

Analyse des correspondances multiples parcimonieuses

Julie Le Borgne

▶ **To cite this version:**

Julie Le Borgne. Analyse des correspondances multiples parcimonieuses. Sciences du Vivant [q-bio]. 2020. dumas-02968178

HAL Id: dumas-02968178
https://dumas.ccsd.cnrs.fr/dumas-02968178

Submitted on 15 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2019-2020 Spécialité : Agronome Spécialisation (et option éventuelle) : Science des données	**Mémoire de fin d'études** ☑ d'ingénieur de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ de master de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ d'un autre établissement (étudiant arrivé en M2)

Analyse des Correspondances Multiples Parcimonieuses

Par : Julie LE BORGNE

Soutenu à Rennes le 07/09/2020

Devant le jury composé de :

Président :

Maître de stage : Vincent Guillemot

Enseignant référent : David Causeur

Autres membres du jury (Nom, Qualité)

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

Ce document est soumis aux conditions d'utilisation
«Paternité-Pas d'Utilisation Commerciale-Pas de Modification 4.0 France»
disponible en ligne http://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr

Remerciements

Je tiens à remercier Vincent Guillemot pour son suivi au long de ces 6 mois et pour m'avoir permis de réaliser ce stage à l'institut Pasteur. Je souhaite également remercier Hervé Abdi et Ju-Chi Yu qui m'ont accompagnée dans ce projet. Aussi, je tiens à remercier l'ensemble des équipes du Hub pour leur accueil ainsi qu'aux stagiaires avec qui j'ai pu échanger. Enfin, je voudrais remercier mes camarades et professeurs de Master sans qui je n'aurais pas pu prendre part à ce projet.

Table des matières

1 Introduction **1**
1.1 Le HUB de Biostatistique et de Bioinformatique 1
1.2 Objectifs et réalisations 1
1.3 La parcimonie dans une méthode d'analyse multivariée 2

2 Notations et notions **3**
2.1 Notations mathématiques 3
2.2 Décomposition en valeurs singulières (SVD) 3
2.3 Obtention de la SVD avec l'algorithme des moindres carrés alternés 4

3 Méthodes **5**
3.1 Méthode de la CSVD, une SVD parcimonieuse 5
3.1.1 Projection dans $\mathscr{B}_1(c)$ et principe du seuillage doux 5
3.1.2 Contraintes de parcimonie de la CSVD 6
3.1.3 Projection dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ 7
3.1.4 Projection dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1) \cap \mathbf{X}^{\perp}$ avec POCS 7
3.2 Méthode existante : Analyse des Correspondances Multiples 7
3.2.1 Codage disjonctif complet 8
3.2.2 Matrice des probabilités centrée 8
3.2.3 Décomposition en valeurs singulières généralisée 8
3.2.4 Coordonnées de L'ACM 9
3.2.5 Contributions des variables et des observations 9
3.3 Méthode de l'ACM parcimonieuse 10
3.3.1 La norme de groupe 10
3.3.2 Contraintes de parcimonie de la CGSVD 10
3.3.3 L'algorithme de l'ACM parcimonieuse 11
3.3.4 Projections de l'ACMP 12

4 Résultats **12**
4.1 Analyse d'un questionnaire sur le Maroilles 13
4.2 Analyse de données génétiques : ADNI 15

5 Conclusion et perspectives **19**

Annexe : Projection dans la boule-ℓ_1 de rayon c **20**

Références **22**

Abréviations

ACP	Analyse des Composantes Principales
ACM	Analyse des Correspondances Multiples
ACMP ou SMCA	Analyse des Correspondances Multiples Parcimonieuse
SVD	*Singular Value Decomposition*
GSVD	*Generalised Singular Value Decomposition*
CSVD	*Constrained Singular Value Decomposition*
CGSVD	*Constrained Generalised Singular Value Decomposition*
ALS	*Alternating Least Squares*

1 Introduction

L'institut Pasteur est une fondation privée à but non lucratif. Son objectif est de contribuer au développement de nouvelles approches diagnostiques, préventives et thérapeutiques des maladies infectieuses à travers la recherche biomédicale, l'enseignement et le développement d'initiatives de santé publique. Un des atouts majeur de la recherche pasteurienne est la mise en œuvre d'approches pluridisciplinaires s'appuyant par exemple sur la bioinformatique ou les biostatitiques. L'institut est un centre international de recherche biomédicale, présent dans le monde entier à travers le réseau International des Instituts Pasteur, qui comporte 32 établissements répartis sur tous les continents. L'institut Pasteur de Paris fait partie de ce réseau d'instituts et regroupe aujourd'hui plus de 2500 collaborateurs dans 130 unités de recherches différentes.

1.1 Le HUB de Biostatistique et de Bioinformatique

Le Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), crée en 2015, et devenu le département de biologie computationnelle (DBC) en 2019, est une structure qui a pour rôle le soutien des chercheurs de l'institut Pasteur pour le traitement, l'analyse et la modélisation des grandes quantités de données générées par les équipes de recherche présentes sur le campus. Le Hub de Bioinformatique et Biostatistique est la plate-forme de services du DBC. Les ingénieurs du Hub apportent un soutien en bioinformatique et biostatistiques aux équipes de recherche du campus, participent au développement d'outils d'analyse de données et forment les personnels des Institut Pasteur à Paris. Il est aujourd'hui composé de 52 ingénieurs répartis en 7 groupes d'expertise. Mon tuteur de stage le Dr. Vincent Guillemot, est rattaché au groupe d'expertise STATS du Hub, qui traite les questions liées aux statistiques.

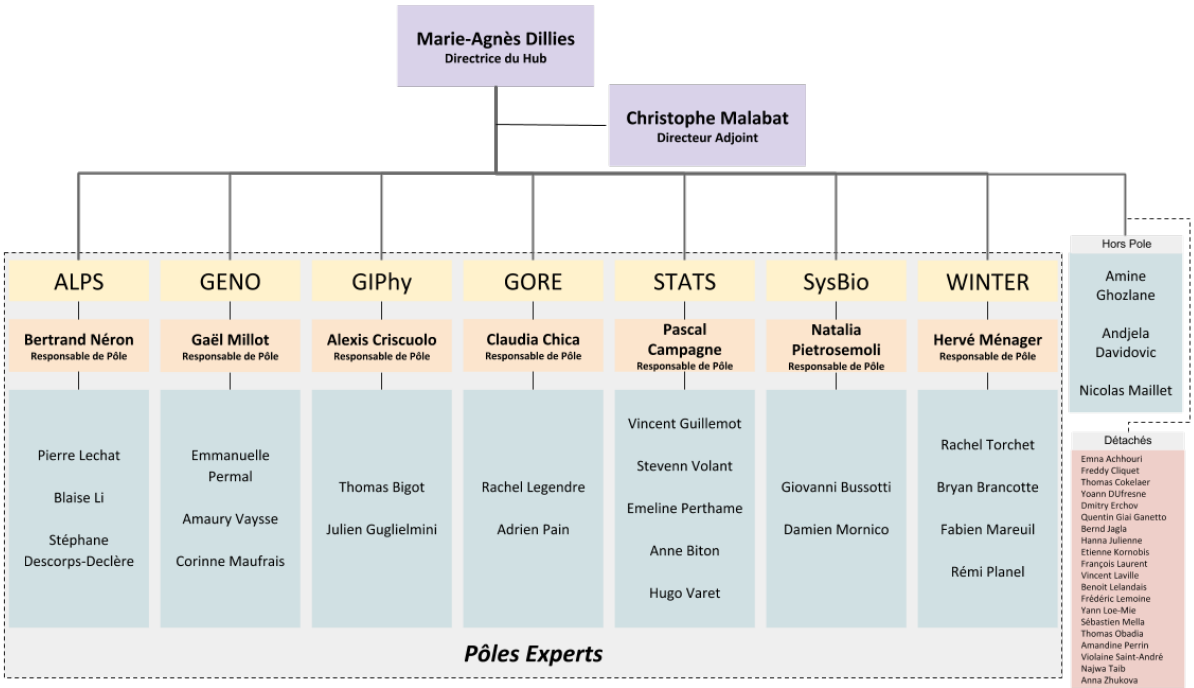


FIGURE 1 – Organigramme du HUB de biostatistiques et bioinformatique.

1.2 Objectifs et réalisations

Au cours de ce stage, j'ai programmé un package R qui implémente une méthode d'analyse de données développée par mon tuteur de stage, Vincent Guillemot : l'Analyse des Correspondances Multiples Parcimonieuse (ACMP). Ce package est disponible publiquement à partir de la page suivante : `https://github.com/JulieLB/SMCA`. J'ai comparé

l'ACMP à l'Analyse des Correspondances Multiples (ACM) sur deux jeux de données : un jeu de données de réponses à un questionnaire sur la connaissance du Maroilles et un jeu de données génétiques lié à la maladie d'Alzheimer. Enfin, j'ai participé à la vérification de propriétés classiques de l'ACM parcimonieuse : équivalence distributionnelle, formules de transitions et propriétés barycentriques.

1.3 La parcimonie dans une méthode d'analyse multivariée

La parcimonie est, en science, un principe selon lequel on utilise le minimum de causes pour expliquer un phénomène. En anglais, *sparse*, que l'on va traduire par « parcimonieux », signifie épars, clairsemé ou rare [1]. Ainsi, dans les domaines de l'informatique et de l'analyse numérique, une matrice parcimonieuse est une matrice dans laquelle la plupart des éléments sont des zéros. Rendre une matrice parcimonieuse permet de simplifier la résolution de calculs complexes en informatique, mais aussi la compréhension des données en statistiques.

Les analyses factorielles sont des méthodes statistiques permettant d'obtenir une visualisation informative des données étudiées et dont les résultats peuvent s'exprimer sous forme de matrices. L'ajout de parcimonie dans ce type d'analyse revient à introduire des zéros dans les matrices de résultats, ce qui améliore leur interprétabilité, en particulier pour l'analyse de données de grande dimension. Ainsi, les analyses factorielles parcimonieuses se révèlent être de puissantes méthodes d'analyse de données.

En effet, en ACM, méthode d'analyse factorielle exploratoire de données qualitatives, l'interprétation est grandement facilitée lorsque les variables et les observations contribuent soit beaucoup, soit très peu à une dimension (ce qui revient à avoir une valeur proche ou égale à zéro dans la matrice de résultat). Lors de l'analyse de données en grande dimension, on retrouve rarement un tel schéma : le plus souvent, de nombreuses observations et variables contribuent à chaque dimension.

Afin de faciliter l'analyse de données qualitatives de grande dimension, il est intéressant d'introduire de la parcimonie dans l'algorithme de l'ACM. Cependant, il existe peu de versions parcimonieuses de l'ACM dans la littérature. Par exemple, [Mori et al., 2016] proposent de rendre l'ACM parcimonieuse grâce à des techniques de rotations similaires aux rotations VARIMAX et [Saporta et al., 2012] introduisent de la parcimonie de groupe en écrivant l'ACM comme une régression avec une contrainte imposée sur les coefficients.

En Analyse des Composantes Principales (ACP), méthode d'analyse factorielle exploratoire de données quantitatives, il est possible d'obtenir un tel schéma par rotation des dimensions (VARIMAX) ou par introduction de parcimonie ([Guillemot et al., 2019], [Witten et al., 2009]). Afin de généraliser cet algorithme de parcimonie pour l'appliquer à l'ACM, nous devons prendre en compte les contraintes suivantes :

1. Des contraintes de poids et de masse imposées sur les rangs et les colonnes ;
2. Une contrainte de groupe qui impose qu'un groupe entier de colonnes représentant une variable soit sélectionné ou écarté en même temps par la parcimonie ;
3. La conservation de l'orthogonalité entre les dimensions pour une meilleure interprétabilité.

Ce document présente une généralisation de la SVD (*Singular Value Decomposition*) parcimonieuse de [Guillemot et al., 2019] qui prend en compte les contraintes énoncées. Pour commencer, nous présenterons les notations mathématiques utilisées et certaines notions essentielles à la compréhension de l'algorithme parcimonieux, puis nous nous pencherons

1. `https://www.wordreference.com/enfr/sparse`

sur les méthodes utilisées pour effectuer une ACM classique et une ACM parcimonieuse. Enfin, nous illustrerons cette nouvelle méthode avec une analyse d'un jeu de données issu d'un sondage à propos du Maroilles ainsi qu'avec une analyse de données génétiques de personnes diagnostiquées sur la maladie d'Alzheimer ainsi que de sujets sains.

2 Notations et notions

2.1 Notations mathématiques

Les matrices sont notées par une lettre majuscule en gras et les vecteurs par une lettre minuscule en gras. Les matrices, vecteurs et éléments issus de la même matrice sont notés avec la même lettre (ex : $\mathbf{X}$, $\mathbf{x}$, x). La transposée est notée $^\top$ et l'opération inverse $^{-1}$. La matrice identité est notée $\mathbf{I}$, les matrices ou vecteurs composés de uns sont notés $\mathbf{1}$, et les matrices et vecteurs de zéros sont notés $\mathbf{0}$. S'il est appliqué à une matrice carré, l'opérateur « diag$\{\cdot\}$ » renvoie le vecteur des éléments de la diagonale de cette matrice. S'il est appliqué à un vecteur, il renvoie une matrice diagonale avec les éléments du vecteur sur sa diagonale. L'espace orthogonal à un vecteur ou une matrice est noté $^\perp$, il s'exprime : $\mathbf{A}^\perp = \{\mathbf{x} \in E, \mathbf{x} \perp \mathbf{A}\}$. La norme ℓ_1 d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ s'exprime $\|\mathbf{x}\|_1 = \sum_i^n |x_i|$. La norme ℓ_2, ou norme Euclidienne, d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ s'exprime $\|\mathbf{x}\|_2 = \sqrt{\sum_i^n x_i^2}$.

La normalisation d'un vecteur $\mathbf{x}$ peut être effectuée de deux manières équivalentes :

(1) Division de $\mathbf{x}$ par sa norme Euclidienne : $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$;

(2) Projection de $\mathbf{x}$ dans la boule ℓ_2 : $proj_{\mathscr{B}_2(1)}(\mathbf{x})$, avec $\mathscr{B}_2(1) = \{\mathbf{x} \in E, \|\mathbf{x}\|_2 \leq 1\}$

La table de données analysée avec une ACM est un tableau multivarié, où chaque observation est décrite par plusieurs variables qualitatives. La matrice présente I rangs (nombre d'observations) et K colonnes (nombre de variables). Chaque variable k possède J_k modalités, et la somme des J_k, notée J, est le nombre total de modalités. La somme de toutes les entrées de la table, noté N, vaut $I \times K$.

2.2 Décomposition en valeurs singulières (SVD)

La SVD est l'outil principal utilisé pour les analyses multivariées, en particulier les méthodes factorielles comme l'ACP ou l'ACM. L'intérêt de cette méthode est de générer des combinaisons linéaires orthogonales et optimales des variables et des observations. Ces combinaisons permettent de définir les dimensions de l'ACP ou de l'ACM.

La SVD utilise la décomposition en valeurs propres d'une matrice semi-définie positive (matrice symétrique dont les valeurs propres sont positives ou nulles) pour obtenir une décomposition similaire applicable à toute matrice rectangulaire. Ainsi, la SVD génère une décomposition en trois matrices simples d'une matrice rectangulaire $\mathbf{X}$ de dimension $I \times J$ [Abdi, 2007].

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^\top$$

Avec :

- $\mathbf{P}$ la matrice $I \times R$ (où $R = min(I, J)$) des vecteurs propres de la matrice $\mathbf{X}\mathbf{X}^\top$. Les colonnes de $\mathbf{P}$ sont les vecteurs singuliers à gauche de $\mathbf{X}$.
- $\mathbf{Q}$ la matrice $J \times R$ des vecteurs propres de la matrice $\mathbf{X}^\top\mathbf{X}$. Les colonnes de $\mathbf{Q}$ sont les vecteurs singuliers à droite de $\mathbf{X}$.
- $\mathbf{\Delta}$ la matrice diagonale $R \times R$ contenant les valeurs singulières de $\mathbf{X}$. $\mathbf{\Delta} = \mathbf{\Lambda}^{1/2}$ avec $\mathbf{\Lambda}$ les valeurs propres de $\mathbf{X}\mathbf{X}^\top$ et $\mathbf{X}^\top\mathbf{X}$.

Notons que $\mathbf{P}$ et $\mathbf{Q}$ sont telles que : $\mathbf{P}^\top\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$. On dit que $\mathbf{P}$ et $\mathbf{Q}$ sont des matrices orthogonales. Cette propriété illustre l'orthogonalité des combinaisons linéaires de la SVD et par extension, l'orthogonalité des dimensions des analyses multivariées.

2.3 Obtention de la SVD avec l'algorithme des moindres carrés alternés

La puissance itérée est une méthode qui permet de trouver le premier vecteur propre et la première valeur propre d'une matrice semi-définie positive. Cette méthode consiste à multiplier un vecteur initial quelconque par la matrice de données, suivie d'une étape de normalisation (sous forme de projection dans $\mathscr{B}_2(1)$), jusqu'à convergence vers le premier vecteur propre de la matrice considérée.

Dans le cas de la SVD, nous utiliserons un algorithme dit « des moindres carrés alternés » ou ALS (*Alternating Least Squares*), présenté dans l'algorithme 1. Cet algorithme permet de retrouver le premier vecteur singulier à droite et à gauche d'une matrice rectangulaire. Le principe est le même que la puissance itérée, sauf que l'on multiple alternativement la matrice de données par le vecteur de droite puis celui de gauche jusqu'à convergence. La puissance itérée fait donc l'approximation que $proj_{\mathscr{B}_2(1)}(\mathbf{X}^\top \mathbf{p}^{(s)}) \to \mathbf{q}_1$ et $proj_{\mathscr{B}_2(1)}(\mathbf{X}\mathbf{q}^{(s+1)}) \to \mathbf{p}_1$ lorsque $s \to +\infty$, avec $\mathbf{p}_1$ et $\mathbf{q}_1$ les premiers vecteurs singuliers de $\mathbf{X}$. La première valeur singulière δ_1 pourra être calculée à partir de $\mathbf{p}_1$ et $\mathbf{q}_1$ ($\delta_1 = \mathbf{p}_1^\top \mathbf{X}\mathbf{q}_1$).

Data: $\mathbf{X}$, ε
Result: Premier triplet singulier de $\mathbf{X}$
$\mathbf{p}^{(0)}$ and $\mathbf{q}^{(0)}$ are randomly initialized;
$\delta^{(0)} \leftarrow 0$;
$\delta^{(1)} \leftarrow \mathbf{p}^{(0)\top}\mathbf{X}\mathbf{q}^{(0)}$;
$s \leftarrow 0$;
while $\left|\delta^{(s+1)} - \delta^{(s)}\right| \geq \varepsilon$ **do**
  $\mathbf{p}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}\mathbf{q}^{(s)}, \mathscr{B}_2(1))$;
  $\mathbf{q}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}^\top\mathbf{p}^{(s+1)}, \mathscr{B}_2(1))$;
  $\delta^{(s+1)} \leftarrow \mathbf{p}^{(s+1)\top}\mathbf{X}\mathbf{q}^{(s+1)}$;
  $s \leftarrow s + 1$;
end

Algorithm 1: Algorithme de l'ALS.

Pour obtenir les vecteurs et valeurs singuliers suivants, on peut utiliser deux méthodes différentes : la déflation de la matrice de données $\mathbf{X}$ ou la projection des vecteurs singuliers dans l'espace orthogonal aux triplets singuliers déjà calculés. Dans les deux cas, cette étape assure l'orthogonalité des vecteurs singuliers entre eux.

La déflation

Le théorème de la déflation énonce que le premier triplet singulier de la $i+1^{\text{ème}}$ matrice déflatée est le $i+1^{\text{ème}}$ triplet singulier de la matrice originale. En effet, le principe de la déflation est de retirer de la matrice de données l'information portée par la valeur et les vecteurs singuliers obtenus. Soit l'étape de déflation après le calcul du $i^{\text{ème}}$ triplet singulier, avec $i \in [1, R]$:

$$\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i - \delta_i \mathbf{p}_i \mathbf{q}_i^\top$$

On pourra alors appliquer ALS sur la nouvelle matrice $\mathbf{X}_{i+1}$ afin d'extraire le triplet singulier de la dimension suivante, qui sera orthogonal au premier.

Projection dans un espace orthogonal

Cette méthode consiste à projeter le résultat de chaque itération dans l'intersection des espaces de la boule ℓ_2 et de l'espace orthogonal aux vecteurs de droite ou de gauche

calculés précédemment, $\mathbf{P}^{\perp}$ et $\mathbf{Q}^{\perp}$. Cela permet d'assurer la normalisation des vecteurs, mais aussi son orthogonalité avec les vecteurs propres des dimensions précédentes. On obtient alors :

$$\begin{cases} \text{proj}_{\mathscr{B}_2(1)\cap\mathbf{P}^{\perp}}(\mathbf{X}\mathbf{q}^{(s)}) \to \mathbf{p} \\ \text{proj}_{\mathscr{B}_2(1)\cap\mathbf{Q}^{\perp}}(\mathbf{X}^{\top}\mathbf{p}^{(s+1)}) \to \mathbf{q} \end{cases} \quad \text{lorsque } s \to +\infty \tag{1}$$

avec $\mathbf{p}$ et $\mathbf{q}$ les vecteurs singuliers d'une dimension quelconque.

Un vecteur est projeté dans la boule ℓ_2 lorsqu'il est normalisé et un vecteur est projeté dans $\mathbf{X}^{\perp}$ en le multipliant par la matrice $\mathbf{I} - \mathbf{X}^{\top}\mathbf{X}$. Afin de projeter les vecteurs à droite et à gauche dans l'intersection de ces deux espaces, on peut simplement projeter dans l'un puis l'autre.

3 Méthodes

Dans cette section, nous présenterons deux méthodes existantes : la décomposition en valeurs singulières sous contraintes (CSVD), qui est une version parcimonieuse de la SVD, et l'ACM, méthode d'analyse de données qualitatives qui repose sur la SVD. Ensuite, nous nous intéresserons à la méthode que j'ai developpée au cours de mon stage : l'ACM parcimonieuse (ACMP). Cette méthode, qui permet d'introduire de la parcimonie dans l'ACM, utilise une généralisation de la CSVD.

3.1 Méthode de la CSVD, une SVD parcimonieuse

La CSVD de [Guillemot et al., 2019] est une méthode de décomposition de matrices qui introduit de la parcimonie à l'aide d'une contrainte sur les rangs et sur les colonnes d'une matrice tout en gardant l'orthogonalité des dimensions.

3.1.1 Projection dans $\mathscr{B}_1(c)$ et principe du seuillage doux

D'après l'annexe 5, la projection d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ dans $\mathscr{B}_1(c) = \{\mathbf{x} \in E, \|\mathbf{x}\|_1 \leq c\}$, avec c une constante positive, repose sur la formule :

$$\text{proj}_{\mathscr{B}_1(c)}(\mathbf{x}) = S(\mathbf{x}, \lambda^*) \text{ avec } \lambda^* \in \mathbb{R}^+ \text{ tel que } \|S(\mathbf{x}, \lambda^*)\|_1 = c \tag{2}$$

où S est la fonction de seuillage doux qui s'exprime :

$$\forall i \in [1, n],\ [S(\mathbf{x}, \lambda)]_i = \begin{cases} 0 \text{ si } |x_i| < \lambda \\ x_i - \lambda \text{ si } x_i > \lambda \\ x_i + \lambda \text{ si } x_i < -\lambda \end{cases} \tag{3}$$

Cette opération crée de la parcimonie dans le vecteur projeté grâce aux propriétés de la fonction de seuillage doux. En effet, sur la figure 2, on peut voir que les éléments dont la valeur est dans l'intervalle $[-\lambda, \lambda]$ sont mis à zéro. En d'autres termes, le vecteur projeté est plus parcimonieux que le vecteur d'origine. Dans l'équation 2 et la figure 2, on peut voir que λ^* est inversement proportionnel à c, donc plus c est petit (car $c \geq 0$), plus le vecteur projeté sera parcimonieux.

Ainsi, dans l'algorithme de la CSVD, la propriété de parcimonie est une conséquence de la contrainte ℓ_1 sur les vecteurs singuliers. Cette contrainte est implémentée par la projection de ces vecteurs dans la boule-ℓ_1. Pour une matrice $\mathbf{X}$ de taille $I \times J$, la contrainte de parcimonie d'une dimension ℓ s'applique par projection des vecteurs à droite et à gauche

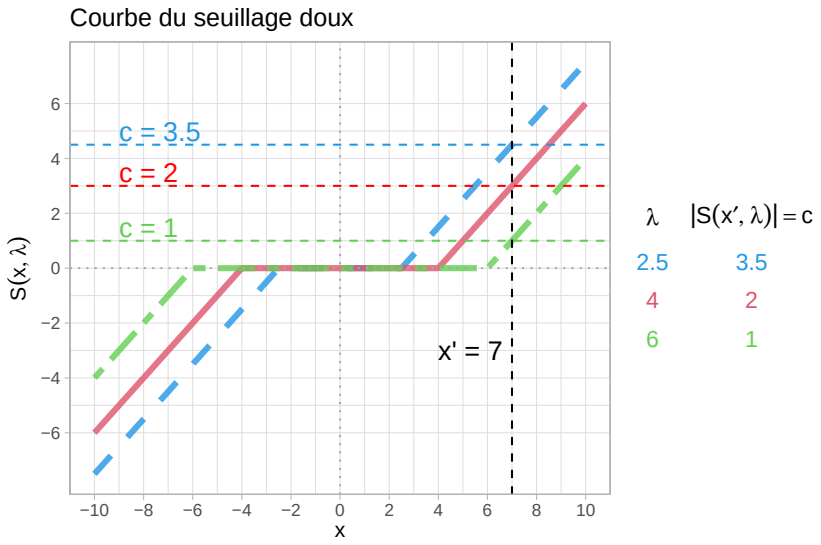


FIGURE 2 – Exemple de courbes de seuillage doux comparant les valeurs prises par λ^* et c pour $\mathbf{x}' = 7$ sous la contrainte $\|S(\mathbf{x}', \lambda^*)\|_1 = c$, avec c valant 3.5, 2 et 1. Les valeurs de λ^* correspondantes sont respectivement égales à 2.5, 4 et 6.

respectivement dans $\mathscr{B}_1(c_{1,\ell})$ et $\mathscr{B}_1(c_{2,\ell})$ où $c_{1,\ell}$ et $c_{2,\ell}$ sont des constantes positives exprimant le degré de contrainte appliqué (plus leur valeur est grande, moins la décomposition sera parcimonieuse, voir table 1).

TABLE 1 – Valeur des contraintes dans le cas d'une projection dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ d'après [Guillemot et al., 2019]

c_1	c_2	Degré de parcimonie
$1+\epsilon_1$	$1+\epsilon_2$	Élevé
$1/2\sqrt{I}$	$1/2\sqrt{J}$	Moyen
$2/3\sqrt{I}$	$2/3\sqrt{J}$	Faible
$\sqrt{I}$	$\sqrt{J}$	Aucune

3.1.2 Contraintes de parcimonie de la CSVD

La CSVD résout donc le problème d'optimisation suivant :

$$\underset{\mathbf{P},\mathbf{\Delta},\mathbf{Q}}{\arg\min} \tfrac{1}{2} \left\| \mathbf{X} - \mathbf{P}\mathbf{\Delta}\mathbf{Q}^\top \right\|_2^2 \text{ avec } \begin{cases} \mathbf{P}^\top\mathbf{P} = \mathbf{I} \\ \mathbf{Q}^\top\mathbf{Q} = \mathbf{I} \end{cases}, \text{ et } \forall \ell = 1, ..., R \begin{cases} \|\mathbf{p}_\ell\|_1 \leq c_{1,\ell} \\ \|\mathbf{q}_\ell\|_1 \leq c_{2,\ell} \end{cases} \tag{4}$$

La SVD classique permet de trouver la solution exacte de $\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^\top$. Contrairement à la SVD, la solution de l'équation 4 par la CSVD est une approximation ($\mathbf{X} \simeq \mathbf{P}\mathbf{\Delta}\mathbf{Q}^\top$). On dit du résultat d'une CSVD qu'il est composé de pseudo-valeurs singulières et de pseudo-vecteurs singuliers.

Avec une contrainte de parcimonie implémentée par une projection des vecteurs singuliers dans la boule-ℓ_1, la méthode de déflation ne permet plus d'assurer l'orthogonalité des dimensions. On choisit alors la méthode de projection des vecteurs dans l'espace orthogonal aux vecteurs singuliers déjà estimés. D'après les équations 1 et 4, pour toute dimension ℓ, on obtient les pseudo-vecteurs singuliers ainsi :

$$\begin{cases} \text{proj}_{\mathscr{B}_1(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^\perp}(\mathbf{X}\mathbf{q}_\ell^{(s)}) \to \mathbf{p}_\ell \\ \text{proj}_{\mathscr{B}_1(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^\perp}(\mathbf{X}^\top\mathbf{p}_\ell^{(s+1)}) \to \mathbf{q}_\ell \end{cases} \quad \text{lorsque } s \to +\infty \tag{5}$$

3.1.3 Projection dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$

L'algorithme de [Gloaguen et al., 2017] calcule la projection exacte d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ dans l'espace $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ et repose sur la formule suivante :

$$\text{proj}_{\mathscr{B}_1(c)\cap\mathscr{B}_2(1)}(\mathbf{x}) = S(\mathbf{x}, \lambda^*) \text{ avec } \lambda^* \in \mathbb{R}^+ \text{ tel que } \frac{\|S(\mathbf{x}, \lambda^*)\|_1}{\|S(\mathbf{x}, \lambda^*)\|_2} = c \tag{6}$$

D'après [Guillemot et al., 2019], pour que l'intersection des espaces $\mathscr{B}_1(c)$ et $\mathscr{B}_2(1)$ existe, la valeur de c doit être comprise dans l'intervalle $[1; \sqrt{n}]$.

3.1.4 Projection dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1) \cap \mathbf{X}^\perp$ avec POCS

Afin de projeter les vecteurs à droite et à gauche dans les espaces $\mathscr{B}_1(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^\perp$ et $\mathscr{B}_1(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^\perp$, on utilise la méthode de la projection dans un ensemble convexe (POCS) [Combettes, 1993]. On distingue deux composantes : la projection dans l'intersection de la boule-ℓ_1 et de la boule-ℓ_2 et la projection dans l'espace orthogonal aux triplets singuliers déjà estimés. La méthode POCS consiste à projeter alternativement le vecteur dans ces deux espaces jusqu'à convergence, comme illustré sur la figure 3.

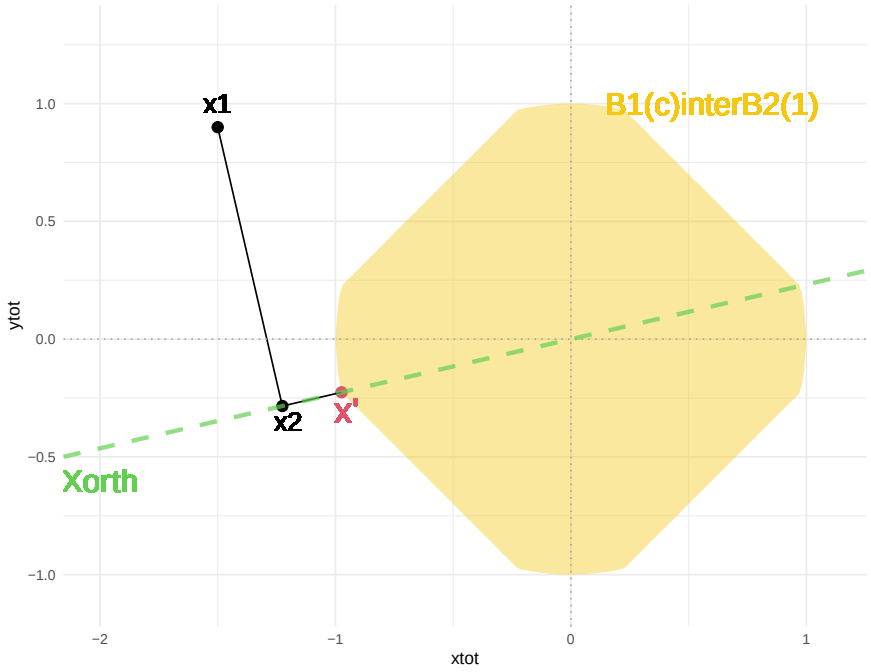


FIGURE 3 – Exemple de la projection d'un vecteur $\mathbf{x}$ dans l'intersection de $\mathbf{X}^\perp$ (hyperplan de dimension 1 représenté par la droite verte) et de $\mathscr{B}_1(1.2) \cap \mathscr{B}_2(1)$ (espace représenté en jaune) dans un espace à deux dimensions à l'aide de l'algorithme POCS. On a $\mathbf{x}' = proj_{\mathscr{B}_1(1.2)\cap\mathscr{B}_2(1)\cap\mathbf{X}^\perp}(\mathbf{x})$.

3.2 Méthode existante : Analyse des Correspondances Multiples

L'ACM est une technique d'analyse de données multivariées qualitatives ([Saporta, 2011], [Abdi and Valentin, 2007]). En particulier, il s'agit de données dans lesquelles chaque observation est décrite par plusieurs variables qualitatives corrélées entre elles. Ainsi, l'ACM peut être perçue comme une généralisation de l'ACP, appliquée sur des données catégorielles plutôt que quantitatives. Le premier objectif de l'ACM est d'extraire les informations importantes de ces données sous la forme de nouvelles variables orthogonales que l'on appelle des dimensions. Le second est d'observer les schémas de similarité des observations et des variables.

3.2.1 Codage disjonctif complet

Pour effectuer une ACM, les variables de la table de données sont recodées à l'aide d'un codage disjonctif complet : chaque variable est représentée par un groupe de variables binaires.

Dans un jeu de données catégorielles, chaque variable qualitative possède plusieurs modalités. Par exemple, considérons les données représentées dans la table 2a : ce jeu de données contient deux variables catégorielles, dont une variable « Porte des lunettes » qui possède les modalités « Oui » et « Non ». Chacune de ces modalités sera codée sous la forme d'une variable binaire qui aura pour entrée un 1 si l'individu correspondant à la ligne possède cette modalité et un 0 sinon. La matrice de données joignant toutes les variables binaires, appelée tableau disjonctif, possède autant de colonnes qu'il y a de modalités (table 2b).

TABLE 2 – Transformation des deux variables en codage disjonctif complet

(a) Variables qualitatives

Lunettes	Age
Oui	1
Oui	3
Non	3

(b) Tableau Disjonctif

Lunettes		Age		
Oui	Non	1	2	3
1	0	1	0	0
1	0	0	0	1
0	1	0	0	1

3.2.2 Matrice des probabilités centrée

La première étape de l'analyse est de calculer la matrice des probabilités centrée à partir du tableau disjonctif. La matrice est normalisée par le nombre total de mesures N et double centrée à l'aide de masses (chaque individu a la même masse) et de poids (chaque variable binaire a un poids inversement proportionnel à sa fréquence).

Soit $\mathbf{Y}$ le tableau disjonctif, on obtient $\mathbf{X}$ la matrice des probabilités centrée :

$$\mathbf{X} = \frac{1}{N}\mathbf{Y} - \mathbf{r}\mathbf{c}^\top$$

Avec :

- $\mathbf{r} = \mathbf{Y1} \times N^{-1}$ le vecteur des probabilités totales sur les rangs, qui correspond aux masses des individus ;
- $\mathbf{c} = \mathbf{Y}^\top\mathbf{1} \times N^{-1}$ le vecteur des probabilités totales sur les colonnes, qui correspond aux poids des variables binaires.

3.2.3 Décomposition en valeurs singulières généralisée

L'ACM repose sur la décomposition en valeurs singulières généralisée (GSVD) de la matrice des probabilités centrée, qui permet de calculer les combinaisons linéaires optimales et orthogonales des variables et des observations, tout en intégrant les masses et les poids comme contraintes [Abdi, 2007]. $\mathbf{X}$ une matrice $I \times J$ se décompose ainsi :

$$\mathbf{X} = \tilde{\mathbf{P}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{Q}}^\top \text{ Avec } \tilde{\mathbf{P}}^\top\mathbf{D_r}\tilde{\mathbf{P}} = \tilde{\mathbf{Q}}^\top\mathbf{D_c}\tilde{\mathbf{Q}} = \mathbf{I} \text{ où } \mathbf{D_r} = \text{diag}\{\mathbf{r}\} \text{ et } \mathbf{D_c} = \text{diag}\{\mathbf{c}\} \quad (7)$$

Contrairement à une SVD, les vecteurs singuliers généralisés $\tilde{\mathbf{P}}$ et $\tilde{\mathbf{Q}}$ sont orthogonaux sous les contraintes imposées par $\mathbf{D_r}$ et $\mathbf{D_c}$, les matrices diagonales de poids et de masse.

Une façon d'obtenir une telle décomposition est de décomposer en éléments singuliers la matrice $\mathbf{X}$ pondérée par $\mathbf{D_r}$ et $\mathbf{D_c}$:

$$\begin{aligned}\mathbf{D_r}^{-1/2}\mathbf{X}\mathbf{D_c}^{-1/2} &= \mathbf{D_r}^{-1/2}\tilde{\mathbf{P}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{Q}}^\top\mathbf{D_c}^{-1/2} \text{ d'après (7)} \\ &= \mathbf{P}\mathbf{\Delta}\mathbf{Q}^\top \text{ avec } \begin{cases}\mathbf{P} = \mathbf{D_r}^{-1/2}\tilde{\mathbf{P}} \\ \mathbf{Q} = \mathbf{D_c}^{-1/2}\tilde{\mathbf{Q}} \\ \mathbf{\Delta} = \tilde{\mathbf{\Delta}}\end{cases}\end{aligned} \tag{8}$$

$$\text{Or, } \begin{cases}\mathbf{P}^\top\mathbf{P} = \tilde{\mathbf{P}}^\top\mathbf{D_r}^{-1}\tilde{\mathbf{P}} = \mathbf{I} \\ \mathbf{Q}^\top\mathbf{Q} = \tilde{\mathbf{Q}}^\top\mathbf{D_c}^{-1}\tilde{\mathbf{Q}} = \mathbf{I}\end{cases}$$

La matrice pondérée $\mathbf{D_r}^{-1/2}\mathbf{X}\mathbf{D_c}^{-1/2}$ peut donc être décomposée en éléments singuliers $\mathbf{P}, \mathbf{Q}$ et $\mathbf{\Delta}$ avec $\mathbf{P}^\top\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$. Cette décomposition peut être calculée avec l'algorithme de l'ALS, contrairement à la GSVD de $\mathbf{X}$. Les matrices contenant les triplets singuliers généralisés sont calculées ainsi :

$$\tilde{\mathbf{P}} = \mathbf{D_r}^{1/2}\mathbf{P} \text{ et } \tilde{\mathbf{Q}} = \mathbf{D_c}^{1/2}\mathbf{Q} \text{ et } \tilde{\mathbf{\Delta}} = \mathbf{\Delta} \tag{9}$$

3.2.4 Coordonnées de L'ACM

On peut calculer les coordonnées des observations et des variables sur les nouveaux axes (les dimensions) à l'aide des matrices singulières généralisées, $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}$ et $\tilde{\mathbf{\Delta}}$. Les matrices contenant ces coordonnées sont appelées $\mathbf{F}$, pour les coordonnées des observations, et $\mathbf{G}$, pour celles des variables. On les représente sous forme de matrices dont chaque élément est la corrélation entre une observation (en ligne) et une composante (en colonne) pour $\mathbf{F}$ et la corrélation entre une variable et une composante pour $\mathbf{G}$. On calcule $\mathbf{F}$ et $\mathbf{G}$ tel que :

$$\begin{cases}\mathbf{F} = \mathbf{D_r}^{-1}\tilde{\mathbf{P}}\mathbf{\Delta} = \mathbf{D_r}^{-1/2}\mathbf{P}\mathbf{\Delta} \\ \mathbf{G} = \mathbf{D_c}^{-1}\tilde{\mathbf{Q}}\mathbf{\Delta} = \mathbf{D_r}^{-1/2}\mathbf{Q}\mathbf{\Delta}\end{cases} \tag{10}$$

3.2.5 Contributions des variables et des observations

La contribution est un outil d'interprétation de l'ACM qui permet d'identifier les observations ou variables importantes pour une dimension. La contribution d'un rang i à une dimension l et d'une colonne j à une dimension l sont respectivement obtenues par :

$$ctr_{i,l} = r_i\frac{f_{i,l}^2}{\lambda_l} \text{ et } ctr_{j,l} = c_i\frac{g_{i,l}^2}{\lambda_l} \text{ où } \lambda_l = \delta_l^2 \tag{11}$$

On pourra également calculer les contributions d'une variable à une dimension en sommant la contribution des colonnes représentant des modalités issues de cette variable. La contribution d'une variable k de modalités m_j à une dimension l est obtenue par :

$$ctr_{k,l} = \sum_j^{m_k} ctr_{j,l} \tag{12}$$

La corrélation de la variable k avec une dimension l est :

$$\eta_{k,l}^2 = J\sum_j^{m_k} ctr_{j,l}\lambda_l \tag{13}$$

3.3 Méthode de l'ACM parcimonieuse

En ACM parcimonieuse, on utilise la décomposition en valeurs singulière généralisée sous contraintes (CGSVD) (algorithme généralisé de la CSVD de [Guillemot et al., 2019], c'est-à-dire avec incorporation de matrices de poids et de masses et d'une normalisation par groupe dans la décomposition) à la place de la GSVD pour calculer les nouveaux axes. La CGSVD impose une contrainte de groupe sur les rangs et sur les colonnes, qui induit de la parcimonie.

3.3.1 La norme de groupe

Dans certaines situations, il existe une partition naturelle des rangs ou des colonnes, dont les éléments sont réunis sous forme de groupes. Par exemple, en ACM, les colonnes représentant les modalités peuvent être regroupées par variable. Il semble alors logique de sélectionner ou de retirer simultanément les éléments d'un même groupe. En appliquant une contrainte de groupe lors de la décomposition d'une matrice de données, on s'assure d'une unité de parcimonie sur les éléments d'un même groupe.

Pour cela, on peut appliquer une contrainte induisant de la parcimonie de groupe de nature similaire à la contrainte ℓ_1 sur les éléments d'un groupe par une norme de groupe, notée $\|.\|_{\mathcal{G}}$. On définit :

- $\mathcal{G}$ une partition non chevauchante d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ telle que : $\mathcal{G} = \{\iota_k, k = 1, ..., K\}$ où K est le nombre de groupes et ι_k les indices des éléments du $k^{\text{ème}}$ groupe ;
- $\mathbf{x}_{\iota_k}$ le sous vecteur du $k^{\text{ème}}$ groupe, soit un vecteur composé des éléments de ce groupe.

Il s'agit de construire la norme de groupe sur le modèle d'une norme-$\ell_{p,p'}$:

$$\|\mathbf{x}\|_{\mathcal{G}} = \left\| \begin{bmatrix} \|\mathbf{x}_{\iota_1}\|_{p'} \\ \vdots \\ \|\mathbf{x}_{\iota_K}\|_{p'} \end{bmatrix} \right\|_p$$

Afin d'établir une norme de groupe qui induit de la parcimonie sur un vecteur contraint par cette norme (par exemple, $\|\mathbf{x}\|_{\mathcal{G}} \leq c$), on peut utiliser les propriétés de la fonction de seuillage doux de la même manière que dans la CSVD. Ainsi, on choisit d'utiliser une norme-$\ell_{1,2}$ (on prend norme-ℓ_p = norme-ℓ_1 et norme-$\ell_{p'}$ = norme-ℓ_2).

La norme de groupe, ou norme-$\ell_{1,2}$, est alors l'équivalent de la norme-ℓ_1 d'un vecteur composé de la norme-ℓ_2 des sous-vecteurs $\mathbf{x}_{\iota_k}$, avec $k = 1, ..., K$:

$$\|\mathbf{x}\|_{\mathcal{G}} = \sum_{k}^{K} \|\mathbf{x}_{\iota_k}\|_2$$

La contrainte de groupe s'applique sous la forme de la projection à chaque itération des vecteurs singuliers dans la boule-$\ell_{\mathcal{G}}$, notée $\mathscr{B}_{\mathcal{G}}(.)$. Elle s'exprime sous la forme $\mathscr{B}_{\mathcal{G}}(c) = \{\mathbf{x} \in E, \|\mathbf{x}\|_{\mathcal{G}} \leq c\}$, avec c une constante positive.

3.3.2 Contraintes de parcimonie de la CGSVD

Ainsi, la CGSVD résout le problème d'optimisation suivant :

$$\underset{\tilde{\mathbf{P}},\tilde{\mathbf{\Delta}},\tilde{\mathbf{Q}}}{\arg\min} \tfrac{1}{2} \left\| \mathbf{X} - \tilde{\mathbf{P}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{Q}}^\top \right\|_2^2 \text{ avec } \begin{cases} \tilde{\mathbf{P}}^\top {\mathbf{D_r}}^{-1} \tilde{\mathbf{P}} = \mathbf{I} \\ \tilde{\mathbf{Q}}^\top {\mathbf{D_c}}^{-1} \tilde{\mathbf{Q}} = \mathbf{I} \end{cases}, \text{ et } \forall \ell = 1, ..., R \begin{cases} \left\| {\mathbf{D_r}}^{-1/2} \mathbf{p}_\ell \right\|_{\mathcal{G}_r} \leq c_{1,\ell} \\ \left\| {\mathbf{D_c}}^{-1/2} \mathbf{q}_\ell \right\|_{\mathcal{G}_c} \leq c_{2,\ell} \end{cases} \tag{14}$$

où $c_{1,\ell}$ et $c_{2,\ell}$ sont des constantes positives et $\mathcal{G}_r$ et $\mathcal{G}_c$ les partitions respectivement sur les rangs et sur les colonnes. La partition sur les colonnes $\mathcal{G}_c$ correspond généralement à un groupement des colonnes par variable. Sur l'exemple du tableau 2b, on aura le groupement $\mathcal{G}_c = \{\iota_1 = (1,2), \iota_2 = (3,4,5)\}$ et donc $\mathbf{x}_{\iota_1} = (Lunettes_{Oui}, Lunettes_{Non})$ et $\mathbf{x}_{\iota_2} = (Age_1, Age_2, Age_3)$. La partition sur les rangs peut correspondre à un groupement des individus par la modalité prise sur une variable indépendante (par exemple, la ville d'origine). Pour l'exemple du tableau 2b, on pourra regrouper les trois individus par leur catégorie d'âge : $\mathcal{G}_c = \{\iota_1 = (1), \iota_3 = (2,3)\}$.

De la même manière que pour la CSVD, la solution de l'équation 14 par la CGSVD est une approximation ($\mathbf{X} \simeq \tilde{\mathbf{P}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{Q}}^\top$, $\tilde{\mathbf{P}}$ et $\tilde{\mathbf{Q}}$ sont les pseudo-vecteurs singuliers généralisés avec une contrainte de groupe). Contrairement à la GSVD utilisée en ACM classique, le résultat d'une CGSVD est donc composé de pseudo-valeurs et pseudo-vecteurs singuliers.

3.3.3 L'algorithme de l'ACM parcimonieuse

Data: $\mathbf{X}$, $\mathbf{r}$, $\mathbf{c}$, $\mathcal{G}_c$, $\mathcal{G}_r$
Parameters: ε, R, c_1, c_2
Result: Sparse MCA of $\mathbf{X}$
Define $\mathbf{P} = \sqrt{\mathbf{r}}$;
Define $\mathbf{Q} = \sqrt{\mathbf{c}}$;
$\mathbf{D_r} \leftarrow diag(\mathbf{r})$;
$\mathbf{D_c} \leftarrow diag(\mathbf{c})$;
$\mathbf{X} \leftarrow {\mathbf{D_r}}^{-1/2}\mathbf{X}{\mathbf{D_c}}^{-1/2}$;
for $\ell = 1, \ldots, R$ **do**
  $\mathbf{p}^{(0)}$ and $\mathbf{q}^{(0)}$ are randomly initialized;
  $\delta^{(0)} \leftarrow 0$;
  $\delta^{(1)} \leftarrow {\mathbf{p}^{(0)}}^\top \mathbf{X}\mathbf{q}^{(0)}$;
  $s \leftarrow 0$;
  while $\left|\delta^{(s+1)} - \delta^{(s)}\right| \geq \varepsilon$ **do**
    $\mathbf{p}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}\mathbf{q}^{(s)}, \mathscr{B}_{\mathcal{G}_r}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^\perp)$;
    $\mathbf{q}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}^\top\mathbf{p}^{(s+1)}, \mathscr{B}_{\mathcal{G}_c}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^\perp)$;
    $\delta^{(s+1)} \leftarrow {\mathbf{p}^{(s+1)}}^\top \mathbf{X}\mathbf{q}^{(s+1)}$;
    $s \leftarrow s + 1$;
  end
  $\mathbf{\Delta} \leftarrow \text{vect}\left(\mathbf{\Delta}, \delta^{(s+1)}\right)$;
  $\mathbf{P} \leftarrow \text{vect}\left(\mathbf{P}, \mathbf{p}^{(s+1)}\right)$;
  $\mathbf{Q} \leftarrow \text{vect}\left(\mathbf{Q}, \mathbf{q}^{(s+1)}\right)$;
end
$\mathbf{P} \leftarrow {\mathbf{D_r}}^{1/2}\mathbf{P}$;
$\mathbf{Q} \leftarrow {\mathbf{D_c}}^{1/2}\mathbf{Q}$;
$\mathbf{F} \leftarrow {\mathbf{D_r}}^{-1}\mathbf{P}\mathbf{\Delta}$;
$\mathbf{G} \leftarrow {\mathbf{D_c}}^{-1}\mathbf{Q}\mathbf{\Delta}$

Algorithm 2: Algorithme général de l'ACM parcimonieuse.

L'algorithme de l'ACMP est présenté dans l'Algorithme 2. On utilise les moindres carrés alternés dont la composante clé est la projection des vecteurs à droite et à gauche dans l'intersection des espaces de la boule définie par une contrainte de groupe (la boule-$\ell_\mathcal{G}$), la boule-ℓ_2, et l'espace orthogonal aux pseudo-triplets singuliers déjà estimés.

3.3.4 Projections de l'ACMP

Comme pour l'algorithme de la CSVD, on utilise la méthode POCS [Combettes, 1993] pour projeter les vecteurs à droite et à gauche dans les espaces $\mathscr{B}_{\mathcal{G}_r}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^\perp$ et $\mathscr{B}_{\mathcal{G}_c}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^\perp$ avec deux composantes : la projection dans l'intersection de la boule de groupe et de la boule-ℓ_2 et la projection dans l'espace orthogonal aux triplets singuliers déjà estimés.

Projection dans l'intersection de la boule de groupe et de la boule ℓ_2

Nous nous intéressons à un algorithme de projection de $\mathbf{x}$, un vecteur fixé de $\mathbb{R}^n$ divisé en K groupes non chevauchants, dans l'intersection de la boule $\ell_\mathcal{G}$ de rayon c et de la boule ℓ_2 de rayon 1 issu de [Guillemot et al., 2020]. Cette méthode généralise la projection dans l'espace $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ [Gloaguen et al., 2017].

On note $\mathcal{G} = \{\iota_k, k = 1, ..., K\}$ une partition de l'ensemble des entiers allant de 1 à n, où K est le nombre de groupes, ι_k les indices des éléments du $k^{\text{ème}}$ groupe et n est égal à I ou J selon que l'on considère des groupes d'individus ou des groupes de variables. On a :

$$\text{proj}_{\mathscr{B}_\mathcal{G}(c) \cap \mathscr{B}_2(1)}(\mathbf{x}) = S_\mathcal{G}(\mathbf{x}, \lambda^*) \text{ avec } \lambda^* \in \mathbb{R}^+ \text{ tel que } \frac{\|S_\mathcal{G}(\mathbf{x}, \lambda^*)\|_\mathcal{G}}{\|S_\mathcal{G}(\mathbf{x}, \lambda^*)\|_2} = c \tag{15}$$

où $S_\mathcal{G}$ est une généralisation de la fonction de seuillage doux à la norme de groupe, qui s'exprime :

$$\forall k \in [1, K], \ [S_\mathcal{G}(\mathbf{x}, \lambda)]_{\iota_k} = \begin{cases} \mathbf{0} \text{ si } \|\mathbf{x}_{\iota_k}\|_2 < \lambda \\ \left(1 - \dfrac{\lambda}{\|\mathbf{x}_{\iota_k}\|_2}\right) \mathbf{x}_{\iota_k} \text{ sinon} \end{cases} \tag{16}$$

où $\mathbf{x}_{\iota_k}$ est un vecteur contenant les éléments du $k^{\text{ème}}$ groupe. Soit $\mathbf{v}$ le vecteur contenant les normes ℓ_2 des sous-vecteurs $\mathbf{x}_{\iota_k}$ pour $k = 1, ..., K$:

$$\mathbf{v} = \begin{bmatrix} \|\mathbf{x}_{\iota_1}\|_2 \\ \vdots \\ \|\mathbf{x}_{\iota_K}\|_2 \end{bmatrix}$$

Alors, d'après les formules de $S_\mathcal{G}$ et S des équations 3 et 16, on obtient :

$$\forall \lambda \in \mathbb{R}^+, \frac{\|S_\mathcal{G}(\mathbf{x}, \lambda)\|_\mathcal{G}}{\|S_\mathcal{G}(\mathbf{x}, \lambda)\|_2} = \frac{\|S(\mathbf{v}, \lambda)\|_1}{\|S(\mathbf{v}, \lambda)\|_2} \tag{17}$$

Projeter $\mathbf{x}$ dans $\mathscr{B}_\mathcal{G}(c) \cap \mathscr{B}_2(1)$ revient donc à résoudre la même équation que pour projeter $\mathbf{v}$ dans $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$, ce qui peut être effectué avec l'algorithme présenté dans [Gloaguen et al., 2017].

4 Résultats

Afin d'illustrer la méthode de l'ACM parcimonieuse, nous analyserons un jeu de données issu d'un sondage à l'aide de l'ACM classique et de l'ACMP. Cette analyse nous permettra de comparer les résultats des deux analyses et de montrer comment les interpréter. Nous analyserons également des données génétiques avec l'ACM et l'ACMP, ce qui nous permettra de montrer l'intérêt de cette méthode pour l'analyse de ce type de données.

4.1 Analyse d'un questionnaire sur le Maroilles

Dans cette section, nous analyserons le jeu de données 'cheese' issu du package R4SPISE2018. La version des données utilisée est le jeu de données nettoyé (cleandata) importé dans le package SMCA.

TABLE 3 – Extrait du jeu de données *cheese*

Sex	Age	City	Know	C01	C05	...
f	1	Angers	2	2	1	...
f	3	Angers	2	1	1	...
m	3	Lille	2	2	2	...

cheese est un jeu de données issu d'un sondage à propos du Maroilles. Les observations sont les personnes interrogées.

Le sondage est composé de deux ensembles de questions :

- Le premier ensemble évalue la connaissance de la personne sondée sur le sujet. Ces questions ont été résumées en une seule variable 'Know', qui indique la connaissance de la personne sur une échelle de 1 à 4 selon le nombre de bonnes réponses aux questions (table 3).
- Le deuxième ensemble évalue le comportement ou l'opinion de la personne sondée par rapport au Maroilles industriel et artisanal. La personne doit répondre à chaque question sur une échelle de 1 à 4 (1 étant « je suis totalement d'accord » et 4 « je ne suis pas du tout d'accord »). Les réponses n'étant pas très équilibrées, elles ont été binarisées (avec $(1, 2) \longrightarrow 1$ « je suis d'accord » et $(3, 4) \longrightarrow 2$ « je ne suis pas d'accord »). Seules les questions pouvant avoir une réponse binaire ont été gardées (C01, C05, C07, C10, C11, C18).

Certaines informations sur la personne sondée ont été récoltées : leur sexe, leur âge (variable catégorielle avec quatre classes d'âge) et leur ville d'origine (Lille ou Angers).

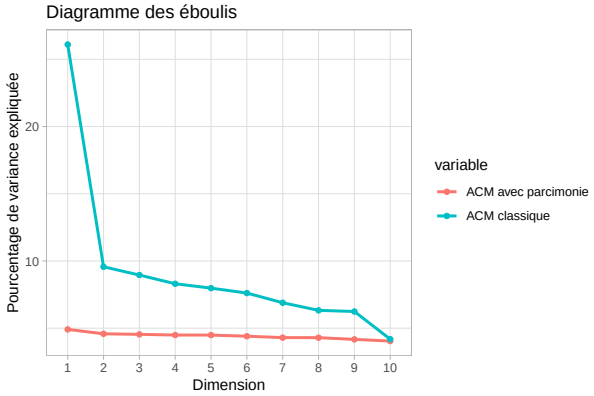
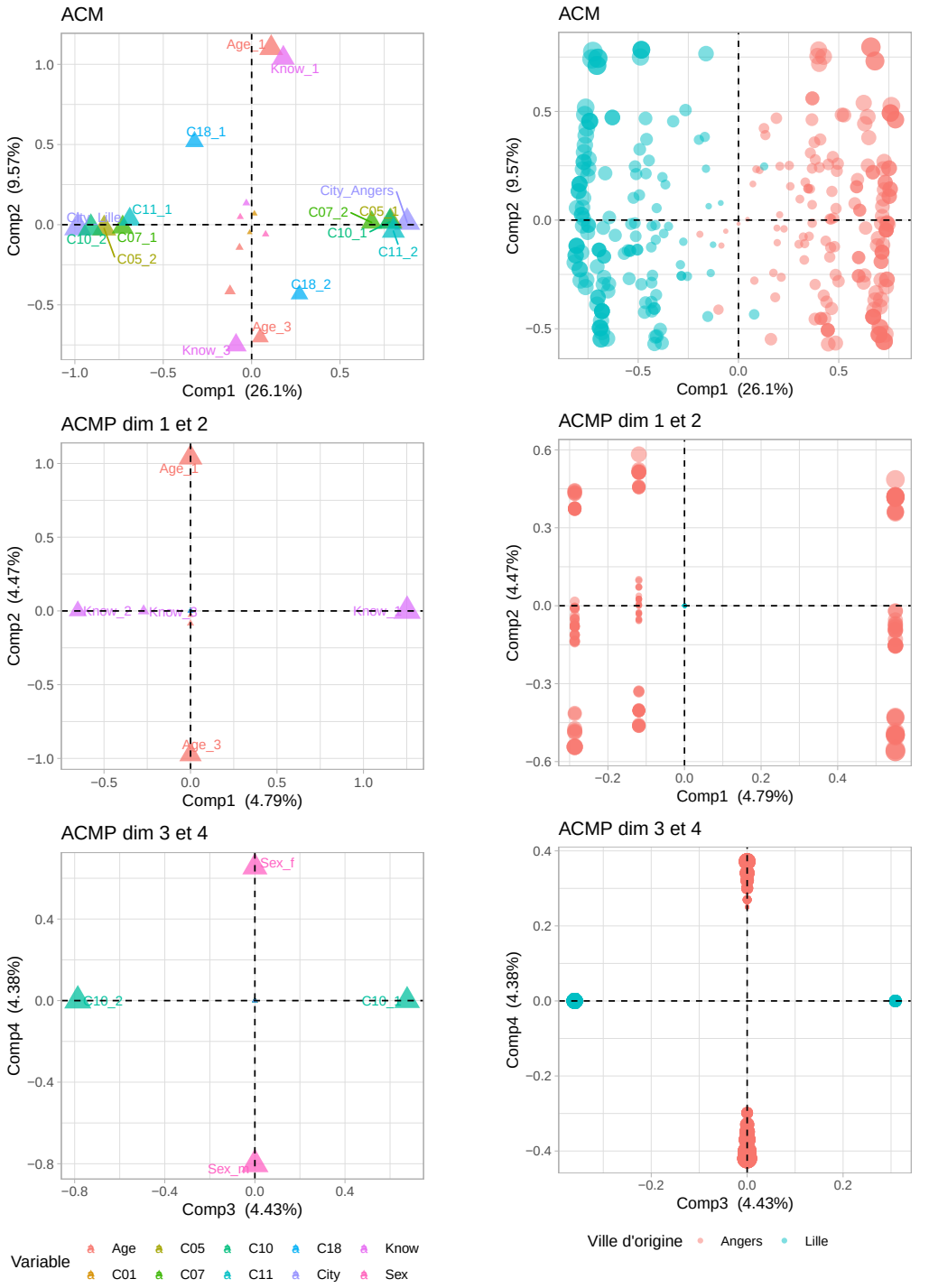


FIGURE 4 – Graphe des pourcentages de variance expliquées par les dimensions de l'ACMP (en rouge) et de l'ACM classique (en bleu).

Une ACM classique, ainsi qu'une ACMP sont appliquées sur le jeu de données. Pour l'ACMP, on regroupe les rangs par la ville d'origine des individus et les colonnes par variable. On donne à l'ACMP une contrainte de groupe maximale sur les rangs et sur les colonnes ($c_1 = c_2 = 1$) qui impose qu'un seul groupe d'individus et une seule variable sont sélectionnés par dimension.

Le diagramme des éboulis de la figure 4 montre que la première dimension de l'ACM classique capture une grande partie de la variabilité des données, tandis que pour l'ACM parcimonieuse, on observe des pseudo-valeurs propres presque toutes égales et bien plus



(a) Graphe des modalités. (b) Graphes des individus.

FIGURE 5 – Graphe des modalités et des individus de l'ACM sur les deux premières dimensions et de l'ACM parcimonieuse sur les quatre premières dimensions (plus un point est gros, plus l'individu ou la modalité contribue à la dimension).

faibles. Nous commenterons donc les quatre premières dimensions de l'ACMP et seulement les deux premières de l'ACM.

La première dimension obtenue avec l'ACM classique affichée sur la figure 5 nous montre que la ville d'origine est la première source de variabilité de ce jeu de données (le graphe des individus de l'ACM de la figure 5b illustre l'opposition entre les individus d'Angers et ceux de Lille). Un des intérêts de l'ACMP est de montrer la variabilité intra villes, car pour chaque dimension, un seul groupe d'individus est sélectionné (sur la figure 5b, on n'observe que des individus d'Angers sur les trois premières dimensions et de Lille sur la quatrième).

Le graphe des modalités (figure 5a) de l'ACM classique nous montre en première dimension une opposition d'opinion concernant le Maroilles (les variables C10, C05 ,C07 et C11 sont bien représentées) entre les habitants des deux villes et en deuxième dimension une opposition de connaissance liée à l'âge pour tous les individus.

Les résultats de l'ACM parcimonieuse (figures de l'ACMP de 5a) montrent que l'âge, puis la connaissance ('Knowledge') et le genre sont des composantes importantes de la structure du groupe d'Angers, tandis que le groupe de ville d'origine Lille, associé à la quatrième dimension, est structuré par une variable d'opinion (C10). Ces résultats représentent le cas extrême de parcimonie appliquée sur une ACM et pourraient être plus complexes avec une contrainte de parcimonie moins élevée.

4.2 Analyse de données génétiques : ADNI

Afin de montrer l'intérêt que pourrait avoir l'ACMP pour l'analyse de données génomiques, nous analyserons dans cette partie un jeu de données issu du projet ADNI (*Alzheimer's Disease Neuroimaging Initiative*) [Mueller et al., 2005], avec l'ACMP. ADNI est une étude multicentrique conçue pour développer des biomarqueurs de différentes natures permettant la détection précoce et le suivi de la maladie d'Alzheimer.

L'analyse de données génétiques avec l'ACP demande de recoder les SNPs (*Single Nucleotide Polymorphism*), qui sont des variables catégorielles, en variables numériques, par exemple avec la méthode du dosage allélique [Foulkes, 2009]. Cependant, avec cette technique, on suppose que l'effet du variant génétique est linéaire, ce qui n'est pas toujours le cas. L'ACMP pourrait permettre d'analyser des données génétiques sans cette contrainte.

Le jeu de données étudié est composé de 791 sujets qui ont été diagnostiqués sur la maladie d'Alzheimer (voir [Association. and Association., 2013]) et dont le génome a été séquencé. Il existe cinq diagnostics différents :

- *CN* (*Cognitively normal*), les individus témoins ;
- *SMC* (*Significant Memory Concern*), les individus ayant des troubles de mémoire ;
- *EMCI* (*Early Mild Cognitive Impairement*), les individus aux premiers stades de troubles cognitifs ;
- *LMCI* (*Late Mild Cognitive Impairement*), les individus avec des troubles cognitifs avancés ;
- *AD* (*Alzheimer's disease*), les individus atteints de la maladie d'ALzheimer.

Les 134 variables génétiques sont des SNPs. Un SNP, ou polymorphisme d'un seul nucléotide, est une paire de base du génome qui varie entre les individus d'une même espèce. Chaque élément est donc un facteur composé de deux lettres parmi A, T, C et G (bases de l'ADN). Les facteurs rares ont été regroupés (*Minor Allele Frequency*) pour qu'ils n'aient pas trop de poids dans notre analyse [Beaton et al., 2016]. Cet ensemble de SNPs est un extrait de la base de données génomiques ADNI. Seuls les SNPs associés avec les gènes

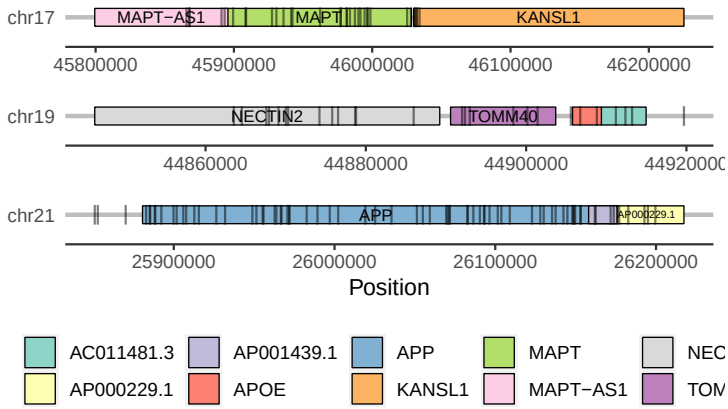
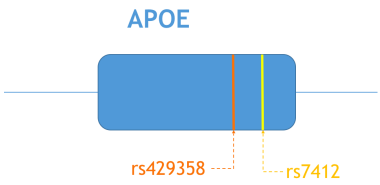
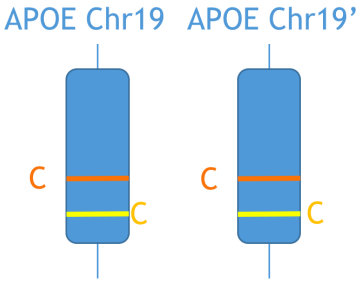


FIGURE 6 – Schéma des gènes présents dans le jeu de données. La position des SNP est indiquée par une barre.

MAPT, APP, ApoE et TOMM40 ont été sélectionnés (figure 6) car ces gènes sont considérés comme des contributeurs potentiels de plusieurs pathologies de la maladie d'Alzheimer d'après [Beaton et al., 2016].



(a) Schéma du gène ApoE et de deux SNP pouvant porter un facteur de risque de la maladie d'ALzheimer, [Roses et al., 2016].



(b) Combinaison allélique qui augmente le risque d'Alzheimer d'après [Cariaso and Lennon, 2011].

SNP	Option 1	Option 2
rs429358	*T*	*C*
rs7412	*T*	*C*

(c) Tableau des allèles possibles des SNP de ApoE.

FIGURE 7 – Le gène ApoE (chromosome 19).

Nous tenterons de voir si des marqueurs identifiés comme facteur de risque de la maladie d'Alzheimer, comme certains variants génétiques de ApoE (voir la figure 7), peuvent être observés avec l'ACMP.

Une ACM classique, ainsi qu'une ACMP sont appliquées sur le jeu de données. Pour l'ACMP, on regroupe les individus par leur diagnostic et les colonnes par SNP. On donne à l'ACMP une contrainte de groupe élevée sur les rangs et sur les colonnes ($c_1 = 0.06 \times \sqrt{I}$ et $c_2 = 0.07 \times \sqrt{J}$) afin d'observer une séparation la plus franche possible entre les diagnostics. Dans le cas de données génétiques, la plus grande part de variabilité est souvent liée à la dérive génétique entre les populations, d'après [Coop et al., 2009]. Nous avons donc fait le choix de ne pas mettre de parcimonie sur les deux premières dimensions (comme on peut le voir sur la figure 8) pour qu'elle capte en grande partie cette information. Ces résultats sont disponibles sur l'application shiny suivante : `https://jleborgne.shinyapps.io/smca_adni/` .

Il s'agit d'identifier les dimensions présentant un groupe d'individus malades bien repré-

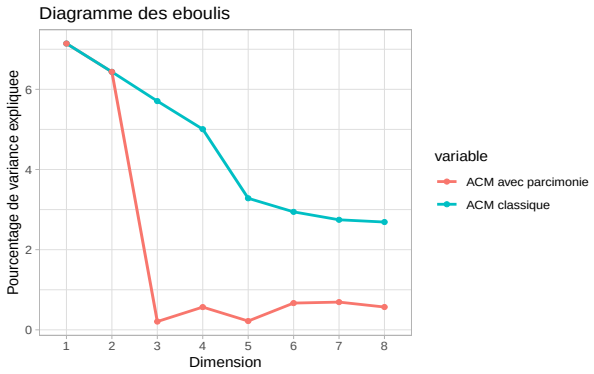
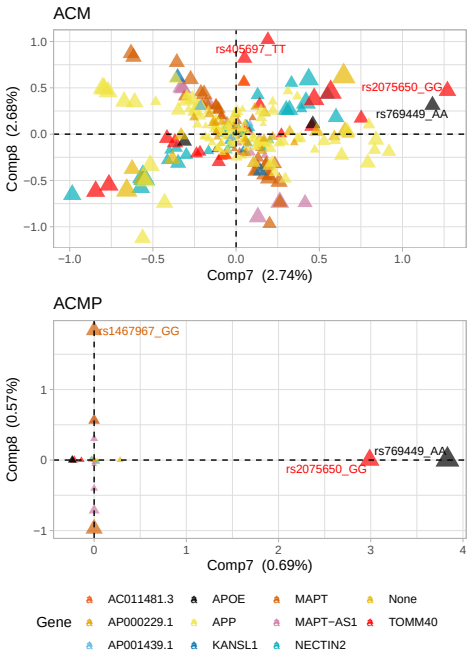


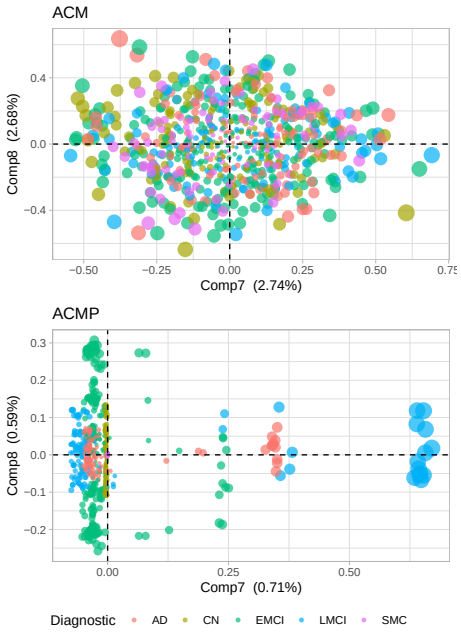
FIGURE 8 – Graphe des pourcentages de variance expliquée par les dimensions de l'ACMP (en rouge) et de l'ACM classique (en bleu) des données ADNI.

senté, puis de regarder quels allèles, issus de quels gènes, lui sont associés. Dans cette analyse, on a pu observer une telle situation sur la dimension 7. En effet, sur la figure 9b, on peut voir un groupe de *AD* et de *LMCI* (les deux diagnostics les plus sévères) se distinguer sur la dimension 7 de l'ACMP. Sur le graphe des modalités de la figure 9a de l'ACMP, deux modalités sont bien représentées ; l'une est issue du gène APOE, en noir, et l'autre de TOMM40, en rouge (ces gènes se trouvent tous deux sur le chromosome 19 et sont très liés, [Cariaso and Lennon, 2011]). La figure 9c montre également que les gènes APOE et TOMM40 sont bien représentés sur la dimension 7. Les modalités bien représentées sur la dimension 7 sont les combinaisons GG du SNP *rs2075650* et AA de *rs769449*. D'après [Cariaso and Lennon, 2011], le nucléotide G sur *rs2075650* est un facteur de risque de troubles cognitifs et le nucléotide A sur *rs769449* est un facteur de risque de la maladie d'Alzheimer. Notons que ce phénomène est également observable sur l'ACM classique de la figure 9, mais on ne peut pas voir de lien clair entre les diagnostics et les gènes car beaucoup d'individus et de variables sont bien représentés sur cette dimension.

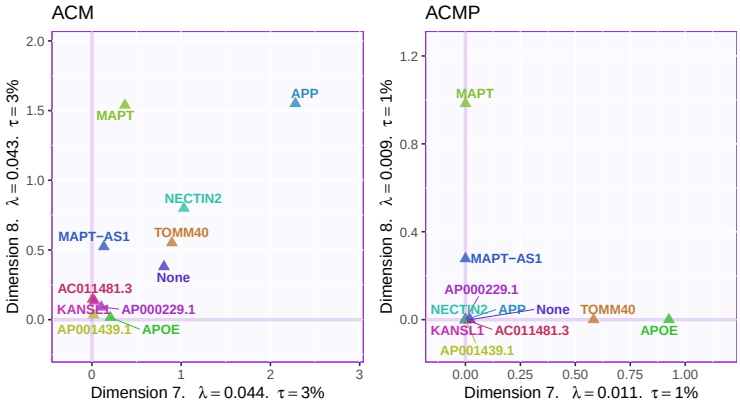
Cependant, si l'on observe la figure 9d de l'ACMP, qui montre les barycentres et un ellipsoïde de confiance de chaque groupe d'individus, on peut voir que les barycentres des groupes *AD* et *LMCI* sur la dimension 7 sont proches de zéro. L'association avec les gènes APOE et TOMM40 ne concerne qu'une petite partie des individus *AD* et *LMCI*.



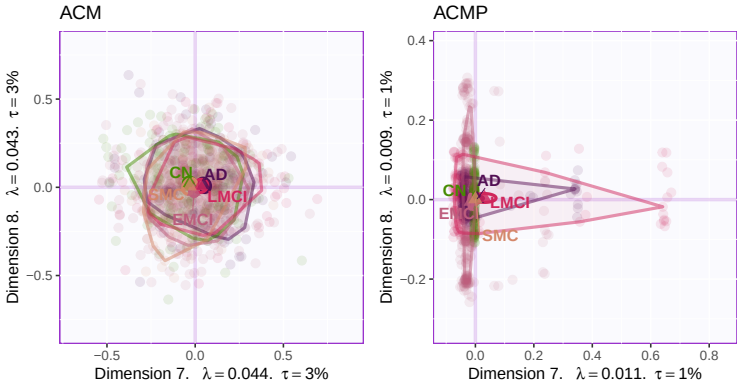
(a) Graphe des modalités



(b) Graphes des Individus



(c) Graphe des Gènes (somme des η^2 des SNPs).



(d) Graphe des individus de l'ACM et de l'ACMP avec les barycentres (et leur intervalle de confiance) des groupes de diagnostics.

FIGURE 9 – Graphes des individus et des modalités (versions des SNP) de l'ACM et de l'ACM parcimonieuse sur les dimensions 7 et 8.

5 Conclusion et perspectives

L'ACMP est une version parcimonieuse de l'ACM qui incorpore dans la GSVD une contrainte de groupe sur les modalités par variable ou sur les individus. Cette nouvelle méthode d'analyse exploratoire de données qualitatives peut simplifier l'interprétation des dimensions factorielles ou encore révéler une vue plus complète de l'analyse.

Les perspectives possibles de ce projet seront de prendre en compte des groupes chevauchants ou une structure de groupe hiérarchique sur les variables ou sur les rangs, ce qui peut être utile notamment pour l'analyse de données SNP, souvent liés à des gènes qui sont structurés en réseaux de régulation.

L'ACM est une méthode statistique largement utilisée grâce à ses propriétés permettant une visualisation informative des données. Nous avons donc tenté de retrouver des propriétés équivalentes dans la version parcimonieuse. Par exemple, nous avons retrouvé des propriétés barycentriques en ACMP, grâce à un équivalent de la formule de transition. Ces propriétés assurent la symétrie entre les observations et les variables dans notre analyse. Cependant, d'autres propriétés importantes ne semblent plus vérifiées, comme l'équivalence distributionnelle ou la projection d'éléments supplémentaires, ce qui peut poser de nombreux problèmes dans l'utilisation de cette méthode. Des travaux sont en cours pour transposer ces propriétés dans le cadre de l'ACMP.

Ce stage m'a permis d'acquérir certaines compétences comme la compréhension du développement de méthode statistiques, mais aussi ses limites, notamment avec la perte de certaines propriétés de l'ACM.

Cette expérience m'a permis de me confronter au monde de la recherche bio-informatique, et m'a donné envie de continuer dans cette voie sur une thèse. L'objet de ma thèse ne portera pas sur la programmation de nouvelles méthodes mais plutôt de l'analyse de données génétiques. En ce sens, mon stage m'apporte un autre point de vue sur l'analyse de ce type de données et une capacité à me confronter à de nouvelles méthodes.

Annexe : Projection dans la boule-ℓ_1 de rayon c

On cherche l'expression de la projection d'un vecteur $\mathbf{x} \in \mathbb{R}^n$ dans $\mathscr{B}_1(c) = \{\mathbf{x} \in E, \|\mathbf{x}\|_1 \leq c\}$, avec c une constante positive.

Si $\|\mathbf{x}\|_1 \leq c, proj_{\mathscr{B}_1(c)}(\mathbf{x}) = \mathbf{x}$. Sinon, la projection consiste à résoudre le problème suivant :

$$proj_{\mathscr{B}_1(c)}(\mathbf{x}) = \underset{\mathbf{y}\in\mathbb{R}^n}{\arg\min} \tfrac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \text{ tel que } \|\mathbf{y}\|_1 \leq c \tag{18}$$

On note $\mathbf{x}^*$ la solution de l'équation 18. Pour trouver $\mathbf{x}^*$, on passe par la minimisation du Lagrangien associé :

$$\begin{aligned} \mathscr{L}(\mathbf{y}; \lambda) &= \tfrac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 - \lambda(\|\mathbf{y}\|_1 - c) \\ &= \tfrac{1}{2} \sum_{i=1}^{n} (x_i - y_i)^2 + \lambda \sum_{i=1}^{n} (|y_i| - c) \end{aligned} \tag{19}$$

On minimise $\mathscr{L}$ par rapport à $\mathbf{y}$, puis la solution $\mathbf{x}^*$ sera obtenue pour la valeur de λ telle que $\|\mathbf{x}\|_1 \leq c$. Cette valeur de λ sera notée λ^*.

Soit les dérivées partielles de $\mathscr{L}$ par rapport à $y_i, \forall i \in [1, n]$:

$$\frac{\partial \mathscr{L}}{\partial y_i}(\mathbf{y}, \lambda) = (y_i - x_i) + \lambda \frac{\partial |y_i|}{\partial y_i}$$

où

$$\frac{\partial |y_i|}{\partial y_i} = \begin{cases} -1 \text{ si } y_i < 0 \\ 1 \text{ si } y_i > 0 \end{cases} \quad \text{et } \frac{\partial |y_i|}{\partial y_i} \in [-1, 1] \text{ si } y_i = 0 \tag{20}$$

On veut résoudre :

$$\frac{\partial \mathscr{L}}{\partial y_i} = 0 \tag{21}$$

On distingue trois cas :

- Si $\underline{y_i > 0}$: $x_i^* = x_i - \lambda^*$
- Si $\underline{y_i < 0}$: $x_i^* = x_i + \lambda^*$
- Si $\underline{y_i = 0}$: $x_i^* = 0$

Ce troisième cas n'est vérifié que si $|x_i| < \lambda$. En effet, d'après les équations 21 et 20 :

$$y_i = 0 \Leftrightarrow -1 \leq \frac{\partial |y_i|}{\partial y_i} \leq 1 \Leftrightarrow -1 \leq \frac{x_i}{\lambda} \leq 1 \Leftrightarrow x_i \in [-\lambda, \lambda] \tag{22}$$

Ainsi,

$$\forall i \in [1, n], x_i^* = \begin{cases} 0 \text{ si } |\mathbf{x}_i| < \lambda^* \\ \mathbf{x}_i - \lambda^* \text{ si } \mathbf{x}_i > \lambda^* \\ \mathbf{x}_i + \lambda^* \text{ si } \mathbf{x}_i < -\lambda^* \end{cases} \quad \text{avec } \lambda^* \text{ tel que } \|x_i^*\|_1 \leq c \tag{23}$$

On reconnait l'expression de la fonction de seuillage doux, on a donc $\mathbf{x}^*(\lambda^*) = S(\mathbf{x}, \lambda^*)$.

Références

[Abdi, 2007] Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition (gsvd). In Salkind, N., editor, *Encyclopedia of Measurement and Statistic*, pages 907–912. Sage, Thousand Oaks (CA).

[Abdi and Valentin, 2007] Abdi, H. and Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*.

[Association. and Association., 2013] Association., A. P. and Association., A. P. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5*. American Psychiatric Association Arlington, VA, 5th ed. edition.

[Beaton et al., 2016] Beaton, D., Dunlop, J., and Abdi, H. (2016). Partial least squares correspondence analysis : A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, 21(4) :621–651. Place : US Publisher : American Psychological Association.

[Cariaso and Lennon, 2011] Cariaso, M. and Lennon, G. (2011). SNPedia : a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*, 40(D1) :D1308–D1312.

[Combettes, 1993] Combettes, P. (1993). The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2) :182–208.

[Coop et al., 2009] Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., and Pritchard, J. K. (2009). The Role of Geography in Human Adaptation. *PLoS Genetics*, 5(6) :e1000500.

[Foulkes, 2009] Foulkes, A. S. (2009). *Applied Statistical Genetics with R*. Springer New York.

[Gloaguen et al., 2017] Gloaguen, A., Guillemot, V., and Tenenhaus, A. (2017). An efficient algorithm to satisfy ℓ_1 and ℓ_2 constraints. In *49èmes Journées de statistique*, Avignon, France.

[Guillemot et al., 2019] Guillemot, V., Beaton, D., Gloaguen, A., Löfstedt, T., Levine, B., Raymond, N., Tenenhaus, A., and Abdi, H. (2019). A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE*, 14(3) :e0211463.

[Guillemot et al., 2020] Guillemot, V., Le Borgne, J., Gloaguen, A., Tenenhaus, A., Saporta, G., Chollet, S., Beaton, D., and Abdi, H. (2020). Sparse Multiple Correspondence Analysis. In *Accepté dans les 52èmes Journées de statistique*, Nice, France.

[Mori et al., 2016] Mori, Y., Kuroda, M., and Makino, N. (2016). Sparse Multiple Correspondence Analysis. In *Nonlinear Principal Component Analysis and Its Applications*, pages 47–56. Springer Singapore, Singapore. Series Title : SpringerBriefs in Statistics.

[Mueller et al., 2005] Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., and Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *NEUROIMAGING CLINICS OF NORTH AMERICA*, 15(4) :869+.

[Roses et al., 2016] Roses, A., Sundseth, S., Saunders, A., Gottschalk, W., Burns, D., and Lutz, M. (2016). Understanding the genetics of apoe and tomm40 and role of mitochondrial structure and function in clinical pharmacology of alzheimer's disease. *Alzheimer's & Dementia*, 12(6) :687 – 694.

[Saporta, 2011] Saporta, G. (2011). *Probabilités, Analyse des Données et Statistique*. Technip, Paris, France, 3rd edition.

[Saporta et al., 2012] Saporta, G., Bernard, A., and Guinot, C. (2012). A generalisation of sparse PCA to multiple correspondence analysis. In *ERCIM 2012*, Oviedo, Spain.

[Witten et al., 2009] Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3) :515–534.

l'institut Agro agriculture • alimentation • environnement	Diplôme : Ingénieur Spécialité : Agronome Spécialisation / option : Data Science Enseignant référent : David Causeur

Auteur(s) : Julie Le Borgne Date de naissance* : 24/01/1997	Organisme d'accueil : Institut Pasteur Adresse : 25 Rue du Dr Roux, 75015 Paris
Nb pages : 19 Annexe(s) : 1	
Année de soutenance : 2020	Maître de stage : Vincent Guillemot

Titre français : Analyse des Correspondances Multiples Parcimonieuses

Titre anglais : Sparse Multiple Correspondance Analysis

Résumé (1600 caractères maximum) :

Ce rapport de stage présente l'Analyse des Correspondances Multiples Parcimonieuse (ACMP), l'équivalent parcimonieux de l'Analyse des Correspondances Multiples (ACM). L'ACM est une méthode largement utilisée pour l'analyse exploratoire de données catégorielles. L'ajout de parcimonie dans une ACM améliore son interprétabilité, en particulier pour l'analyse de données de grande dimension. Basée sur la décomposition en valeurs singulières (SVD), l'ACM pourra être rendue parcimonieuse en généralisant l'algorithme de la SVD contrainte (CSVD). Plus particulièrement, la CSVD nécessite deux propriétés supplémentaires : prendre en compte les matrices de masse et de poids caractéristiques de l'ACM (poids des individus et fréquence des modalités) et sélectionner des groupes entiers de variables (un groupe étant constitué du codage disjonctif complet d'une variable catégorielle) ou d'observations.

Nous illustrons l'ACMP avec deux jeux de données : le premier est un questionnaire sur la perception et la connaissance du Maroilles et le deuxième un jeu de données génétiques extrait d'une étude sur la maladie d'Alzheimer. Ces deux exemples nous permettent de montrer l'intérêt de la contrainte de parcimonie pour l'interprétation des dimensions estimées, ainsi que pour l'exploration des liens entre individus et modalités.

Abstract (1600 caractères maximum) :

This report presents the Sparse Multiple Correspondence Analysis (SMCA), the sparse equivalent of Multiple Correspondence Analysis (MCA). MCA is a widely used method for exploratory analysis of categorical data. Adding sparsity to an MCA improves its interpretability, especially for the analysis of large data. Based on singular value decomposition (SVD), MCA can be made sparse by generalizing the constrained SVD algorithm (CSVD). More specifically, CSVD requires two additional properties: to take into account the mass and weight matrices that are characteristic of the MCA (weight of individuals and frequency of modalities) and to select entire groups of variables (one group consisting of the complete disjunctive coding of a categorical variable) or of observations.

We illustrate the SMCA with two sets of data: the first is a questionnaire on perception and knowledge of Maroilles and the second is a set of genetic data taken from a study on Alzheimer's disease. These two examples allow us to show the interest of the sparsity constraint for the interpretation of the estimated dimensions, as well as for the exploration of the links between individuals and modalities.

Mots-clés : Parcimonie, Analyse Multivariée, Correspondances Multiples, Alzheimer

Key Words: Sparsity, Multivariate Analysis, Multiple Correspondence, Alzheimer

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*

Document à intégrer au mémoire