiduelle (Res.df)	p-value
		CLVR		
0	0.500	0.000	390.60 (332)	-
1	0.761	0.637	334.04 (331)	< 0.0001
2	0.803	0.615	306.85 (330)	< 0.0001
3	0.817	0.637	295.86 (329)	0.0009
4	0.823	0.670	291.97 (328)	0.0486
5	0.847	0.769	275.35 (327)	< 0.0001
6	0.853	0.813	269.88 (326)	0.0193
7	0.858	0.802	267.83 (325)	0.1537
8	0.860	0.802	266.44 (324)	0.2369
9	0.861	0.824	265.75 (323)	0.4059
10	0.861	0.802	265.13 (322)	0.4294
		CLVlar		
0	0.500	0.000	390.60 (332)	-
1	0.753	0.516	332.64 (331)	< 0.0001
2	0.821	0.769	296.24 (330)	< 0.0001
3	0.844	0.802	280.03 (329)	< 0.0001
4	0.861	0.802	266.88 (328)	0.0002
5	0.867	0.813	259.70 (327)	0.0073
6	0.870	0.813	257.76 (326)	0.1630
7	0.919	0.879	198.05 (325)	< 0.0001
8	0.925	0.868	191.00 (324)	0.0079
9	0.935	0.890	177.33 (323)	0.0002
10	0.936	0.890	175.89 (322)	0.2297

Nous avons appliqué et comparé les méthodes de régression PCR, PLS, *Sparse* PLS et l'approche proposée, CLVlar. Pour ce jeu de données, qui comporte 575 variables, la stratégie CLVlar est préférée à la stratégie CLVR en raison de son temps de calcul plus court.

La Figure 4.6 montre les RMSEP de validation croisée par la méthode du *leave-one-out* pour la régression sur composantes principales (PCR), la régression PLS (PLSR), la régression *Sparse* PLS (SPLSR) et la régression CLVlar.

Au niveau de la performance de prédiction, nous pouvons observer que les méthodes PCR et PLSR sont moins performantes que la régression *Sparse* PLS qui ne donne du poids qu'à un sous-ensemble de variables. La régression basée sur la procédure CLVlar, qui combine classification et sélection de groupes de variables, présente une performance en prédiction assez comparable à la régression *Sparse* PLS, un peu meilleure ou un peu moins bonne en fonction du jour considéré. La liste des variables sélectionnées en groupes, au fur et à mesure des itérations de la procédure CLVlar, figure dans le Tableau 4.7.

Pour les trois méthodes que sont la régression PLS, la régression *Sparse* PLS et la régression CLVlar, le nombre de composantes est choisi sur la base des résultats de la validation croisée. Pour le jour 17, on a retenu $K = 4$ composantes pour la PLSR, $\eta = 0.6$ et $K = 2$ pour la *Sparse* PLSR, $K = 3$ pour CLVlar. Pour le jour 22, on a retenu $K = 3$ pour la PLSR, $\eta = 0.8$ et $K = 4$ pour la *Sparse* PLSR, $K = 7$ pour CLVlar. Les coefficients estimés de ces trois modèles sont représentés dans la Figure 4.7. Nous pouvons observer que les coefficients non nuls de CLVlar sont moins nombreux alors que pour la régression *Sparse* PLS un nombre beaucoup plus important de variables ont des coefficients non nuls. De plus, pour le jour 17, presque toutes les variables sélectionnées par CLVlar le sont aussi par *Sparse* PLS. Pour le jour 22, on observe une différence de sélection de variables entre CLVlar et *Sparse* PLS. Par exemple, des variables de grand temps de rétention (supérieure à T400) ont été sélectionnées par *Sparse* PLS alors qu'elles sont moins souvent incluses dans CLVlar.

TABLE *4.7 – Groupes de variables extraites en utilisant la stratégie CLVlar. Chaque variable, ou ion, est identifiée par sa Masse et son Temps de rétention : (MxTy).*

groupes	variables
	Jour 17
G1	M277T42 ; M176T43 ; M170T44 ; M192T44 ; M154T45-1 ; M203T46 ; M132T48 ; M204T59 ; M211T91
G2	M130T381
G3	M121T42 ; M146T53 ; M282T391 ; M395T458 ; M303T689 ; M330T743 ; M329T744 ; M419T1042 ; M866T1047 ; M375T1051 ; M822T1052 ; M778T1056 ; M279T1410
G4	M137T42 ; M139T42
G5	M661T1010 ; M662T1010 ; M663T1010
G6	M380T905
G7	M264T54
G8	M363T384 ; M403T400 ; M402T400 ; M446T413 ; M447T413 ; M550T419 ; M490T423 ; M594T426 ; M534T431 ; M638T433 ; M578T438 ; M682T439 ; M726T444 ; M622T445 ; M770T450 ; M666T452
G9	M363T904
G10	M319T339 ; M358T386
	Jour 22
G1	M176T43 ; M170T44 ; M192T44 ; M265T46 ; M227T46 ; M114T46-1 ; M114T46-2 ; M114T46-3 ; M114T46-5 ; M132T48 ; M116T49 ; M271T68 ; M160T227 ; M247T468
G2	M289T58
G3	M357T57
G4	M332T462 ; M520T1074
G5	M565T1054
G6	M409T401 ; M461T409 ; M545T966 ; M822T1052
G7	M332T134
G8	M202T36 ; M124T36 ; M185T36 ; M184T36 ; M165T36 ; M203T36 ; M175T36 ; M203T36-1 ; M184T36-1 ; M99T36 ; M487T37 ; M165T39 ; M143T39 ; M154T45 ; M152T46 ; M263T56 ; M204T59 ; M123T65 ; M186T69 ; M232T132 ; M265T331 ; M206T468 ; M131T468 ; M590T1022 ; M565T1054 ; M520T1074 ; M465T1097 ; M464T1097 ; M420T1106
G9	M344T642
G10	M363T904

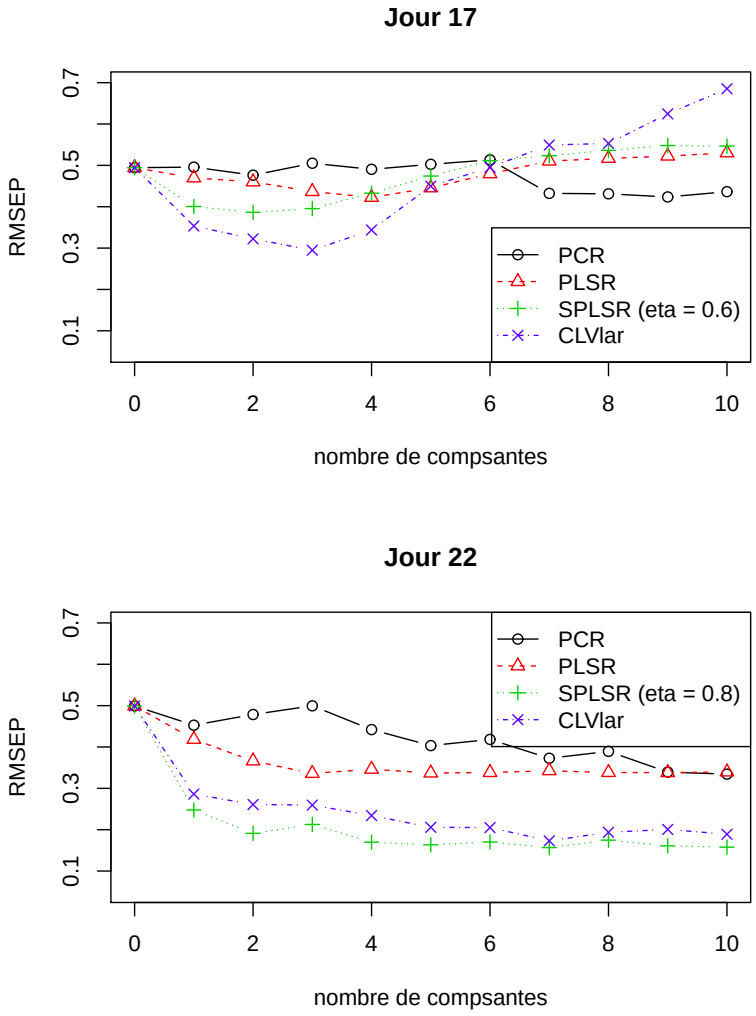

FIGURE *4.6 – La racine carrée de l'erreur quadratique moyenne de prédiction (RMSEP) estimées par validation croisée leave-one-out pour les méthodes PCR, PLSR , Sparse PLSR et CLVlar.*

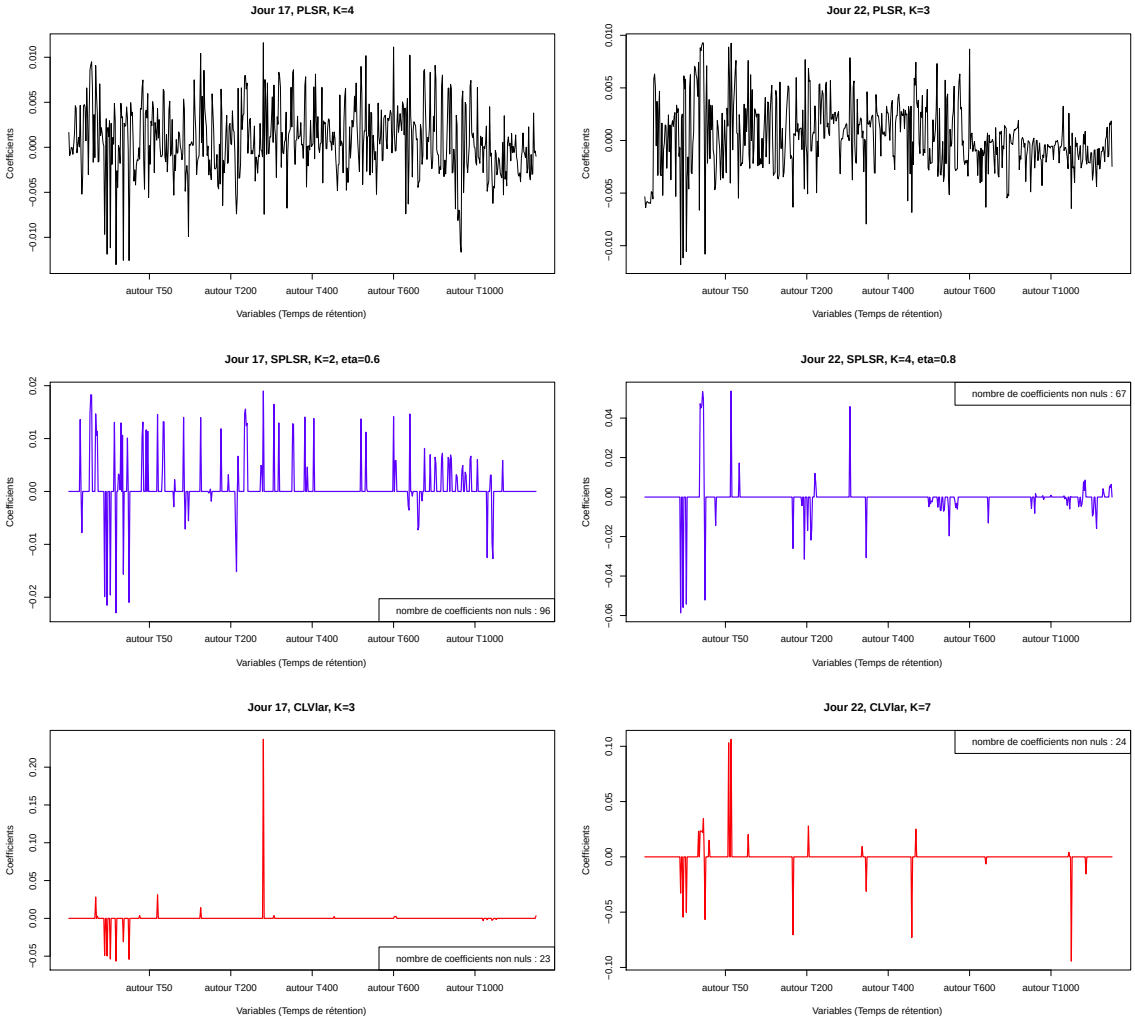

FIGURE *4.7 – Coefficients de régression pour les modèles PLSR, SPLSR et CLVlar. Les 575 variables sont ordonnées par leurs Temps de rétention.*

4.5 Conclusion

Dans ce chapitre, deux stratégies (CLVR et CLVlar) combinant la classification des variables et la régression ont été explorées. Des groupes de variables homogènes (au sens de l'unidimensionalité) sont extraits, un par un, en sélectionnant et en suivant certaines branches d'un arbre hiérarchique. L'arbre hiérarchique est obtenu à l'aide de la méthode CLV. Ici la classification CLV est dite supervisée car une variable de réponse $\mathbf{y}$ existe et elle est utilisée dans la construction et/ou la lecture de l'arbre. Chaque groupe de variables extrait, séquentiellement, est représenté par une variable latente, ou méta-variable, qui est utilisée comme nouveau prédicteur dans un modèle de régression construit progressivement.

Comparée à d'autres méthodes qui combinent classification de variables et modélisation, l'approche CLV supervisée permet de chercher des groupes de variables "directionnels" ìntégrant des variables corrélées entre elles que ce soit positivement ou négativement. De plus, un critère de variation locale de l'homogénéité des groupes est proposé pour rechercher un compromis entre la réduction de la dimensionalité du problème et la perte d'homogénéité interne lorsque l'on considère des groupes candidats de grande taille croissante. Le nombre optimal de groupes n'est pas un paramètre à définir a priori. A la place, un paramètre pour le contrôle de la variation locale relative du critère est à choisir. Par expérience, une valeur entre 5 % et 10% est suggérée.

La première stratégie (CLVR) proposée consiste à intégrer la variable de réponse dans la classification. A chaque itération, un arbre hiérarchique est construit avec les variables qui ne sont pas encore extraites. La procédure permet de lister tous les groupes intéressants, du point de vue de l'explication de la réponse. La seconde stratégie (CLVlar) a été proposée pour pallier le problème du temps de calcul nécessaire avec la première stratégie, lorsque les données sont de grande dimension. Le gain de temps de calcul réside dans le fait que l'arbre hiérarchique est construit une seule fois. Au niveau de la performance de prédiction, les deux approches sont assez comparables à la régression *Sparse* PLS.

Chapitre 5

ClustVarLV : un package R

5.1 Introduction

Pour la mise en œuvre de la méthode de classification de variables CLV et les extensions proposées dans les chapitre 3 et 4, nous avons développé un package, appelé **ClustVarLV** (Vigneau et Chen, 2014), sous logiciel R (R Core Team, 2014).

Pour la classification d'observations, de nombreuses fonctions R sont déjà disponibles. Nous pouvons citer les fonctions les plus souvent utilisées : `hclust`, `kmeans` dans le package **stats** (R Core Team, 2014) , et `agnes`, `pam` dans le package **cluster** (Maechler *et al.*, 2013). Ces fonctions sont basées sur des algorithmes classiques de la classification hiérarchique ou du partitionnement. D'autres fonctions, utilisant d'autres algorithmes, sont aussi disponibles, par exemple, pour réaliser du *Fuzzy Clustering* dans le package **cluster** et **e1071** (Meyer *et al.*, 2014), pour réaliser des classifications doubles permettant de segmenter simultanément les observations et les variables dans le package **biclust** (Kaiser *et al.*, 2013). Les packages **FunCluster** (Henegar, 2012) et **ORIClust** (Liu *et al.*, 2012) ont été développés pour des applications en bioinformatique, **sparcl** (Witten et Tibshirani, 2013) pour la classification d'observations en grande dimension ($p \gg n$) etc ...

Chavent *et al.* (2012) ont adapté la méthode CLV dans le cas de données mixtes (variables quantitatives et qualitatives ou un mélange des deux) et ont développé un package **ClustOfVar** (Chavent *et al.*, 2013). A la différence de **ClustVarLV**, ce

package n'offre pas la possibilité de former des "groupes locaux", de choisir le type de standardisation à appliquer aux variables, ni d'intégrer des variables externes.

Le package **ClustVarLV** couvre cinq aspects que l'on peut résumer comme suit :

1. Classification de variables basée sur un algorithme de classification hiérarchique ou un algorithme du partitionnement
   - fonctions concernées : `CLV`, `CLV_kmeans`
2. Choix du type de groupes : groupes directionnels ou groupes locaux
   - fonctions concernées : `CLV`, `CLV_kmeans`
   - argument concerné : method
3. Classification en tenant compte de variables externes
   - fonctions concernesé : `CLV`, `CLV_kmeans`, `LCLV`
   - arguments concernés : Xu et Xr
4. Classification en tenant compte de la présence éventuelle de variables atypiques ou de variables de bruit
   - fonctions concernées : `CLV_kmeans`
   - arguments concernés : strategy et rho
5. Classification de variables supervisée et régression
   - fonction concernée : `CLVR`
   - arguments concernés : strategy et gamma

5.2 Exemple illustratif

Nous illustrons le package avec le jeu de données RMN considéré dans § 3.5.1. Ce jeu de données contient deux zones spectrales mesurées sur 150 observations de jus d'orange mélangé avec du jus de clémentine. Les données sont disponibles dans le package. Nous pouvons récupérer la zone spectrale de 0.5 à 2.3 ppm contenant 180 variables spectrales avec les lignes de code suivantes :

```
R> library(ClustVarLV)
R> data(authen_NMR)
R> X <- authen_NMR$Xz2
```

Les spectres sont représentés (Figure 3.4) en utilisant les lignes de code suivantes :

```
R> xlab <- as.numeric(colnames(X))
R> plot(xlab, X[1,], type="l", xlab="ppm", ylab="",
+ ylim=c(14.8,15.8), xlim=rev(range(xlab)))
R> for (i in (1:nrow(X))) lines(xlab,X[i,])
```

5.2.1 Les fonctions basiques

Afin de réaliser une classification de 180 variables spectrales, en fonction de leur structure de variance-covariance, nous pouvons construire un arbre hiérarchique avec la fonction `CLV`. Dans cet exemple, des "groupes locaux" sont envisagés. Une transformation pareto est effectuée :

```
R> pareto <- function(x){x=sqrt(sd(x))}
R> X <- scale(X, center=T, scale=apply(X,2,pareto))
R> res <- CLV(X, method=2, sX=F, graph = TRUE)
[1] "initial value of the criterion :  12.39"
```

La valeur 12.39 correspond à la valeur du critère CLV initial lorsque chaque variable forme un groupe à elle seule. Notons que les résultats obtenus avec la fonction `CLV` est un objet de classe clv. Une description de cet objet est disponible en utilisant la fonction `print.clv` :

```
R> print(res)


number of variables:  180
number of observations:  150
measure of proximity: covariance
consolidation for K in c(20:2)
$tabres: results of the clustering
$partitionK or [[K]]: partition for K clusters
     $clusters: groups membership (before and after consolidation)
     $comp: latent components of the clusters (after consolidation)
```

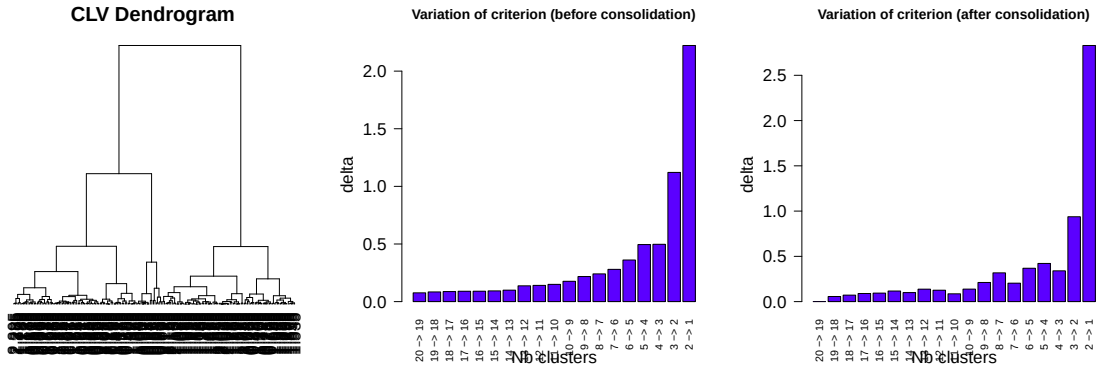

FIGURE *5.1 – Sortie graphique de la fonction* `CLV`.

Les sorties graphiques de la fonction `CLV` sont montrées dans la Figure 5.1. En plus de dendrogramme, les variations du critère de CLV avant et après la consolidation sont présentées. Une forte variation du critère de classification est observée quand on passe de 3 à 2 groupes. Une partition en 3 groupes de variables est donc retenue ici. Pour avoir plus de détails, nous pouvons utiliser la fonction `descrip_gp` avec un nombre, K, de groupes fixé.

```
R> descrip_gp(res, X, K=3)
$number
 1  2  3
61 82 37
$groups
$groups[[1]]
      cor in group cor next group
1.355         0.84          -0.22
1.335         0.71          -0.12
1.255         0.67          -0.08
 ...           ...            ...
$groups[[2]]
      cor in group cor next group
2.165         0.89          -0.16
```

```
2.005          0.88           -0.25
1.985          0.88           -0.31
 ...           ...            ...
$groups[[3]]
      cor in group cor next group
0.845          0.84           -0.06
0.835          0.84           -0.19
0.825          0.81           -0.18
 ...           ...            ...
$cormatrix
      Comp1 Comp2 Comp3
Comp1  1.00 -0.48 -0.37
Comp2 -0.48  1.00 -0.25
Comp3 -0.37 -0.25  1.00
```

Dans les sorties de la fonction `descrip_gp`, nous présentons (i) le nombre de variables dans chaque groupe, (ii) le coefficient de corrélation de chaque variable avec sa variable latente associée, et la variable latente du groupe le plus proche. (iii) la matrice de corrélation entre les variables latentes de groupes. Cela permet d'avoir une vision globale du partitionnement des variables. Une représentation graphique des groupes (Figure 5.2) sur la base de l'analyse en composantes principales est aussi possible en utilisant la fonction `gpmb_on_pc`.

```
R> gpmb_on_pc(res, X, K = 3, axeh = 1, axev = 2, v_symbol = TRUE)
R> gpmb_on_pc(res, X, K = 3, axeh = 2, axev = 3, v_symbol = TRUE)
```

5.2.2 Classification en présence de variables atypiques ou de variables de *noise*.

Comme présenté dans le Chapitre 3, nous avons mis en œuvre les stratégies "$K+1$" et "*Sparse* LV" associées à l'algorithme du partitionnement de la fonction `CLV_kmeans`. Un exemple est donné par les lignes de code suivantes :

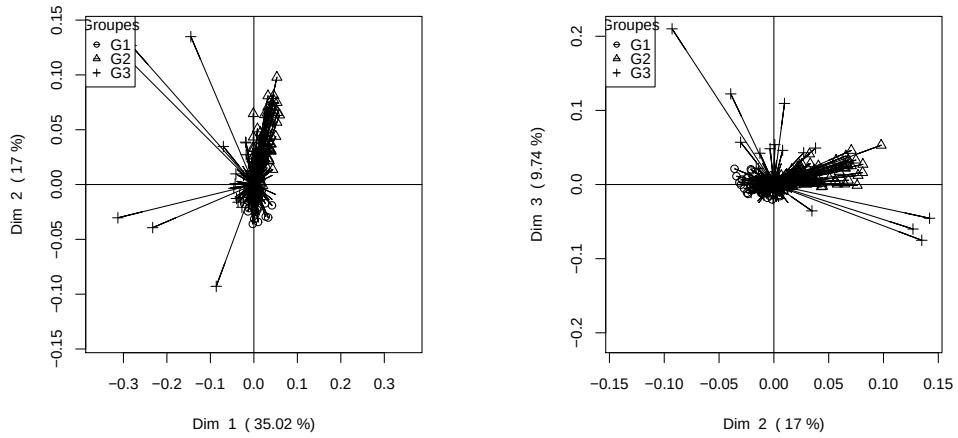

FIGURE *5.2 – Sortie graphique de la fonction* `gpmb_on_pc`.

```
R> K <- 3
R> rho = 0.5
R> init <- res[[K]]$clusters[1,]
R> res_kp1 <- CLV_kmeans(X, method = 2, sX =F, init =init,
R> + strategy = "kplusone", rho=rho)
R> res_splv <- CLV_kmeans(X, method = 2, sX =F, init =init,
R> + strategy = "sparselv", rho=rho)
```

Dans cet exemple, la partition en 3 groupes obtenue par la coupure du dendrogramme est prise comme initialisation de `CLV_kmeans`. La stratégie ("$K+1$" ou "*Sparse* LV") et le seuil de corrélation (ρ compris entre 0 et 1) doivent être indiqués. Nous pouvons consulter les variables qui sont affectées dans le *noise cluster* (groupe 0 en sortie de la fonction) ou qui sont associées un *loading* nul avec les lignes de code suivantes :

```
R> which(res_kp1$clusters[2,]==0)   #noise cluster
R> res_splv$sloading                #loading sparse
```

Les deux fonctions descriptives `descrip_gp` et `gpmb_on_pc` peuvent aussi être utilisées suite à la classification avec identification d'un *noise cluster*. Par exemple :

```
R> gpmb_on_pc(res_kp1, X, K = 3, axeh = 1, axev = 2, v_symbol = T)
R> gpmb_on_pc(res_kp1, X, K = 3, axeh = 2, axev = 3, v_symbol = T)
```

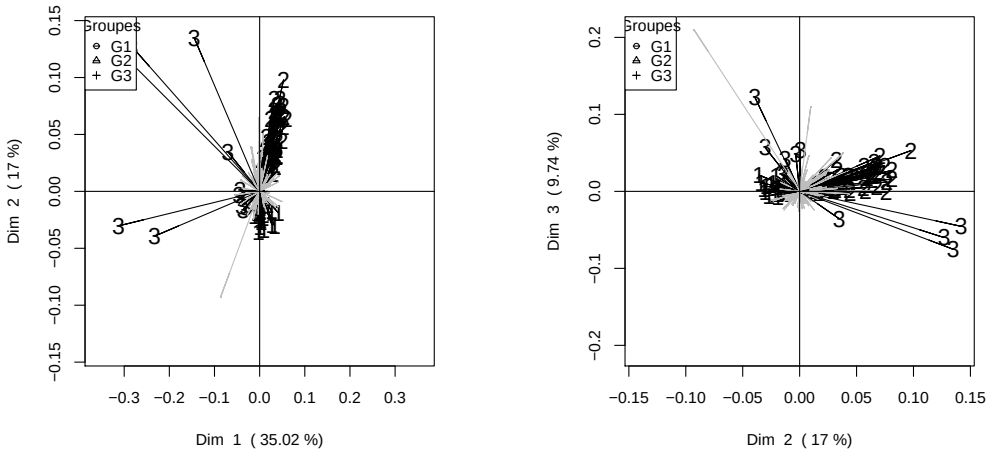

FIGURE *5.3 – Sortie graphique de la fonction* `gpmb_on_pc`.

Les variables en gris et non numérotées (Figure 5.3) correspondent aux variables placées dans le *noise cluster*.

5.2.3 CLV supervisée

Dans le chapitre 4, nous avons proposé deux méthodes de classification CLV dites supervisées (CLVR et CLVlar) qui permettent de chercher séquentiellement des groupes de variables pertinents vis-à-vis d'une variable de réponse $\mathbf{y}$. Une séquence de groupes peut être identifiée en utilisant la fonction `CLVR` .

```
R> library(lars)
R> gamma = 0.05
R> res.clvr <- CLVR(X, y, strategy = "clvr",
+ gamma = gamma, nbgp=10)
R> res.clvlar <- CLVR(X, y, strategy = "clvlar",
+ gamma = gamma, nbgp=10)
```

L'argument `nbgp` donne un nombre maximal de groupes à extraire. Les variables latentes correspondent à la première composante principale de chaque groupe de variables retenu. Des modèles de régression linéaire (ou de régression logistique) intégrant petit-à-petit ces variables latentes peuvent être ajustés en utilisant les fonctions `lm` ou `glm` du package **stats**.

5.3 Conclusion

Le package **ClustVarLV** offre un moyen opérationnel sous R, accessible à tout utilisateur, qui permet de mettre en œuvre la méthode CLV pour la classification de variables autour de variables latentes.

La version 1.2 de ce package intègre les fonctions de base. Une version 1.3, avec mise en ligne prévue au cours de l'été 2014, intégrera de nouvelles fonctions (notamment `CLVR` pour la classification supervisée) et des fonctions enrichies (notamment `CLV_kmeans` pour la classification avec procédure de mise à l'écart de certaines variables). Pour compléter le volet lié à la classification supervisée, de nouvelles sorties graphiques et des outils pour la construction des modèles de régression, avec validation croisée, sont envisagés.

Chapitre 6

Conclusion

L'analyse de données en grande dimension, notamment avec un grand nombre de variables doit faire face à de nombreuses difficultés et nécessite des stratégies de réduction de la dimensionalité et de sélection pour faciliter l'interprétation des résultats. Dans l'analyse exploratoire, les méthodes *sparse* PCA ont été proposées. Elles consistent à chercher des composantes principales dont les *loadings* de certaines variables sont nuls. Dans l'analyse prédictive, la pénalité de norme L^1 qui est utilisée dans la méthode Lasso effectue une sorte de sélection de variables et améliore ainsi à la fois la qualité de prédiction et l'interprétation du modèle. Cependant, la présence d'une structure de groupe au sein de l'ensemble de variables est aussi une caractéristique importante des données en grande dimension. L'exploration de cette structure peut être intégrée dans les approches d'analyses exploratoire et prédictive pour améliorer la compréhension des données.

Le point de départ de ce travail a été la classification de variables autour de variables latentes (CLV). Elle consiste à regrouper les variables et à représenter chaque groupe par sa variable latente. Nous avons suggéré de profiter des avantages de la méthode CLV qui sont, entre autres, le partitionnement en groupe directionnel et la réduction de la dimensionalité en utilisant la variable latente de groupe.

Les principales contributions de ce travail de thèse sont les suivantes :

– Nous avons proposé une modification de la méthode CLV pour la situation où il existe de nombreuses variables atypiques, isolées, et de bruit. Avec la stratégie

"$K + 1$", ces variables ne seront pas affectées dans les groupes principaux de sorte qu'une partition "propre" avec des groupes homogènes est obtenue. Avec la stratégie "*Sparse* LV", les variables atypiques n'auront pas de contribution aux variables latentes des groupes en utilisant les *sparse loadings*. Dans ce cas, la méthode proposée peut être considérée comme une alternative à *Sparse PCA*. Nous avons appliqué ces deux approches à des jeux de données simulées et réelles pour lesquels ils existaient à la fois des groupes de variables et des variables éloignées de la structure. Une amélioration de l'interprétation des groupes a été observée en identifiant les variables atypiques. Bien que les résultats obtenus avec les deux stratégies soient très similaires, la stratégie "$K+1$" est plus adaptée si l'objectif principal est de créer des groupes de variables. La stratégie "*Sparse* LV" est à privilégier si l'utilisateur cherche surtout à extraire des composantes latentes comme en analyse factorielle.

– Nous avons également exploré une méthode CLV dite supervisée en tenant compte d'une variable à prédire. Les modèles traditionnels comme la régression pas-à-pas ou la régression Lasso souffrent généralement de problèmes de stabilité en ne choisissant qu'une variable parmi un groupe de variables, alors que, des informations très utiles peuvent être apportées par des variables corrélées associées en groupe. La stratégie générale consiste à rechercher des groupes de variables pertinents vis-à-vis de la réponse et à les représenter par une méta-variable qui est la variable latente CLV. Le modèle de régression est ensuite construit sur ces variables latentes. La première stratégie proposée, CLVR, consiste à intégrer la variable de réponse dans des procédures de classification CLV répétées itérativement. La seconde stratégie, CLVlar, consiste à effectuer une recherche de groupes sur l'arbre hiérarchique en utilisant la procédure LARS. Ces deux approches permettent d'identifier les groupes pertinents. Cependant, en termes de temps de calcul, CLVlar est préférable en présence d'un très grand nombre de variables. Comparées à d'autres méthodes combinant la classification et la régression (Park *et al.*, 2007; Hastie *et al.*, 2001b), nos approches montrent deux intérêts : (i) elles permettent d'identifier des "groupes directionnels" de variables en tenant compte simultanément des corrélations positives et négatives. (ii) la

taille des groupes est contrôlée par un paramètre d'homogénéité qui est moins difficile à choisir que le nombre de groupes lui-même. En termes de qualité de prédiction et d'interprétation sur les données utilisées, les approches testées se sont montrées plus performantes que la Régression sur Composantes Principales, la régression PLS et assez comparables à la régression *Sparse* PLS.

– Nous avons finalisé un package `R` appelé **ClustVarLV** pour la mise en œuvre de la méthode de CLV et les approches proposées. L'algorithme de classification hiérarchique avec consolidation (`CLV`) et l'algorithme de partitionnement de type *k-means* `CLV_kmeans` ont été implémentés. **ClustVarLV** présente plusieurs intérêts : (i) la classification de variables en groupes directionnels et groupes locaux. (ii) la classification en tenant compte des variables externes. (iii) la classification en présence de variables atypiques, et (iv) la classification de variables supervisée.

Annexe A

Liste des travaux

Posters

- Chen, M., Qannari, E. M., Cariou, V., and Vigneau, E. (2012). Sparse Regularization for High Dimensional Data. *AFRODATA, 2nd African-European Conference on Chemometrics*, November 19-23 , Stellenbosch, South Africa.
- El Ghaziri, A., Chen, M., Vigneau, E. (2013). A strategy for identifying "outlier" variables in a big data set. *Chimiométrie*, 25-26 Septembre, Brest, France.
- Chen, M., and Vigneau, E. (2014). Segmentation of consumers while discarding irrelevant information. *Sensometrics 2014*, July 29 - August 1, Chicago, USA.
- Chen, M., Vigneau, E., Navez, B., Cottet V. (2014). Segmentation of consumers : new approaches for discarding irrelevant information. *EuroSense, 6th European Conference on Sensory and Consumer Research*, September 7-10, Copenhagen, Denmark.

Communications orales

- Chen, M., and Vigneau, E. (2013). Clustering and selection of variables. *Chimiométrie*, 25-26 Septembre, Brest, France.
- Chen, M., and Vigneau, E. (2014). Clustering of variables in noisy data set. *13èmes Journées Agro-Industrie et Méthodes Statistiques*, (25) 26-28 Mars, Rabat, Maroc (récompensé par le prix jeune chercheur).

– Vigneau, E., Chen, M., et Qannari, E. M. (2014). ClustVarLV : un package pour la classification de variables autour de variables latentes. *Troisièmes Rencontres R*, 25-27 Juin, Montpellier, France.

Publications

– Vigneau, E., Charles, M., and Chen, M. (2014). External preference segmentation with additional information on consumers : A case study on apples. *Food Quality and Preference*, 32, 83-92.

– Chen, M., and Vigneau, E., Supervised Clustering of Variables. *Advances in Data Analysis and Classification* (submitted 15-10-2013, revised 02-04-2014).

– Chen, M., and Vigneau, E., Two strategies for finding relevant clusters of variables : adding a noise cluster and clustering around sparse latent variables. *Computational Statistics and Data Analysis* (submitted the 20-06-2014).

Annexe B

Proposition

Notons n le nombre d'observations, p le nombre de variables, K le nombre de groupes, ρ le paramètre et $U = [\delta_{kj}]$ la matrice des indicateurs liée à une partition de variables où le terme générique $\delta_{kj} = 1$ si la variables $\mathbf{x}_j$ appartient au groupe G_k et $\delta_{kj} = 0$ sinon.

Proposition 1. *Pour un nombre de groupes K donné, il est évident que si $\rho = 0$, les nouveaux critères T_{new} et S_{new} reviennent aux critères ordinaires T et S. Chaque variable est attribuée à un seul groupe (U est une matrice stochastique). Par contre,*

(i) Si $\rho = 1$, toutes les variables sont attribuées dans le groupe de bruit. Alors $\forall(j,k)$, $\delta_{kj} = 0$.

(ii) Si ρ augmente, la variabilité totale du groupe de bruit va augmenter. Si de plus les variables sont centrées réduites, alors la taille de groupe de bruit va augmenter.

Démonstration. Considérons le critère T_{new} (3.6) pour groupes directionnels. Il dépend d'une partition U associée au nombre de groupes, K, et au paramètre, ρ. Nous avons :

$$T_{new}(U;K,\rho) = \underbrace{\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\; n\; \mathrm{cov}^2\left(\mathbf{x}_j,\mathbf{c}_k\right)}_{T(U)} + \rho^2 \underbrace{\sum_{j=1}^{p}(1-\sum_{k}\delta_{kj})\; \mathrm{var}\left(\mathbf{x}_j\right)}_{A(U)}$$

sous la contrainte $\|\mathbf{c}_k\| = n\; \mathrm{var}\left(\mathbf{c}_k\right) = 1$.

(i) Si $\rho = 1$, alors

$$\begin{aligned} T_{new}(U; K, \rho = 1) &= \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj}\ n\ \mathrm{cov}^2\ (\mathbf{x}_j, \mathbf{c}_k) + \sum_{j=1}^{p} (1 - \sum_{k} \delta_{kj})\ \mathrm{var}\ (\mathbf{x}_j) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj}\ \mathrm{var}\ (\mathbf{x}_j)\ \ cor^2(\mathbf{x}_j, \mathbf{c}_k) + \sum_{j=1}^{p} (1 - \sum_{k} \delta_{kj})\ \mathrm{var}\ (\mathbf{x}_j) \\ &\leqslant \sum_{j=1}^{p} \mathrm{var}\ (\mathbf{x}_j) \end{aligned} \tag{B.1}$$

Ainsi, $\max_U \{T_{new}(U; K, \rho = 1)\} = \sum_{j=1}^{p} \mathrm{var}\ (\mathbf{x}_j)$ est obtenu, pour les variables $\mathbf{x}_j$ telles que $\max_k\{cor^2(\mathbf{x}_j, \mathbf{c}_k)\} < 1$, si $\sum_k \delta_{kj} = 0$. Ces variables sont attribuées au groupe de bruit.
S'il existe une variable $\mathbf{x}_j$ telle que $\max_k\{cor^2(\mathbf{x}_j, \mathbf{c}_k)\} = 1$, l'optimum sera atteint soit en affectant cette variable au groupe de bruit, soit en l'affectant au groupe qui a une variable latente parfaitement corrélée avec elle. Ce cas est fort peu probable. Si cette situation est rencontrée, la règle adoptée est d'affecter cette variable au groupe de bruit.

(ii) Soit $U_1 = [\delta^1_{kj}]$ une partition optimale en K groupes lorsque $\rho = \rho_1$ et $U_2 = [\delta^2_{kj}]$ une partition optimale en K groupes lorsque $\rho = \rho_2$, alors :

$$\begin{aligned} \max_U \{T_{new}(U; K, \rho_1)\} &= T_{new}(U_1; K, \rho_1) \\ &= T(U_1) + \rho_1^2 A(U_1) \\ \text{en particulier} \quad &\geqslant T(U_2) + \rho_1^2 A(U_2) \\ \Rightarrow T(U_1) + \rho_1^2 (A(U_1) - A(U_2)) &\geqslant T(U_2) \end{aligned} \tag{B.2}$$

De même,

$$\begin{aligned}
\max_U \{T_{new}(U; K, \rho_2)\} &= T_{new}(U_2; K, \rho_2) \\
&= T(U_2) + \rho_2^2 A(U_2) \\
\text{en particulier} \quad &\geqslant T(U_1) + \rho_2^2 A(U_1) \\
\Rightarrow T(U_1) + \rho_2^2(A(U_1) - A(U_2)) &\leqslant T(U_2) \qquad \text{(B.3)}
\end{aligned}$$

avec (B.2) et (B.3), nous avons :

$$\begin{aligned}
T(U_1) + \rho_1^2 \left(A(U_1) - A(U_2)\right) &\geqslant T(U_1) + \rho_2^2 \left(A(U_1) - A(U_2)\right) \\
\Rightarrow (\rho_1^2 - \rho_2^2)(A(U_1) - A(U_2)) &\geqslant 0 \qquad \text{(B.4)}
\end{aligned}$$

A partir de (B.4), si $\rho_1 > \rho_2$, alors $A(U_1) \geqslant A(U_2)$. La variabilité totale du groupe de bruit augmente avec le paramètre ρ.

Remarque. La même démonstration peut être faite pour les groupes locaux en considérant le critère :

$$S_{new}(U; K, \rho) = \underbrace{\sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \sqrt{n} \operatorname{cov}(\mathbf{x}_j, \mathbf{c}_k)}_{S(U)} + \rho \underbrace{\sum_{j=1}^{p} (1 - \sum_{k} \delta_{kj}) \sqrt{\operatorname{var}(\mathbf{x}_j)}}_{B(U)}$$

□

Bibliographie

Al-Zoubi, M., Al-Dahoud, A., and Yahya, A. A. 2010. New Outlier Detection Method Based on Fuzzy Clustering. *WSEAS Transactions on Information Science and Applications*, **7**(5), 681–690. 26

Au, W. H., Chan, K. C., Wong, A. K., and Wang, Y. 2005. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Trans Comput Biol Bioinform*, **2**(2), 83–101. 53

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. 2006. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, **101**, 119–137. 7

Barnes, R. J., Dhanoa, M. S., and Lister, Susan J. 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.*, **43**(5), 772–777. 68

Berget, I., Mevik, B., Vebø, H., and Næs, T. 2005. A strategy for finding relevant clusters ; with an application to microarray data. *Journal of Chemometrics*, **19**(9), 482–491. 24, 29

Bernaards, C. A., and Jennrich, R. I. 2003. Orthomax rotation and perfect simple structure. *Psychometrika*, **68**(4), 585–588. 3

Bondell, H. D., and Reich, B. J. 2008. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, **64**(1), 115–123. 52

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. 2013. Correlated variables in regression : Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, **143**(11), 1835 – 1858. 11, 55

Centner, V., Massart, D., de Noord, O. E., de Jong, S., Vandeginste, B. M., and Sterna, C. 1996. Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*, **68**(21), 3851–3858. 7

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. 2012. Clustofvar : An r package for the clustering of variables. *Journal of Statistical Software*, **50**(1), 1–16. 85

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. 2013. *ClustOfVar : Clustering of variables*. R package version 0.8. 85

Chun, H., and Keles, S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **72**(1), 3–25. 7, 52, 72

Courant, F., Pinel, G., Bichon, E., Monteau, F., Antignac, J., and B., Le Bizec. 2009. Development of a metabolomic approach based on liquid chromatography-high resolution mass spectrometry to screen for clenbuterol abuse in calves. *Analyst*, **134**(8), 1637–1646. 76

Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., Liu, J., Misko, T. P., Perrin, R. J., Bales, K. R., Soares, H., Fagan, A. M., and Holtzman, D. M. 2011. Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer's Disease Diagnosis and Prognosis. *PLoS ONE*, **6**(4), e18850. 74, 75

Cuesta-Albertos, J. A., Gordaliza, A., and Matran, C. 1997. Trimmed k-means : An attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576. 26

Dahl, T., and Næs, T. 2009. Identifying outlying assessors in sensory profiling using fuzzy clustering and multi-block methodology. *Food Quality and Preference*, **20**(4), 287 – 294. 24, 29

Dave, R. N. 1991. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, **12**(11), 657–664. 25, 26, 27, 29

Dave, R. N., and Krishnapuram, R. 1997. Robust clustering methods : a unified view. *Fuzzy Systems, IEEE Transactions on*, **5**(2), 270–293. 26

Dhillon, I. S., Marcotte, E. M., and Roshan, U. 2003. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, **19**(13), 1612–9. 11

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics*, **32**, 407–499. 17, 20

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25), 14863–14868. 52

El Ghaoui, L., Viallon, V., and Rabbani, T. 2010. *Safe Feature Elimination in Sparse Supervised Learning*. Tech. rept. UC/EECS-2010-126. Electrical Engineering and Computer Sciences Department, University of California at Berkeley, Berkeley. 6

Enki, D. G., Trendafilov, N. T., and Jolliffe, I. T. 2012. A clustering approach to interpretable principal components. *Journal of applied statistics*, **40**(3), 583–599. 11

Ester, M., Kriegel, H., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of Knowledge Discovery in Databases*, **96**, 226–231. 26

Fan, J., and Lv, J. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **70**(5), 849–911. 6

Filzmoser, P., Liebmann, B., and Varmuza, K. 2009. Repeated double cross validation. *Journal of Chemometrics*, **23**(4), 160–171. 72

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. 2010. A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**(2-3), 89–109. 26

Hasegawa, K., Miyashita, Y., and Funatsu, K. 1997. GA Strategy for variable selection in QSAR studies : GA-based PLS analysis of calcium channel antagonists. *Journal of Chemical Information and Computer Sciences*, **37**(2), 306–310. 7

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, **1**(2), RESEARCH0003. 4

Hastie, T., Tibshirani, R., and Friedman, J. 2001a. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA : Springer New York Inc. 6

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. 2001b. Supervised harvesting of expression trees. *Genome Biol*, **2**(1), RESEARCH0003. 53, 55, 94

Henegar, C. 2012. *FunCluster : Functional Profiling of Microarray Expression Data*. R package version 1.09. 85

Hesterberg, T., Choi, N., Meier, L., and Fraley, C. 2008. Least angle and L1 penalized regression : A review. *Statistics Surveys*, **2**, 61–93. 6

Hoerl, A., and Kennard, R. 1970. Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67. 6

Huang, J., Ma, S., and Zhang, C.-H. 2008. Adaptive Lasso for sparse highdimensional regression models. *Statistica Sinica*, **18**, 1603–1618. 20

Jacob, C. C., Dervilly-Pinel, G., Biancotto, G., Monteau, F., and Le Bizec, B. 2014. Global urine fingerprinting by LC-ESI(+)-HRMS for better characterization of metabolic pathway disruption upon anabolic practices in bovine. *Metabolomics*, 1–14. 44

Jolliffe, I., Trendafilov, N., and Uddin, M. 2003. A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**(3), 531–547. 4, 28

Jolliffe, I.T. 2002. *Principal Component Analysis (second edition).* New York : Springer-Verlag New York, Inc. 3

Kaiser, S., Santamaria, R., Tatsiana, Khamiakova, Sill, M., Theron, R., Quintales, L., and Leisch., F. 2013. *biclust : BiCluster Algorithms.* R package version 1.0.2. 85

Kuhn, M., and Johnson, K. 2014. *AppliedPredictiveModeling : Functions and Data Sets for 'Applied Predictive Modeling'.* R package version 1.1-5. 74

Le Cao, K. A., Rossouw, D., Robert-Granie, C., and Besse, P. 2008. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article 35. 7, 52

Liu, T., Lin, N., Shi, N., and Zhang, B. 2012. *ORIClust : Order-restricted Information Criterion-based Clustering Algorithm.* R package version 1.0-1. 85

Ma, S., and Huang, J. 2007. Clustering threshold gradient descent regularization : with applications to microarray studies. *Bioinformatics*, **23**(4), 466–72. 53

Ma, S., Song, X., and Huang, J. 2007. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8**, 60. 53

Maechler, M., Rousseeuw, ., Struyf, A., Hubert, M., and Hornik, K. 2013. *cluster : Cluster Analysis Basics and Extensions.* R package version 1.14.4. 85

Mallows, C. L. 1973. Some Comments on C p. *Technometrics*, **15**(4), 661–675. 6

Marghny, M. H., and Taloba, Ahmed I. 2011. Outlier Detection using Improved Genetic K-means. *International Journal of Computer Applications*, **28**(11), 33–36. Published by Foundation of Computer Science, New York, USA. 26

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. 2012. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, **118**(0), 62–69. 7

Meinshausen, N. 2007. Relaxed Lasso. *Computational Statistics and Data Analysis*, **52**(1), 374 – 393. 20

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. 2014. *e1071 : Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3. 85

Naes, T., and Kowalski, B. 1989. Predicting sensory profiles from external instrumental measurements. *Food Quality and Preference*, **1**(4-5), 135–147. 68

Park, M. Y., Hastie, T., and Tibshirani, R. 2007. Averaged gene expressions for regression. *Biostatistics*, **8**(2), 212–27. 53, 54, 55, 94

R Core Team. 2014. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 85

Revelle, W., and Rocklin, T. 1979. Very Simple Structure : An Alternative Procedure For Estimating The Optimal Number Of Interpretable Factors. *Multivariate Behavioral Research*, **14**(4), 403–414. 3

Sarle, W. S. 1990. *The VARCLUS Procedure. SAS/STAT User's Guide, 4th Edition*. Cary, NC : SAS Institute, Inc. 3, 10

Shen, H., and Huang, J. Z. 2008. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**(6), 1015–1034. 4, 29

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**(1), 267–288. 6, 16, 17, 52

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. 2012. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **74**(2), 245–266. 6

Tolosi, L., and Lengauer, T. 2011. Classification with correlated features : unreliability of feature ranking and solutions. *Bioinformatics*, **27**(14), 1986–94. 53

Vigneau, E. 2012. *Clustering of variables around Latent Variables with external information on the samples or/and the variables.* 2nd African-European Conference on Chemometrics, Stellenbosch, South Africa. 15

Vigneau, E., and Chen, M. 2014. *ClustVarLV : Clustering of variables around Latent Variables*. R package version 1.2. 85

Vigneau, E., and Qannari, E. M. 2003. Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131–1150. 3, 10, 13

Vigneau, E., and Qannari, E.M. 2002. Segmentation of consumers taking account of external data. A clustering of variables approach. *Food Quality and Preference*, **13**(7-8), 515–521. 15

Vigneau, E., and Thomas, F. 2012. Model calibration and feature selection for orange juice authentication by 1H NMR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **117**(0), 22–30. 39, 42, 72

Vigneau, E., Sahmer, K., Qannari, E. M., and Bertrand, D. 2005. Clustering of variables to analyze spectral data. *Journal of Chemometrics*, **19**(3), 122–128. 70

Vigneau, E., Charles, M., and Chen, M. 2014. External preference segmentation with additional information on consumers : A case study on apples. *Food Quality and Preference*, **32, Part A**(0), 83–92. 15

Wang, Y., and Wu, Q. 2012. Sparse PCA by iterative elimination algorithm. *Advances in Computational Mathematics*, **36**(1), 137–151. 4

Witten, Daniela M., and Tibshirani, Robert. 2013. *sparcl : Perform sparse hierarchical clustering and sparse k-means clustering.* R package version 1.0.3. 85

Wold, H. 1966. *Estimation of Principal Components and Related Models by Iterative Least squares*. New York : Academic Press. Pages 391–420. 6, 52

Yousef, M., Jung, S., Showe, L. C., and Showe, M. K. 2007. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics*, **8**, 144. 53

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **68**(1), 49–67. 6

Zou, H. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429. 20

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **67**(2), 301–320. 6, 19, 52, 72

Zou, H., Hastie, T., and Tibshirani, R. 2006. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286. 4, 25, 28

Abstract

With the development of high-throughput analysis techniques, researchers have adopted systematic approaches to describe simultaneously a large number of variables. However, one of the important challenges lies in the difficulty to summarise and interpret this enormous quantity of information.

We adopt a clustering of variables approach (CLV) which allows us to highlight disjunctive structures, and therefore, reduce the dimensionality of the problem and facilitate the interpretation of the data at hand. However, in order to further improve the relevance of such approaches, two directions of investigation are proposed.

The first direction involves filtering the data by setting aside atypical variables or variables associated with noise. For this purpose, a strategy to create an additional group of variables, called noise cluster, and a strategy based on the definition of sparse latent variables are proposed and compared.

The second direction concerns the development of a clustering of variables procedure directed to the explanation of a response variable. The implementation of iterative algorithms provides a sequence of group latent variables with good predictive performance. These latent variables are also easy to interpret since each predictive component is associated with a subset of variables assumed to have a one-dimensional structure.

Résumé

Avec le développement des techniques d'analyse à haut débit, les chercheurs ont adopté des démarches de profilage systémique qui permettent l'analyse descriptive simultanée d'un grand nombre de variables. Une des difficultés réside dans la synthèse et l'interprétation de ces nombreuses informations.

Nous adoptons ici une approche de classification de variables (CLV) qui permet de mettre en lumière des structures disjonctives pour la réduction de la dimensionnalité du problème, facilitant ainsi l'interprétation des données. Cependant, afin d'améliorer davantage la pertinence de ce type d'approches, deux directions d'investigation sont proposées.

La première consiste à filtrer les données de sorte à écarter les variables isolées ou associées à du bruit de fond. Une stratégie qui consiste à créer un groupe supplémentaire de variables, appelé "noise cluster", ainsi qu'une stratégie fondée sur la définition de variables latentes de groupe creuses (ou *sparse*) sont proposées et comparées.

La seconde direction d'investigation est le développement d'une procédure de classification de variables dirigée vers l'explication d'une variable de réponse. Un algorithme itératif de classification/extraction est proposé. Il fournit une séquence de variables latentes de groupes ayant de bonnes performances en prédiction. Elles sont également simples à interpréter dans la mesure où chaque composante prédictrice n'est associée qu'à un sous-ensemble de variables exploratoires conçu pour avoir une structure pratiquement unidimensionnelle.