Développement de méthodes statistiques nécessaires à l'analyse de données génomiques : Application à l'influence du polymorphisme génétique sur les caractéristiques cutanées individuelles et l'expression du vieillissement cutané

Anne Bernard, Saporta Gilbert

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS | le cnam

École Doctorale Informatique, Télécommunications et Électronique (EDITE)

Laboratoire CEDRIC

THÈSE DE DOCTORAT

présentée par : **Anne BERNARD**

soutenue le : **20 décembre 2013**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline / Spécialité : **Informatique**

Développement de méthodes statistiques nécessaires à l'analyse de données génomiques

application à l'influence du polymorphisme génétique sur les caractéristiques cutanées individuelles et l'expression du vieillissement cutané

THÈSE CO-DIRIGÉE PAR

SAPORTA Gilbert	*Professeur, CNAM, Paris*
GUINOT Christiane	*Docteur/HDR, Université François Rabelais, Tours*

RAPPORTEURS

BESSE Philippe	*Professeur, INSA, Toulouse*
SABATIER Robert	*Professeur, Université de Montpellier*

EXAMINATEURS

ABDI Hervé	*Professeur, The University of Texas at Dallas, USA*
LATREILLE Julie	*Docteur, Chanel, Paris*
TENENHAUS Arthur	*Professeur, SUPELEC, Gif-sur-Yvette*
ZAGURY Jean-François	*Professeur, CNAM, Paris*

"À mes parents,"

Remerciements

Je tiens tout d'abord à remercier très chaleureusement mes deux directeurs de thèse, le Professeur Gilbert Saporta et le Docteur Christiane Guinot pour leur soutien, leur disponibilité et l'aide précieuse qu'ils m'ont apportée tout au long de cette thèse. L'intérêt qu'ils ont porté à mon travail et la confiance qu'ils m'ont accordée m'ont permis de m'épanouir durant ces trois années agréables et passionnantes. Je tiens à leur adresser, ainsi qu'aux Professeurs Hervé Abdi et Arthur Tenenhaus, toute ma reconnaissance pour leurs conseils et les connaissances statistiques qu'ils m'ont permis d'acquérir ainsi que le temps précieux qu'ils m'ont consacré, même à plus de 8 000 km de distance. Travailler à leur côté a été une chance et un réel plaisir.

Je remercie également très sincèrement les Professeurs Philippe Besse et Robert Sabatier qui m'ont fait l'honneur d'accepter d'être les rapporteurs de ce travail. Je les remercie pour le temps qu'ils ont accordé à la relecture de ce manuscrit et pour l'intérêt qu'ils lui ont accordé.

Un grand merci au Professeur Jean-François Zagury pour le réel enthousiasme dont il a toujours fait part et pour ses précieux conseils et un grand merci à toute son équipe, en particulier Sigrid, Cédric, Taoufik, Vincent et Christiane pour leur gentillesse et leur aide.

Je tiens à remercier chaleureusement TOUTE l'équipe du CE.R.I.E.S. pour leur soutien, leur gentillesse et pour ces quatre années de bonheur passées à leur côtés. Un grand merci à Julie Latreille, Emmanuelle Mauger et la petite dernière Frédérique Soppelsa pour le plaisir qu'elles m'ont donné à venir travailler tous les jours à leurs côtés.

Et puis à côté de cela il y a les ami(e)s, en particulier Sandra, Emilie, Hanna, Julie et Claude qui ont toujours cru en moi et m'ont soutenu malgré la distance. Des amitiés longues et sincères qui sont une vraie source d'énergie au quotidien. Des ami(e)s et bien plus... Merci à Pierre-Hakim pour son soutien au quotidien, merci de m'avoir permis d'avancer et de me dépasser chaque jour un peu plus.

Mes pensées les plus profondes vont à ma famille, ma grand-mère, et en particulier à mes parents qui sont une vraie source de bonheur, ma plus grande fierté et la raison de ces aboutissements.

Toutes les personnes que j'ai eu la chance de croiser et qui m'ont entourées durant ces trois années m'ont permis de passer des moments formidables et d'apprécier chaque jour la chance que j'avais d'être aussi bien "tombée". Merci à toutes ces personnes pour leur gentillesse, leur humilité et leur si grand savoir qui m'a toujours impressionné...

Résumé

Les nouvelles technologies développées ces dernières années dans le domaine de la génétique ont permis de générer des bases de données de très grande dimension, en particulier de Single Nucleotide Polymorphisms (SNPs), ces bases étant souvent caractérisées par un nombre de variables largement supérieur au nombre d'individus. L'objectif de ce travail a été de développer des méthodes statistiques adaptées à ces jeux de données de grande dimension et permettant de sélectionner les variables les plus pertinentes au regard du problème biologique considéré. Dans la première partie de ce travail, un état de l'art présente différentes méthodes de sélection de variables non supervisées et supervisées pour 2 blocs de variables et plus. Dans la deuxième partie, deux nouvelles méthodes de sélection de variables non supervisées de type "sparse" sont proposées : la Group Sparse Principal Component Analysis (GSPCA) et l'Analyse des Correspondances Multiples sparse (ACM sparse). Vues comme des problèmes de régression avec une pénalisation group LASSO elles conduisent à la sélection de blocs de variables quantitatives et qualitatives, respectivement. La troisième partie est consacrée aux interactions entre SNPs et dans ce cadre, une méthode spécifique de détection d'interactions, la régression logique, est présentée. Enfin, la quatrième partie présente une application de ces méthodes sur un jeu de données réelles de SNPs afin d'étudier l'influence possible du polymorphisme génétique sur l'expression du vieillissement cutané au niveau du visage chez des femmes adultes. Les méthodes développées ont donné des résultats prometteurs répondant aux attentes des biologistes, et qui offrent de nouvelles perspectives de recherches intéressantes.

Mots clés : sélection de variables, ACP sparse, ACM, SNP-SNP interactions, régression logique, méthodes multiblocs, méthodes sparse non supervisées.

Abstract

New technologies developed recently in the field of genetic have generated high-dimensional databases, especially SNPs databases. These databases are often characterized by a number of variables much larger than the number of individuals. The goal of this dissertation was to develop appropriate statistical methods to analyse high-dimensional data, and to select the most biologically relevant variables.

In the first part, I present the state of the art that describes unsupervised and supervised variables selection methods for two or more blocks of variables. In the second part, I present two new unsupervised "sparse" methods: Group Sparse Principal Component Analysis (GSPCA) and Sparse Multiple Correspondence Analysis (Sparse MCA). Considered as regression problems with a group LASSO penalization, these methods lead to select blocks of quantitative and qualitative variables, respectively. The third part is devoted to interactions between SNPs. A method employed to identify these interactions is presented: the logic regression. Finally, the last part presents an application of these methods on a real SNPs dataset to study the possible influence of genetic polymorphism on facial skin aging in adult women. The methods developed gave relevant results that confirmed the biologist's expectations and that offered new research perspectives.

Keywords: feature selection, sparse PCA, MCA, SNP-SNP interactions, logic regression, multiblocks methods, unsupervised sparse methods.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

ACP Analyse en Composantes Principales 50

ADN Acide DésoxyriboNucléique 213

AFC Analyse Factorielle des Correspondances 183

AFCM Analyse Factorielle des Correspondances Multiples 183

AIC Akaïke's Information criterion 37

AID Automatic Interaction Detection 110

AUC Area Under Curve 125

AVE Average Variance Explained 161

BIC Bayesian Information Criterion 38

BioGRID Biological General Repository for Interaction Datasets 111

bp base pairs 215

CART Classification And Regression Tree 110

CCA Canonical Correlation Analysis 67

CCPPRB Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale 222

CE.R.I.E.S. CEntre de Recherche et d'Investigation Epidémiologique et Sensorielle . . 131

CHAID CHi-squared Automatic Interaction Detector 122

CNAM Conservatoire National des Arts et Métiers 131

CNIL Commision Nationale de l'Informatique et des Libertés 222

CNV Copy Number Variation 225

CRAN The Comprehensive R Archive Network

CP Composante Principale 36

Introduction

Depuis quelques années la génomique a été bouleversée grâce à des outils technologiques récemment commercialisés : les puces de génotypage. L'utilisation de ces puces connaît un essor croissant car elles permettent de cribler et de mesurer l'information contenue non plus dans 1, 2, ou 10 gènes mais dans les 25 000 gènes du génome. Souvent utilisées dans le domaine de la cancérologie, ces nouveaux supports ont permis de caractériser et d'analyser différents processus biologiques et biophysiques fondamentaux.

Le CE.R.I.E.S. mène depuis 2002 une série d'études sur les relations qui lient le polymorphisme génétique aux caractéristiques et à l'expression du vieillissement cutané. Une approche "gènes candidats" a d'abord été adoptée et différents gènes ont été étudiés (Elfakir et al. [2009], Latreille et al. [2009], Latreille et al. [2011],Le Clerc et al. [2012], Ezzedine et al. [2012], Jdid et al. [2013]). En 2010 une étude d'association génome entier a été mise en place. Les puces de génotypage ont été utilisées et des analyses SNP par SNP ont été réalisées. Les premiers résultats ont fait l'objet d'une publication en 2012 (Le Clerc et al. [2012]). Ces analyses "classiques" de données GWAS (Genome Wide Association Study) permettent de mettre en évidence les effets les plus importants. Cependant les effets plus modérés mais combinés de gènes ne peuvent pas être mis en évidence. Différentes approches peuvent être envisagées pour prendre en compte cet effet multifactoriel. Une approche non supervisée qui vise à structurer et résumer l'information contenue dans la base de données comme l'ACP (Analyse en Composantes Principales) et l'ACM (Analyse des Correspondances Multiples), ou une approche supervisée à visée explicative telles que les méthodes de régression.

Dans ce contexte de données de très grande dimension où le nombre de variables est largement supérieur au nombre d'individus, le problème spécifique de la sélection de variables se pose et nécessite une approche particulière. Les deux approches envisagées sont adaptées à ce contexte. Le but dans un premier temps est d'utiliser des méthodes non supervisées pénalisées, c'est-à-dire "sparse", telles que l'ACP sparse, afin de sélectionner les variables les plus pertinentes dans le problème posé. Les variables considérées dans cette thèse étant catégorielles on appliquera une pénalisation de type group LASSO à l'ACM dans le but d'obtenir une version parcimonieuse de cette méthode et de simplifier l'interprétation des résultats. Une fois l'approche exploratoire réalisée une approche explicative est envisagée. Les données pouvant être structurées par blocs, des méthodes telles que la Regularized Generalized Canonical Correlation Analysis (RGCCA) ou des régressions group LASSO peuvent être utilisées. Répondre à l'objectif est d'autant plus difficile que, dans la majorité des cas, c'est l'interaction entre plusieurs gènes qui permet d'expliquer un phénomène précis. Il faut pour cela considérer des méthodes de détection d'interactions telle que la régression logique et intégrer les résultats obtenus aux deux approches proposées. Une fois les signaux (gènes) intéressants identifiés, il est important d'apporter une interprétation biologique afin de mesurer la pertinence des résultats par rapport au problème biologique considéré. L'interaction avec les biologistes et leur contribution dans ce travail est indispensable pour valider les méthodes développées et comprendre les processus biologiques mis en exergue grâce à ces approches.

L'objectif de ce travail de recherche est donc de développer de nouvelles méthodes statistiques nécessaires à l'analyse de données génétiques et plus particulièrement de données de SNPs afin d'étudier l'influence du polymorphisme génétique sur les caractéristiques et l'expression du vieillissement cutané, ainsi que d'interagir avec des biologistes afin d'interpréter les résultats obtenus et les pistes biologiques envisagées.

La thèse est composée de quatre parties :

La première partie considère le problème de sélection de variables dans le cas supervisé et non supervisé. Un état de l'art de plusieurs méthodes parcimonieuses ("sparse")

existantes est réalisé dans le cas unibloc (contexte exploratoire non supervisé) où une présentation de plusieurs versions "sparse" de l'ACP est réalisée ; le cas à 2 blocs (contexte supervisé) avec une présentation de plusieurs versions "sparse" de la régression PLS ; et le cas à K blocs (contexte supervisé multiblocs) en considérant la méthode group LASSO et sa version sparse (SGL), ainsi que la RGCCA et sa version sparse (SGCCA).

La deuxième partie traite le problème de sélection de variables dans le cadre non supervisé de données quantitatives structurées par blocs ainsi que de données catégorielles. Deux nouvelles méthodes sont proposées. La première, nommée GSPCA, est une méthode d'exploration parcimonieuse de données quantitatives structurées par blocs. Elle permet la sélection de groupes de variables quantitatives sur chacun des axes. La deuxième, nommée ACM sparse, est une méthode d'analyse de données parcimonieuse dans le cas où les données sont catégorielles mais pas nécessairement structurées par blocs. Contrairement aux méthodes de sélection de variables dans le cas supervisé, ces deux méthodes ne permettent pas une sélection globale mais une sélection axe par axe qui facilite l'interprétation des résultats.

La troisième partie présente l'intérêt de détecter des interactions entre polymorphismes, et deux approches différentes sont décrites. La première est une approche biologique basée sur les interactions déjà connues dans la littérature et stockées dans des bases de données accessibles. Elle permet d'avoir une vision globale des interactions connues et des voies biologiques intéressantes qui en ressortent. La deuxième est une approche statistique. La détection d'interactions entre SNPs est possible grâce à la régression logique qui est explicitée dans cette troisième section.

Enfin, la quatrième partie présente les applications de ces méthodes sur des données génétiques réelles (SNPs). Une présentation du contexte de l'étude et des pré-traitements effectués est réalisée. Dans un premier temps, une approche exploratoire utilisant les nouvelles méthodes développées est considérée puis une approche supervisée est réalisée. Les interactions sont ensuite détectées et prises en compte par la suite dans les différentes ap-

proches. Les résultats sont comparés à l'aide de critères biologiques et des interprétations sont proposées grâce à la collaboration avec les biologistes.

Nous conclurons ce travail par un bilan et des perspectives de travail futur.

Notations et rappels

Les matrices sont désignées par des lettres majuscules à caractères gras (exemple : $\mathbf{X}$), les vecteurs par des lettres minuscules à caractères gras (exemple : $\mathbf{q}$), les éléments des vecteurs et des matrices par des lettres minuscules en italique avec les indices appropriés si nécessaire (exemple : $x_{i,j}$ est un élément de $\mathbf{X}$). L'opérateur du signe est dénoté sign(). La matrice identité est définie par $\mathbf{I}$, le vecteur colonne rempli de chiffres uns par $\mathbf{1}$. Le rang d'une matrice est noté rank(), la transposée d'une matrice^T, et l'inverse^{-1}. Appliqué à une matrice carrée, l'opérateur diagonale, noté diag(), prend les éléments de la diagonale de la matrice et les stocke dans un vecteur colonne ; lorsqu'il est appliqué à un vecteur, l'opérateur diag stocke les éléments du vecteur sur la diagonale d'une matrice. L'opérateur trace dénoté tr() calcule la somme des éléments diagonaux d'une matrice carrée.

La norme L_1 est notée $||.||_1$ et définie de la manière suivante :

$$||\mathbf{x}||_1 = \sum_{i=1}^{I} |x_i| \tag{1}$$

La norme L_2 est notée $||.||_2$ et définie de la manière suivante :

$$||\mathbf{x}||_2 = \sqrt{\mathbf{x}\mathbf{x}^T} = \left(\sum_{i=1}^{I} |x_i|^2 \right)^{1/2} \tag{2}$$

La norme L_2 $\mathbf{W}$-généralisée est notée $||.||_{\mathbf{W}}$ et définie comme suit :

$$||\mathbf{x}||_{\mathbf{W}} = \sqrt{\mathbf{x}\mathbf{W}\mathbf{x}^T} \tag{3}$$

avec $\mathbf{W}$ une matrice définie positive.

Pour toutes matrices $\mathbf{A}$ et $\mathbf{B}$ carrées, les propriétés suivantes de la trace sont vérifiées (Schott [2005]) :

Propriété 1.

$$\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \tag{4}$$

Propriété 2.

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}((\mathbf{AB})^T) = \mathrm{tr}(\mathbf{B}^T\mathbf{A}^T) = \mathrm{tr}(\mathbf{BA}) \tag{5}$$

Propriété 3. *Cas particulier de la proriété 2 pour* $\mathbf{B}$=$\mathbf{A}^T$

$$\mathrm{tr}(\mathbf{AA}^T) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) \tag{6}$$

Le produit standard entre matrices est dénoté par une simple juxtaposition ou par $\times$ quand il est nécessaire de l'expliciter (exemple : $\mathbf{XY} = \mathbf{X} \times \mathbf{Y}$ est le produit des matrices $\mathbf{X}$ et $\mathbf{Y}$). Les matrices de rang-1 sont notées $\mathbf{X}^{(1)}$ et les matrices de rang-L notées $\mathbf{X}^{(L)}$. Si les données sont structurées en K tables de données recueillies sur les mêmes observations, chaque table est appelée un bloc. Les données de chaque bloc sont stockées dans une matrice rectangulaire de dimension $I \times J_{[k]}$ notée $\mathbf{X}_{[k]}$ (en général centrée réduite), où I est le nombre d'observations et $J_{[k]}$ le nombre de variables collectées sur les mêmes observations pour la k-ème table. Le nombre total de variables est appelé J (*i.e*, $J = \sum_{k=1}^{K} J_{[k]}$). Les blocs de variables sont considérés comme des sous-matrices de matrices plus larges et sont représentés entre crochets, séparés par des barres verticales. Par exemple, les K blocs $\mathbf{X}_{[k]}$, chacun de dimensions I lignes par $J_{[k]}$ colonnes, sont concaténés dans une matrice notée $\mathbf{X}$ de dimensions $I \times J$:

$$\mathbf{X} = \left[\mathbf{X}_{[1]}|\ldots|\mathbf{X}_{[k]}|\ldots|\mathbf{X}_{[K]}\right] \tag{7}$$

De plus, appliqué à une matrice structurée par blocs, l'opérateur diag() produit une matrice carrée bloc diagonale.

Propriétés des dérivées de matrices

Si l'on considère la matrice $\mathbf{X}$ et des matrices $\mathbf{A}$ et $\mathbf{B}$, carrées, constantes et indépendantes de $\mathbf{X}$, les propriétés suivantes sont vérifiées (Magnus and Neudecker [1988], Petersen and Pedersen [2006]) :

Propriété 4.

$$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} = \frac{\operatorname{tr}(\partial \mathbf{X})}{\partial \mathbf{X}} \tag{8}$$

Propriété 5.

$$\frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{X} \tag{9}$$

Propriété 6.

$$\frac{\partial \operatorname{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{XA})}{\partial \mathbf{X}} = \mathbf{A}^T \tag{10}$$

Propriété 7.

$$\frac{\partial \operatorname{tr}(\mathbf{AXB})}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{B}^T \tag{11}$$

Lorsque qu'une fonction n'est pas différentiable en un point, on calcule alors le sous-gradient, noté ∂, qui généralise la dérivée de fonctions non différentiables. Les sous-gradients sont généralement utilisés dans le cadre d'étude de fonctions convexes ou d'optimisation convexe.

Propriété 8. *Le sous-gradient de la norme* L_2 $\|\mathbf{w}\|_2$*, avec* $\mathbf{w}$ *un vecteur* $\in \mathbb{R}^J$ *vaut :*

$$\partial \|\mathbf{w}\|_2 = \begin{cases} \dfrac{\mathbf{w}}{\|\mathbf{w}\|_2} & si \quad \mathbf{w} \neq 0 \\ \in \{\mathbf{z} |\ \|\mathbf{z}\|_2 \leq 1\} & si \quad \mathbf{w} = 0 \end{cases} \tag{12}$$

avec $\mathbf{z}$ un vecteur $\in \mathbb{R}^J$.

Décomposition en valeurs singulières

La décomposition en valeurs singulières a de très nombreuses applications utiles en statistiques multivariées (Greenacre [1984], Jessup and Sorensen [1994], Strang [2003], Abdi [2007], Yanai et al. [2011], Golub and Van Loan [2012]). Toute matrice $\mathbf{X}$ de rang L, de dimensions $I \times J$, peut être décomposée en trois matrices $\mathbf{P}$, $\mathbf{\Delta}$, $\mathbf{Q}$ comme suit :

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T, \tag{13}$$

où $\mathbf{P}$ $(I \times L)$ et $\mathbf{Q}$ $(J \times L)$ sont orthonormales ((i.e., chaque colonne a une norme de 1, et deux colonnes différentes sont orthogonales) et $\mathbf{\Delta}$ $(L \times L)$ est une matrice diagonale où

chacun de ses éléments δ_j $(j = 1, \ldots, L)$ est appelé valeur singulière. Les valeurs singulières sont les racines carrées des valeurs propres des matrices $\mathbf{X}^T\mathbf{X}$ et $\mathbf{X}\mathbf{X}^T$. On pose $\mathbf{F}$=$\mathbf{P}\boldsymbol{\Delta}$, les Composante Principales (CPs), en anglais "factor scores", et $\mathbf{Q}$ les "loadings" correspondants, ou vecteurs propres (Saporta [2006]).

Chapitre 1

Méthodes sparses supervisées et non supervisées : état de l'art

Dans de nombreux domaines, en particulier en biologie, les méthodes d'analyse multivariée, et plus particulièrement les méthodes de régression, se sont révélées très utiles pour l'identification de liens éventuels entre un ou plusieurs facteurs et un phénomène d'intérêt tels que la survenue d'une maladie ou la valeur d'une variable biologique par exemple. Ces méthodes permettent de "modéliser" le phénomène étudié. Toutefois une sélection de variables est souvent nécessaire au préalable, en particulier lorsque l'on dispose d'un nombre de variables explicatives J important pour modéliser la variable d'intérêt. Le but de cette sélection est en général d'améliorer la connaissance du phénomène de causalité entre les descripteurs et la variable à prédire, ou encore d'améliorer la qualité de la prédiction.

Dans le cadre de cette sélection de variables plusieurs approches sont envisageables. La première consiste à utiliser des algorithmes de sélection de modèle par sélection de variables (de type "backward", "forward", "stepwise", ...) en minimisant des critères pénalisés (Cp, AIC, BIC). De nombreux critères de choix de modèles ont été proposés pour la régression linéaire multiple. En pratique, les critères les plus utilisés sont la statistique du F de Fisher pour comparer des séquences de modèles emboités ; le R^2 qui croît à mesure de l'introduction de variables dans le modèle et qui peut servir à comparer deux modèles avec le même nombre de variables ; le Cp de Mallows [1973] qui est une estimation de l'erreur quadratique moyenne de prévision ; le "Akaïke's Information criterion (AIC)" (Akaike [1974]) qui représente un compromis entre le biais, diminuant avec le nombre de paramètres libres,

et la parcimonie (i.e., la volonté de décrire les données avec le plus petit nombre de paramètres possibles) ; le "Bayesian Information Criterion (BIC)" (Sawa [1978]) qui vise la sélection de variables statistiquement significatives dans le modèle ; ou encore le PRESS de Allen (Allen [1971]), qui est un autre type de critère issu de la validation croisée, appelé aussi Leave One Out (LOO) qui permet de comparer les capacités prédictives de deux modèles. Lorsque le nombre de variables J est grand, il est difficile de pouvoir explorer tous les modèles possibles afin de sélectionner "le meilleur" au sens de l'un des critères cités ci-dessus. Pour pallier ce problème, différentes stratégies (qui doivent être choisies en fonction de l'objectif recherché et de la valeur de J) ont été proposées. Les méthodes de type pas à pas consistent à considérer d'abord un modèle faisant intervenir un certain nombre de variables explicatives, puis procèdent par élimination ou ajout successif de variables. La méthode ascendante (forward selection) consiste à ajouter une variable (la plus significative) au modèle à chaque pas, la procédure s'arrêtant lorsque toutes les variables sont introduites dans le modèle. La méthode descendante (backward elimination) démarre du modèle complet et à chaque étape élimine du modèle la variable la moins significative. Enfin, la méthode stepwise est une combinaison de ces deux méthodes et introduit une étape d'élimination de variables après chaque étape de sélection. Cependant lorsque J est grand la méthode descendante est inexploitable voire impossible si le nombre de variables J est supérieur au nombre d'individus I. Lorsque $J > I$, l'estimateur des moindres carrés ordinaires n'existe pas et on se retrouve dans un cas de multicolinéarité. Des méthodes telles que la régression ridge ou Partial Least Squares (PLS) permettent une réduction de dimension tout en conservant l'ensemble des variables, ce qui est souvent perçu comme un avantage. Cependant dans des contextes de grande dimension ($J \gg I$), de telles combinaisons deviennent ininterprétables.

Pour y remédier une autre approche consiste à utiliser des algorithmes plus récents appelés "sparse". Ils permettent la sélection de modèles par pénalisation et régularisation (ridge, LASSO, elastic net) en produisant des combinaisons "sparse" de variables, c'est-à-dire avec un grand nombre de coefficient nuls. Les méthodes telles que LASSO, elastic net ou sparse PLS permettent de créer de la parcimonie dans les résultats afin de faciliter

leur interprétation. Les méthodes qui seront présentées et utilisées tout au long de ce manuscrit sont en grande partie basées sur ces algorithmes. Les algorithmes de sélection par pénalisation s'utilisent surtout dans le cas supervisé (régression multiple) mais comme certaines méthodes non supervisées telles que l'ACP et l'ACM peuvent être vues comme un problème de type régression, ces algorithmes peuvent s'appliquer.

La première grande partie de ce chapitre est consacrée aux méthodes "sparse" en régression linéaire, puis dans un deuxième temps pour des données unibloc (un seul bloc de données $\mathbf{X}$) dans un contexte non supervisé (panorama de plusieurs versions de l'ACP sparse). Par la suite nous nous placerons dans le cas supervisé pour l'analyse de 2 blocs de variables (un bloc de variables explicatives $\mathbf{X}$ et un bloc de variable(s) à expliquer) et une présentation de différentes méthodes de régression PLS dans leur version "sparse" sera réalisée.

Dans certains domaines, en particulier en biologie et en génomique, les données de très grandes dimensions peuvent être structurées par blocs (blocs de gènes, blocs de protéines, blocs de SNPs, etc). Il est donc important de savoir comment traiter ces données, les analyser ainsi que de sélectionner des blocs de variables pour faciliter l'interprétation des résultats. Pour ce faire, la deuxième partie de cette section est dédiée aux méthodes dites "sparse" dans le cas où les données sont à priori structurées par blocs, pour 3 blocs ou plus. Une présentation de la régression group LASSO et de sa version "sparse" est réalisée dans le cadre d'une régression avec une seule variable à expliquer, et une méthode multiblocs (RGCCA) pouvant considérer K blocs de variables ($K > 3$) est présentée dans sa version standard et "sparse".

1.1 Méthodes sparse

1.1.1 Méthodes de régularisation

Soit un modèle linéaire classique à J prédicteurs $\mathbf{x}_1, ..., \mathbf{x}_J$. La réponse prédite par un tel modèle prend alors la forme :

$$\widehat{\mathbf{y}} = \widehat{\beta}_0 + \sum_{j=1}^{J} \widehat{\beta}_j \mathbf{x}_j, \tag{1.1}$$

où les coefficients de régression $\widehat{\beta}_j$ et l'ordonnée à l'origine $\widehat{\beta}_0$ sont estimés par le critère usuel des moindres carrés (Gauss [1855]) :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{1.2}$$

c'est-à-dire en recherchant les $\boldsymbol{\beta}$ qui minimisent l'erreur quadratique entre les valeurs prédites et les valeurs observées. La précision de ces estimateurs étant fonction du nombre d'observations et de la matrice $\mathbf{X}$, il devient clair que lorsque le nombre de prédicteurs J est grand, des problèmes de sur-ajustement surviennent.

1.1.1.1 Régression ridge et LASSO

Une solution au problème de surajustement consiste à pénaliser les termes du modèle en considérant soit une pénalisation de norme L_2 (régression ridge, Hoerl and Kennard [1988]), soit une pénalisation de norme L_1 (régression LASSO, Tibshirani [1996]) :

$$\widehat{\boldsymbol{\beta}}_\lambda^{ridge} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \qquad \text{avec } \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{J} \beta_j^2 \tag{1.3}$$

$$\widehat{\boldsymbol{\beta}}_\lambda^{LASSO} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \qquad \text{avec } \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\beta_j|. \tag{1.4}$$

On considère la réponse $\mathbf{y}$ centrée et la matrice $\mathbf{X}$ standardisée (moyenne nulle, variance égale à 1) car on perd la propriété d'invariance par changement d'échelle dès que l'on abandonne les moindres carrés ordinaires.

En présence de multicolinéarité, la régression régularisée de type ridge permet de meilleures prédictions que le modèle des moindres carrés ordinaires (Ordinary Least Squares (OLS)) classique, grâce à un meilleur compromis entre biais et variance. L'équation (1.3) peut s'écrire comme la somme des carrés des résidus pénalisée (Penalized Residual Sum of Squares (PRSS)) de la façon suivante :

$$\begin{aligned} PRSS(\boldsymbol{\beta}^{ridge}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2 \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}, \end{aligned} \tag{1.5}$$

pour $\lambda > 0$. Il existe donc une solution unique (Hoerl and Kennard [2000]) que l'on obtient en calculant la dérivée de PRSS($\boldsymbol{\beta}^{ridge}$) par rapport à $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y}^T\mathbf{y}) = 0 \tag{1.6}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}) = -2\mathbf{X}^T\boldsymbol{\beta} \tag{1.7}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{1.8}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}) = 2\lambda\boldsymbol{\beta} \tag{1.9}$$

Ainsi on obtient :

$$\frac{\partial PRSS(\boldsymbol{\beta}^{ridge})}{\partial \boldsymbol{\beta}^{ridge}} = -2\mathbf{X}^T\boldsymbol{\beta} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \tag{1.10}$$

$$= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}. \tag{1.11}$$

La solution de PRSS($\boldsymbol{\beta}^{ridge}$) est donc obtenue en annulant la dérivée.

$$\begin{aligned} \frac{\partial PRSS(\boldsymbol{\beta}^{ridge})}{\partial \boldsymbol{\beta}^{ridge}} = 0 \quad &\Leftrightarrow \quad -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0 \\ &\Leftrightarrow \quad -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = 0 \\ &\Leftrightarrow \quad \boldsymbol{\beta}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}) = \mathbf{X}^T\mathbf{y} \\ &\Leftrightarrow \quad \boldsymbol{\beta}_\lambda^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \text{ (si l'inverse existe)} \end{aligned} \tag{1.12}$$

La solution est indexée par le paramètre λ qui contrôle la régularisation. Ainsi pour chaque λ il y a une solution. Lorsque λ tend vers 0, on obtient la solution des moindres carrés, si elle existe, lorsqu'il tend vers l'infini, on obtient un $\boldsymbol{\beta}_\lambda^{ridge}$ nul.

Le choix du λ peut être fait de plusieurs façons. Dans leur article, Hoerl and Kennard [1970] tracent $\boldsymbol{\beta}_\lambda^{ridge}$ en fonction de λ et choisissent le λ pour lequel les coefficients sont à peu près stables. Cette méthode non objective a souvent été critiquée. Une autre possibilité consiste à estimer le nombre de degrés de liberté *df* de la manière suivante :

$$df = \mathrm{tr}\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-\right)^{-1}\mathbf{X}^T\right) \tag{1.13}$$

où $\mathbf{W} = \mathrm{diag}(|\boldsymbol{\beta}_{\lambda,j}^{ridge}|)$ et $\mathbf{W}^-$ indique une pseudo-inverse. La pratique standard consiste désormais à utiliser la validation croisée, ce qui implique de chercher la valeur de λ qui

minimise l'erreur quadratique moyenne. Si la taille de l'échantillon le permet, il sera découpé en deux : un ensemble d'apprentissage et un ensemble de test. L'approche de validation croisée K-fold (K-fold cross validation) la plus commune est en quatre étapes :

i) Partitionner l'ensemble d'apprentissage $\mathcal{T}$ en K groupes de même dimension ($K = 10$ est très souvent employé, McLachlan et al. [2005]). On suppose $\mathcal{T} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K)$,

ii) Pour chaque $k = 1, 2, \ldots, K$, ajuster le modèle $\hat{f}_{-k}^{(\lambda)}(\mathbf{x})$ à l'ensemble d'apprentissage en excluant le k-ème groupe,

iii) Calculer les valeurs ajustées pour les observations contenues dans l'échantillon considéré,

iv) Calculer l'erreur de validation croisée pour le k-fold :

$$(CV_{Error})_k^{(\lambda)} = |\mathcal{T}_k|^{-1} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}_k} (\mathbf{y} - \hat{f}_{-k}^{(\lambda)}(\mathbf{x}))^2, \tag{1.14}$$

avec $|\mathcal{T}_k|$ le cardinal de l'ensemble $\mathcal{T}_k$.

L'erreur globale de validation croisée est ensuite définie comme :

$$(CV_{Error})^{(\lambda)} = K^{-1} \sum_{k=1}^{K} (CV_{Error})_k^{(\lambda)}. \tag{1.15}$$

Le λ sélectionné est celui qui minimise $(CV_{Error})^{(\lambda)}$ et il est noté λ^*. Le modèle choisi $\hat{f}^{(\lambda^*)}(\mathbf{x})$ est calculé sur la totalité de l'ensemble d'apprentissage et l'erreur est calculée en appliquant $\hat{f}^{(\lambda^*)}(\mathbf{x})$ sur l'ensemble de test.

Remarque Si la taille de l'échantillon ne permet pas d'utiliser la validation croisée K-fold, plusieurs solutions sont possibles. La première méthode ("testset validation") consiste à diviser l'échantillon de taille I en échantillon d'apprentissage et échantillon de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. La seconde possibilité est l'utilisation du LOO. Cette méthode est un cas particulier de la validation croisée K-fold où $K = I$, c'est-à-dire que l'on apprend sur $(I - 1)$ observations puis on valide le modèle sur la Ième observation et l'on répète cette opération I fois.

La régression ridge permet de contourner les problèmes de colinéarité même en présence d'un nombre important de variables explicatives ($J > I$). Cependant elle conserve tous

les prédicteurs dans le modèle ce qui peut rendre difficile l'interprétation des résultats. D'autres approches par régularisation permettent également une sélection, c'est le cas de la régression Least Absolute Shrinkage and Selection Operator (LASSO). Contrairement à la régression ridge, la régression LASSO permet la sélection automatique des variables en imposant la nullité de certains coefficients selon les valeurs du paramètre de pénalisation λ. Elle correspond à la minimisation d'un critère des moindres carrés avec une pénalité de type L_1 (et non plus L_2 comme dans la régression ridge). La fonction optimisée est l'erreur empirique mesurée par $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ à laquelle on ajoute un terme de régularisation, $\lambda\|\boldsymbol{\beta}\|_1$, correspondant à la norme L_1 de $\boldsymbol{\beta}$ pondérée par un paramètre λ positif. Si $\lambda = 0$, on retrouve l'estimateur des moindres carrés, quand il existe, où aucun coefficient n'est nul et par conséquent aucune sélection n'est réalisée. En revanche, plus la valeur de λ est élevée, plus le nombre de coefficients nuls augmente. La solution obtenue avec la méthode LASSO est dite parcimonieuse ("sparse" en anglais) car elle comporte beaucoup de coefficients nuls. L'équation (1.4) peut s'écrire de la manière suivante :

$$PRSS(\boldsymbol{\beta}^{lasso}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1. \tag{1.16}$$

Contrairement à la régression ridge, il n'existe pas de forme analytique de la solution $\widehat{\boldsymbol{\beta}}_\lambda^{lasso}$ dans le cas général. Il faut alors utiliser des techniques de programmation quadratique car c'est un problème d'optimisation convexe. Si l'on construit le chemin de solutions du vecteur $\boldsymbol{\beta}$ en fonction du paramètre λ, on s'aperçoit qu'il est linéaire par morceaux. Ainsi, on peut construire une suite de $J+1$ valeurs croissantes de $\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_J$ telle que $\widehat{\boldsymbol{\beta}}_{\lambda_0}^{lasso}$ soit l'estimateur des moindres carrés, $\widehat{\boldsymbol{\beta}}_{\lambda_1}^{lasso}$ ait exactement un coefficient nul, $\widehat{\boldsymbol{\beta}}_{\lambda_2}^{lasso}$ deux coefficients nuls, etc. jusqu'à $\widehat{\boldsymbol{\beta}}_{\lambda_J}^{lasso}$ égal au vecteur nul. La détermination de cette suite de valeurs de λ et des $\boldsymbol{\beta}$ correspondants se fait par l'algorithme 1 Least Angle Regression (LARS) proposé par Efron et al. [2004] et implémenté dans le package "glmnet" (Friedman et al. [2010]) du logiciel R (R Development Core Team [2008]).

Le LASSO présente toutefois des inconvénients dans plusieurs situations :

1. Dans le cas $J \gg I$ (nombre de prédicteurs supérieur au nombre d'observations), le LASSO ne sélectionne que I prédicteurs au maximum.

Algorithm 1 Algorithme LARS

Initialisation. Commencer avec tous les coefficients β_j égaux à 0.

Etape 1. Trouver le prédicteur $\mathbf{x}_j$ le plus corrélé à $\mathbf{y}$.

Etape 2. Augmenter le β_j dans le sens du signe de la corrélation avec $\mathbf{y}$. Prendre comme résidus $r = \mathbf{y} - \hat{\mathbf{y}}$. S'arrêter lorsque d'autres prédicteurs $\mathbf{x}_k$ sont autant corrélés à r que $\mathbf{x}_j$.

Etape 3. Augmenter (β_j, β_k) dans leur direction conjointe des moindres carrés jusqu'à ce qu'un autre prédicteur $\mathbf{x}_m$ soit aussi corrélé avec le résidu r.

Etape 4. Continuer jusqu'à ce que tous les prédicteurs soient dans le modèle.

2. Si un groupe contient des prédicteurs très corrélés entre eux, le LASSO tend à sélectionner uniquement un seul prédicteur dans le groupe et ce prédicteur est un prédicteur quelconque du groupe.

1.1.1.2 Régularisation elastic net

Une solution proposée plus récemment par Zou and Hastie [2005] consiste à utiliser une combinaison convexe des régressions ridge et LASSO afin de pallier les limitations du LASSO. L'estimateur elastic net est de la forme :

$$\widehat{\boldsymbol{\beta}}^{elasticnet} = \arg\min_{\boldsymbol{\beta}} L(\lambda_1, \lambda_2, \boldsymbol{\beta}), \tag{1.17}$$

où :

$$\begin{aligned} L(\lambda_1, \lambda_2, \boldsymbol{\beta}) &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1, \end{aligned} \tag{1.18}$$

avec $\alpha = \dfrac{\lambda_2}{(\lambda_1 + \lambda_2)}$, $\|\boldsymbol{\beta}\|_2^2$ et $\|\boldsymbol{\beta}\|_1$ définis comme en (1.3) et (1.4). Fixer α à 0.5 revient à donner autant d'importance à la contrainte L_1 qu'à la contrainte L_2.

Les deux paramètres de régularisation, λ_1 et λ_2, permettent d'une part la sélection de variables et d'autre part d'autoriser dans le cas $I \leqslant J$, la sélection de plus de I variables. Le LASSO est donc le cas particulier $\lambda_2 = 0$ de l'elastic net, et la ridge celui où $\lambda_1 = 0$. La solution elastic net est linéaire par morceaux. Si λ_2 est fixé, un algorithme nommé Least Angle Regression-Elastic Net (LARS-EN) résout efficacement tout le chemin de solution

de l'elastic net. Il est basé sur l'algorithme LARS de Efron et al. explicité précédemment. Seuls les coefficients non-nuls sont conservés à chaque étape du LARS-EN. L'algorithme converge généralement rapidement, en particulier dans le cas où $J \gg I$.

La figure 1.1 présente l'illustration en 2-dimensions de la géométrie de la régression ridge, LASSO et elastic net dans le cas où $\alpha = 0.5$. Les singularités que l'on observe aux sommets pour le LASSO et l'elastic net permettent la sparsité (sélection de variables avec coefficients nuls). Ce que l'on ne retrouve pas dans le cas de la régression ridge car aucun coefficient n'est mis à zéro. Par ailleurs, les bords sont convexes et cette convexité varie en fonction du α choisi dans l'elastic net.

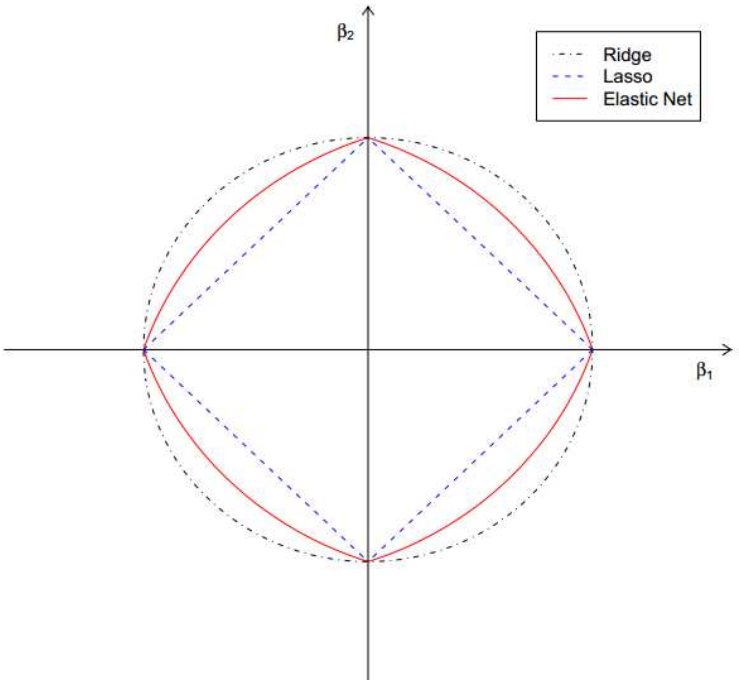

FIGURE 1.1 – Illustration en 2-dimensions de la géométrie ridge (petits pointillés noirs), LASSO (pointillés bleus, forme de losange) et elastic net pour $\alpha = 0.5$ (courbe continue rouge). Elastic net est un compromis entre LASSO et ridge, il y a des singularités au sommet et les bords sont convexes. La convexité varie avec α (figure tirée de Zou [2005]).

Le choix des paramètres λ_1 et λ_2 dans le cas de l'elastic net est fait par validation croisée. Des valeurs de (λ_1, λ_2) sont balayées sur une grille, et pour chaque point, une validation croisée est réalisée afin de trouver le couple (λ_1, λ_2) minimisant l'erreur de prédiction.

Exemple d'application

Les données proviennent d'une étude sur le cancer de la prostate (Stamey et al. [1989]) qui a examiné la corrélation entre le niveau d'antigène spécifique de la prostate (prostate specific antigen : PSA) et un certain nombre de mesures cliniques chez les hommes sur le point de recevoir une prostatectomie radicale. Les données ont été analysées par Tibshirani [1996] et Zou and Hastie [2005]. Les 97 patients sont décrits au moyen de 8 variables cliniques :

1. le logarithme du volume de la tumeur (lcavol),
2. le logarithme du poids de la prostate (lweight),
3. l'âge,
4. le logarithme du taux d'hyperplasie bénigne de la prostate (lbph),
5. l'invasion des vésicules séminales, réponse oui/non (SVI),
6. le logarithme de la pénétration capsulaire (LCP),
7. le score de Gleason (gleason) et
8. le pourcentage de Gleason grade 4/5 (pgg45).

La variable svi est une variable binaire et la variable gleason est une variable catégorielle. Le but de l'étude est de prédire la variable réponse $\mathbf{y}$ étant le logarithme de PSA (lpsa).

La figure 1.2 présente les diagrammes de dispersion des variables (représentant les graphiques pour chaque paire de variables). Quelques corrélations avec lpsa sont évidentes (lcavol par exemple) mais un bon modèle prédictif ne peut être construit sur la simple analyse de cette matrice.

La matrice des corrélations entre les prédicteurs est donné dans la table 1.1 et présente de fortes corrélations. Par exemple, lcavol et lcp sont fortement liées à la variable réponse lpsa et avec chacune des autres variables. Il faudra donc prendre en compte les effets conjointement pour mettre en exergue des relations entre la variable réponse et les prédicteurs. De la même manière que dans Zou and Hastie [2005], les prédicteurs ont été standardisés (variance unitaire) et le jeu de données a été séparé de manière aléatoire en un ensemble d'apprentissage (67 patients) et un ensemble test (30 patients).

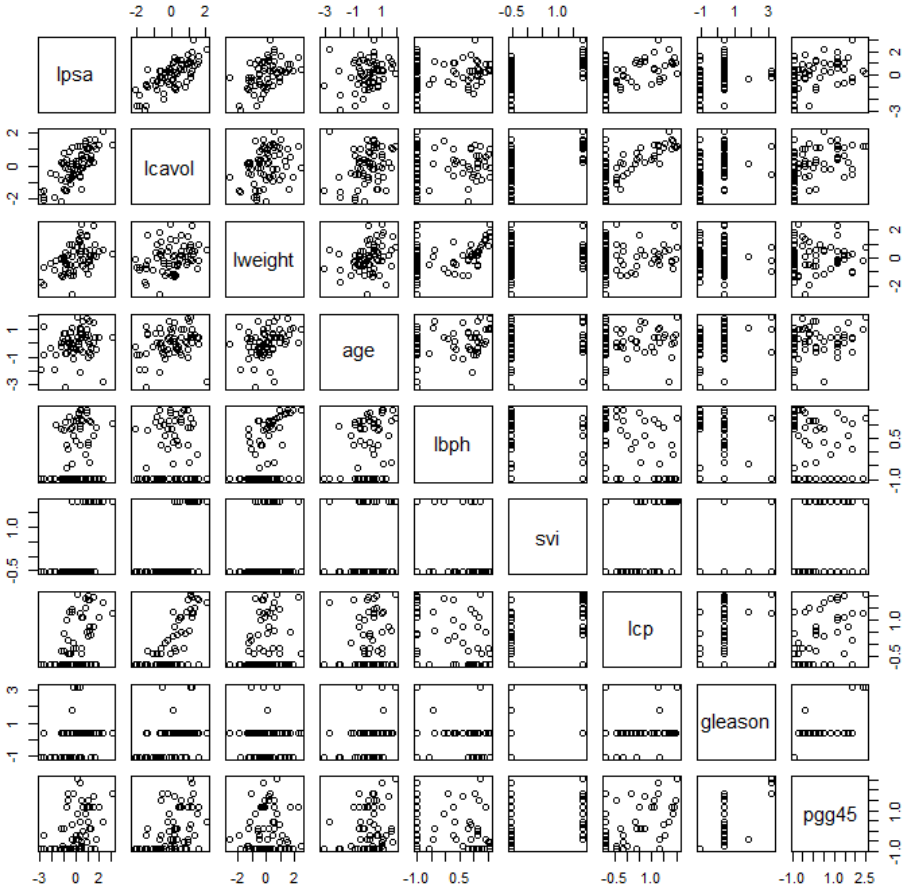

FIGURE 1.2 – Données cancer de la prostate : diagrammes de dispersion (en anglais, "scatter plots"). La première ligne correspond à la variable à expliquer, les autres lignes correspondent aux prédicteurs.

TABLE 1.1 – Données cancer de la prostate : tableau des corrélations entre prédicteurs.

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1,000							
lweight	0,300	1,000						
age	0,286	0,317	1,000					
lbph	0,063	0,437	0,287	1,000				
svi	0,593	0,181	0,129	-0,139	1,000			
lcp	0,692	0,157	0,173	-0,089	0,671	1,000		
gleason	0,426	0,024	0,366	0,033	0,307	0,476	1,000	
pgg45	0,483	0,074	0,276	-0,030	0,481	0,663	0,757	1,000

Dans un premier temps une estimation par les moindres carrés a été réalisée sur l'ensemble d'apprentissage. Le prédicteur lcavol présente l'effet le plus fort, ainsi que svi et lweight. L'erreur moyenne de prédiction sur l'ensemble test vaut 0.545. Par la suite la régression ridge, le LASSO et l'elastic net ont été appliqués au jeu de données. Pour chacune de ces méthodes un paramètre λ est à déterminer. Celui-ci est choisi en minimisant l'erreur de prédiction estimée calculée par validation croisée 10-fold. La validation croisée est effectuée sur l'ensemble d'apprentissage et l'ensemble test permet de juger de la performance du modèle sélectionné. Les performances de ces méthodes ont été comparées dans la table 1.2. Les deux dernières lignes de la table fournissent l'erreur moyenne de prédiction et l'écart type estimé sur l'ensemble test.

TABLE 1.2 – Données cancer de la prostate : comparaison de différentes méthodes.

	Moindres carrés	Ridge	Lasso	Elastic net	PCR	PLS
(Intercept)	2,480	2,452	2,570	2,482	2,523	2,512
lcavol	0,680	0,442	0,565	0,503	0,570	0,436
lweight	0,305	0,256	0,199	0,215	0,323	0,360
age	-0,141	-0,049	**0,000**	**0,000**	-0,153	-0,021
lbph	0,210	0,171	0,029	0,083	0,216	0,243
svi	0,305	0,238	0,115	0,173	0,322	0,259
lcp	-0,288	-0,001	**0,000**	**0,000**	-0,050	0,085
gleason	-0,021	0,042	**0,000**	**0,000**	0,228	0,006
pgg45	0,267	0,137	0,016	0,007	-0,063	0,008
Erreur Test	0,586	0,546	0,494	0,505	0,526	0,656
Ecart-type	0,184	0,165	0,159	0,165	0,121	0,180

La figure 1.3 présente le chemin de régularisation de la régression ridge, LASSO et elastic net. La régression ridge ne fait aucune sélection de variables, elles sont donc toutes dans le modèle final (voir table 1.2). Le LASSO et elastic net retiennent les variables lcavol, lweigth, lbph, svi et pgg45 dans le modèle final. L'elastic net est un compromis entre la régression ridge qui n'écarte aucune variable, et le LASSO. L'avantage des méthodes de sélection par pénalisation LASSO et elastic net réside dans le fait qu'elles réalisent une sélection de variables dans le modèle de départ en fixant certains coefficients à zéro facilitant alors l'interprétation des résultats.

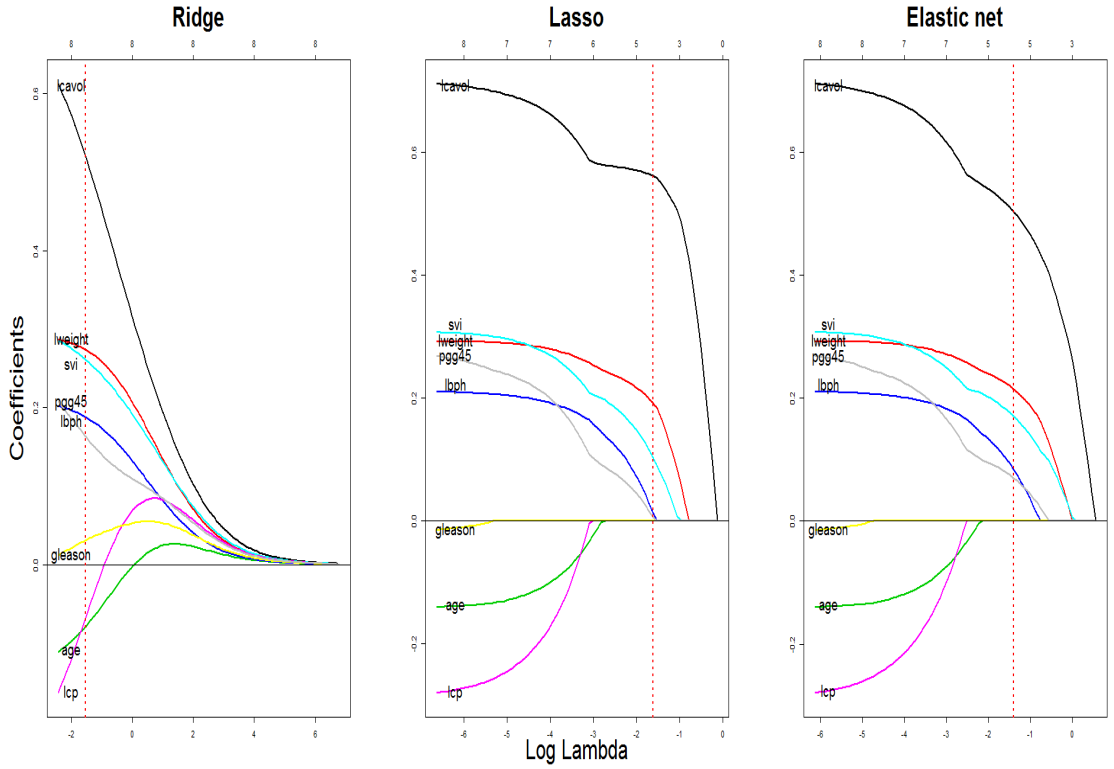

FIGURE 1.3 – Données du cancer de la prostate : chemin de régularisation de la régression ridge (à gauche), du LASSO (au centre) et de l'elastic net pour $\alpha = 0.5$ (à droite). Les traits en pointillés rouges représentent la valeur du λ optimal caculé par validation croisée. Les variables qui ne sont pas à zéro pour cette valeur de λ sont sélectionnées.

Des méthodes de sélection de modèle projection sur composantes orthogonales ont également été réalisées. La régression PLS (qui sera présentée dans la section 1.1.3.1) et la régression sur composantes principales. La validation croisée a permis de retenir 7 composantes pour la régression en composantes principales (on ne réduit pas beaucoup la complexité par rapport au modèle initial) et 2 composantes pour la PLS. Les résultats obtenus se situent dans les colonnes 5 et 6 de la table 1.2. Ces deux méthodes ont tendance à produire des résultats similaires à la régression ridge. Deux autres modèles ont également été obtenus par des méthodes de sélection de variables forward et backward. Les résultats ne sont pas présentés ici mais les deux modèles sélectionnés donnent exactement les mêmes résultats et sont plus parcimonieux que le modèle initial.

1.1.2 Les ACP Sparse

On peut approcher le problème de sélection de variable dans un contexte non supervisé par des techniques de réduction de dimension, telle que l'Analyse en Composantes Principales (ACP) dans sa version pénalisée, ou "sparse". En ACP, chaque composante principale est une combinaison linéaire de l'ensemble des variables de départ (Saporta [2006]). L'interprétation des résultats est donc difficile dans un contexte de données de grande dimension. Des techniques d'ACP avec rotation sont souvent utilisées afin d'aider à l'interprétation des composantes principales (Jolliffe [1995]), telle que la rotation varimax (Kaiser [1958]). Le but de cette technique est de trouver des "loadings" spécifiques des axes et plus simples d'interprétation (Thurstone [1947], Cattell [1978], Abdi [2003]). Vines [2000], par exemple, considère les composantes principales et restreint les "loadings" à prendre un petit ensemble de valeurs proches de 0, 1 ou -1. Cependant la structure obtenue ne contient pas de coefficients exactement à zéro. Une alternative a alors été proposée par de nombreux auteurs : l'ACP sparse. Le but de l'ACP sparse est d'obtenir des composantes facilement interprétables en imposant une contrainte à l'ACP pour obtenir des composantes avec des "loadings sparse" (nombre important de "loadings" nuls). Nous développerons ici plusieurs approches de la version pénalisée de l'ACP ; celle de Jolliffe et al. [2003], celle de Zou et al. [2006] et enfin celle de Shen and Huang [2008].

1.1.2.1 SCoTLASS de Jolliffe et al. [2003]

L'idée principale de la "sparsification" en ACP est de choisir des combinaisons linéaires des variables mesurées qui maximisent successivement la variance, comme en ACP, mais en imposant des contraintes supplémentaires qui vont sacrifier de la variance au bénéfice de l'interprétabilité. La contrainte supplémentaire impose une borne sur la somme des valeurs absolues des "loadings" sur chaque composante. Contrairement à la rotation, cette technique permet de fixer des "loadings" exactement à zéro. Soit $\mathbf{X}$ une matrice $I \times J$, où I et J sont le nombre d'observations et le nombre de variables, respectivement. L'ACP cherche les combinaisons linéaires $\mathbf{Xq}_j$ $(j = 1, \ldots, J)$ des variables initiales de $\mathbf{X}$, de sorte

que la variance soit maximale :

$$\mathbf{q}_j^T(\mathbf{X}^T\mathbf{X})\mathbf{q}_j, \tag{1.19}$$

sous la contrainte :

$$\mathbf{q}_j^T\mathbf{q}_j = 1. \tag{1.20}$$

La méthode Simplified Component Technique-LASSO (SCoTLASS) proposée par Jolliffe et al. [2003] réalise cette maximisation sous la contrainte supplémentaire :

$$\sum_{k=1}^{J} |q_{jk}| \leq t, \tag{1.21}$$

pour un paramètre de régularisation t, où q_{jk} est le k-ème élément du vecteur $\mathbf{q}_j$ $(j = 1, ..., J)$. Le résultat obtenu dépendra donc du choix du paramètre de régularisation t. En effet :

- $\forall t \geq \sqrt{J}$, on retrouve l'ACP,
- $\forall t \leq 1$, il n'y a pas de solution et,
- pour $t = 1$, un seul coefficient non nul q_{jk} pour chaque j.

On cherche alors t, tel que $1 < t < \sqrt{J}$ en faisant décroître t depuis $\sqrt{J}$. Cette méthode présente cependant quelques inconvénients car le choix de t, essentiel dans cette méthode, est très délicat. Par ailleurs le problème posé n'étant pas convexe, l'obtention d'un maximum global n'est pas évidente et induit des temps de calcul très élevés.

1.1.2.2 Sparse PCA de Zou et al. [2006]

La méthode proposée par Zou et al. [2006] aborde l'ACP sparse comme un problème de régression pénalisée. On suppose $\mathbf{X}$ centrée. On utilise la décomposition en valeurs singulières de $\mathbf{X}$ (Singular Value Decomposition (SVD)).

ACP comme un problème de régression

L'ACP peut être vue comme un problème de régression. Chaque composante principale d'une ACP est une combinaison linéaire de l'ensemble des variables de départ. On considère la SVD de $\mathbf{X}$. Sachant que les composantes principales sont des combinaisons de variables

(dont les coefficients sont les "loadings"), on souhaite retrouver les valeurs de ces "loadings". Les estimateurs des moindres carrés ordinaires $\hat{\boldsymbol{\beta}}$ sont obtenus en minimisant le problème suivant :

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{f}_\ell - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{1.22}$$

avec $\mathbf{f}_\ell$ étant la ℓ-ème composante principale. La solution est donnée par :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{f}_\ell \tag{1.23}$$

Dans le cas de données de très grandes dimensions, $\text{rank}(\mathbf{X}) < J$. On considère alors la matrice pseudo-inverse de Moore-Penrose $\mathbf{X}^+$ définie par :

$$\mathbf{X}^+ = \mathbf{Q}\boldsymbol{\Delta}^+\mathbf{P}^T \tag{1.24}$$

où $\boldsymbol{\Delta}^+$ une matrice diagonale contenant l'inverse des éléments non-nuls de la diagonale de $\boldsymbol{\Delta}$. L'expression (1.23) devient :

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^+\mathbf{f}_\ell \tag{1.25}$$

Sachant que $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, on peut en déduire que le vecteur $\mathbf{Q}^T\mathbf{q}_\ell$ est la $\ell-$ème colonne de la matrice identité de dimension $L \times L$, si $\mathbf{q}_\ell$ est le ℓ-ème loading.
Ainsi $\mathbf{Q}^T\mathbf{q}_\ell = [0, \ldots, 1, \ldots, 0]^T$ de dimension $L \times 1$. Sachant que $\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$:

$$\begin{aligned} \mathbf{X}\mathbf{q}_1 &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{q}_\ell \\ &= \mathbf{F}\mathbf{Q}^T\mathbf{q}_\ell \\ &= \mathbf{F}\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{f}_i \end{aligned} \tag{1.26}$$

La solution devient :

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{X}^+\mathbf{f}_\ell \\ &= \mathbf{X}^+\mathbf{X}\mathbf{q}_\ell \\ &= \mathbf{q}_\ell. \end{aligned} \tag{1.27}$$

Ainsi, le ℓ-ème loading $\mathbf{q}_\ell$ peut être retrouvé en régressant les composantes principales sur les J variables car chaque colonne de $\mathbf{F}$ est combinaison linéaire des J variables. Ainsi la

SVD et donc l'ACP peuvent effectivement être vues comme un problème de type régression.

Version sparse de l'ACP

La Sparse Principal Component Analysis (SPCA) définie par Zou et al. [2006] utilise le fait que l'ACP puisse être considérée comme un problème de régression pour introduire une pénalité de type elastic net dans le critère de régression (1.22). Cela permet de produire de la sparsité au niveau des "loadings" et d'obtenir des composantes plus facilement interprétables. Le critère de la SPCA s'écrit :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Z}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1, \tag{1.28}$$

avec $\mathbf{Z}_i = \mathbf{X}\mathbf{q}_i$. On développe l'expression de la façon suivante :

$$\begin{aligned}\|\mathbf{Z}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1 &= \|\mathbf{X}\mathbf{q}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1 \\ &= (\mathbf{q}_i - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\mathbf{q}_i - \boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1.\end{aligned} \tag{1.29}$$

L'algorithme SPCA est décrit comme l'algorithme 2.

Algorithm 2 Algorithme SPCA de Zou et al. [2006]

Initialisation. Poser $\mathbf{A} = \mathbf{Q}[, 1:k]$, les "loadings" des k premières composantes principales.

Etape 2. Pour $\mathbf{A} = [\alpha_1, \ldots, \alpha_k]$ fixé, résoudre le problème elastic net suivant pour $j = 1, \ldots, J$:

$$\boldsymbol{\beta}_j = \arg\min_{\boldsymbol{\beta}} (\alpha_j - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\alpha_j - \boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1 \tag{1.30}$$

Etape 3. Pour $\mathbf{B} = [\beta_1, \ldots, \beta_j]$ fixé, calculer la SVD de $\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ et poser $\mathbf{A} = \mathbf{P}\mathbf{Q}^T$.

Etape 4. Répéter les étapes 2 et 3 jusqu'à convergence.

Normalisation. $\hat{\mathbf{q}}_j = \dfrac{\beta_j}{\|\beta_j\|}$, pour $j = 1, \ldots, k$.

Les paramètres λ_1 et λ_2 sont des paramètres de régularisation. Lorsque $I > J$, $\lambda_1 = 0$ peut être le choix par défaut car une régularisation n'est pas nécessaire étant donné que nous ne nous plaçons pas dans le cas où le nombre de variables et supérieur à celui des individus.

De manière générale, λ_1 est un nombre positif et petit pour surmonter les problèmes de colinéarité dans la matrice $\mathbf{X}$. En principe, plusieurs valeurs de λ_2 peuvent être testées successivement pour aider au choix des paramètres de réglage puisque l'algorithme ci-dessus converge rapidement. L'algorithme LARS-EN permet d'obtenir une séquence d'approximation "sparse" de chaque composante principale et les valeurs de λ_2 correspondantes. Par conséquent, le λ_2 choisi sera celui qui donne un bon compromis entre la variance et la "sparsité". Cette méthode permet alors de fixer certains "loadings" à zéro afin de réduire le nombre de variables explicatives et faciliter ainsi l'interprétation des composantes "sparse" résultantes.

Propriétés et variance ajustée

En ACP, les composantes principales ne sont pas corrélées et les "loadings" sont orthogonaux. En effet si l'on considère la SVD de $\mathbf{X} = \mathbf{F}\mathbf{Q}^T$ avec $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ et $\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$ la matrice de covariance de $\mathbf{X}$, on a alors $\mathbf{F} = \mathbf{X}\mathbf{Q}$ et $\mathbf{F}^T\mathbf{F} = \mathbf{Q}^T\mathbf{X}^T\mathbf{X}\mathbf{Q} = \mathbf{Q}^T\mathbf{\Sigma}$ qui est donc diagonale. Ainsi les composantes sont non corrélées et les loadings sont orthogonaux. Dans le cas de l'ACP avec rotation par exemple, on a $\mathbf{Q}_{\text{rot}} = \mathbf{X}\mathbf{B} = \mathbf{X}\mathbf{Q}\mathbf{T} = \mathbf{F}\mathbf{T}$ avec $\mathbf{T}$ la matrice de rotation, alors $\mathbf{B}^T\mathbf{B} = \mathbf{T}^T\mathbf{Q}^T\mathbf{Q}\mathbf{T} = \mathbf{T}^T\mathbf{T} = \mathbf{I}$ si la rotation est orthogonale. Par ailleurs, $\mathbf{Q}_{\text{rot}}^T\mathbf{Q}_{\text{rot}} = \mathbf{T}^T\mathbf{Q}\mathbf{Q}\mathbf{T} = \mathbf{T}^T\mathbf{\Delta}^2\mathbf{T}$ qui est non diagonale. La propriété de non corrélation des composantes n'est donc pas satisfaite dans ce cas là. La simultanéité de ces propriétés est caractéristique de l'ACP et ne peut donc pas être retrouvée en ACP sparse. Jolliffe et al. [2003] ayant forcé l'orthogonalité des "loadings" pour la méthode ScoTLASS, la propriété de non corrélation a alors été sacrifiée. La SPCA n'impose aucune des deux propriétés. Si $\hat{\mathbf{Z}}$ désigne les composantes principales modifiées ("sparse") alors la variance totale expliquée par $\hat{\mathbf{Z}}$ est $\text{tr}(\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})$ qui est cependant trop optimiste si les $\hat{\mathbf{Z}}$ sont corrélées. Zou et al. [2006] ont proposé une nouvelle formule pour calculer la variance totale expliquée par $\hat{\mathbf{Z}}$ qui prend en compte les corrélations entre les $\mathbf{X}$. Ils utilisent la régression par projection afin de supprimer la dépendance linéaire entre les composantes corrélées. On pose $\hat{\mathbf{Z}}_{j.1,\ldots,j-1}$ les résidus après ajustement de $\hat{\mathbf{Z}}_j$ sur $\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_{j-1}$, c'est-à-dire :

$$\hat{\mathbf{Z}}_{j.1,\ldots,j-1} = \hat{\mathbf{Z}}_j - \mathbf{H}_{j.1,\ldots,j-1}\hat{\mathbf{Z}}_j, \tag{1.31}$$

où $\mathbf{H}_{j.1,\ldots,j-1}$ est la matrice de projection de $\{\mathbf{Z}_i\}_i^{j-1}$. Alors la variance ajustée de $\hat{\mathbf{Z}}_j$ est

$\left\|\hat{\mathbf{Z}}_{j.1,\ldots,j-1}\right\|_2^2$ et la variance totale expliquée est définie comme $\sum_{j=1}^{k}\left\|\hat{\mathbf{Z}}_{j.1,\ldots,j-1}\right\|_2^2$.

1.1.2.3 Sparse PCA-rSVD de Shen and Huang [2008]

La Sparse Principal Component Analysis via SVD régularisée (SPCA-rSVD) développée par Shen and Huang [2008] est une autre approche de l'ACP pénalisée. Elle utilise le lien entre l'ACP et la décomposition en valeurs singulières d'une matrice afin d'extraire des composantes principales en résolvant un problème d'approximation de matrice de rang inférieur découlant des travaux de décomposition matricielle de Eckart and Young [1936]). De la même manière que précédemment, $\mathbf{X}$ est une matrice de rang L avec $\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ où $\mathbf{P}$ $(I \times L)$ et $\mathbf{Q}$ $(J \times L)$ sont orthonormales et $\boldsymbol{\Delta}$ $(L \times L)$ est la matrice diagonale des valeurs singulières. On pose $\mathbf{F}$=$\mathbf{P}\boldsymbol{\Delta}$ les composantes principales et $\mathbf{Q}$ les "loadings" correspondants.

Approximation par une matrice de rang inférieur

Une des propriétés importantes de la SVD est qu'elle permet une bonne reconstitution de la matrice originale (en termes des moindres carrés) par une matrice de rang inférieur (Higham [1988]). La meilleure matrice d'approximation de rang 1 $\mathbf{X}^{(1)}$ de $\mathbf{X}$ est la solution du problème d'optimisation suivant :

$$\arg\min_{\mathbf{X}^{(1)}} \left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_2^2 = \arg\min_{\mathbf{X}^{(1)}} \left(\mathrm{tr}\left((\mathbf{X} - \mathbf{X}^{(1)})^T(\mathbf{X} - \mathbf{X}^{(1)})\right)\right) \tag{1.32}$$

où

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T \tag{1.33}$$

La démonstration de cette égalité est fournie en annexe E.

Les meilleures matrices d'approximation de rang 1 pour les composantes suivantes peuvent être obtenues de manière séquentielle via l'approximation de rang 1 des matrices résiduelles. Les paires $(\delta_k \mathbf{p}_k, \mathbf{q}_k)$, $k > 1$, fournissent les meilleures approximations de rang 1 des matrices résiduelles correspondantes. Par exemple, $\delta_2 \mathbf{p}_2 \mathbf{q}_2^T$ est la meilleure approximation de rang-1 de la première matrice déflatée $\mathbf{X}^{\perp_1} = \mathbf{X} - \delta_1 \mathbf{p}_1 \mathbf{q}_1^T$. $\mathbf{X}^{\perp_1}$ est le complément

orthogonal de $\mathbf{X}$ et si la même procédure est répétée sur $\mathbf{X}^{\perp_1}$, le premier vecteur singulier de $\mathbf{X}^{\perp_1}$ sera le second de la matrice $\mathbf{X}$. L'approximation de matrices de rang inférieur peut être généralisée au rang-L. Le problème d'optimisation devient alors :

$$\underset{\mathbf{X}^{(L)}}{\arg\min} \left\| \mathbf{X} - \mathbf{X}^{(L)} \right\|_2^2 = \underset{\mathbf{X}^{(L)}}{\arg\min} \left(\mathrm{tr} \left((\mathbf{X} - \mathbf{X}^{(L)})^T (\mathbf{X} - \mathbf{X}^{(L)}) \right) \right) \tag{1.34}$$

et la solution est :

$$\mathbf{X}^{(L)} \equiv \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^T \equiv \sum_{\ell=1}^{L} \mathbf{f}_\ell \mathbf{q}_\ell^T, \tag{1.35}$$

avec $\mathbf{X}^{(L)}$ la matrice d'approximation de rang-L la plus proche de $\mathbf{X}$. Une démonstration (pour la norme $L-2$ est proposée dans Golub and Van Loan [1983] à la page 72.

SVD régularisée

Dans le cas de données de grandes dimensions, le nombre de "loadings" non nuls est très élevé et l'interprétation des résultats reste difficile. Une solution à ce problème est d'introduire de la "sparsité" afin d'éliminer certains "loadings". Shen and Huang [2008] ont adapté la SVD pour calculer des matrices d'approximation de rang inférieur d'une matrice sous diverses contraintes introduisant des pénalités. La SVD peut être perçue comme un problème de type régression, ainsi pour créer de la "sparsité" sur les "loadings" $\mathbf{q}$, un seuillage est imposé sur les coefficients de régression grâce à une fonction de pénalisation dans le problème d'optimisation (1.32). Cependant, le vecteur des "loadings" $\mathbf{q}$ doit être de longueur unitaire afin d'obtenir une représentation unique. Cette contrainte rend une application d'une pénalisation sur $\mathbf{q}$ inappropriée. Pour surmonter cette difficulté, on réécrit $\mathbf{f}\mathbf{q}^T = \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T$, où $\tilde{\mathbf{f}}$ et $\tilde{\mathbf{q}}$ sont des versions ré-échelonnées de $\mathbf{f}$ et $\mathbf{q}$ tels que $\tilde{\mathbf{f}} = \delta\tilde{\mathbf{p}}$ avec $\tilde{\mathbf{p}}$ de longueur unitaire et $\tilde{\mathbf{q}}$ libre de toute contrainte de norme. Le problème d'optimisation devient :

$$\underset{\tilde{\mathbf{f}},\tilde{\mathbf{q}}}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_2^2 + P_\lambda(\tilde{\mathbf{q}}) \tag{1.36}$$

où $P_\lambda(\tilde{\mathbf{q}})$ est une fonction de pénalisation et λ est un paramètre de régularisation.

La fonction de pénalisation fixe certains éléments de $\tilde{\mathbf{q}}$ à zéro, ce qui produit de la "sparsité' au sein des "loadings" et des variables sont supprimées. L'algorithme permettant de résoudre le problème (1.36) est un algorithme itératif qui minimise $\left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_2^2 + P_\lambda(\tilde{\mathbf{q}})$

en fonction de $\tilde{\mathbf{f}}$ et $\tilde{\mathbf{q}}$ sous la contrainte $\|\tilde{\mathbf{f}}\|_2 = \delta$. Dans un premier temps, pour $\tilde{\mathbf{q}}$ fixé, le $\tilde{\mathbf{f}}$ solution de (1.36) est :

$$\tilde{\mathbf{f}} = \mathbf{X}\tilde{\mathbf{q}}/\|\mathbf{X}\tilde{\mathbf{q}}\|_2. \tag{1.37}$$

A présent, pour un $\tilde{\mathbf{f}}$ fixé, on réécrit $\|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\|_2^2 + P_\lambda(\tilde{\mathbf{q}})$ de la manière suivante :

$$\begin{aligned} \|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\|_2^2 + P_\lambda(\tilde{\mathbf{q}}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \tilde{f}_i\tilde{q}_j)^2 + \sum_{j=1}^{J} P_\lambda(|\tilde{q}_j|) \\ &= \sum_{j=1}^{J}\left(\sum_{i=1}^{I}(x_{ij} - \tilde{f}_i\tilde{q}_j)^2 + P_\lambda(|\tilde{q}_j|)\right) \end{aligned} \tag{1.38}$$

On peut donc optimiser chacune des composantes séparément. En développant le carré et sachant que $\sum_{i=1}^{I} \tilde{f}_i^2 = \delta^2$, on obtient :

$$\begin{aligned} \sum_{i=1}^{I}(x_{ij} - \tilde{f}_i\tilde{q}_j)^2 &= \sum_{i=1}^{I} x_{ij}^2 - 2\sum_{i=1}^{I} x_{ij}\tilde{f}_i\tilde{q}_j + \sum_{i=1}^{I}\tilde{f}_i^2\tilde{q}_j^2 \\ &= \sum_{i=1}^{I} x_{ij}^2 - 2(\mathbf{X}^T\tilde{\mathbf{f}})_j\tilde{q}_j + \delta^2\tilde{q}_j^2 \end{aligned} \tag{1.39}$$

Ainsi le $\tilde{q}_j$ optimal minimise $\delta^2\tilde{q}_j^2 - 2(\mathbf{X}^T\tilde{\mathbf{f}})_j\tilde{q}_j + P_\lambda(|\tilde{q}_j|)$ et dépend du choix de P_λ. La pénalité peut être par exemple la pénalisation LASSO, une pénalisation impliquant un seuillage doux ("soft thresholding") ou fort ("hard thresholding"). Ces deux derniers seuillages sont définis de la manière suivante :

$$P_\lambda(x)^{hard} = \begin{cases} 0 & \text{si} \quad |x| \leq \lambda \\ x & \text{si} \quad |x| > \lambda \end{cases} \tag{1.40}$$

$$P_\lambda(x)^{soft} = \begin{cases} 0 & \text{si} \quad |x| < \lambda \\ x - \lambda & \text{si} \quad x \geq \lambda \\ x + \lambda & \text{si} \quad x \leq -\lambda \end{cases} \tag{1.41}$$

et présentés dans la figure 1.4. En conclusion, le $\tilde{\mathbf{q}}$ minimisant (1.36) est obtenu en appliquant une règle de seuillage h_λ au vecteur $\mathbf{X}^T\tilde{\mathbf{f}}$ composante par composante.

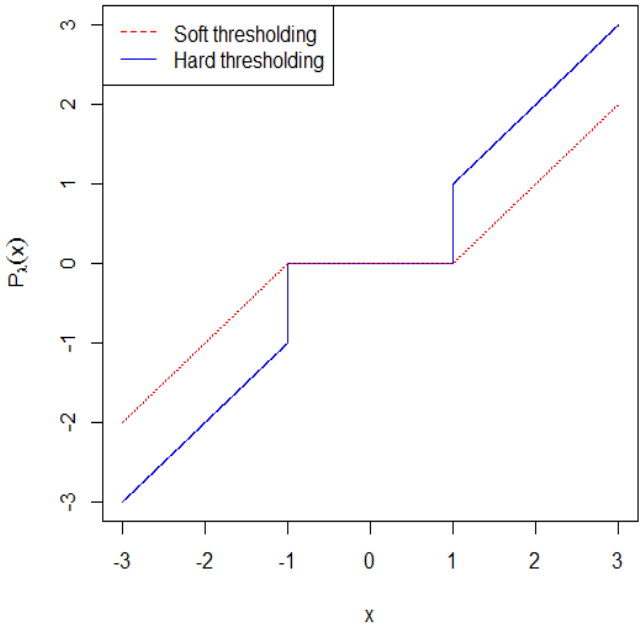

FIGURE 1.4 – Représentation des pénalisations "soft thresholding" (en pointillés rouges) et "hard thresholding" (en bleu).

L'algorithme itératif SPCA-rSVD est l'algorithme 3.

Algorithm 3 Algorithme sPCA-rSVD

Initialisation. Application de la SVD sur $\mathbf{X}$ pour obtenir la meilleure approximation de rang-1 de $\mathbf{X}$, avec $\mathbf{f}^*\mathbf{q}^*$ où $\|\mathbf{f}^*\|_2 = \delta$ et $\|\mathbf{q}^*\|_2 = 1$. On pose $\tilde{\mathbf{q}}_{old} = \mathbf{q}^*$ et $\tilde{\mathbf{f}}_{old} = \mathbf{f}^*$.

Itération.

a) $\tilde{\mathbf{q}}_{new} = h_\lambda(\mathbf{X}^T\tilde{\mathbf{f}}_{old})$,

b) $\tilde{\mathbf{f}}_{new} = \dfrac{\mathbf{X}\tilde{\mathbf{q}}_{new}}{\|\mathbf{X}\tilde{\mathbf{q}}_{new}\|_2}$.

Réitération. On répète l'étape 2 en remplaçant $\tilde{\mathbf{p}}_{old}$ et $\tilde{\mathbf{q}}_{old}$ par $\tilde{\mathbf{p}}_{new}$ et $\tilde{\mathbf{q}}_{new}$ juqu'à convergence.

Standardisation. On standardise $\tilde{\mathbf{q}}_{new}$ final : $\mathbf{q} = \tilde{\mathbf{q}}_{new}/\|\tilde{\mathbf{q}}_{new}\|_2$ les nouveaux "loadings sparse".

Il est défini seulement pour des vecteurs de dimension 1. Pour obtenir les "loadings sparse" sur les dimensions > 1, il faut appliquer cet algorithme sur l'approximation de rang-1 des matrices résiduelles. L'algorithme ne fait intervenir que la matrice $\mathbf{Q}$, la méthode est donc applicable dans un contexte de données de très grandes dimensions ($J \gg I$) et le temps de calcul est relativement court.

Propriétés et variance ajustée

La sélection de variables s'effectue axe par axe. Comme nous l'avons vu précédemment, les propriétés de non corrélation des composantes et l'orthogonalité des "loadings" sont perdues en ACP sparse. Shen and Huang [2008] donnent une nouvelle définition de la variance expliquée par les composantes principales en réponse à la perte de ces propriétés. On considère $\mathbf{Q}_k = [\mathbf{q}_1, \ldots, \mathbf{q}_k]$ la matrice des k premiers "loadings sparse" et $\mathbf{X}_k$ la projection de $\mathbf{X}$ sur le sous espace k-dimensionnel créé par les k premiers "loadings sparse" telles que

$$\mathbf{X}_k = \mathbf{X}\mathbf{Q}_k(\mathbf{Q}_k^T\mathbf{Q}_k)^{-1}\mathbf{Q}_k^T. \tag{1.42}$$

La variance totale expliquée par les k premières composantes est $\mathrm{tr}(\mathbf{X}_k^T\mathbf{X}_k)$ et la variance ajustée de la k-ème composante

$$\mathrm{tr}(\mathbf{X}_k^T\mathbf{X}_k) - \mathrm{tr}(\mathbf{X}_{k-1}^T\mathbf{X}_{k-1}). \tag{1.43}$$

On peut également définir le pourcentage cumulé de variance expliquée (Cumulative Percentage of Explained Variance (CPEV))par les premières k CPs comme étant

$$\mathrm{tr}(\mathbf{X}_k^T\mathbf{X}_k)/\,\mathrm{tr}(\mathbf{X}^T\mathbf{X}). \tag{1.44}$$

Le CPEV peut être utilisé pour déterminer le nombre de composantes nécessaires.

1.1.3 Méthodes PLS et Sparse PLS

Plaçons-nous à présent dans un cadre très général de méthodes d'analyse des données dans un contexte supervisé permettant d'étudier l'effet d'un bloc de variables observées sur les mêmes individus sur un bloc d'une ou plusieurs variables réponses.

La régression PLS (Partial Least Squares) proposée par Wold et al. [1983], peut être utilisée dans ce contexte (pour une presentation synthetique et historique voir Tenenhaus [1998]). Elle permet de relier un bloc de variables à expliquer $\mathbf{X}$, à un bloc de variables explicatives $\mathbf{Y}$ et ce, même dans un contexte de données de grandes dimensions ($J > I$). On obtient les composantes PLS par application successive de l'analyse factorielle inter-batteries de Tucker [1958]. L'algorithme Nonlinear Iterative PArtial Least Squares (NIPALS) permet le traitement de données manquantes. La méthode NIPALS, développée par Wold [1966], permet d'étudier un seul bloc de variables (K=1). Elle conduit à l'analyse en composantes principales lorsqu'il n'y a aucune donnée manquante, et fonctionne également lorsqu'il y en a.

1.1.3.1 Méthode PLS : Partial Least Squares

La régression PLS (Wold et al. [1983], Wold et al. [2001]) permet de modéliser la liaison entre un bloc de variables à expliquer $\mathbf{Y}$ et un bloc de variables explicatives $\mathbf{X}$ (Tenenhaus [1998]). La méthode consiste à remplacer une matrice des données prédictives $\mathbf{X}$ par une nouvelle matrice, dérivée de $\mathbf{X}$, que l'on désigne par $\mathbf{T}$, comprenant le même nombre de lignes (observations) que $\mathbf{X}$, mais un nombre de colonnes H très inférieur à J. On impose, de plus, que les colonnes de la matrice $\mathbf{T}$ soient des combinaisons linéaires des variables d'origine. Sous forme matricielle, la relation peut s'écrire $\mathbf{T} = \mathbf{XW}$ avec $\mathbf{W}$ la matrice $(J \times H)$ des coefficients définissant les combinaisons linéaires. $\mathbf{T}$ est donc une nouvelle matrice de composantes orthogonales dont les colonnes $\mathbf{t}_h$ sont obtenues par combinaison linéaire des variables d'origine. Le calcul des composantes $\mathbf{T}$ se fait en tenant compte des variables à prédire $\mathbf{Y}$. Plus précisément, on cherche à effectuer une double modélisation correspondant aux deux relations :

$$\begin{aligned} \mathbf{X} &= \mathbf{TA}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{TB}^T + \mathbf{F} \end{aligned} \tag{1.45}$$

avec $\mathbf{A}$ $(J \times H)$ et $\mathbf{B}$ $(Q \times K)$ les matrices des "loadings" telles que leurs colonnes soient, respectivement :

$$\mathbf{a}_h = \mathbf{X}_{h-1}^T \mathbf{t}_h / (\mathbf{t}_h^T \mathbf{t}_h) \tag{1.46}$$
$$\mathbf{b}_h = \mathbf{Y}_{h-1}^T \mathbf{t}_h / (\mathbf{t}_h^T \mathbf{t}_h)$$

et $\mathbf{E}$ $(I \times J)$ et $\mathbf{F}$ $(I \times Q)$, soient les matrices des résidus associées à la prédiction de $\mathbf{X}$ et de $\mathbf{Y}$, respectivement, $h = 1, \ldots, H$.

Suivant le type de variable réponse considérée, on parlera de régression PLS1 (une seule variable $\mathbf{y}$ à expliquer) et de régression PLS2 (plusieurs variables à expliquer).

Régression PLS1

La régression PLS1 consiste à relier une seule variable à expliquer $\mathbf{y}$ à un bloc de variables explicatives $\mathbf{X}$. Le tableau $\mathbf{X}$ est composé de J colonnes notées $\mathbf{x}_j$, centrées et réduites. L'algorithme de la PLS1 débute par la recherche de m composantes orthogonales $\mathbf{t}_h = \mathbf{X}\mathbf{a}_h$ bien explicatives de leur propre bloc et corrélées à $\mathbf{y}$. (Le nombre m est obtenu par validation croisée.) Pour ce faire, pour chaque $h = 1, \ldots, m$, on recherche des composantes $\mathbf{t}_h = \mathbf{X}\mathbf{a}_h$ maximisant le critère

$$\mathrm{Cov}(\mathbf{X}\mathbf{a}_h, \mathbf{y}) \tag{1.47}$$

sous la contrainte $\|\mathbf{a}_h\| = 1$ et sous contrainte d'orthogonalité entre $\mathbf{t}_h$ et les composantes précédentes $\mathbf{t}_1, \ldots, \mathbf{t}_{h-1}$.

Dans un deuxième temps une régression de $\mathbf{y}$ sur les composantes PLS $\mathbf{t}_h$ est réalisée et la régression en fonction de $\mathbf{X}$ est exprimée.

Régression PLS2

La régression PLS2 consiste à relier un bloc $\mathbf{Y}$ contenant Q variables à expliquer à un bloc de variables explicatives $\mathbf{X}$. L'algorithme de la PLS2 débute par la recherche de m composantes orthogonales $\mathbf{t}_h = \mathbf{X}\mathbf{a}_h$ et m composantes $\mathbf{u}_h = \mathbf{Y}\mathbf{b}_h$, bien corrélées entre elles et explicatives de leur propre bloc. Pour ce faire, pour chaque $h = 1, \ldots, m$, on recherche des composantes $\mathbf{t}_h = \mathbf{X}\mathbf{a}_h$ et $\mathbf{u}_h = \mathbf{Y}\mathbf{b}_h$ maximisant le critère

$$\mathrm{Cov}(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h) \tag{1.48}$$

sous des contraintes de norme et d'orthogonalité entre $\mathbf{t}_h$ et les composantes précédentes $\mathbf{t}_1, \ldots, \mathbf{t}_{h-1}$. Dans un deuxième temps une régression de $\mathbf{Y}$ sur les composantes PLS $\mathbf{t}_h$ est réalisée et la régression en fonction de $\mathbf{X}$ est exprimée.

Choix du nombre de composantes PLS

Le nombre de composantes PLS est généralement déterminé par validation croisée. On définit pour la h-ème composante les critères du PREdiction Sum of Squares (PRESS) et du Residual Sum of Squares (RSS) par les formules suivantes :

$$PRESS_h = \sum_{i=1}^{I} (\mathbf{y}_i - \hat{\mathbf{y}}_{h(-i)})^2, \tag{1.49}$$

avec $\hat{\mathbf{y}}_{h(-i)}$ la prévision de $\mathbf{y}_i$ à partir du modèle à h composantes estimé sans l'observation i. On choisit donc h tel que PRESS_h soit le plus petit possible. Par ailleurs,

$$RSS_h = \sum_{i=1}^{I} (\mathbf{y}_i - \hat{\mathbf{y}}_{hi})^2, \tag{1.50}$$

avec $\hat{\mathbf{y}}_{hi}$ la prévision de $\mathbf{y}_i$ à partir du modèle estimé à h composantes.

On définit le pouvoir prédictif de la h-ème composante par :

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}} \quad \text{avec } RSS_0 = \sum_{i=1}^{I} (\mathbf{y}_i - \bar{\mathbf{y}})^2. \tag{1.51}$$

On peut finalement mesurer le pouvoir prédictif du modèle par le critère :

$$Q^2 = 1 - \prod_{h=1}^{H} \frac{PRESS_h}{RSS_{h-1}}. \tag{1.52}$$

La régression PLS permet de faire face au problème où J (nombre de variables) $\gg I$ (nombre d'individus), cependant elle n'effectue pas de sélection de variables, alors que dans un contexte de données de très grande dimension, il y a une forte probabilité qu'un grand nombre de variables ne soient pas impliquées dans le phénomène étudié ou encore que beaucoup de variables soient corrélées entre elles. Il serait alors évidemment plus intéressant d'effectuer une sélection de variables auparavant.

1.1.3.2 Sparse PLS-SVD de Lê Cao et al. [2008]

La version pénalisée de la PLS proposée par Lê Cao et al. [2008] combine l'intégration et la sélection de variables simultanément. Une fonction de "seuillage doux" et une SVD

sont combinées à la PLS. Dans un premier temps, une brève introduction du principe de l'approche PLS-SVD est exposée (aussi souvent appelée PLS correlation en imagerie cérébrale, cf. Krishnan et al. [2011]) afin de mieux comprendre l'approche sparse PLS définie par Lê Cao et al. [2008].

On considère $\mathbf{X}$ la matrice des prédicteurs et $\mathbf{Y}$ la matrice $I \times Q$ des variables à expliquer. Nous reprenons ici les même notations que celles définies dans la partie notations pour définir la décomposition en valeurs singulières avec $\mathbf{M} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ où $\mathbf{P}$ $(I \times L)$ et $\mathbf{Q}$ $(I \times L)$ sont orthonormales et $\boldsymbol{\Delta}$ $(L \times L)$ est une matrice diagonale. On rappelle que les valeurs singulières sont les racines carrées des valeurs propres des matrices $\mathbf{M}^T\mathbf{M}$ et $\mathbf{M}\mathbf{M}^T$. Les colonnes de $\mathbf{P}$ et $\mathbf{Q}$ correspondent aux "loadings" PLS de $\mathbf{X}$ et $\mathbf{Y}$ si $\mathbf{M} = \mathbf{X}^T\mathbf{Y}$.

PLS-SVD et SPLS-SVD

En PLS-SVD la décomposition SVD de $\mathbf{M} = \mathbf{X}^T\mathbf{Y}$ est réalisée une seule fois, et pour chaque dimension h, $\mathbf{M}$ est directement déflatée par sa matrice d'approximation de rang-1 ($\mathbf{M}_h = \mathbf{M}_{h-1} - \delta_h \mathbf{p}_h \mathbf{q}_h^T$). Comme précédemment énoncé dans le paragraphe 1.1.2.3, Shen and Huang [2008] ont proposé une approche sparse PCA utilisant la SVD de $\mathbf{M} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ et en pénalisant les "loadings" $\mathbf{q}_j$ de l'ACP. La SPLS-SVD s'appuie sur l'algorithme développé par Shen and Huang [2008] en considérant la pénalisation comme une pénalisation de type "seuillage doux". Dans le cas précis de la SPLS-SVD, l'important est de pouvoir pénaliser à la fois les vecteurs "loadings" $\mathbf{p}_j$ et $\mathbf{q}_j$ afin de réaliser une sélection de variables dans les deux blocs de données en même temps. Une propriété importante de la PLS est l'interprétabilité des "loadings" comme mesure relative de l'importance des variables dans le modèle (Wold et al. [2004]). Le problème d'optimisation devient alors :

$$\underset{\mathbf{p},\mathbf{q}}{\arg\min} \|\mathbf{M} - \mathbf{p}\mathbf{q}^T\|_F^2 + P_{\lambda_1}(\mathbf{p}) + P_{\lambda_2}(\mathbf{q}). \tag{1.53}$$

avec P_λ une fonction de "seuillage doux" comme définie dans l'expression (1.41).

Ce problème peut se résoudre de manière itérative en remplaçant $\mathbf{X}$ par $\mathbf{M}$ dans les étapes d'itération de l'algorithme défini dans le paragraphe 1.1.2.3 :

$$\mathbf{q}_{\text{new}} = P_{\lambda_1}(\mathbf{M}_{h-1}^T \mathbf{p}_{\text{old}}) \tag{1.54}$$

$$\mathbf{p}_{\text{new}} = P_{\lambda_2}(\mathbf{M}_{h-1} \mathbf{q}_{\text{old}}) \tag{1.55}$$

L'algorithme Sparse Partial Least Squares (SPLS) détaillé ci-après (algorithme 4) est basé sur l'algorithme PLS introduit dans le paragraphe 1.1.3.1 et sur le calcul de la décomposition SVD de $\mathbf{M}$ pour chaque dimension. Dans le cas où il n'y a pas de contrainte

Algorithm 4 Algorithme SPLS-SVD

Initialisation. $\mathbf{X}_0 = \mathbf{X}$ $\mathbf{Y}_0 = \mathbf{Y}$ Pour $h = 1, ..., H$:

Étape 1. On pose $\mathbf{M}_{h-1} = \mathbf{X}_{h-1}^T \mathbf{Y}_{h-1}$
Étape 2. On décompose $\mathbf{M}_{h-1}$ et on extrait la première paire de vecteurs singuliers $\mathbf{p}_{\text{old}} = \mathbf{p}_h$ et $\mathbf{q}_{\text{old}} = \mathbf{q}_h$

Étape 3. Jusqu'à convergence de $\mathbf{p}_{\text{new}}$ et $\mathbf{q}_{\text{new}}$:

i. $\mathbf{p}_{\text{new}} = P_{\lambda_2}(\mathbf{M}_{h-1}\mathbf{q}_{\text{old}})$, normalisation de $\mathbf{p}_{\text{new}}$
ii. $\mathbf{q}_{\text{new}} = P_{\lambda_1}(\mathbf{M}_{h-1}\mathbf{p}_{\text{old}})$, normalisation de $\mathbf{q}_{\text{new}}$
iii. $\mathbf{p}_{\text{old}} = \mathbf{p}_{\text{new}}$ et $\mathbf{q}_{\text{old}} = \mathbf{q}_{\text{new}}$

Étape 4.
$\mathbf{t}_h = \mathbf{X}_{h-1}\mathbf{p}_{\text{new}} / \mathbf{p}_{\text{new}}^T \mathbf{p}_{\text{new}}$
$\mathbf{w}_h = \mathbf{Y}_{h-1}\mathbf{q}_{\text{new}} / \mathbf{q}_{\text{new}}^T \mathbf{q}_{\text{new}}$

Étape 5.
$\mathbf{a}_h = \mathbf{X}_{h-1}\mathbf{t}_h / \mathbf{t}_h^T \mathbf{t}_h$
$\mathbf{b}_h = \mathbf{Y}_{h-1}\mathbf{t}_h / \mathbf{t}_h^T \mathbf{w}_h$

Étape 6.
$\mathbf{X}_h = \mathbf{X}_{h-1} - \mathbf{t}_h \mathbf{a}_h^T$
$\mathbf{Y}_h = \mathbf{Y}_{h-1} - \mathbf{t}_h \mathbf{b}_h^T$

de sparsité ($\lambda_1 = \lambda_2 = 0$), on obtient alors les mêmes résultats que dans une PLS classique.

Choix des paramètres de pénalisation

Les deux paramètres de pénalisation (λ_1 et λ_2) peuvent être choisis simultanément en calculant l'erreur de prédiction Root Mean Squared Error Prediction (RMSEP) par validation croisée (K-fold cross validation), ce pour chaque dimension h. Lors du calcul "optimal"

des paramètres de pénalisation en optimisant le pouvoir prédictif du modèle, il se peut que le nombre de variables restantes soit encore trop élevé pour permettre aux experts du domaine d'interpréter correctement les résultats. Pour y remédier Lê Cao et al. [2008] proposent de fixer le nombre de coefficients non nuls dans chacun des vecteurs des "loadings" $\mathbf{p}_h$ et $\mathbf{q}_h$, pour chaque dimension, afin de répondre de manière plus précise aux besoins des experts dans le cas de données de grande dimension.

1.1.3.3 Sparse PLS de Chung and Keles [2010]

Contrairement à la sparse PLS définit par Lê Cao et al. [2008], Chung and Keles [2010] imposent la sparsité durant la construction des vecteurs directeurs, afin de produire des variables latentes dépendant d'un nombre réduit de prédicteurs (Chung and Keles [2010]). On considère la matrice $\mathbf{X}$ des prédicteurs, la matrice $\mathbf{Y}$ des variables réponses et la matrice $\mathbf{W}$ de dimension $J \times H$ $(1 \leq H \leq \min(I, J))$ des vecteurs directeurs. L'objectif principal de la PLS est de trouver ces vecteurs directeurs. Le h-ème vecteur directeur $\mathbf{w}_h$ est obtenu en résolvant le problème d'optimisation suivant :

$$\begin{aligned} &\arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{M}\mathbf{w}, \\ \text{ssc.} \quad &\mathbf{w}^T\mathbf{w} = 1 \quad \text{et } \mathbf{w}^T\mathbf{S}_{\mathbf{XX}}\widehat{\mathbf{w}}_\ell = 0 \quad \ell = 1, \ldots, k-1 \end{aligned} \tag{1.56}$$

avec $\mathbf{M} = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ et $\mathbf{S}_{\mathbf{XX}}$ la matrice de covariance des prédicteurs.

La SPLS développée par Chung and Keles [2010] introduit une sélection de variable dans la PLS en résolvant le problème de minimisation suivant, au lieu de la formule originale de la PLS (1.56) :

$$\arg\min_{\mathbf{w}} \left(-k\mathbf{w}^T\mathbf{M}\mathbf{w} + (1-k)(\mathbf{c}-\mathbf{w})^T\mathbf{M}(\mathbf{c}-\mathbf{w}) + \lambda_1\|\mathbf{c}\|_1 + \lambda_2\|\mathbf{c}\|_2\right), \tag{1.57}$$

sous la contrainte $\mathbf{w}^T\mathbf{w} = 1$, où $\mathbf{M} = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$.

La pénalité L_1 impose la sparsité à un vecteur directeur $\mathbf{c}$ proche de la solution initiale $\mathbf{w}$. La pénalisation L_2 permet de prendre en compte la singularité de la matrice $\mathbf{M}$. L'introduction du paramètre k permet de contrôler la partie concave de la fonction $\mathbf{w}^T\mathbf{M}\mathbf{w}$. Le poids w sera donné par le c optimal (pour la résolution en c et en w).

Dans le cas où $c = w$, on retrouve le problème du SCoTLASS de Jolliffe et al. [2003] (section 1.1.2.1), k étant positif il n'a aucune influence sur l'optimum, modulo la pénalité sur la norme L_2 qui sera utile pour déterminer la solution (en pratique, on prendra $\lambda_2 = 0$, ce qui supprimera cette différence). La résolution de (1.57) se fait de manière itérative et alternativement à c ou w fixé, bien que la fonction d'objectif ait un terme concave non nul mais considéré comme négligeable quand k est petit. A c fixé, l'optimisation en w peut s'écrire comme un problème aux moindres carrés avec contraintes, qui se résout à l'aide des multiplicateurs de Lagrange. A w fixé, ce problème est équivalent à celui de l'elastic net (Zou and Hastie [2005]) et peut être résolu efficacement via l'algorithme LARS décrit section 1.1.1 (Efron et al. [2004]). Si la réponse $\mathbf{Y}$ est univariée, la solution de (1.57) est un vecteur directeur à seuillage doux :

$$\hat{\mathbf{c}} = \left(|\mathbf{Z}| - \frac{\lambda_1}{2}\right)_+ \text{sign}(\mathbf{Z}), \tag{1.58}$$

avec $\mathbf{Z} = \mathbf{X}^T\mathbf{Y}/\|\mathbf{X}^T\mathbf{Y}\|$ et $(x)_+ = \max(0, x)$.

Choix des paramètres de pénalisation

Bien que la formulation du problème (1.57) nous laisse penser qu'il y ait quatre paramètres de régularisation (k, λ_1, λ_2, H), il n'y a en réalité en SPLS que deux paramètres clés à déterminer : le paramètre de régularisation λ_1 et le nombre de composantes à conserver H.

Lorsque $\mathbf{Y}$ est univariée, la solution ne dépend pas de k. Pour $\mathbf{Y}$ multivariée, le choix d'un $k < 1/2$ permet d'éviter les problèmes de solution locale. Cependant plusieurs valeurs de k peuvent être testées. Par ailleurs, si on fait tendre λ_2 vers l'infini, la solution ne dépendra que de λ_1. Il faut donc trouver des critères de choix pour déterminer λ_1 et H. Si l'on considère un vecteur directeur à seuillage doux $\tilde{\mathbf{w}}$:

$$\tilde{\mathbf{w}} = \left(|\hat{\mathbf{w}}| - \eta \max_{1\leq i\leq I} |\hat{\mathbf{w}}_i|\right) \mathbf{I}\left(|\hat{\mathbf{w}}| \geq \eta \max_{1\leq i\leq I} |\hat{\mathbf{w}}_i|\right) \text{sign}(\hat{\mathbf{w}}), \quad 0 \leq \eta \leq 1. \tag{1.59}$$

Ici, η joue le rôle du paramètre de régularisation λ_1 dans la formule (1.58). Le paramètre η est sélectionné par validation croisée pour chaque vecteur directeur. En revanche, il n'y a pas de paramètre de régularisation pour chacun de ces vecteurs pour des raisons de temps de calcul. Cette approche ne permet pas de déterminer un minimum unique pour le critère

de validation croisée car différentes combinaisons de régularisation des vecteurs directeurs peuvent mener à la même prédiction $\mathbf{Y}$. Le choix du nombre de composantes H est quant à lui déterminé par validation croisée comme dans la PLS originale.

1.2 Méthodes multiblocs

Dans de nombreux domaines d'application, comme en biologie, en génétique ou encore en neurosciences, le cas de données explicatives organisées en blocs est souvent rencontré. Ces données peuvent être considérées comme K blocs de variables $\mathbf{X}_{[1]}, \ldots, \mathbf{X}_{[K]}$, où chaque bloc $\mathbf{X}_{[k]}$ (avec $k = 1, \ldots, K$) représente un jeu de p_k variables observées sur le même jeu de I individus. Lorsque la variable réponse est unique, des méthodes de régression multiblocs telle que la régression group LASSO peuvent être utilisées. Cette méthode est présentée dans la section suivante ainsi que sa version pénalisée dans le cas où l'on souhaiterait faire de la sélection de variables au sein du bloc.

L'étude des relations entre deux blocs de variables (sans considérer de manière explicite de variable à expliquer) peut être réalisée à l'aide d'une analyse canonique des corrélations (Canonical Correlation Analysis (CCA)). Cependant dans le cadre de données génomiques par exemple, le nombre de variables (gènes, SNPs, expression de gènes) est généralement supérieur au nombre d'individus et la CCA ne peut être utilisée directement. Une version régularisée a donc été proposée (González et al. [2009]).

Dans le cas où l'on souhaite étudier les liens entre plusieurs blocs de variables observées sur le même ensemble d'individus (trois ou plus), Tenenhaus and Tenenhaus [2011] ont proposé une méthode nommée "Regularized Generalized Canonical Correlation Analysis" (RGCCA) qui combine la puissance des méthodes d'analyse de données multiblocs (maximisation de critères bien définis) et la flexibilité de la PLS Path modeling (PLS-PM) (l'utilisateur décide quels blocs sont connectés et lesquels ne le sont pas) proposée par Wold [1985], Lohmöller [1989],Krämer [2007] et Vinzi [2010]. Un des points principaux de cette méthode, et qui la différencie d'autres méthodes basées sur la maximisation de fonctions de corrélations ou encore de fonctions de corrélations et covariances, repose sur le fait que tous

les blocs ne sont pas nécessairement connectés. La RGCCA et sa version sparse SGCCA (sélection de variables au sein des blocs) sont présentées sections 1.2.3 et 1.2.4.

1.2.1 Estimateur group LASSO

L'estimateur group LASSO, introduit par Yuan and Lin [2005] est une extension du LASSO pour la sélection de groupes de variables dans le cas où une seule variable réponse est à expliquer. Il considère le problème général de régression avec une fonction de pénalité étant un intermédiaire entre la pénalité L_1 utilisée dans le LASSO et la pénalité L_2 utilisée en régression ridge pour la sélection de groupes de variables.

Soit $\mathbf{y}$ une variable réponse $I \times 1$, et $\mathbf{X}$ une matrice de variables quantitatives composée de K sous-matrices $\mathbf{X}_{[k]}$, $k = 1, ..., K$, chacune de dimensions $I \times J_{[k]}$, avec $J_{[k]}$ le nombre de variables dans le groupe k. On définit $\beta_{[k]}$, le vecteur de coefficients de longueur $J_{[k]}$, $k = 1, ..., K$. On considère le problème général de régression avec K groupes :

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} + \boldsymbol{\varepsilon} \tag{1.60}$$

où $\boldsymbol{\varepsilon}$ le terme d'erreur. Étant données des matrices définies positives $\mathbf{W}_{[1]}, \ldots, \mathbf{W}_{[K]}$, l'estimateur group LASSO est défini comme la solution de :

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} \right\|_{\mathbf{W}}^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_{[k]}\|_{\mathbf{W}_{[k]}} \tag{1.61}$$

où $\lambda \geq 0$ est le paramètre de régularisation. On rappelle que $\|.\|_{\mathbf{W}}$ est la norme L_2 $\mathbf{W}$-généralisée définie comme suit :

$$\|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}\mathbf{W}\mathbf{x}^T} \tag{1.62}$$

Il y a plusieurs choix possibles pour la matrice $\mathbf{W}$. Dans Yuan and Lin [2005], leur choix s'est porté sur $\mathbf{W}_{[k]} = J_{[k]}\mathbf{I}$. Le group LASSO peut alors s'écrire :

$$\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} \right\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{J_{[k]}} \|\boldsymbol{\beta}_{[k]}\|_2, \tag{1.63}$$

Le minimum est obtenu lorsque la dérivée de la fonction est nulle. Cependant la norme L_2 n'est pas différentiable en 0. Nous pouvons donc utiliser les conditions d'optimalité de

Karush-Kuhn-Tucker pour l'optimisation convexe (Kuhn and Tucker [1951]). On obtient alors :

$$-\mathbf{X}_{[k]}^T(y-\mathbf{X}\boldsymbol{\beta}) = \lambda\sqrt{J_{[k]}}\mathbf{s}_{[k]}, \qquad k=1,\dots,K \tag{1.64}$$

avec $-\mathbf{X}_{[k]}^T(y-\mathbf{X}\boldsymbol{\beta})$ le gradient de $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ et où chaque $\mathbf{s}_{[k]} \in \partial\|\boldsymbol{\beta}_{[k]}\|_2$, c'est-à-dire,

$$\mathbf{s}_{[k]} = \begin{cases} \boldsymbol{\beta}_{[k]}/\|\boldsymbol{\beta}_{[k]}\|_2 & \text{si} \quad \boldsymbol{\beta}_{[k]} \neq 0 \\ \{\mathbf{z}\in\mathbb{R}^{J_{[k]}} : \|\mathbf{z}\|_2 \leq 1\} & \text{si} \quad \boldsymbol{\beta}_{[k]} = 0 \end{cases} \tag{1.65}$$

Ainsi une condition nécessaire et suffisante pour que $\boldsymbol{\beta}$ soit solution de (1.63) est la suivante :

$$-\mathbf{X}_{[k]}^T(y-\mathbf{X}\boldsymbol{\beta}) = \lambda\sqrt{J_{[k]}}\frac{\boldsymbol{\beta}_{[k]}}{\|\boldsymbol{\beta}_{[k]}\|_2} \qquad \text{si} \quad \boldsymbol{\beta}_{[k]} \neq 0 \tag{1.66}$$

$$\left\|-\mathbf{X}_{[k]}^T(y-\mathbf{X}\boldsymbol{\beta})\right\|_2 \leq \lambda\sqrt{J_{[k]}} \qquad \text{si} \quad \boldsymbol{\beta}_{[k]} = 0 \tag{1.67}$$

Ainsi, si $\boldsymbol{\beta}_{[k]} = 0$: $\left\|-\mathbf{X}_{[k]}^T(y-\mathbf{X}\boldsymbol{\beta})\right\|_2 \leq \lambda\sqrt{J_{[k]}}$. En revanche, si $\boldsymbol{\beta}_{[k]} \neq 0$, alors :

$$\boldsymbol{\beta}_{[k]} = \left(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]} - \frac{\lambda\sqrt{J_{[k]}}}{\|S_k\|}\right)_+ S_{[k]}, \tag{1.68}$$

où

$$S_{[k]} = \mathbf{X}_{[k]}^T(y - \sum_{j\neq k}\mathbf{X}_{[j]}\boldsymbol{\beta}_{[j]}) \tag{1.69}$$

$$= \mathbf{X}_{[k]}^T(y - \mathbf{X}\boldsymbol{\beta}_{-[k]}), \tag{1.70}$$

avec $\boldsymbol{\beta}_{-[k]} = (\boldsymbol{\beta}_{[1]}^T,\dots,\boldsymbol{\beta}_{[k-1]}^T, 0^T, \boldsymbol{\beta}_{[k+1]}^T,\dots,\boldsymbol{\beta}_{[K]}^T)$.

Si de plus, $\mathbf{X}_{[k]}$ est orthonormale ($\mathbf{X}_{[k]}^T\mathbf{X}_{[k]} = \mathbf{I}$), alors la solution de (1.66) s'écrit :

$$\boldsymbol{\beta}_{[k]} = \left(1 - \frac{\lambda\sqrt{J_{[k]}}}{\|S_k\|}\right)_+ S_{[k]}. \tag{1.71}$$

La solution de (1.63) peut être obtenue en appliquant de manière itérative (1.71) à $k=1,\dots,K$. L'algorithme est très stable et converge rapidement mais les temps de calculs augmentent lorsque le nombre de variables considérées croît. Les programmes sont utilisables et disponibles sur R dans le package "grplasso" (Meier [2013]).

Le group LASSO agit donc de la même manière que le LASSO mais au sein du groupe : en fonction de la valeur de λ, un groupe entier de prédicteurs sortira du modèle. Dans le cas où $K = 1$ (un seul groupe de variables), on retrouve le LASSO. Cependant, le group LASSO ne produit pas de sparsité au sein d'un groupe. En effet, si un groupe de paramètres est non nul, chaque paramètre du groupe sera non nul. Pour créer de la sparsité au sein d'un groupe et au niveau individuel, une pénalité plus générale, la sparse group LASSO, a été proposée par Simon et al. [2013], le but étant de sélectionner des groupes de variables et des prédicteurs au sein de ces groupes.

1.2.2 Estimateur Sparse group LASSO

Dans le cas de données de grande dimension, la méthode Sparse Group LASSO (SGL) permet une sélection de prédicteurs au sein des groupes de variables sélectionnés. Le critère est le suivant :

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_{[k]} \boldsymbol{\beta}_{[k]}\|^2 + (1-\alpha)\lambda \sum_{k=1}^{K} \sqrt{J_{[k]}} \|\boldsymbol{\beta}_{[k]}\|_2 + \alpha\lambda \|\boldsymbol{\beta}\|_1 \tag{1.72}$$

où $\alpha \in [0,1]$ est une combinaison convexe des pénalités LASSO et group LASSO (pour $\alpha = 0$ on retrouve la pénalité group LASSO, et pour $\alpha = 1$, celle du LASSO). Le vecteur $\boldsymbol{\beta} = (\boldsymbol{\beta}_{[1]}, ..., \boldsymbol{\beta}_{[K]})$ est le vecteur entier des paramètres. Cette pénalisation est proche de celle de l'elastic net commentée section 1.1.1 mais diffère du fait que la pénalité $\|.\|_2$ n'est pas différentiable en 0, ainsi des groupes entiers sont mis à zéro. L'expression (1.72), qui est une somme de fonctions convexes, est donc convexe. Il existe une solution optimale que l'on obtient à l'aide des équations du sous-gradient. Simon et al. [2013] ont proposé un algorithme efficace du gradient ("accelerated generalized gradient descent") pour ajuster le modèle. Les programmes et les algorithmes sont disponibles dans le package "SGL" du logiciel R (Noah et al. [2013]).

1.2.3 RGCCA: Regularized Generalized Canonical Correlation Analysis

Afin d'analyser les liens entre plusieurs blocs de variables (trois ou plus) observées sur le même ensemble d'individus, Tenenhaus and Tenenhaus [2011] ont proposé une méthode nommée RGCCA (Regularized Generalized Canonical Correlation Analysis) qui permet

de trouver des combinaisons de blocs de variables telles que : les composantes du bloc expliquent correctement leur propre bloc, et que les composantes des blocs connectés soient très corrélées. Cette méthode vise à extraire l'information partagée par les K blocs de variables, en tenant compte d'un graphe de connexions entre les blocs déterminé à priori.

Dans un contexte de données de très grande dimension, ou en présence de multicolinéarité entre les blocs, les méthodes basées sur la corrélation conduisent à de fausses relations entre les blocs. Cela donne une impression de liens entre des blocs qui s'avèrent ne pas être valides lorsqu'ils sont examinés objectivement. La RGCCA constitue une version régularisée de diverses méthodes basées sur la corrélation et rend possible l'analyse de blocs de données mal conditionnés. Elle permet la mise en place d'un continuum entre les critères basés sur la corrélation et la covariance.

On considère K blocs de variables $\mathbf{X}_{[1]}, \ldots, \mathbf{X}_{[K]}$, une matrice de "design" (conception des liens entre blocs) $\mathbf{C} = c_{kj}$ (avec $c_{kj} = 1$ si les blocs $\mathbf{X}_{[k]}$ et $\mathbf{X}_{[j]}$ sont liés, 0 sinon) qui fait office de graphe de connexions entre blocs, une fonction g et des constantes de régularisation $\tau_1, \ldots, \tau_K$ comprises entre 0 et 1. La RGCCA est définie comme le problème d'optimisation suivant :

$$\begin{aligned} &\underset{\mathbf{a}_1,\ldots,\mathbf{a}_K}{\arg\max} \sum_{k,j=1,k\neq j}^{K} c_{kj} g\left(\mathrm{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}, \mathbf{X}_{[j]}\mathbf{a}_{[j]})\right) \\ &\text{sous contrainte} \quad \tau_k \|\mathbf{a}_{[k]}\|^2 + (1-\tau_k)\,\mathrm{Var}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}) = 1, \quad k = 1, \ldots, K. \end{aligned} \tag{1.73}$$

Dans ce problème d'optimisation, g peut être définie comme l'identité, $g(x) = x$ (schéma de Horst proposé dans Krämer [2007]), la valeur absolue, $g(x) = |x|$ (schéma centroïde proposé dans Wold [1985]) ou encore la fonction carrée, $g(x) = x^2$ (schéma factoriel décrit dans Lohmöller [1989]). Le schéma de Horst pénalise la corrélation négative entre les composantes alors que les schémas centroïde et factoriel présentent des alternatives permettant à deux composantes d'être corrélées négativement. Le vecteur $\mathbf{a}_k$ fait référence au vecteur des poids externes et le vecteur $\mathbf{y}_{[k]} = \mathbf{X}_{[k]}\mathbf{a}_{[k]}$ à la composante externe (qui résume le bloc). La composante interne (qui tient compte des relations entre blocs) est définie de la

manière suivante :

$$\mathbf{z}_{[k]} = \sum_{j \neq k} e_{kj} \mathbf{y}_{[k]}. \tag{1.74}$$

Les poids internes e_{kj} sont choisis parmi les trois schémas présentés précédemment :

$$\begin{aligned} \text{Horst :} \quad & e_{kj} = c_{kj} \\ \text{Centroïde :} \quad & e_{kj} = c_{kj} \operatorname{sign}(\operatorname{Cor}(\mathbf{y}_{[j]}, \mathbf{y}_{[k]})) \\ \text{Factoriel :} \quad & e_{kj} = c_{kj} \operatorname{Cov}(\mathbf{y}_{[j]}, \mathbf{y}_{[k]}). \end{aligned} \tag{1.75}$$

Les équations de stationnarité sont obtenues en annulant les dérivées du Lagrangien associé au problème d'optimisation (1.73) (Tenenhaus and Tenenhaus [2011] p.259). On obtient alors la solution suivante :

$$\mathbf{a}_k = \frac{\left[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}_{[k]}^T \mathbf{X}_{[k]}\right]^{-1} \mathbf{X}_{[k]}^T \mathbf{z}_{[k]}}{\sqrt{\mathbf{z}_{[k]}^T \mathbf{X}_{[k]} \left[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}_{[k]}^T \mathbf{X}_{[k]}\right]^{-1} \mathbf{X}_{[k]}^T \mathbf{z}_{[k]}}}. \tag{1.76}$$

Choix du paramètre de régularisation τ

Les poids externes dépendent du paramètre de régularisation τ choisi. La terminologie suivante est inspirée de l'approche PLS (Tenenhaus et al. [2005]) : la situation correspondant à $\tau_k = 0$ est appelée "mode B", celle correspondant à $\tau_k = 1$ est appelé "nouveau mode A" et celle où $0 < \tau_k < 1$ est appelé "mode Ridge".

Prenons le cas où $\tau_k = 0$ (mode B). La contrainte de normalisation devient $\operatorname{Var}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}) = 1$ et l'équation de stationnarité s'écrit :

$$\mathbf{a}_{[k]} = I^{\frac{1}{2}} \left[\mathbf{z}_{[k]}^T \mathbf{X}_{[k]} \left(\mathbf{X}_{[k]}^T \mathbf{X}_{[k]}\right)^{-1} \mathbf{X}_{[k]}^T \mathbf{z}_{[k]}\right]^{-\frac{1}{2}} \left(\mathbf{X}_{[k]}^T \mathbf{X}_{[k]}\right)^{-1} \mathbf{X}_{[k]}^T \mathbf{z}_{[k]}. \tag{1.77}$$

Ce vecteur des poids externes $\mathbf{a}_{[k]}$ est proportionnel au vecteur des coefficients de régression dans la régression multiple de $\mathbf{z}_{[k]}$ sur $\mathbf{X}_{[k]}$. En revanche, il est important de souligner qu'en raison de l'inversion de la matrice de covariance intrabloc, le calcul du vecteur des poids externes ne peut être appliqué à un bloc de variables mal conditionné.

Dans le cas où $\tau_k = 1$ (nouveau mode A), la contrainte de normalisation devient $\|\mathbf{a}_{[k]}\| = 1$ et l'équation de stationnarité s'écrit :

$$\mathbf{a}_{[k]} = \mathbf{X}_{[k]}^T \mathbf{z}_{[k]} / \|\mathbf{X}_{[k]}^T \mathbf{z}_{[k]}\|. \tag{1.78}$$

La composante externe $\mathbf{y}_{[k]} = \mathbf{X}_{[k]}\mathbf{a}_{[k]}$ est la première composante PLS dans la régression PLS de la composante interne $\mathbf{z}_{[k]}$ sur le bloc $\mathbf{X}_{[k]}$. Dans le "mode A" original de l'approche PLS, les poids externes sont calculés de la même façon que dans (1.78) mais normalisée de manière à ce que l'expression $\mathbf{y}_{[k]} = \mathbf{X}_{[k]}\mathbf{a}_{[k]}$ soit standardisée. Le "nouveau mode A" réduit la matrice de covariance de l'intrabloc à l'identité, ce qui est très utile pour des données de grande dimension car il évite l'inversion de la matrice de covariance intrabloc.

Lorsque l'utilisateur veut favoriser la stabilité (variance élevée) comparée à la corrélation, $\tau = 1$ est le choix naturel. Si en revanche il veut donner la priorité à la corrélation entre $\mathbf{y}_{[k]} = \mathbf{X}_{[k]}\mathbf{a}_{[k]}$ et les composantes voisines, alors $\tau_k = 0$ est le choix le plus approprié. Pour un compromis entre la variance et la corrélation, le paramètre τ_k peut être déterminé à l'aide de la formule de Schäfer and Strimmer [2005]. Cette estimation automatique du paramètre permet de se rapprocher du critère de corrélation même en cas de forte multicolinéarité ou lorsque le nombre de variables est largement supérieur au nombre d'individus.

Algorithme RGCCA

Il n'y a pas de solution analytique au problème d'optimisation (1.73). Un algorithme à convergence monotone basé sur la modification de l'algorithme PLS de Wold [1985] a été proposé par Tenenhaus and Tenenhaus [2011] (algorithme 5) et est décrit ci-après.

Cependant, l'algorithme proposé comporte deux limitations :

1. Il n'existe aucune preuve que l'algorithme converge vers un point fixe,
2. Il n'existe aucune garantie que l'algorithme converge vers un optimum global.

Représentation graphique de la RGCCA

En RGCCA, les conventions graphiques sont similaires à celles de la PLS-Path modeling. Chaque bloc $\mathbf{X}_{[k]}$ est représenté par une ellipse et chaque variable par un rectangle. Les ellipses contiennent les noms des blocs, et les rectangles contiennent les noms des variables. Chaque variable est reliée à son bloc par une flèche. Pour le "nouveau mode A" ($\tau_k = 1$), la flèche va de l'ellipse vers le rectangle pour symboliser le fait que chaque poids externe est calculé par régression simple des variables du bloc sur la composante du bloc interne (voir figure 1.5). Pour le "mode B" ($\tau_k = 0$), la flèche part du rectangle et va vers l'ellipse pour

Algorithm 5 Algorithme PLS pour la RGCCA

Initialisation.

1. Choix arbitraire des valeurs initiales des K vecteurs $\tilde{\mathbf{a}}^0_{[1]}, \ldots, \tilde{\mathbf{a}}^0_{[K]}$.
2. Calcul des vecteurs des poids externes normalisés :

$$\mathbf{a}^0_{[k]} = \frac{[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{X}_{[k]}]^{-1}\tilde{\mathbf{a}}^0_{[k]}}{\sqrt{(\tilde{\mathbf{a}}^0_{[k]})^T[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{X}_{[k]}]^{-1}\tilde{\mathbf{a}}^0_{[k]}}}.$$

Pour $s = 1, \ldots,$ jusqu'à convergence
Pour $k = 1, \ldots, K$

Calcul de la composante interne $\mathbf{X}_{[k]}$.
Calcul de la composante interne en fonction du schéma choisi :

$$\begin{aligned}\mathbf{z}^s_{[k]} = &\sum_{j<k} c_{kj} w\,\mathrm{Cov}(\mathbf{X}_{[k]}\mathbf{a}^s_{[k]}, \mathbf{X}_{[j]}\mathbf{a}^{s+1}_{[j]})\mathbf{X}_{[j]}\mathbf{a}^{s+1}_{[j]} \\ &+ \sum_{j>k} c_{kj} w[\mathrm{Cov}(\mathbf{X}_{[k]}\mathbf{a}^s_{[k]}, \mathbf{X}_{[j]}\mathbf{a}^s_{[j]})]\mathbf{X}_{[j]}\mathbf{a}^s_{[j]}.\end{aligned}$$

$w(x) = 1$ pour le schéma de Horst, x pour le schéma factoriel et $\mathrm{sign}(x)$ pour le schéma centroïde.

Calcul du vecteur des poids externes pour le bloc $\mathbf{X}_{[k]}$. Calcul du vecteur des poids externes :

$$\mathbf{a}^{s+1}_{[k]} = \frac{[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{X}_{[k]}]^{-1}\mathbf{X}^T_{[k]}\mathbf{z}^s_{[k]}}{\sqrt{(\mathbf{z}^s_{[k]})^T\mathbf{X}_{[k]}[\tau_k \mathbf{I} + (1-\tau_k)\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{X}_{[k]}]^{-1}\mathbf{X}^T_{[k]}\mathbf{z}^s_{[k]}}}.$$

symboliser le fait que le vecteur des poids externes pour un bloc est calculé par régression multiple de la composante du bloc interne sur les variables du bloc. Enfin pour le "mode Ridge", les doubles flèches sont utilisées pour symboliser la continuité entre le "nouveau mode A" et le "mode B". Deux blocs connectés sont reliés par une ligne. La figure résultante obtenue est appelée un modèle. Deux sous-modèles sont également considérés : le modèle externe concerne les relations entre les variables des blocs et la composante de leur bloc, et le modèle interne qui tient compte des relations entre les composantes du bloc.

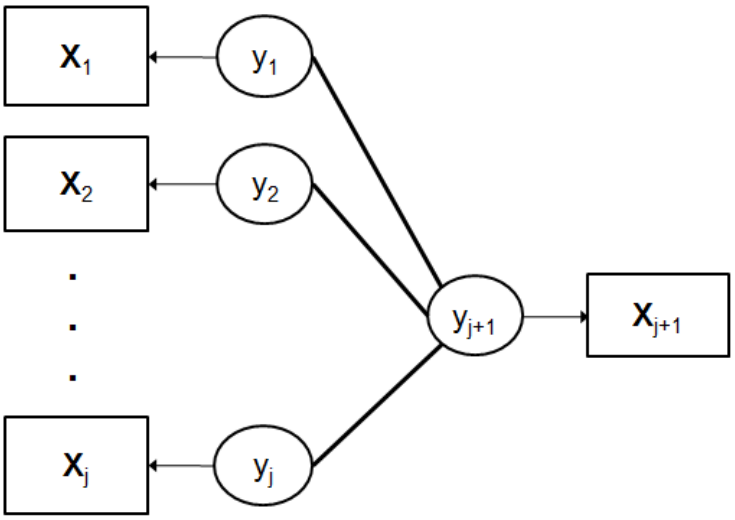

FIGURE 1.5 – Représentation d'un modèle RGCCA "nouveau mode A".

1.2.4 SGCCA: Sparse Generalized Canonical Correlation Analysis

Les données biomédicales, sont connues pour être des mesures parcimonieuses. Afin de tenir compte de cette parcimonie et d'améliorer l'interprétation du modèle RGCCA, une solution est d'identifier les sous-ensembles de variables de chaque bloc qui sont impliqués dans la relation entre les blocs connectés. Cette étape de sélection peut être obtenue en ajoutant, dans le problème d'optimisation RGCCA, une pénalisation "sparse". Une telle étape de sélection de variables est obtenue en appliquant une pénalisation L_1 sur les vecteurs des poids externes $\mathbf{a}_{[1]}, \ldots, \mathbf{a}_{[K]}$ qui induit un modèle RGCCA "sparse" donnant lieu à la Sparse Generalized Canonical Correlation Analysis (SGCCA) présentée dans le papier soumis de Tenenhaus et al. [2013].

On considère la RGCCA avec tous les τ_k égaux à 1, ce qui signifie que les contraintes sont appliquées sur la longueur des $\mathbf{a}_{[k]}$. Un pénalisation L_1 est appliquée à $\mathbf{a}_{[1]}, \ldots, \mathbf{a}_{[K]}$ ce qui conduit au problème d'optimisation suivant :

$$\begin{aligned} &\underset{\mathbf{a}_{[1]},\ldots,\mathbf{a}_{[K]}}{\arg\max} \sum_{k,j=1,k\neq j}^{K} c_{kj}\, g\left(\mathrm{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}, \mathbf{X}_{[j]}\mathbf{a}_{[j]})\right) \\ &\text{ssc.} \quad \left\|\mathbf{a}_{[k]}\right\|_2 = 1 \text{ et } \|\mathbf{a}_{[k]}\|_1 \leq s_k, \quad k = 1, \ldots, K. \end{aligned} \tag{1.79}$$

La solution au problème d'optimisation (1.79) est de la forme :

$$\mathbf{a}_{[k]} = \frac{S(\frac{1}{I}\mathbf{X}_{[k]}^T\mathbf{z}_{[k]}, \lambda_{1k})}{\|S(\frac{1}{I}\mathbf{X}_{[k]}^T\mathbf{z}_{[k]}, \lambda_{1k})\|_2}, \tag{1.80}$$

où

$$\mathbf{z}_{[k]} = \sum_{k,j=1,k\neq j}^{K} c_{kj} w(\frac{1}{I}\mathbf{a}_{[k]}^T\mathbf{X}_{[k]}^T\mathbf{X}_{[j]}\mathbf{a}_{[j]})\mathbf{X}_{[j]}\mathbf{a}_{[j]} \tag{1.81}$$

avec

$$\mathbf{z}_{[k]} = \begin{cases} \sum_{k,j=1,k\neq j}^{K} c_{kj}\mathbf{X}_{[j]}\mathbf{a}_{[j]} & \text{pour le schéma de Horst} \\ \sum_{k,j=1,k\neq j}^{K} c_{kj}\,\text{sign}[\text{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}, \mathbf{X}_{[j]}\mathbf{a}_{[j]})]\mathbf{X}_{[j]}\mathbf{a}_{[j]} & \text{pour le schéma factoriel} \\ \sum_{k,j=1,k\neq j}^{K} c_{kj}\,\text{sign}[\text{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}, \mathbf{X}_{[j]}\mathbf{a}_{[j]})]\mathbf{X}_{[j]}\mathbf{a}_{[j]} & \text{pour le schéma centroïde} \end{cases} \tag{1.82}$$

S est l'opérateur de seuillage doux tel que $S(a,\lambda) = \text{sign}(a)\max(0, |a| - \lambda)$ et λ_{1k} est choisi de telle manière que $\|\mathbf{a}_{[k]}\|_1 \leq s_k$.

En considérant l'équation (1.80), l'algorithme 6 itératif a été proposé par Tenenhaus et Tenenhaus afin de répondre au problème d'optimisation (1.79). La convergence monotone de l'algorithme est garantie par le fait que :

$$\sum_{k,j=1;k\neq j}^{K} c_{kj} g\left(\text{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}^s, \mathbf{X}_{[j]}\mathbf{a}_{[j]}^s)\right) \leq \sum_{k,j=1;k\neq j}^{K} c_{kj} g\left(\text{Cov}(\mathbf{X}_{[k]}\mathbf{a}_{[k]}^{s+1}, \mathbf{X}_{[j]}\mathbf{a}_{[j]}^{s+1})\right) \tag{1.83}$$

Cet algorithme est très stable et atteint généralement la convergence au bout de quelques itérations. Par ailleurs, il permet également de traiter facilement les données manquantes en écartant les éléments manquants dans le calcul des produits scalaires. A la fin de l'algorithme, une composante est obtenue par bloc. Il est possible de calculer plusieurs composantes orthogonales pour chaque bloc en utilisant une technique de déflation. Les matrices résiduelles sont calculées par régression des blocs originaux sur la composante du précédent bloc. Le nombre de composantes par bloc peut aussi varier d'un bloc à l'autre en choisissant de ne pas déflater tous les blocs.

Algorithm 6 Algorithme Sparse Generalized Canonical Correlation Analysis

Initialisation.

1. K blocs $\mathbf{X}_{[1]}, \ldots, \mathbf{X}_{[K]}$, K contraintes L_1 $s_1, \ldots, s_K$, une matrice de design $\mathbf{C}$.
2. Choix arbitraire des K vecteurs normalisés $\mathbf{a}^0_{[1]}, \ldots, \mathbf{a}^0_{[K]}$.

Pour $s = 1, \ldots,$ jusqu'à convergence
Pour $k = 1, \ldots, K$

Calcul des composantes internes

$$\mathbf{z}^s_{[k]} = \sum_{k=1}^{k-1} c_{kj} w \operatorname{Cov}(\mathbf{X}_{[k]}\mathbf{a}^s_{[k]}, \mathbf{X}_{[j]}\mathbf{a}^{s+1}_{[j]})\mathbf{X}_{[j]}\mathbf{a}^{s+1}_{[j]} + \sum_{j=k+1}^{K} c_{kj} w[\operatorname{Cov}(\mathbf{X}_{[k]}\mathbf{a}^s_{[k]}, \mathbf{X}_{[j]}\mathbf{a}^s_{[j]})]\mathbf{X}_{[j]}\mathbf{a}^s_{[j]}.$$

Calcul du vecteur des poids externes

$$\mathbf{a}^s_{[k]} = \frac{S(\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{z}_{[k]}, \lambda_{1k})}{\|S(\frac{1}{I}\mathbf{X}^T_{[k]}\mathbf{z}_{[k]}, \lambda_{1k})\|_2}$$

où S est l'opérateur de seuillage doux définit plus haut, et $\lambda_{1k} = 0$ si $\|\mathbf{a}^s_{[k]}\|_1 \leq s_k$ ou λ_{1k} choisi de sorte que $\|\mathbf{a}^s_{[k]}\|_1 = s_k$.

Stabilité de la sélection de variables

A chaque étape de l'algorithme, différentes variables sont sélectionnées. La performance prédictive peut être associée à un indicateur de stabilité ou au nombre de variables qui contribuent à la construction des composantes. Pour évaluer la stabilité des sélections, l'indicateur Fleiss'k défini dans Fleiss [1971] peut être utilisé. Pour chaque variable, le nombre de fois où la variable est sélectionnée ou non au cours de 10 itérations est enregistré. Ces fréquences sont résumées dans le score de Fleiss'k qui est une mesure d'accord entre les 10 itérations. Le score est toujours inférieur à 1, et plus la valeur de k est élevée, plus le modèle est stable. Des exemples d'application sont présentés dans le papier soumis de Tenenhaus et al. [2013].

Chapitre 2

Nouvelles approches multiblocs sparse non supervisées

Les approches non supervisées permettent l'exploration des données et l'analyse de liens possibles entre les variables. Dans le cadre de données génétiques, ces approches sont souvent utilisées dans un but exploratoire. Des applications de l'ACP et/ou de la SVD à des données de gène d'expression ont été publiées par exemple dans Alter et al. [2000], Holter et al. [2000], Holter et al. [2001], Raychaudhuri et al. [2000], Troyanskaya et al. [2001] , Yeung and Ruzzo [2001] et Yeung et al. [2002].

Comme nous l'avons vu au chapitre précédent, dans un contexte de données de très grande dimension la notion de "sparsité" est très importante pour faciliter l'interprétation des résultats. Dans le cas de l'ACP plusieurs versions pénalisées ont été développées. Cependant dans le cas où les données sont structurées par blocs, aucune méthode non supervisée de sélection de variables n'avait encore été proposée. La méthode GSPCA a été développée au cours de ce travail de thèse afin de permettre la sélection de blocs de variables quantitatives dans le cas non supervisé. Cette nouvelle méthode est une extension de la SPCA-rSVD de Shen and Huang [2008] (voir section 1.1.2.3) pour le cas où les données sont structurées par blocs.

Dans le cas de données qualitatives l'ACM est très fréquemment utilisée pour explorer les liens entre les variables. Cependant aucune version "sparse" de cette méthode n'avait été développée à ce jour. L'ACM étant un cas particulier de l'ACP pour des blocs de variables indicatrices, une version "sparse" de l'ACM (l'ACM sparse) est proposée dans ce

chapitre comme une extension de la méthode GSPCA. Elle permet la sélection de variables qualitatives et facilite l'interprétation des résultats obtenus avec l'ACM. Ces deux nouvelles méthodes sont présentées dans ce chapitre et leur utilité et pertinence seront démontrées à partir d'exemples illustratifs. Le contexte multiblocs étant la clé de ces deux développements, des rappels et des notions sur les données multiblocs sont présentés au préalable dans le paragraphe qui suit.

2.1 Rappels

2.1.1 SVD sur une matrice structurée par blocs

Quand $\mathbf{X}$ est composée de K sous-matrices (voir partie notations equation (7)), il est possible d'écrire la SVD par "blocs" de la manière suivante :

$$\begin{aligned}\mathbf{X} &= \left[\mathbf{X}_{[1]}|\ldots|\mathbf{X}_{[k]}|\ldots|\mathbf{X}_{[K]}\right] \\ &= \left[\mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[1]}^T|\ldots|\mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[k]}^T|\ldots|\mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[K]}^T\right] \\ &= \mathbf{P}\mathbf{\Delta}\left[\mathbf{Q}_{[1]}^T|\ldots|\mathbf{Q}_{[k]}^T|\ldots|\mathbf{Q}_{[K]}^T\right],\end{aligned} \tag{2.1}$$

où $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T|\ldots|\mathbf{Q}_{[k]}^T|\ldots|\mathbf{Q}_{[K]}^T\right]^T$ la matrice des vecteurs singuliers avec $\mathbf{Q}_{[k]}^T$ les sous-matrices, chacune de dimensions $L \times J_{[k]}$ (voir figure 2.1). Les matrices $\mathbf{P}$ et $\mathbf{\Delta}$ ne sont pas structurées par blocs. Quand les blocs $\mathbf{X}_{[1]}, \ldots \mathbf{X}_{[K]}$ sont constitués d'une seule variable (colonne), la SVD sur la matrice par bloc est écrite de la même manière que la SVD. Il est important ici de souligner que les sous-blocs ne sont pas obtenus à partir des vecteurs propres d'autres sous-blocs. La SVD ici est simplement décomposée sur les colonnes en fonctions de groupes (blocs de colonnes) définis à priori.

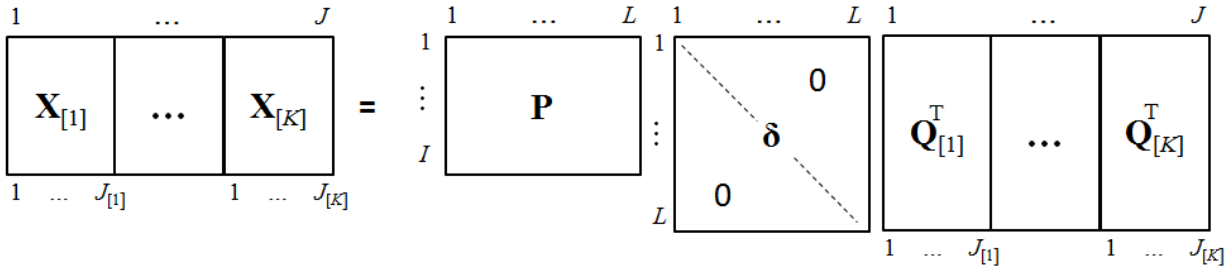

FIGURE 2.1 – SVD sur une matrice $\mathbf{X}$ structurée par blocs.

2.1.2 Propriétés de la SVD sur une matrice structurée par blocs

Dans ce contexte, la SVD permet aussi la meilleure reconstitution de rang inférieur de la matrice originale structurée par blocs. La meilleure matrice d'approximation de rang 1 $\mathbf{X}^{(1)}$ de $\mathbf{X}$ est la solution au problème (1.32) défini en section 1.1.2.3. Mais selon la propriété 2 du chapître notations, on peut écrire $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_2^2$ de la manière suivante :

$$\begin{aligned}\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_2^2 &= \operatorname{tr}\left((\mathbf{X}-\mathbf{f}_1\mathbf{q}_1^T)^T(\mathbf{X}-\mathbf{f}_1\mathbf{q}_1^T)\right) \\ &= \|\mathbf{X}\|_2^2 - 2\operatorname{tr}\left(\mathbf{q}_1\mathbf{f}_1^T\mathbf{X}\right)+\delta_1^2 \\ &= \|\mathbf{X}\|_2^2 - 2\sum_{k=1}^{K}\operatorname{tr}\left(\mathbf{q}_{1,[k]}\mathbf{f}_1^T\mathbf{X}_{[k]}\right)+\delta_1^2,\end{aligned} \tag{2.2}$$

où $\mathbf{q}_1^T=(\mathbf{q}_{1,[1]},\ldots,\mathbf{q}_{1,[k]},\ldots,\mathbf{q}_{1,[K]})^T$ est la première ligne de $\mathbf{Q}^T$ avec $\mathbf{q}_{1,[k]}$ un vecteur de dimension $J_{[k]}\times 1$ (voir figure 2.2). Les matrices $\mathbf{q}_{1,[k]}\mathbf{q}_{1,[k]}^T$ et $\mathbf{q}_{1,[k]}\mathbf{f}_1^T\mathbf{X}_{[k]}$ sont des matrices $J_{[k]}\times J_{[k]}$.

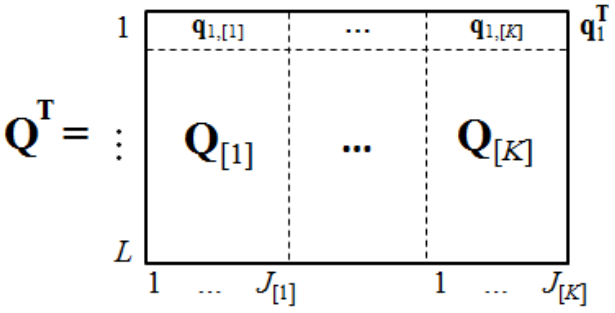

FIGURE 2.2 – Détails de la matrice $\mathbf{Q}$ dans la SVD par blocs.

2.1.3 ACP comme une SVD d'une matrice structurée par blocs

L'ACP peut être définie à partir d'une SVD avec $\mathbf{F}=\mathbf{P}\boldsymbol{\Delta}$ la matrice des composantes principales et $\mathbf{Q}$ la matrice des "loadings". Lorsque $\mathbf{X}$ est structurée en K blocs, l'ACP de $\mathbf{X}$ peut être définie comme la SVD sur une matrice par blocs (cf. section 2.1.1) avec :

$$\begin{aligned}\mathbf{X} &= \left[\mathbf{X}_{[1]}|\ldots|\mathbf{X}_{[k]}|\ldots|\mathbf{X}_{[K]}\right] \\ &= \mathbf{P}\boldsymbol{\Delta}\left[\mathbf{Q}_{[1]}^T|\ldots|\mathbf{Q}_{[k]}^T|\ldots|\mathbf{Q}_{[K]}^T\right] \\ &= \mathbf{F}\left[\mathbf{Q}_{[1]}^T|\ldots|\mathbf{Q}_{[k]}^T|\ldots|\mathbf{Q}_{[K]}^T\right].\end{aligned} \tag{2.3}$$

Dans ce cas, la matrice des "loadings" est $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T | \ldots | \mathbf{Q}_{[k]}^T | \ldots | \mathbf{Q}_{[K]}^T\right]^T$ avec $\mathbf{Q}_{[k]}^T$ un bloc de dimensions $L \times J_{[k]}$, et $\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}$ la matrice des composantes principales.

2.1.4 SVD généralisée ou GSVD

La décomposition en valeurs singulières généralisée GSVD (Generalized Singular Value Decomposition) généralise la SVD en introduisant les contraintes suivantes :

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \qquad \text{avec} \qquad \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}. \tag{2.4}$$

La matrice $\mathbf{M}$ est une matrice positive définie de dimensions $I \times I$ représentant les contraintes imposées sur les lignes de $\mathbf{X}$ (Lebart et al. [1977], Greenacre [1984], Abdi [2007]). Si $\mathbf{M}$ est diagonale, les éléments de $\mathbf{M}$ sont appelés les masses. La matrice $\mathbf{W}$ est une matrice définie positive de dimensions $J \times J$ représentant les contraintes imposées sur les colonnes de $\mathbf{X}$. Si $\mathbf{W}$ est diagonale, les éléments de $\mathbf{W}$ sont appelés les poids.

2.1.5 Propriétés de la GSVD

Tout comme la SVD, la GSVD permet la meilleure reconstitution de la matrice originale par une matrice de rang inférieure. La meilleure matrice d'approximation de rang 1 de $\mathbf{X}$ est la solution de la minimisation du carré pondéré de la norme $\mathbf{W}$-généralisée pondérée par les masses $\mathbf{M}$:

$$\underset{\mathbf{X}^{(1)}}{\arg\min} \left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 = \underset{\mathbf{X}^{(1)}}{\arg\min} \left(\operatorname{tr}\left(\mathbf{M}^{1/2}(\mathbf{X} - \mathbf{X}^{(1)})\mathbf{W}(\mathbf{X} - \mathbf{X}^{(1)})^T\mathbf{M}^{1/2}\right)\right), \tag{2.5}$$

où

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T. \tag{2.6}$$

La démonstration est fournie en annexe E.

2.1.6 GSVD appliquée à une matrice structurée par blocs

On peut écrire la GSVD par blocs quand la matrice $\mathbf{X}$ est structurée en K blocs. Dans ce cas on retrouve l'expression (2.4) mais avec $\mathbf{M}$ une matrice $I \times I$, $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T | \ldots | \mathbf{Q}_{[k]}^T | \ldots | \mathbf{Q}_{[K]}^T\right]^T$ et $\mathbf{W}$ une matrice $J \times J$ définie positive qui peut être écrite sous la forme :

$$\mathbf{W} = \operatorname{diag}\left(\left[\mathbf{W}_{[1]}^T | \ldots | \mathbf{W}_{[k]}^T | \ldots | \mathbf{W}_{[K]}^T\right]\right), \tag{2.7}$$

avec $\mathbf{W}_{[k]}$ de dimensions $J_{[k]} \times J_{[k]}$ et $\mathbf{W}$ une matrice bloc diagonale dont les blocs sont conformes à ceux de la matrice $\mathbf{Q}$.

2.1.7 Propriétés d'une GSVD appliquée à une matrice structurée par blocs

De la même manière que dans 2.1.5, la meilleure matrice d'approximation de rang 1 $\mathbf{X}^{(1)}$ de $\mathbf{X}$ structurée par blocs est la solution au problème d'optimisation (2.5). Cependant à partir de l'équation (E.9) explicitée en annexe, l'expression $\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2$ peut s'écrire en fonction des blocs (définis à priori) de la manière suivante :

$$\begin{aligned}\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 &= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \mathbf{X}^{(1)})\mathbf{W}(\mathbf{X} - \mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right) \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\delta \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2 \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\delta \sum_{k=1}^{K} \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}_1\mathbf{q}_{1,[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2,\end{aligned} \tag{2.8}$$

où $\mathbf{q}_1^T = (\mathbf{q}_{1,[1]}, \ldots, \mathbf{q}_{1,[k]}, \ldots, \mathbf{q}_{1,[K]})^T$ la première ligne de $\mathbf{Q}^T$ avec $\mathbf{q}_{1,[k]}$ un vecteur de dimension $J_{[k]} \times 1$. Cette écriture est très utile en vue de son utilisation dans les méthodes sparse qui seront définies par la suite.

2.1.8 Analyse des correspondances comme GSVD

L'analyse des correspondances est une méthode statistique de visualisation d'associations entre lignes et colonnes d'une table de contingence $\mathbf{N}$ de dimensions $I \times J$. $\mathbf{N}$ est une matrice de nombres positifs (on suppose qu'aucune ligne ou colonne n'est entièrement nulle). On définit $\mathbf{X}$ comme la matrice déduite de $\mathbf{N}$ telle que :

$$\mathbf{X} = \frac{1}{I}\mathbf{N}. \tag{2.9}$$

On définit :

- $\mathbf{r} = \mathbf{X}\mathbf{1}$ le vecteur des proportions marginales des lignes
- $\mathbf{c} = \mathbf{X}^T\mathbf{1}$ le vecteur des proportions marginales des colonnes
- $\mathbf{D}_\mathbf{r} = \operatorname{diag}(\mathbf{r})$

- $\mathbf{D_c} = \text{diag}(\mathbf{c})$.

Pour la GSVD de $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{r}\mathbf{c}^T$ sous les contraintes $\mathbf{M}$ et $\mathbf{W}$ on pose $\mathbf{M} = {\mathbf{D_r}}^{-1}$ et $\mathbf{W} = {\mathbf{D_c}}^{-1}$. En conclusion, la GSVD de $\overline{\mathbf{X}}$ est :

$$\overline{\mathbf{X}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \quad \text{avec} \quad \mathbf{P}^T{\mathbf{D_r}}^{-1}\mathbf{P} = \mathbf{Q}^T{\mathbf{D_c}}^{-1}\mathbf{Q} = \mathbf{I}. \tag{2.10}$$

La matrice des coordonnées des profils lignes $\mathbf{F}$ et des coordonnées des profils colonnes $\mathbf{G}$ sont, respectivement :

$$\begin{aligned} \mathbf{F} &= {\mathbf{D_r}}^{-1}\mathbf{P}\boldsymbol{\Delta} \\ \mathbf{G} &= {\mathbf{D_c}}^{-1}\mathbf{Q}\boldsymbol{\Delta}. \end{aligned} \tag{2.11}$$

2.1.9 De l'analyse des correspondances à l'analyse des correspondances multiples

Supposons le tableau original de données catégorielles de dimensions $I \times J$. L'ACM convertit les variables catégorielles en matrice de variables indicatrices où les données catégorielles ont été recodées en variables binaires. Si la k-ème variable possède $J_{[k]}$ catégories, la matrice indicatrice correspondante aura $J = \sum_{k=1}^{K} J_{[k]}$ colonnes. Ainsi, l'ACM est obtenue par une analyse des correspondances standard sur des blocs de matrices indicatrices. La sPCA-rSVD définie en section 1.1.2.3 peut alors être étendue au cas de variables indicatrices structurées par blocs pour permettre la sélection de variables dans le cas de l'ACM.

2.1.10 GSVD comme un problème de type régression

Considérons la GSVD de $\mathbf{X}$ avec $\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}$ la matrice des composantes principales. Les estimateurs des moindres carrés ordinaires $\hat{\boldsymbol{\beta}}$ sont obtenus en minimisant la norme $\mathbf{W}$-généralisée, sous les contraintes de masses $\mathbf{M}$:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmin}} \, \|\mathbf{f}_1 - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{W}} \, . \tag{2.12}$$

La solution est donnée par :

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^+_{\mathbf{W}}\mathbf{M}\mathbf{f}_1, \tag{2.13}$$

avec $\mathbf{X}^+_{\mathbf{W}} = \mathbf{Q}\boldsymbol{\Delta}^+_{\mathbf{W}}\mathbf{P}^T = (\mathbf{X}^T\mathbf{M}\mathbf{X})^{-1}\mathbf{W}\mathbf{X}^T$.

Sachant que $\mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}$, on en déduit que $\mathbf{Q}^T\mathbf{W}\mathbf{q}_1$ est la première colonne de la matrice identité de dimensions $L \times L$. Ainsi $\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 = [1, 0, \ldots, 0]^T$ est un vecteur de dimension $L \times 1$. A partir de $\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$, on obtient :

$$\begin{aligned} \mathbf{X}\mathbf{W}\mathbf{q}_1 &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{F}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{F}\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{f}_1. \end{aligned} \tag{2.14}$$

La solution devient :

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{X}_{\mathbf{W}}^{+}\mathbf{M}\mathbf{f}_1 \\ &= \mathbf{X}_{\mathbf{W}}^{+}\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{Q}\boldsymbol{\Delta}_{\mathbf{W}}^{+}\mathbf{P}^T\mathbf{M}\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{Q}\boldsymbol{\Delta}_{\mathbf{W}}^{+}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{q}_1. \end{aligned} \tag{2.15}$$

On en conclut alors que les "loadings" $\mathbf{q}_1$ peuvent être retrouvés en régressant les composantes principales sur les J variables.

2.1.11 GSVD régularisée

Comme nous l'avons vu au paragraphe 1.1.2.3, la SVD peut être régularisée car elle peut être considérée comme un problème de régression. La GSVD pouvant également être considérée de la sorte, elle peut également être régularisée. On considère $\tilde{\mathbf{f}}$ et $\tilde{\mathbf{q}}$ de la même manière que précédemment. Dans le cas général, le problème d'optimisation peut s'écrire :

$$\underset{\tilde{\mathbf{f}},\tilde{\mathbf{q}}}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}), \tag{2.16}$$

avec $\mathbf{W}$ une matrice définie positive, $P_\lambda(\tilde{\mathbf{q}})$ une fonction de pénalisation et λ un paramètre de pénalisation. On peut alors écrire :

$$\begin{aligned} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\,\mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) \\ &\quad + \mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}\tilde{\mathbf{f}}^T\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}) \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\,\mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2\,\mathrm{tr}(\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}) + P_\lambda(\tilde{\mathbf{q}}). \end{aligned} \tag{2.17}$$

Lorsque les données sont divisées en K groupes, chacun de cardinal $J_{[k]}$, Yuan and Lin [2005] ont proposé la pénalisation group LASSO pour créer de la "sparsité" au niveau des blocs. On considère le problème général de régression avec K groupes :

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} + \boldsymbol{\varepsilon}, \tag{2.18}$$

où $\mathbf{y}$ est un vecteur $I \times 1$, $\boldsymbol{\varepsilon}$ le terme d'erreur, $\mathbf{X}_{[k]}$ la matrice de dimensions $I \times J_{[k]}$ correspondant au k-ème bloc, et $\boldsymbol{\beta}_{[k]}$ un vecteur de coefficients de longueur $J_{[k]}$, $k = 1, \ldots, K$. Étant données des matrices définies positives $\mathbf{W}_{[1]}, \ldots, \mathbf{W}_{[K]}$, l'estimateur group LASSO est défini comme la solution de :

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} \right\|^2_{\mathbf{W}} + \lambda \sum_{k=1}^{K} \left\| \boldsymbol{\beta}_{[k]} \right\|_{\mathbf{W}_{[k]}}, \tag{2.19}$$

où $\lambda \geq 0$ est le paramètre de pénalisation. Dans notre cas, l'expression devient :

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]}\mathbf{q}_{[k]} \right\|^2_{\mathbf{W}} + \lambda \sum_{k=1}^{K} \left\| \mathbf{q}_{[k]} \right\|_{\mathbf{W}_{[k]}}, \tag{2.20}$$

où $\left\|\mathbf{q}_{[k]}\right\|_{\mathbf{W}_{[k]}} = \sqrt{\mathbf{q}_{[k]}^T \mathbf{W}_{[k]} \mathbf{q}_{[k]}}$ (voir équation (2.7)).

Dans les paragraphes suivants, les deux nouvelles méthodes développées durant ce travail de thèse sont présentées : la GSPCA pour la sélection de blocs de variables quantitatives et l'ACM sparse pour la sélection de variables qualitatives dans un contexte non supervisé.

2.2 Méthode Group Sparse PCA (GSPCA)

2.2.1 Définition

Soit $\mathbf{X}$ une matrice $I \times J$ de variables quantitatives divisée en K sous-matrices $\mathbf{X}_{[k]}$, $k = 1, ..., K$ comme défini en (7). La SVD de $\mathbf{X}$ est définie de la même manière que dans la partie notations. Dans ce contexte, le problème (2.16) peut être écrit de la manière suivante :

$$\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_2^2 + P_\lambda(\tilde{\mathbf{q}}) = \underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left(\mathrm{tr}\left((\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T)^T (\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T) \right) + P_\lambda(\tilde{\mathbf{q}}) \right). \tag{2.21}$$

La fonction de pénalisation étant additive, $P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]})$, ainsi d'après (2.17) :

$$\arg\min_{\tilde{\mathbf{f}},\tilde{\mathbf{q}}} \left(\|\mathbf{X}\|_2^2 - 2\sum_{k=1}^{K} \operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \sum_{k=1}^{K} \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right). \tag{2.22}$$

Pour trouver la solution optimale au problème de minimisation, un algorithme itératif est utilisé sous la contrainte $\|\tilde{\mathbf{f}}\|_2 = \|\delta\tilde{\mathbf{p}}\|_2 = \delta$ (car $\|\tilde{\mathbf{p}}\|_2 = 1$) .

Dans un premier temps, pour $\tilde{\mathbf{q}}$ fixé, nous cherchons le $\tilde{\mathbf{f}}$ qui minimise le problème. Il peut être obtenu par :

$$\tilde{\mathbf{f}} = \mathbf{X}\tilde{\mathbf{q}} / \|\mathbf{X}\tilde{\mathbf{q}}\|_2 . \tag{2.23}$$

Par la suite, pour $\tilde{\mathbf{f}}$ fixé, on cherche $\tilde{\mathbf{q}}$. Le problème de minimisation (2.22) devient :

$$\arg\min_{\tilde{\mathbf{q}}} \left(\|\mathbf{X}\|_2^2 - 2\sum_{k=1}^{K} \operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \sum_{k=1}^{K} \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right). \tag{2.24}$$

Le terme $\|\mathbf{X}\|_2^2$ ne dépend pas de $\tilde{\mathbf{q}}$ et les composantes $\tilde{\mathbf{q}}$ peuvent être optimisées séparément. Ainsi le $\tilde{\mathbf{q}}_{[k]}$ optimal minimise :

$$R(\tilde{\mathbf{q}}_{[k]}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + P_\lambda(\tilde{\mathbf{q}}_{[k]}). \tag{2.25}$$

et dépend de la forme du $P_\lambda(.)$ choisi.

Les données sont structurées par blocs, la fonction de pénalisation P_λ choisie est donc la pénalisation group LASSO décrite dans l'équation (1.63). En conclusion, le $\mathbf{q}_{[k]}$ cherché minimise :

$$R(\tilde{\mathbf{q}}_{[k]}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \lambda \left\|\tilde{\mathbf{q}}_{[k]}\right\|_2 . \tag{2.26}$$

Par souci de lisibilité on pose :

$$\begin{aligned} u(\tilde{\mathbf{q}}_{[k]}) &= \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) \\ v(\tilde{\mathbf{q}}_{[k]}) &= \lambda \left\|\tilde{\mathbf{q}}_{[k]}\right\|_2 . \end{aligned} \tag{2.27}$$

Et ainsi $R(\tilde{\mathbf{q}}_{[k]}) = u(\tilde{\mathbf{q}}_{[k]}) + v(\tilde{\mathbf{q}}_{[k]})$. Minimiser $R(\tilde{\mathbf{q}}_{[k]})$ revient donc à minimiser chaque élément de la somme (c'est-à-dire les fonctions u et v). Le minimum d'une fonction est

obtenu lorsque sa dérivée est nulle. La fonction u est différentiable en 0 (le gradient noté ∇u peut alors être calculé) ce qui n'est pas le cas pour la fonction v. En effet, la norme L_2 n'étant pas différentiable en 0, c'est le sous-gradient de la fonction v en $\tilde{\mathbf{q}}_{[k]}$, noté $\partial v(\tilde{\mathbf{q}}_{[k]})$, qui est calculé. Le minimum cherché est donc solution de : $\nabla u(\tilde{\mathbf{q}}_{[k]}) + \partial v(\tilde{\mathbf{q}}_{[k]}) = 0$. D'après les propriétés 5 et 6 définies dans la partie notations on obtient :

$$\nabla u(\tilde{\mathbf{q}}_{[k]}) = 2\delta^2 \tilde{\mathbf{q}}_{[k]} - 2\mathbf{X}_{[k]}^T \tilde{\mathbf{f}}. \tag{2.28}$$

L'équation (2.28) étant le résultat de la SVD, les solutions optimales correspondent à $\mathbf{q} = \mathbf{q}_1$ et $\mathbf{f} = \mathbf{f}_1$. En conclusion, à partir de ce résultat et de la propriété 8, on en déduit que :

si $\mathbf{q}_{1,[k]} \neq \mathbf{0}$

$$2\delta_1^2 \tilde{\mathbf{q}}_{1,[k]} - 2\mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1 + \lambda \frac{\mathbf{q}_{1,[k]}}{\left\|\mathbf{q}_{1,[k]}\right\|_2} = 0, \tag{2.29}$$

si $\mathbf{q}_{1,[k]} = \mathbf{0}$

$$\left\|2\mathbf{X}_{[k]}^T \mathbf{f}_1\right\|_2 \leq \lambda. \tag{2.30}$$

Les expressions (2.29) et (2.30) peuvent s'écrire :

$$\mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1 = \delta_1^2 \tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2} \frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_2} \tag{2.31}$$

$$\left\|\mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1\right\|_2 \leq \frac{\lambda}{2} \tag{2.32}$$

En combinant (2.31) et (2.32), on obtient que le minimum est de la forme :

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2\delta_1^2} \frac{1}{\left\|\mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1\right\|_2}\right)_+ \mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1 \tag{2.33}$$

où

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{2.34}$$

La règle de seuillage h_λ peut être définie de la manière suivante :

$$h_\lambda(y) = \left(1 - \frac{\lambda}{2\delta_1^2} \frac{1}{\|y\|_2}\right)_+ y, \tag{2.35}$$

avec $y = \mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1$.

2.2.2 Algorithme

L'algorithme de la GSPCA est décrit ci-après (algorithme 7) :

Algorithm 7 Algorithme GSPCA

Initialisation (Etape 1) : Application de la SVD à $\mathbf{X}$ et obtention de la meilleure approximation de rang 1 de $\mathbf{X}$ $\delta\mathbf{pq} = \mathbf{fq}$ avec $\mathbf{p}$ et $\mathbf{q}$ des vecteurs unitaires.
On fixe $\tilde{\mathbf{q}}^s = \mathbf{q}_1$ et $\tilde{\mathbf{f}}^s = \delta_1\mathbf{p}_1$.

Itération (Etape 2) :

a) $\tilde{\mathbf{q}}^{s+1} = \left[\tilde{\mathbf{q}}^{s+1}_{1,[1]}, \ldots, \tilde{\mathbf{q}}^{s+1}_{1,[K]}\right] = \left[h_\lambda(\mathbf{X}^T_{[1]}\tilde{\mathbf{f}}^s), \ldots, h_\lambda(\mathbf{X}^T_{[K]}\tilde{\mathbf{f}}^s)\right]$;

b) $\tilde{\mathbf{f}}^{s+1} = \mathbf{X}_{[k]}\tilde{\mathbf{q}}^{s+1} / \left\|\mathbf{X}_{[k]}\tilde{\mathbf{q}}^{s+1}\right\|_2$

Réitèrer l'étape 2 en remplaçant $\tilde{\mathbf{f}}^s$ et $\tilde{\mathbf{q}}^s$ par $\tilde{\mathbf{f}}^{s+1}$ et $\tilde{\mathbf{q}}^{s+1}$ jusqu'à convergence.

Standardisation $\tilde{\mathbf{q}}^{s+1}$ finale définie comme $\mathbf{q}=\tilde{\mathbf{q}}^{s+1} / \left\|\tilde{\mathbf{q}}^{s+1}\right\|_2$ correspond au "loading sparse" voulu.

Le critère de convergence dans l'étape de réitération est le suivant :

$$\begin{aligned} \left\|\tilde{\mathbf{f}}^{s+1} - \tilde{\mathbf{f}}^s\right\| &< err \\ \left\|\tilde{\mathbf{q}}^{s+1} - \tilde{\mathbf{q}}^s\right\| &< err \end{aligned} \tag{2.36}$$

avec une marge d'erreur (err) fixée par l'utilisateur (dans l'exemple d'application on choisira $err = 10^{-4}$). L'algorithme converge assez rapidement en pratique. Si on fixe $\lambda = 0$ dans l'algorithme ci-dessus, l'étape 2a revient à faire $\mathbf{q}^{s+1} = \mathbf{X}^T\mathbf{f}^s$ et l'algorithme devient l'algorithme bien connu des moindres carrés alternés pour calculer la SVD. Il converge alors en une seule itération. La procédure itérative de la GSPCA est définie pour des vecteurs unidimensionnels et est utilisée pour obtenir le premier vecteur "sparse" des "loadings" $\mathbf{q}_1$. Les "loadings sparse" suivants $\mathbf{q}_i$ ($i > 1$) peuvent être obtenus séquentiellement via une approximation de rang 1 des matrices des résidus. Un exemple d'application de la GSPCA est présenté dans la section suivante.

2.2.3 Exemple d'application

L'exemple d'application de la Group Sparse PCA concerne un tableau de données sur des crabes bleus. Les crabes bleus ont une valeur commerciale importante en Caroline du Nord et l'apparence des crabes malades dans la région spécifique de Pamlico River est devenue préoccupante depuis 1986. Gemperline et al. [1992] ont fait l'hypothèse que le stress environnemental affaiblissait leur organisme ce qui empêchait leur réponse immunitaire d'éliminer l'infection causée par une bactérie (chitinoclastic bacteria). Cette bactérie a été considérée comme étant la cause des lésions comprises entre 5 et 25 mm dans leur carapace.

Afin de déterminer si cette maladie des crabes était directement liée aux taux d'oligo-éléments trouvés dans leur corps, des tissus de branchies, hépatopancréas et des muscles ont été prélevés sur 48 crabes dont 16 crabes sains venant d'Albemarle Sound, 16 crabes sains venant de Pamlico River et enfin 16 crabes malades venant de Pamlico River. La figure 2.3 illustre la composition du tableau de données. Vingt-cinq oligo-éléments issus des trois tissus ont été analysés pour chacun des crabes. Le tableau de données est composé de $I = 48$ observations (crabes) et $J = 75$ variables quantitatives (25 blocs d'oligo-éléments analysés sur 3 tissus différents). Les données sont naturellement structurées par blocs : 25 blocs de 3 variables sur 48 crabes.

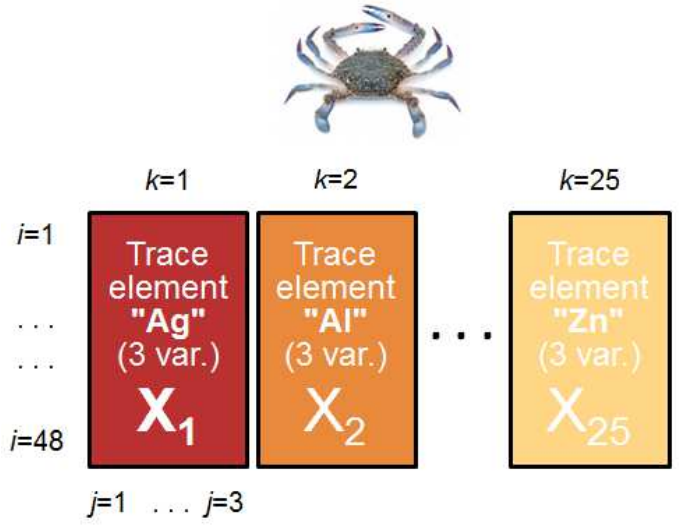

FIGURE 2.3 – Composition du tableau de données des crabes.

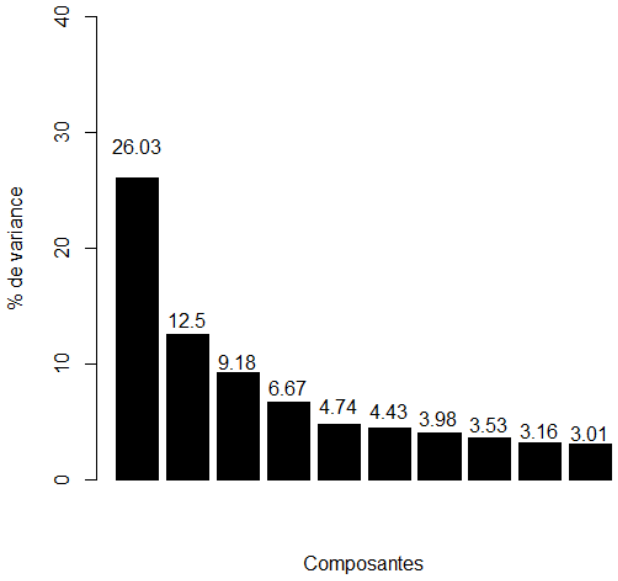

FIGURE 2.4 – Données crabes : Représentation du pourcentage de variance obtenu par ACP pour les premières composantes

Dans un premier temps, une ACP a été réalisée sur ces données afin d'explorer les liens entre les différentes variables et de visualiser la répartition des crabes en fonction de leur provenance et de leur état de santé. Seules les deux premières composantes sont considérées (règle du coude, voir figure 2.4). La figure 2.5 représente la répartition des crabes sur le premier plan (en bleu, les crabes sains d'Albemarle Sound, en vert les crabes sains de Pamlico River et en rouge les crabes malades de Pamlico River). Le premier axe oppose les crabes d'Albemarle Sound aux crabes malades de Pamlico River. Cette dimension est particulièrement liée au taux d'oligo-éléments contenus dans les poumons (et plus particulièrement le cuivre). Cela signifie que les crabes d'Albemarle Sound se distinguent des crabes malades de Pamlico River par un fort taux d'oligo-éléments cuivre dans les poumons et un faible taux des autres oligo-éléments.

Le deuxième axe oppose les crabes sains de Pamlico River (corrélation positive à l'axe 2) aux autres. Les variables qui correspondent aux taux d'oligo-élements potassium et magnésium (corrélation positive) ainsi que cadium (corrélation négative) sont les plus contribu-

tives de cette dimension. Cela signifie que les crabes sains de Pamlico River se distinguent des autres par un fort taux d'oligo-éléments potassium et magnésium et par un faible taux d'oligo-élément cadium.

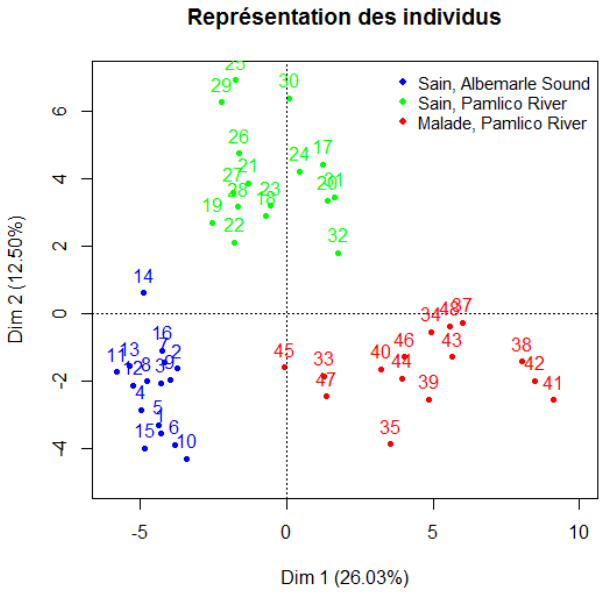

FIGURE 2.5 – Données crabes : Représentation des crabes sur le premier plan (ACP).

Par la suite, l'ACP sparse de Zou et al. [2006] et la GSPCA ont été réalisées sur ce tableau de données afin de sélectionner des variables et des blocs de variables (c'est-à-dire des oligo-éléments), respectivement, sur chacun des axes. Plusieurs valeurs du paramètre de pénalisation λ (comprises entre 0 et 10) ont été testées pour la GSPCA et les deux valeurs λ_1 et λ_2 pour l'ACP sparse ont été choisies de manière à ce que le nombre de variables sélectionnées par axe corresponde à celui de la GSPCA pour chacune des valeurs du λ. La table 2.1 présente le nombre de variables sélectionnées sur la première et la deuxième composante, le temps (CPU Time en secondes) et le nombre d'itérations (pour l'ensemble des deux premières CPs) nécessaires à l'algorithme pour converger, en fonction de la valeur de λ (comprises entre 0 et 10).

L'algorithme converge quel que soit la valeur de λ et comme attendu, le temps mis pour converger est plus important lorsque le nombre d'itérations est élevé. Le temps de convergence est plus long pour l'ACP sparse que pour la GSPCA pour chacune des valeurs de λ. Ceci prouve l'efficacité de cette nouvelle méthode.

TABLE 2.1 – Données crabes : Temps (en secondes) et nombre d'itérations nécessaires à la convergence de l'algorithme en fonction des valeurs de λ pour l'ACP sparse et la GSPCA.

Valeurs de λ	0	1	2	3	4	5	6	7	8	9	10
Nb var. CP1	75	75	72	66	57	42	39	33	21	9	6
Nb var. CP2	75	72	60	45	36	21	18	3	3	0	0
CPU Time SPCA	0.26	0.26	0.24	0.20	11.08	2.20	2.44	0.62	0.40	0.04	0.06
CPU Time GSPCA	0.02	0.10	0.08	0.10	0.16	0.14	0.22	0.08	0.08	0.04	0.04
Nombre itérations sur 2 CPs (GSPCA)	2	33	32	40	57	50	81	30	26	11	16

Les figures 2.6 et 2.7 présentent l'évolution du nombre de variables sélectionnées et le pourcentage cumulé de variance expliquée (CPEV) en fonction du paramètre de pénalisation pour les deux premières dimensions, respectivement. Plus la valeur du λ est élevée, moins il y a de variables sélectionnées. On observe un léger coude sur la figure 2.6 pour $\lambda = 4,5$. Sur la première dimension, 51 variables sont conservées, et CPEV= $21,61\%$, et sur la deuxième, 30 variables sont conservées, et CPEV= $29,46\%$. Si on augmente la valeur du λ à 5, 42 variables sont conservées sur la première dimension pour un CPEV égal à 20,53% et 21 variables sont conservées sur la deuxième pour un CPEV égal à 27,73%. Afin de trouver un compromis entre le nombre de variables sélectionnées et la perte de pourcentage cumulé de variance expliquée, nous choisirons $\lambda = 5$ qui élimine presque la moitié des variables sur la première composante et plus du tiers sur la deuxième, tout en conservant un CPEV proche du CPEV de départ (pour $\lambda = 0$, CPEV= $26,03\%$ pour la première dimension, et CPEV= $38,53\%$ pour la deuxième). Le λ aurait également pu être choisi de manière plus objective par validation croisée (10-fold par exemple) de la même manière que dans Shen and Huang [2008].

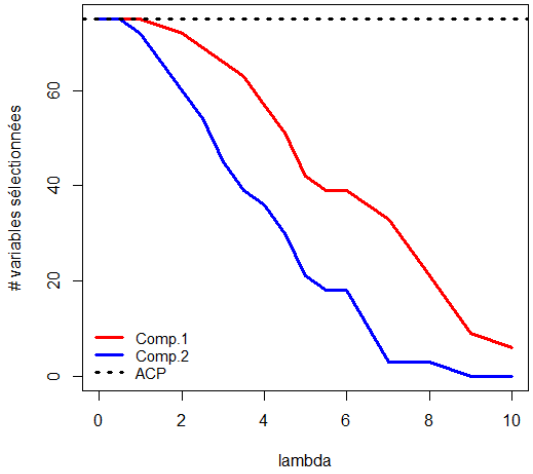

FIGURE 2.6 – Données crabes : Évolution du nombre de variables sélectionnées en fonction du paramètre de pénalisation λ.

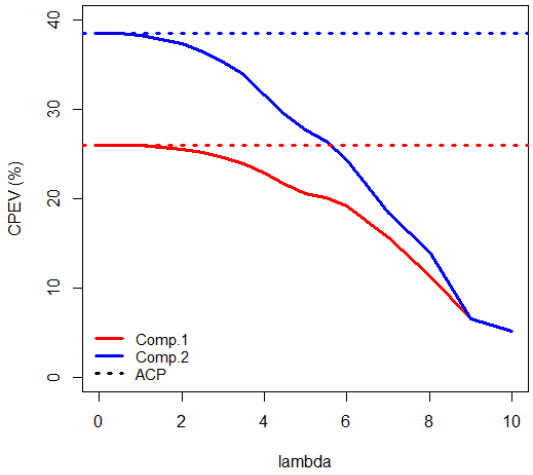

FIGURE 2.7 – Données crabes : Évolution du pourcentage cumulé de variance expliquée en fonction du paramètre de pénalisation λ.

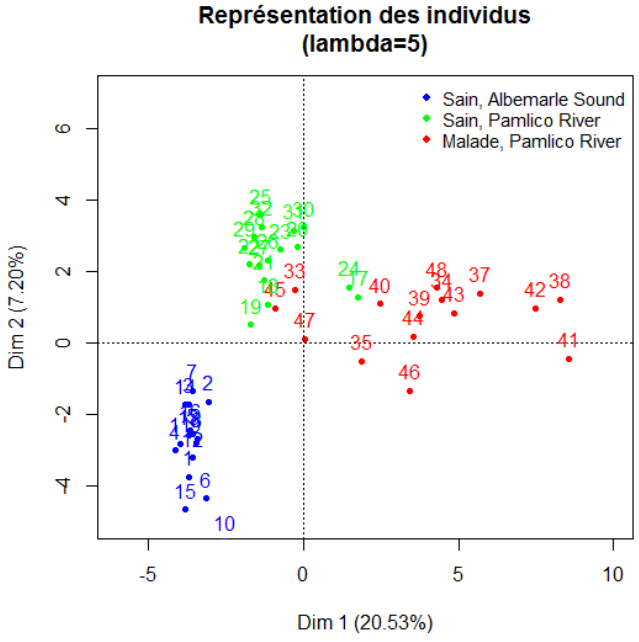

FIGURE 2.8 – Données crabes : Représentation des individus sur le premier plan (GSPCA) pour $\lambda = 5$.

La figure 2.8 montre que malgré l'élimination de certaines variables sur les deux premières composantes de la GSPCA les crabes sont encore bien distincts en fonction de leur provenance. Cela signifie que les variables conservées après GSPCA sont pertinentes et pourraient expliquer un lien possible entre la maladie et le taux de certains oligo-éléments stockés dans les tissus des crabes. En revanche on perd en pouvoir discriminant sur le deuxième axe car un plus grand nombre de blocs de variables a été supprimé. La discrimination entre les crabes sains et malades de Pamlico river est moins flagrante que dans l'ACM standard. Si on choisit une valeur de λ plus basse, 3,5 au lieu de 5 par exemple, alors 39 variables sont conservées sur le deuxième axe et la discrimination des crabes sains et malades est parfaite. Le choix du λ dépend donc du but recherché.

Le tableau 2.2 résume et compare les "loadings" et les variances obtenus à l'aide de l'ACP, de l'ACP sparse de Zou et al. [2006] et de la GSPCA pour les 6 blocs d'oligo-éléments dont les 4 cités précédemment ("cuivre", "cadium", "potassium" et "magnésium"). Pour l'ACP sparse, on fixe $\lambda = 0$ et $\lambda_1 = (0,01; 0,2)$ de manière à ce que chaque approximation "sparse" explique à peu près la même part de variance que les composantes dans l'ACP classique. Pour la GSPCA, le λ est fixé à 5. L'ACP sparse ne sélectionne que certaines variables, sans tenir compte de l'appartenance des variables au bloc correspondant car elle ne considère pas les structures par blocs contrairement à la GSPCA. L'ACP sparse identifie correctement les variables les plus contributives trouvées dans l'ACP pour les deux premières dimensions. La GSPCA sélectionne le bloc entier de variables correspondant et fixe les autres à zéro. La variance ajustée est quasiment la même pour les 3 méthodes mais la GSPCA produit des "loadings sparse" ce qui facilite l'interprétation et la visualisation des résultats. Les résultats de la GSPCA recoupent ceux de l'ACP sparse mais des blocs de variables entiers sont supprimés, ce qui permet de faciliter l'interprétation des résultats et de visualiser plus rapidement les oligo-éléments (et non plus uniquement certaines modalités) potentiellement liés à la maladie des crabes bleus.

TABLE 2.2 – Données crabes : "Loadings" et variances obtenus avec ACP, ACP sparse et GSPCA sur les deux premières dimensions pour 6 des 25 oligo-éléments analysés.

	ACP		**ACP sparse**		**GSPCA**	
Variable	Comp 1	Comp 2	Comp 1	Comp 2	Comp 1	Comp 2
Poumon.Al	0,195	-0,047	0,306	**0,000**	0,277	**0,000**
Hepatopancreas.Al	0,124	-0,051	**0,000**	**0,000**	0,174	**0,000**
Muscle.Al	0,169	-0,022	0,110	-0,006	0,242	**0,000**
Poumon.As	0,185	-0,022	0,397	**0,000**	0,207	**0,000**
Hepatopancreas.As	0,127	-0,113	0,054	**0,000**	0,121	**0,000**
Muscle.As	0,117	0,109	0,067	0,084	0,115	**0,000**
Poumon.Cd	0,003	-0,241	**0,000**	-0,273	**0,000**	-0,373
Hepatopancreas.Cd	-0,108	-0,216	-0,082	-0,142	**0,000**	-0,486
Muscle.Cd	-0,011	-0,108	-0,088	-0,130	**0,000**	-0,244
Poumon.Cu	-0,184	-0,022	-0,291	**0,000**	-0,157	-0,228
Hepatopancreas.Cu	-0,130	-0,085	-0,133	**0,000**	-0,113	-0,239
Muscle.Cu	-0,120	-0,190	**0,000**	-0,273	-0,099	-0,365
Poumon.Mg	0,110	0,216	0,075	0,383	0,066	0,302
Hepatopancreas.Mg	0,114	0,162	0,021	0,075	0,083	0,213
Muscle.Mg	0,177	-0,116	0,024	-0,041	0,131	0,041
Poumon.P	0,083	0,072	**0,000**	**0,000**	**0,000**	**0,000**
Hepatopancreas.P	-0,051	0,237	**0,000**	0,409	**0,000**	**0,000**
Muscle.P	-0,100	0,212	-0,020	0,154	**0,000**	**0,000**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Nb variables sélectionnées	75	75	47	44	42	21
Variance ajustée (%)	26,03	12,50	15,70	9,30	20,53	7,20
Variance cumulée (%)	26,03	38,53	15,70	25,00	20,53	27,73

Al : Aluminium, As : Arsenic, Cd : Cadium, Cu : Cuivre, Mg : Magnesium, P : Phosphore

2.3 Analyse des Correspondances Multiples Sparse (ACM sparse)

L'ACM est un cas particulier de l'ACP pour des blocs de variables indicatrices, c'est pourquoi l'ACM sparse introduite dans cette section est définie comme une extension de la GSPCA et le problème (2.21) peut être généralisé pour l'ACM sparse :

$$\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) \tag{2.37}$$

avec une norme $\mathbf{W}$-généralisée sous la contrainte de masse $\mathbf{M}$.

2.3.1 Définition

Supposons une matrice $I \times K$ de variables catégorielles. Le tableau disjonctif complet correspondant $\mathbf{N}$ est constitué de K sous-matrices (ou blocs) de variables indicatrices $\mathbf{N}_{[k]}$, $k = 1, \ldots, K$, chacune de dimensions $I \times J_{[k]}$. Le nombre total de modalités est $J = \sum_{k=1}^{K} J_{[k]}$. Sélectionner une colonne de la table originale (variables catégorielles) revient à sélectionner un bloc de variables indicatrices dans le tableau disjonctif complet. C'est pourquoi la nouvelle méthode nommée ACM sparse peut être considérée comme une extension de la méthode précédemment présentée (GSPCA) dans le cas de blocs de variables indicatrices. On considère la GSVD pour des blocs de variables indicatrices, comme définie dans l'équation (2.4). Soit $\mathbf{X}$ la matrice définie dans le paragraphe 2.1.8. Dans ce contexte, le problème (2.37) peut être écrit de la manière suivante :

$$\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T \right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) \qquad \text{avec} \qquad \tilde{\mathbf{f}}^T\mathbf{M}\tilde{\mathbf{f}} = \tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}} = 1, \tag{2.38}$$

et sous les contraintes de masse $\mathbf{M}$ on obtient :

$$\begin{aligned} &\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left(\operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T)\mathbf{W}(\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T)^T\mathbf{M}^{\frac{1}{2}} \right) + P_\lambda(\tilde{\mathbf{q}}) \right) = \\ &\underset{\tilde{\mathbf{f}},\tilde{\mathbf{q}}}{\arg\min} \left(\|\mathbf{X}\|_{\mathbf{W}}^2 - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}) + P_\lambda(\tilde{\mathbf{q}}) \right). \end{aligned} \tag{2.39}$$

La fonction de pénalisation étant additive, $P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]})$ et le problème devient :

$$\arg\min_{\tilde{\mathbf{f}},\tilde{\mathbf{q}}} \left(\|\mathbf{X}\|^2_{\mathbf{W}} - 2\sum_{k=1}^{K} \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + \sum_{k=1}^{K} \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right). \tag{2.40}$$

Pour trouver la solution au problème de minimisation, le même principe d'algorithme itératif que celui de la GSPCA est appliqué, sous la contrainte que $\tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = 1$. Dans un premier temps, on considère le problème d'optimisation sous $\tilde{\mathbf{p}}$ pour $\tilde{\mathbf{q}}$ fixé. Le $\tilde{\mathbf{p}}$ minimal est obtenu par :

$$\tilde{\mathbf{p}} = \mathbf{X}\tilde{\mathbf{q}} / \|\mathbf{X}\tilde{\mathbf{q}}\|_{\mathbf{W}}. \tag{2.41}$$

A présent, si on optimise sous $\tilde{\mathbf{q}}$ pour $\tilde{\mathbf{p}}$ fixé, le problème de minimisation (2.40) peut s'écrire :

$$\arg\min_{\tilde{\mathbf{q}}} \left(\|\mathbf{X}\|^2_{\mathbf{W}} + \sum_{k=1}^{K} \left(\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right) \right). \tag{2.42}$$

Le terme $\|\mathbf{X}\|^2_{\mathbf{W}}$ ne dépend pas de $\tilde{\mathbf{q}}$, l'optimisation peut se faire indépendamment des composantes $\tilde{\mathbf{q}}$. Ainsi, le $\tilde{\mathbf{q}}_{[k]}$ optimal minimise :

$$R(\tilde{\mathbf{q}}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}_{[k]}), \tag{2.43}$$

et dépend de la fonction $P_\lambda(.)$ choisie. Les données sont structurées en blocs, la fonction de pénalité P_λ choisie est celle du group LASSO décrite dans le paragraphe 1.2.2 :

$$P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) = \sum_{k=1}^{K} \lambda \left\|\tilde{\mathbf{q}}_{[k]}\right\|_{\mathbf{W}_{[k]}}. \tag{2.44}$$

En conclusion, $\mathbf{q}_{[k]}$ doit minimiser :

$$R(\tilde{\mathbf{q}}_{[k]}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + \lambda \left\|\tilde{\mathbf{q}}_{[k]}\right\|_{\mathbf{W}_{[k]}}. \tag{2.45}$$

De la même manière que dans la section précédente pour la GSPCA, on pose :

$$\begin{aligned} u(\tilde{\mathbf{q}}_{[k]}) &= \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) \\ l(\tilde{\mathbf{q}}_{[k]}) &= -2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) \\ v(\tilde{\mathbf{q}}_{[k]}) &= \lambda \left\|\tilde{\mathbf{q}}_{[k]}\right\|_{\mathbf{W}_{[k]}}. \end{aligned} \tag{2.46}$$

Ainsi $R(\tilde{\mathbf{q}}_{[k]}) = u(\tilde{\mathbf{q}}_{[k]}) + l(\tilde{\mathbf{q}}_{[k]}) + v(\tilde{\mathbf{q}}_{[k]})$ et donc minimiser $R(\tilde{\mathbf{q}}_{[k]})$ à minimiser chaque élément de la somme (c'est-à-dire les fonctions u, l et v). Le minimum d'une fonction est obtenu lorsque sa dérivée est nulle. Les fonctions u et l sont différentiables en 0 mais ce n'est pas le cas de la fonction v étant donné que la norme $\mathbf{W}$-généralisée ne l'est pas en 0. C'est donc le sous-gradient de la fonction v en $\tilde{\mathbf{q}}_{[k]}$ qui est calculé. Le minimum cherché est donc solution de : $\nabla u(\tilde{\mathbf{q}}_{[k]}) + \nabla l(\tilde{\mathbf{q}}_{[k]}) + \partial v(\tilde{\mathbf{q}}_{[k]}) = 0$.

D'après la propriété 2 on a que

$$\begin{aligned} l(\tilde{\mathbf{q}}_{[k]}) &= -2\,\mathrm{tr}\left(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}\right) \\ &= -2\,\mathrm{tr}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}\tilde{\mathbf{f}}^T\mathbf{M}^{\frac{1}{2}}\right) \end{aligned} \tag{2.47}$$

et avec la propriété 6 citée dans la partie notations page 33, on obtient :

$$\begin{aligned} \nabla l(\tilde{\mathbf{q}}_{[k]}) &= -2\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}_{[k]}\mathbf{W}_{[k]}\right)^T\left(\tilde{\mathbf{f}}^T\mathbf{M}^{\frac{1}{2}}\right)^T \\ &= -2\left(\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}\right)\left(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\right) \\ &= -2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}. \end{aligned} \tag{2.48}$$

Par ailleurs

$$\nabla u(\tilde{\mathbf{q}}_{[k]}) = 2\delta^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}. \tag{2.49}$$

Les équations (2.48) et (2.49) étant les résultats de la GSVD, les solutions optimales sont $\mathbf{q} = \mathbf{q}_1$ et $\mathbf{f} = \mathbf{f}_1$. En conclusion, à partir de ce résultat et de la propriété 8, on en déduit que :

si $\mathbf{q}_{1,[k]} \neq \mathbf{0}$

$$-2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1 + 2\delta_1^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]} + \lambda\mathbf{W}_{[k]}\frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_{\mathbf{W}_{[k]}}} = 0 \tag{2.50}$$

et si $\mathbf{q}_{1,[k]} = \mathbf{0}$

$$\left\|-2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \lambda\mathbf{W}_{[k]}. \tag{2.51}$$

Les expressions (2.50) et (2.51) peuvent s'écrire :

$$\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1 = \delta_1^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2}\frac{\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_{\mathbf{W}_{[k]}}} \tag{2.52}$$

$$\left\|\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \frac{\lambda}{2}\mathbf{W}_{[k]}. \tag{2.53}$$

On obtient alors :

$$\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1 = \delta_1^2\tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2}\frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_{\mathbf{W}_{[k]}}} \tag{2.54}$$

$$\left\|\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \frac{\lambda}{2}. \tag{2.55}$$

En combinant (2.54) et (2.55) le minimum cherché est de la forme :

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2}\frac{1}{\left\|\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}}}\right)_+ \frac{1}{\delta_1^2}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1, \tag{2.56}$$

avec

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{2.57}$$

Une règle de seuillage h_λ peut être définie de la manière suivante :

$$h_\lambda(y) = \left(1 - \frac{\lambda}{2}\frac{1}{\|y\|_{\mathbf{W}_{[k]}}}\right)_+ \frac{1}{\delta^2}y \tag{2.58}$$

avec $y = \mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1$.

Dans le cas de l'ACM :

$$\mathbf{M} = {\mathbf{D_r}}^{-1} \qquad \mathbf{W} = {\mathbf{D_c}}^{-1}. \tag{2.59}$$

La norme $\mathbf{W}$-généralisée peut alors être écrite :

$$\begin{aligned} \|\mathbf{X}\|_{\mathbf{W}} &= \sqrt{\mathbf{MXWX}^T} \\ &= \sqrt{{\mathbf{D_r}}^{-1}\mathbf{X}{\mathbf{D_c}}^{-1}\mathbf{X}^T}. \end{aligned} \tag{2.60}$$

Ainsi la norme $\left\|\mathbf{X}_{[k]}^T\mathbf{M}\mathbf{f}_1\right\|_{\mathbf{W}_{[k]}}$ vaut $\left\|\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\mathbf{f}_1\right\|_{\mathbf{D}_{\mathbf{c}_{[k]}}^{-1}}$ et la solution explicite au problème de minimisation devient :

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2}\frac{1}{\left\|\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\tilde{\mathbf{f}}_1\right\|_{\mathbf{D}_{\mathbf{c}_{[k]}}^{-1}}}\right)_+ \frac{1}{\delta_1^2}\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\tilde{\mathbf{f}}_1. \tag{2.61}$$

2.3.2 Algorithme

L'algorithme de l'ACM sparse (algorithme 8) est définit comme suit :

Algorithm 8 Algorithme ACM sparse ou Sparse Multiple Correspondence Analysis (SMCA)

Initialisation (Étape 1) : Application de la GSVD sur $\mathbf{X}$. Calcul de la meilleure approximation de rang 1 de $\mathbf{X}$ $\delta\mathbf{pq} = \mathbf{fq}$ où $\mathbf{p}$ et $\mathbf{q}$ des vecteurs unitaires.
On pose $\mathbf{q}^s = \mathbf{q}_1$ et $\mathbf{f}^s = \delta_1\mathbf{p}_1$.
Itération (Étape 2) :

a) $\mathbf{q}^{s+1} = \left[h_\lambda(\mathbf{X}_{[1]}^T\mathbf{f}^s), \ldots, h_\lambda(\mathbf{X}_{[J]}^T\mathbf{f}^s)\right]$;

b) $\mathbf{f}^{s+1} = \mathbf{X}_{[k]}\tilde{\mathbf{q}}^{s+1} / \left\|\mathbf{X}_{[k]}\tilde{\mathbf{q}}^{s+1}\right\|_{\mathbf{W}}$

Réitérer Étape 2 en remplaçant $\mathbf{f}^s$ et $\mathbf{q}^s$ par $\mathbf{f}^{s+1}$ et $\mathbf{q}^{s+1}$ jusqu'à convergence
Standardisation : $\mathbf{q}^{s+1}$ final est définit comme étant $\mathbf{q}=\mathbf{q}^{s+1} / \left\|\mathbf{q}^{s+1}\right\|_{\mathbf{W}}$ et est le "loading sparse" voulu.

Les vecteurs "sparse" suivants $\mathbf{q}_i$ $(i > 1)$ sont obtenus par approximation de rang 1 des matrices résiduelles.

2.3.3 Exemple d'application

Une application de cette méthode est présentée ici sur un jeu de données bien connu de 27 races de chiens, décrites au moyen de 6 variables qualitatives (Tenenhaus [2007]). On pose $\mathbf{X}$ la matrice des données de dimensions 27×6. Le tableau disjonctif correspondant est constitué de 6 blocs de variables indicatrices avec un total de 16 modalités, soit de dimensions 27×16. Dans un premier temps, une ACM a été réalisée sur ces données afin d'explorer les liens entre les différentes variables et ceux entre les individus. Seules les deux premières composantes sont considérées. La figure 2.9 représente les variables (modalités des variables) sur les deux premières dimensions de l'ACM.

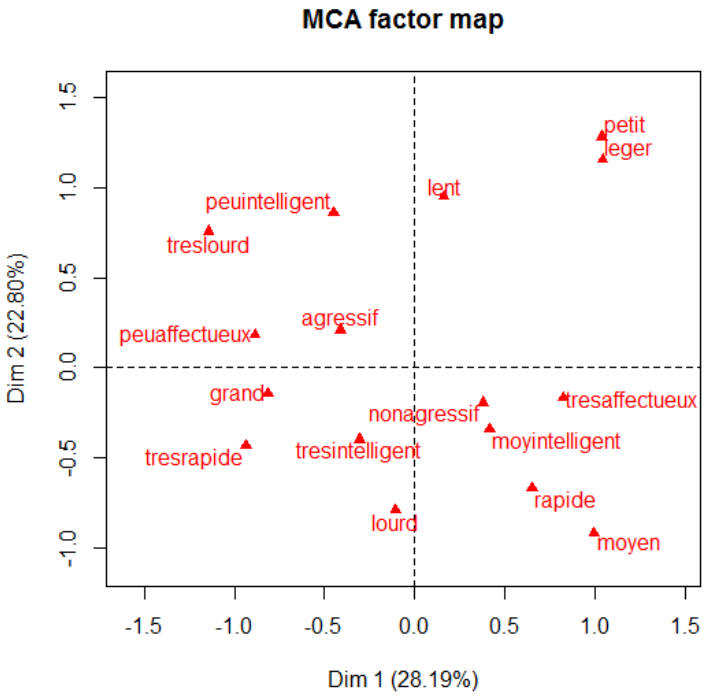

FIGURE 2.9 – Données chiens : Représentation des variables sur les deux premières dimensions de l'ACM.

Sur l'axe 1, les contributions les plus importantes sont celles des modalités liées à l'affection et à la taille. On voit que cet axe discrimine les chiens très affectueux et petits (à droite) des peu affectueux et grands (à gauche). On peut en déduire que les chiens de petite taille ont tendance à être plus affectueux que ceux de grande taille. Sur l'axe 2, les

contributions les plus importantes sont celles du poids de la vélocité et de la taille. L'axe discrimine les chiens légers, petits et lents (en haut) de ceux qui sont lourds et de taille moyenne et rapides (en bas). Les chiens légers sont donc plus souvent des chiens de petite taille mais qui sont plus lents que les chiens plus lourds. Les modalités liées à l'agressivité et l'intelligence ne sont pas bien représentées sur ces deux premières dimensions.

Une approche comparative entre l'ACM et l'ACM sparse est présentée sur les deux premières composantes. Plusieurs valeurs du paramètre de pénalisation λ ont été testées pour l'ACM sparse (comprises entre 0 et 0,5). La table 2.3 présente le nombre de variables sélectionnées sur la première et la deuxième composante, le temps (CPU Time en secondes) et le nombre d'itérations (pour l'ensemble des deux premières CPs) nécessaires à l'algorithme pour converger, en fonction de la valeur de λ. L'algorithme converge très rapidement.

Les figures 2.10 et 2.11 présentent, respectivement, l'évolution du nombre de modalités sélectionnées et celui du pourcentage d'inertie cumulé en fonction du paramètre de régularisation λ choisi dans l'ACM sparse. En pointillés, le résultat de l'ACM. Pour $\lambda = 0,25$, 8 modalités sont sélectionnées sur le premier axe et CPEV$= 23,03\%$. A partir de cette valeur de λ, le CPEV décroit fortement. Nous fixerons donc λ à 0,25 pour la suite de l'analyse. Le pourcentage d'inertie est représenté à titre indicatif étant donné qu'il n'a pas la même signification qu'en ACP (information redondante lors de la construction du tableau de Burt qui sous-estime alors le pourcentage donné). La seule considération du nombre de variables sélectionnées peut alors être suffisant si l'utilisateur à une idée à priori du nombre de variables qu'il souhaite conserver.

TABLE 2.3 – Données chiens : Temps (en secondes) et nombre d'itérations nécessaires à la convergence de l'algorithme en fonction des valeurs de λ pour l'ACM sparse.

Valeurs de λ	0	0.1	0.2	0.25	0.3	0.4	0.5
Nb var. CP1	16	16	11	8	5	2	0
Nb var. CP2	16	16	6	6	6	3	0
Nombre itérations sur 2 CPs	0.02	0.06	0.04	0.06	0.06	0.02	0.02
CPU Time	2	106	71	77	60	7	2

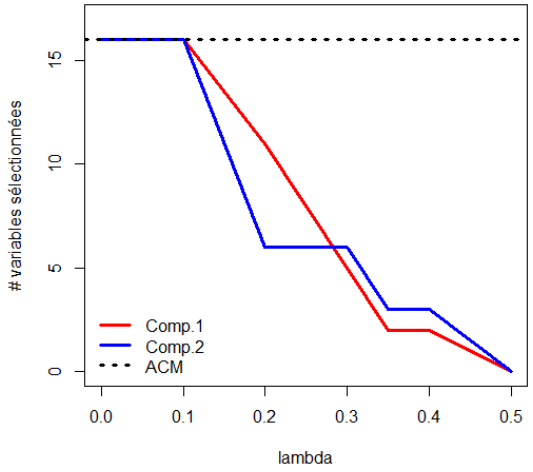

FIGURE 2.10 – Données chiens : Évolution du nombre de modalités sélectionnées en fonction du paramètre de pénalisation λ.

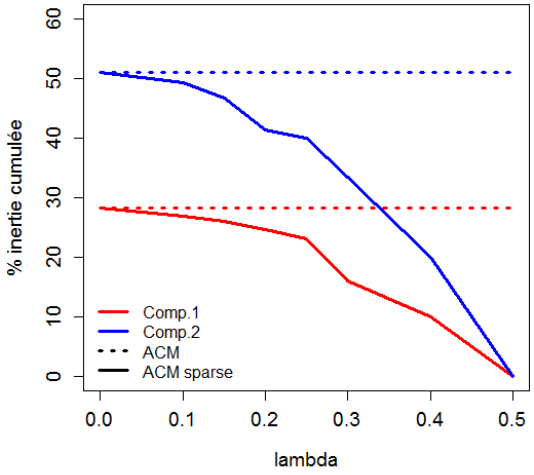

FIGURE 2.11 – Données chiens : Évolution du pourcentage d'inertie cumulé en fonction du paramètre de pénalisation λ.

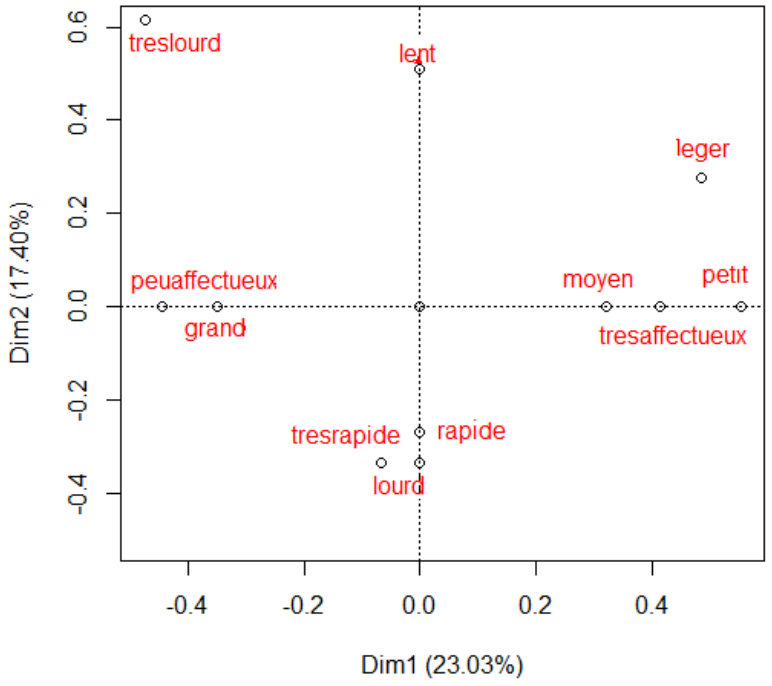

FIGURE 2.12 – Données chiens : Représentation des variables sur les deux premières dimensions de l'ACM sparse.

La figure 2.12 représente les variables sur les deux premières dimensions de l'ACM sparse. Les variables dont les coefficients sont nuls ne sont pas représentées sur la figure. Le tableau 2.4 présente une comparaison des "loadings" obtenus avec l'ACM et l'ACM sparse. L'ACM sparse permet de réduire le nombre de variables sélectionnées par axe, tout en conservant un pourcentage d'inertie cumulé élevé (les variables les plus contributives dans l'ACM sont celles majoritairement conservées dans l'ACM sparse). A partir de la table 2.4 et de la figure 2.12 on remarque que la taille, le poids et l'affection sont les variables les mieux représentées sur l'axe 1 dans l'ACM et ce sont celles qui sont conservées sur l'axe 1 par l'ACM sparse. De la même manière sur l'axe 2, le poids et la rapidité sont les plus contributives dans l'ACM et sont les seules conservées dans l'ACM sparse.

TABLE 2.4 – Données chiens : "Loadings" et variance obtenus avec l'ACM et l'ACM sparse sur les deux premières composantes.

	ACM		**ACM sparse**	
Variable	CP1	CP 2	CP1	CP 2
grand	-0,361	0,071	-0,389	**0,000**
moyen	0,280	0,287	0,226	**0,000**
petit	0,291	-0,400	0,390	**0,000**
leger	0,316	-0,389	0,368	-0,256
lourd	-0,047	0,390	-0,075	0,451
treslourd	-0,294	-0,215	-0,305	-0,479
lent	0,059	-0,383	**0,000**	-0,561
rapide	0,224	0,256	**0,000**	0,282
tres rapide	-0,303	0,156	**0,000**	0,328
moy intelligent	0,173	0,157	**0,000**	**0,000**
peu intelligent	-0,145	-0,309	**0,000**	**0,000**
tres intelligent	-0,086	0,125	**0,000**	**0,000**
peu affectueux	-0,366	-0,084	-0,462	**0,000**
tres affectueux	0,353	0,081	0,445	**0,000**
agressif	-0,170	-0,096	**0,000**	**0,000**
non agressif	0,164	0,093	**0,000**	**0,000**
% inertie	28,19	22,80	23,03	17,40
% inertie cumulé	28,19	50,99	23,03	39,99

L'ACM sparse est donc une extension de la méthode GSPCA pour des données qualitatives. Elle produit de la sparsité au niveau des "loadings" (avec une perte faible du

pourcentage de variance expliquée) ce qui facilite l'interprétation et la compréhension des différents axes. Lorsque le paramètre de régularisation λ est fixé à 0, la GSPCA et l'ACM sparse sont identiques à l'ACP et l'ACM, respectivement, ce qui d'après Zou et al. [2006] est une propriété essentielle d'une "bonne" méthode "sparse". Les deux exemples d'application présentés dans cette section concernent des petits jeux de données, mais ces méthodes prennent tout leur sens dans un contexte de grande dimension, comme nous le verrons par la suite.

2.4 Propriétés de la GSPCA et de l'ACM sparse

Les deux nouvelles méthodes développées au cours de cette thèse possèdent des propriétés importantes pour réaliser de la "sparsité" :

- Sans aucune contrainte de sparsité, la GSPCA et l'ACM sparse reviennent à réaliser une ACP et une ACM, respectivement.
- Ces méthodes sont efficaces dans le cas où I est petit et J est grand.
- Elles permettent d'éviter toute sélection de variables non importantes.

Les propriétés barycentriques de l'ACM sont conservées dans l'ACM sparse pour les individus, mais plus pour les variables (certaines modalités étant mises à zéro). En ACP tout comme en ACM, les composantes principales ne sont pas corrélées et les "loadings" sont orthogonaux. Ces propriétés sont perdues en ACP sparse et en ACM sparse si elles ne sont pas explicitement imposées.

Dans ces deux nouvelles méthodes, la variance expliquée ne peut pas être définie de la même façon qu'en ACP ou en ACM. Dans la GSPCA, la variance expliquée est définie de la même manière que dans la l'ACP sparse de Shen and Huang [2008]. On considère la projection de $\mathbf{X}$ sur le sous espace de dimension k formé des k vecteurs de "loadings" de la manière suivante :

$$\mathbf{X}_{[k]} = \mathbf{X}\mathbf{Q}_{[k]}(\mathbf{Q}_{[k]}^T\mathbf{Q}_{[k]})^{-1}\mathbf{Q}_{[k]}^T, \tag{2.62}$$

avec $\mathbf{Q}_{[k]}$ la matrice des k premiers "loadings sparse". La variance totale expliquée par les k premières composantes est définie comme $\text{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]})$ et la variance ajustée par $\text{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]}) - \text{tr}(\mathbf{X}_{[k-1]}^T\mathbf{X}_{[k-1]})$. Le pourcentage cumulé de variance expliquée (CPEV) par les k premières composantes s'écrit :

$$\text{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]})/\,\text{tr}(\mathbf{X}^T\mathbf{X}). \tag{2.63}$$

Pour l'ACM sparse, une petite modification est apportée. L'ACM est réalisée sur des données en créant des colonnes binaires pour chaque variable avec la contrainte que une et une seule de ces colonnes possède la valeur 1. Ce codage créé des dimensions supplémentaires artificielles car une variable catégorielle est codée par plusieurs colonnes. Par

conséquent, l'inertie (c'est-à-dire la variance) de l'espace solution est artificiellement gonflée et le pourcentage de variance expliquée est largement surestimé (Abdi and Valentin [2007]). Deux corrections sont souvent utilisées pour rectifier ces biais. La première est celle de Benzécri [1979] et la deuxième celle de Greenacre [2010]. Ces corrections tiennent compte du fait que les valeurs propres inférieures à $1/K$ codent pour des dimensions supplémentaires (avec K le nombre de variables catégorielles).

On rappelle que J est le nombre de total de modalités, c'est-à-dire de variables binaires. Si on dénote par λ_ℓ les valeurs propres obtenues par ACM, alors les valeurs propres corrigées, notées λ_ℓ^c sont obtenues de la manière suivante :

$$\lambda_\ell^c = \begin{cases} \left[\left(\frac{J}{J-1}\right)\left(\lambda_\ell - \frac{1}{J}\right)\right]^2 & \text{si } \lambda_\ell > \frac{1}{J} \\ 0 & \text{si } \lambda_\ell \leq \frac{1}{J}. \end{cases} \tag{2.64}$$

Cette formule donne une meilleure estimation de la variance extraite de chaque valeur propre. Normalement, le pourcentage de variance est calculé en divisant chaque valeur propre par la somme des valeurs propres. Cette approche peut être utilisée ici. Cependant l'estimation du pourcentage d'inertie sera trop optimiste.

Une meilleure estimation de la variance a été proposée par Greenacre [2010]. Elle consiste à évaluer plutôt le pourcentage de variance relatif à l'inertie moyenne des blocs non-diagonaux de la matrice de Burt (on rappelle que le tableau de Burt est la matrice $\mathbf{X}^T\mathbf{X}$ de dimensions $J \times J$ associée à $\mathbf{X}$). Cette variance moyenne notée $\bar{\sigma}$ peut être calculée de la manière suivante :

$$\bar{\sigma} = \frac{K}{K-1} \times \left(\sum_\ell \lambda_\ell^2 - \frac{J-K}{K}\right)^2. \tag{2.65}$$

Ainsi, par cette approche, le pourcentage d'inertie serait obtenu par le ratio :

$$\frac{\lambda^c}{\bar{\sigma}} \qquad \text{au lieu de } \frac{\lambda^c}{\sum_\ell \lambda_\ell^c}. \tag{2.66}$$

Chapitre 3

Détection d'interactions SNP-SNP

Dans de nombreux problèmes de régression, le modèle développé ne tient compte que des effets principaux (prédicteurs). Mais souvent, les interactions entre plusieurs prédicteurs peuvent provoquer des différences au niveau de la variable réponse. En génétique, et plus précisément lors de l'analyse de polymorphismes nucléotidiques (en anglais, "Single Nucleotide Polymorphims" :SNPs), la détection d'interactions entre ces signaux est très importante (une définition et une explication des SNPs sont données au chapitre 4 et en annexe B). En effet, certains SNPs sont supposés modifier le risque de développer une maladie. Cependant, il est généralement peu probable qu'un SNP pris individuellement puisse être à l'origine de la susceptibilité ou de la résistance des individus à certaines maladies. En revanche, les interactions entre les SNPs peuvent jouer un rôle important dans l'apparition et/ou le développement de certaines maladies. Les principaux objectifs des études portant sur ces données génétiques sont d'identifier les combinaisons de SNPs conduisant à un risque plus élevé de développer une maladie et de mesurer l'importance de ces interactions.

Une première approche consiste à se servir des bases de données consultables en ligne fournissant des informations sur des interactions connues entre gènes. Cette approche ne permet pas la détection d'interactions mais l'exploration et la visualisation de celles déjà connues impliquant les SNPs considérés. Cela peut donner lieu à des perspectives de recherches intéressantes. Deux de ces bases sont présentées à la section suivante.

Une deuxième approche consiste à détecter des interactions de manière statistique. Il existe de nombreuses approches basées sur des méthodes de classification, telles que

les arbres de classification et de régression (Classification And Regression Tree (CART)) (Breiman et al. [1984]), bagging (Breiman [1996]), les forêts aléatoires (Breiman [2001]), et le Support Vector Machine (SVM) (Cortes and Vapnik [1995]), qui peuvent être appliquées à des données de SNPs et qui permettent de mesurer l'importance des variables prisent séparément. Cependant, la quantification de l'importance des combinaisons de variables n'est pas toujours décrite explicitement. Les arbres de classification ont été développés à l'origine afin de détecter de manière automatique des interactions possibles entre variables. Les méthodes telles que Automatic Interaction Detection (AID) (Morgan and Sonquist [1963]) ou encore CHAID (Kass [1980]) font partie des méthodes d'arbres de classification les plus anciennes. Elles permettent la détection d'interactions en construisant des arbres de décision non-binaires (contrairement à la méthode CART qui calcule des arbres binaires). Un méthode de type régression a été mise au point plus récemment pour résoudre le même type de problème : la régression logique (Ruczinski et al. [2003]). Elle tente d'identifier des combinaisons booléennes de variables binaires pour la prédiction, par exemple, de la maladie pour un individu particulier. En comparaison avec CART ou d'autres procédures de régression (Kooperberg and Ruczinski [2005] ; Witte and Fijal [2001]), la régression logique a montré une bonne performance lorsqu'elle est appliquée à des données de SNPs (Schwender et al. [2004]) et est utilisée fréquemment pour détecter des interactions SNP-SNP (Ruczinski et al. [2001], Kooperberg and Ruczinski [2005], Fritsch and Ickstadt [2007], Chen et al. [2011], Dinu et al. [2012]). Elle sera explicitée dans le paragraphe 3.2 et comparée aux méthodes CHAID et CART à travers un exemple illustratif.

Deux approches seront donc abordées dans ce chapitre : l'approche biologique qui utilise des bases de données fournissant des informations sur les interactions déjà connues, et l'approche statistique qui cherche, sans à priori, des liens possibles entre les gènes pouvant expliquer un phénotype en particulier.

3.1 Approche biologique : interactions biologiques et réseaux connus

Plusieurs bases de données présentant les liens biologiques existants et répertoriés dans la littérature peuvent être trouvées en ligne. La base de données Biological General Repository for Interaction Datasets (BioGRID) par exemple, est une base de données publique créée en 2003 qui archive et diffuse des données sur les interactions génétiques et protéiques provenant de modèles organiques et humains (Stark et al. [2006] ; "http ://www.thebiogrid.org"). BioGRID détient actuellement plus de 700 000 interactions provenant de données à haut débit et d'études axées sur des individus, et provenant de plus de 40 000 publications. L'utilisateur entre le nom de la protéine ou du gène cherché et peut visualiser les interactions connues avec ce gène rapportées dans la littérature.

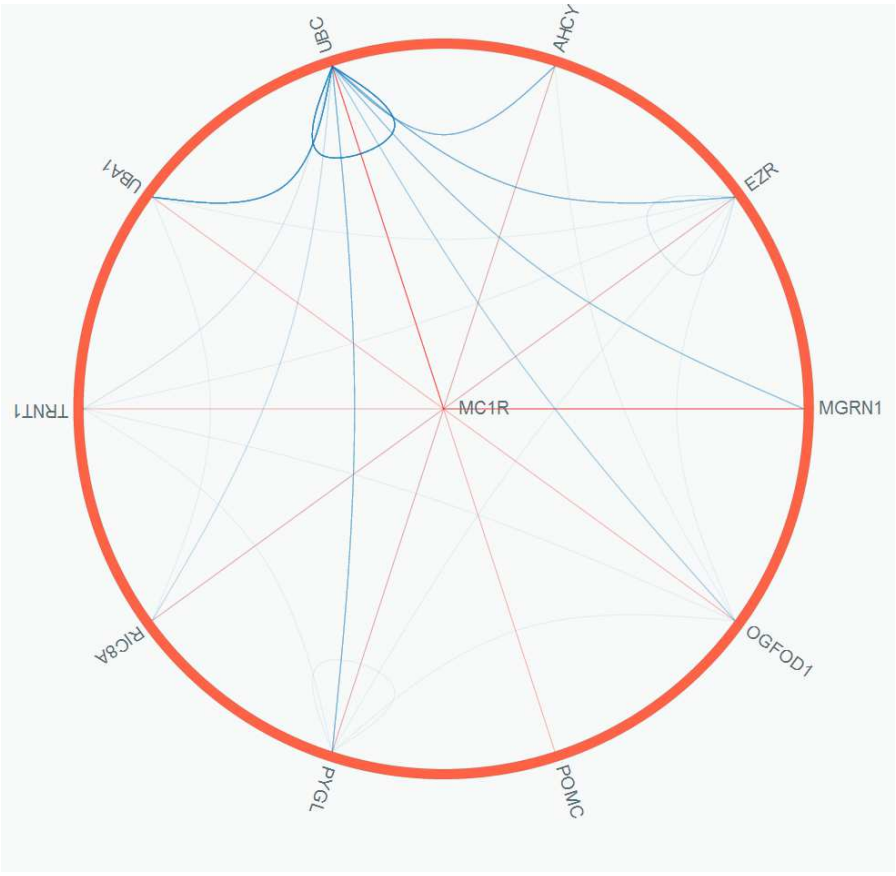

FIGURE 3.1 – Réseau d'interactions pour le gène $MC1R$ obtenu à partir de la base de données BioGRID.

La figure 3.1 présente un exemple de la visualisation obtenue des gènes en interaction avec le gène MelanoCortin 1 Receptor ($MC1R$) impliqué entre autre dans la peau humaine. A partir du nom du gène, on peut tracer le réseau biologique d'interactions connues et ainsi visualiser les liens entre gènes.

Un autre exemple de base de données est la base STRING (Jensen et al. [2009]). Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) est une base de données d'interactions protéiques connues ("http ://string-db.org"). Les interactions comprennent des associations directes (physiques) et indirectes (fonctionnelles) ; et proviennent de quatre sources : le contexte génomique, les expériences à haut-débit, la co-expression et les pré-requis (base de connaissance, littérature et recueil d'informations). STRING intègre les données d'interactions de ces sources pour un grand nombre d'organismes vivants. La base de données couvre actuellement 5 214 234 protéines de 1 133 organismes. Outre les prévisions internes et les transferts, STRING s'appuie également sur de nombreuses ressources maintenues ailleurs telles que PubMed, BioGRID (décrit ci-dessus), Kyoto Encyclopedia of Genes and Genomes (KEGG) ; Kanehisa et al. [2004]), Gene Ontology, etc. Le site permet de générer un réseau d'interactions connues autour d'un gène donné par l'utilisateur (comme montré figure 3.2), mais permet également à l'utilisateur d'entrer un ensemble de gènes ou de protéines afin de visualiser les interactions existantes et connues entre ceux-ci. Chaque noeud correspond à un gène donné en entrée par l'utilisateur. Chaque noeud (gène) peut présenter jusqu'à 8 arêtes (c'est-à-dire les liens fonctionnels) les reliant à d'autres gènes, avec une couleur pour chaque type de preuve : fuchsia pour la mise en évidence des liens par expériences réalisées, rouge pour la fusion de gènes, bleu pour la co-ocurrence, noir pour la co-expression, vert pour les informations contenues dans les bases de données, vert foncé pour les gènes voisins, vert anis pour l'exploration de texte, violet pour l'homologie. La base de données STRING sera utilisée par la suite dans nos analyses pour analyser les interactions entre un ensemble de plusieurs gènes.

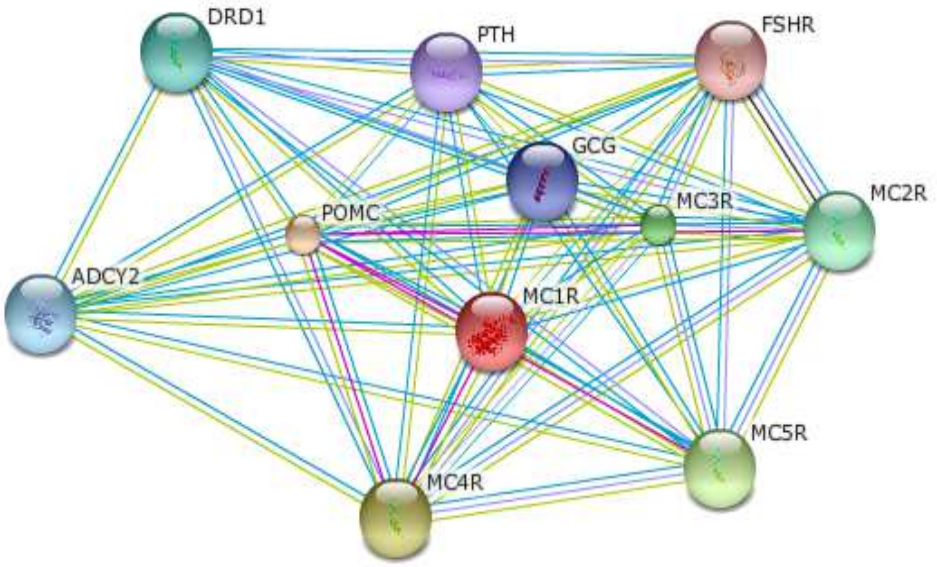

FIGURE 3.2 – Réseau d'interactions pour le gène $MC1R$ obtenu à partir de la base de données STRING (fuchsia : mise en évidence des liens par expériences réalisées, rouge : fusion de gènes, bleu : co-ocurrence, noir : co-expression, vert : informations contenues dans les bases de données, vert foncé : gènes voisins, vert anis : exploration de texte, violet : homologie).

3.2 Approche statistique : régression logique

L'un des principaux objectifs dans les études d'association génétique est la construction de règles de classification comme :

SI le SNP1 est de génotype homozygote de référence (gène qui, chez un individu, est représenté par deux allèles identiques dont la fréquence est la plus élevée) ET le SNP2 est de génotype homozygote variant OU les SNPs 3 ET 4 ne sont pas du génotype homozygote de référence,

ALORS une personne a un risque plus élevé de développer une maladie particulière.

Une procédure mise au point pour résoudre exactement ce type de problème a été proposée par Ruczinski et al. [2003] : la régression logique.

3.2.1 Expressions et régression logique

La régression logique est une méthode de régression adaptative principalement développée pour explorer des interactions d'ordre élevé de données génomiques. Elle est destinée aux situations où la plupart des variables explicatives sont binaires. Étant donné un ensemble de prédicteurs binaires $\mathbf{X}$ (vrai et faux, 0 et 1, ...), le but est de créer de nouveaux et meilleurs prédicteurs de la réponse en tenant compte des combinaisons de ces prédicteurs binaires. Par exemple, si la réponse est binaire (ce qui n'est pas une exigence pour la méthode décrite ici), le but est d'obtenir des règles de décision telles que "si $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, et $\mathbf{X}_4$ sont vraies" ou "$\mathbf{X}_5$ ou $\mathbf{X}_6$ mais pas $\mathbf{X}_7$ sont vraies", alors la réponse est plus susceptible d'être 1. En d'autres termes, nous essayons de trouver des combinaisons booléennes (logiques) de prédicteurs binaires ayant un fort pouvoir prédictif de la variable réponse.

Expressions logiques

Ces combinaisons sont des expressions booléennes logiques telles que $L = (\mathbf{X}_1 \wedge \mathbf{X}_2) \vee \mathbf{X}_3^c$. Les prédicteurs étant binaires, chacune de ces combinaisons sera également binaire. Trois opérateurs peuvent être rencontrés : $\wedge$ (AND), $\vee$ (OR) et ^{c} (NOT). $\mathbf{X}^c$ est appelé conjugué de $\mathbf{X}$. Une expression booléenne peut être générée en combinant de manière itérative deux variables, une variable et une expression booléenne ou deux expressions booléennes, comme indiqué dans l'équation (3.1) :

$$(\mathbf{X}_1 \wedge \mathbf{X}_2^c) \wedge [(\mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_5 \wedge (\mathbf{X}_3^c \vee \mathbf{X}_6))] \tag{3.1}$$

Cette équation peut être lue comme une déclaration "AND", générée à partir des expressions $(\mathbf{X}_1 \wedge \mathbf{X}_2^c)$ et $(\mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_5 \wedge (\mathbf{X}_3^c \vee \mathbf{X}_6))$. L'utilisation de l'interprétation des expressions booléennes permet de représenter une expression booléenne sous la forme d'un arbre binaire, comme représenté sur la figure 3.3. Pour plus de simplicité, seul l'indice de la variable est représenté. Les lettres blanches sur fond noir représentent le conjugué de la variable.

La terminologie (semblable à la terminologie utilisé par Breiman et al. [1984] pour les arbres de décision) et les règles suivantes sont utilisées pour les arbres logiques :

- L'emplacement de chaque élément (variable, conjugué de la variable, opérateurs $\wedge$ et $\vee$)

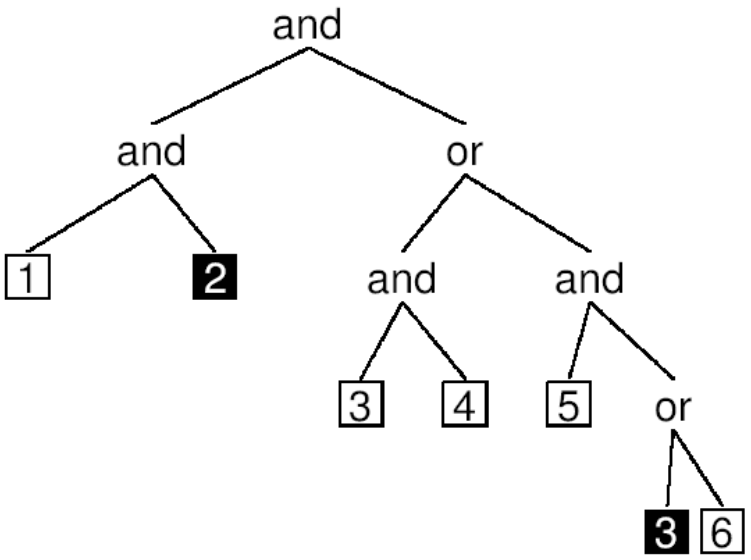

FIGURE 3.3 – Arbre logique représentant l'expression booléenne $(\mathbf{X}_1 \wedge \mathbf{X}_2^c) \wedge [(\mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_5 \wedge (\mathbf{X}_3^c \vee \mathbf{X}_6))]$ (d'après Ruczinski et al. [2003]).

dans l'arbre est un nœud.

- Chaque nœud possède zéro ou deux sous nœuds.
- Les deux sous nœuds d'un nœud sont appelés ses enfants, le nœud lui-même est appelé le parent.
- Le nœud qui n'a pas de parent est appelé la racine.
- Les nœuds sans enfant sont appelés feuilles.
- Les feuilles ne peuvent être occupées que par des lettres ou des conjugués (prédicteurs), tous les autres nœuds sont des opérateurs ($\wedge$ et $\vee$).

Régression logique

On considère $\mathbf{X}_1, ..., \mathbf{X}_k$ l'ensemble des prédicteurs binaires et $\mathbf{y}$ la variable réponse. Si on considère le cas d'un arbre unique, alors on obtient une expression logique unique L. Si la variable réponse est binaire, si L est vraie pour une observation, alors cette observation sera notée 1. En plus de cette approche d'arbre unique, les expressions logiques peuvent aussi fournir des méthodes d'arbres multiples dans lesquelles de nombreuses expressions

logiques L_i $(i = 1, \ldots, I)$ sont construites et combinées par le modèle linéaire généralisé suivant :

$$g\left(E[\mathbf{y}]\right) = \beta_0 + \sum_{i=1}^{I} \beta_i L_i, \tag{3.2}$$

avec β_i, $i = 0, ..., I$, les paramètres et g une fonction de lien. Ce cadre peut comprendre en compte la régression linéaire ($g\left(E[\mathbf{y}]\right) = E[\mathbf{y}]$) ou la régression logistique dans le cas d'une variable réponse binaire ($g\left(E[\mathbf{y}]\right) = \log\left(E[\mathbf{y}]/(1 - E[\mathbf{y}])\right)$). Pour chaque type de modèle, une fonction score est définie et permet d'évaluer la "qualité" du modèle considéré. Pour la régression linéaire, le score peut être la somme des carrés des résidus, et pour la régression logistique, il peut correspondre à la déviance. Le but est donc de trouver les expressions booléennes qui minimisent la fonction score associée au modèle, en estimant simultanément les paramètres β_j et les expressions booléennes L_j.

3.2.2 Recherche du meilleur modèle

Les expressions logiques peuvent être comme on l'a vu, représentées par des arbres logiques qui peuvent être employés pour générer de nouveaux arbres dans la recherche du meilleur modèle.

Mouvements possibles

Ces autres arbres logiques peuvent être générés à partir d'un nombre fini d'opérations telles que la croissance des branches, l'élagage des branches et le changement de feuilles. La figure 3.4 représente les différents types de mouvements possibles pour l'expression logique $L = (S_{11} \wedge S_{21}^c \vee S_{32})$. Les voisins d'un arbre logique sont les arbres qui peuvent être atteints à partir de cet arbre logique par un simple "mouvement". Compte tenu des mouvements possibles, un arbre logique peut être atteint à partir de n'importe quel autre arbre logique à partir d'un nombre fini de mouvements, faisant référence à l'irréductibilité de la théorie des chaînes de Markov.

Cependant, le nombre d'arbres logiques constructibles pour un ensemble donné de variables explicatives est énorme, et il n'existe aucun moyen simple de lister tous les arbres logiques qui donnent des prédictions différentes. Il est donc impossible de procéder à une

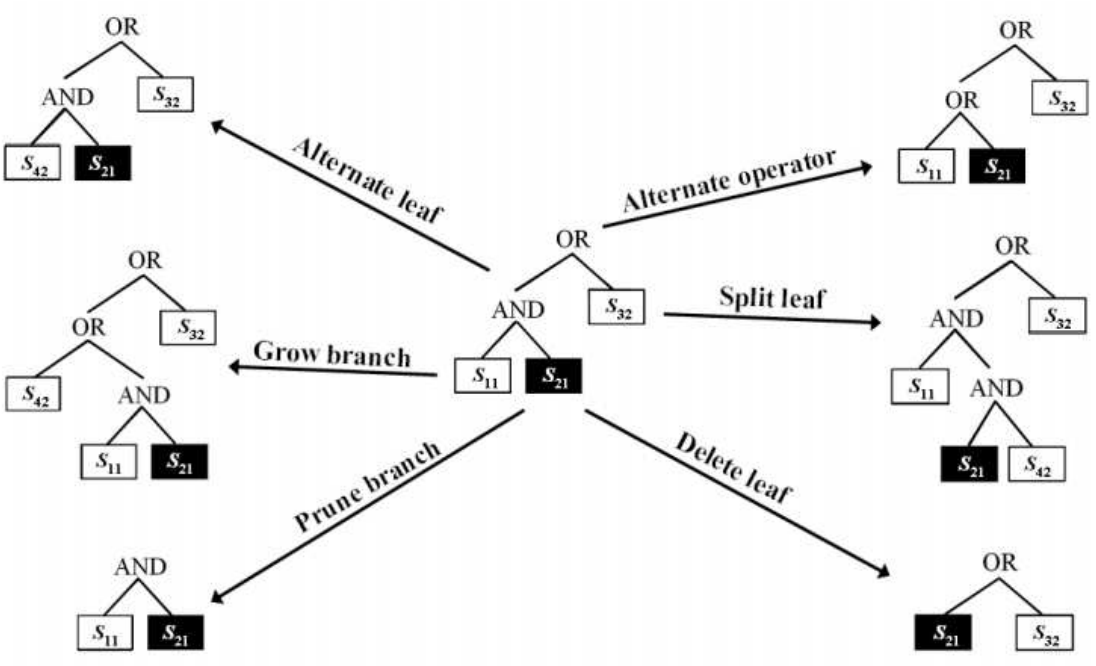

FIGURE 3.4 – Mouvements admissibles dans l'arbre logique pour l'expression logique $L = (S_{11} \wedge S_{21}^c \vee S_{32})$. L'arbre de départ est au centre. Les lettres blanches sur fond noir représentent le conjugué de la variable.

évaluation exhaustive de l'ensemble des différents arbres logiques. Pour ce faire, Ruczinski et al. [2004] proposent d'utiliser un algorithme de recherche stochastique : l'algorithme de recuit simulé.

Remarque. Les algorithmes génétiques et le recuit simulé étant deux techniques d'optimisation travaillant sur les mêmes types de problèmes, les algorithmes génétiques pourraient être également utilisés dans ce contexte (Goldberg et al. [1994]). Ils fournissent des solutions quasi-optimales mais au prix d'un temps de convergence généralement plus long que celui du recuit simulé.

Algorithme de recuit simulé

L'algorithme de recuit simulé est défini sur un espace d'état $\mathcal{S}$, qui est une collection d'états individuels. Chacun de ces états représente une configuration du problème considéré. Les états sont liés par un système de voisinage, et l'ensemble de paires voisines dans $\mathcal{S}$ définit

une structure $\mathcal{M}$ dans l'espace $\mathcal{S} \times \mathcal{S}$. Les éléments de $\mathcal{M}$ sont appelés des mouvements. Deux états s et s' sont adjacents s'ils peuvent être atteints par un unique mouvement.

L'idée de base de l'algorithme de recuit simulé est la suivante :

1. Étant donné un certain état, choisir un mouvement parmi l'ensemble des mouvements possibles, ce qui conduit à un nouvel état.

2. Comparer les scores de l'ancien et du nouvel état. Si le score du nouvel état est mieux que le score de l'ancien état, accepter le mouvement.
   Si le score du nouvel état n'est pas mieux que le score de l'ancien état, accepter le mouvement avec une certaine probabilité.

La probabilité d'acceptation dépend du score des deux états considérés, et d'un paramètre qui reflète à quel endroit de la chaine de recuit on se trouve (ce paramètre est généralement désigné comme la température). Pour toute paire de scores, si l'état proposé a un score moins bon que le précédent, plus on se trouve loin dans la chaine de recuit, plus la probabilité d'acceptation est faible. Il y a différentes options sur la manière de mettre en œuvre l'algorithme de recuit et d'ajuster les modèles logiques. Tous les arbres sont simultanément ajustés dans le modèle simultanément. Cela nécessite, pour des raisons de calcul, de présélectionner le nombre maximum d'arbres t qui peut être choisi arbitrairement grand si nous n'avons aucune idée a priori du nombre maximum d'arbres voulu. En principe, un arbre est sélectionné dans le modèle logique puis un mouvement de déplacement est choisi aléatoirement pour cet arbre. Les paramètres sont ré-ajustés pour le nouveau modèle et le score est déterminé. Celui-ci est ensuite comparé aux scores de ceux des états précédents, comme décrit précédemment.

L'utilisation du recuit simulé permet de trouver un modèle le plus près du meilleur score possible. Dans Ruczinski et al. [2003], le nombre total de feuilles dans les arbres logiques d'un modèle est utilisé comme mesure de la complexité du modèle, qu'ils appellent la taille du modèle. En revanche la détermination de la meilleure taille du modèle est importante. La validation croisée peut être utilisée pour ce faire, et si la taille des données le permet, une approche ensemble test/ensemble d'apprentissage peut être réalisée.

3.2.3 Identification des interactions intéressantes : algorithme logicFS

Dans le cas d'une variable réponse binaire, les méthodes de type Random Forest (RF), par exemple, identifient les variables expliquant les individus "malade" (variable réponse valant 1) comme les variables les plus importantes. Cependant ni les interactions intéressantes ni les génotypes pouvant expliquer les individus "malades" ne sont détectés. Cela peut être considéré comme un inconvénient majeur pour l'analyse des données provenant d'études d'association génétique, étant donné que ce ne sont pas les SNPs pris individuellement mais les interactions entre SNPs qui sont supposés être à l'origine de maladies complexes.

Schwender and Ickstadt [2008] ont donc proposé une procédure, appelée algorithme logicFS (logic Feature Selection), permettant l'identification de variables et d'interactions susceptibles d'expliquer le "statut" d'un individu. Cette approche consiste à utiliser l'algorithme de recuit simulé présenté précédemment et à l'appliquer à différents sous-ensembles des données. Schwender [2007b] montre l'avantage de la méthode logicFS sur un des autres algorithmes de recherche utilisés dans la régression logique basé sur l'approche Monte Carlo par chaîne de Markov : la Monte Carlo (MC) régression logique. Kooperberg and Ruczinski [2005] utilisent cet algorithme sur l'ensemble des données non pas pour trouver le meilleur modèle de régression logique, mais pour obtenir une large collection de modèles presque aussi bons que le meilleur. Cet ensemble est ensuite utilisé pour identifier les combinaisons de variables apparaissant souvent dans ces modèles.

L'utilisation de la logic Feature Selection nécessite la transformation des expressions logiques en forme disjonctive normale (Disjunctive Normal Form (DNF)). L'expression logique donnée en exemple est relativement facile à interpréter, cependant cela devient de plus en plus compliqué d'interpréter ce genre d'expressions lorsqu'elles contiennent de nombreuses variables. Ainsi, Schwender and Ickstadt [2008] ont proposé de convertir ces expressions en DNF, qui est une "OR"-combinaison de "AND"-combinaisons. La DNF de l'arbre logique $L = (\mathbf{X}_1 \wedge \mathbf{X}_2^c) \wedge [(\mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_5 \wedge (\mathbf{X}_3^c \vee \mathbf{X}_6))]$ donné figure 3.3 par

exemple, est donnée par :

$$L = (\mathbf{X}_1 \wedge \mathbf{X}_2^c) \wedge [(\mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_5 \wedge \mathbf{X}_3^c) \vee (\mathbf{X}_5 \wedge \mathbf{X}_6)] \tag{3.3}$$
$$= (\mathbf{X}_1 \wedge \mathbf{X}_2^c \wedge \mathbf{X}_3 \wedge \mathbf{X}_4) \vee (\mathbf{X}_1 \wedge \mathbf{X}_2^c \wedge \mathbf{X}_5 \wedge \mathbf{X}_3^c) \vee (\mathbf{X}_1 \wedge \mathbf{X}_2^c \wedge \mathbf{X}_5 \wedge \mathbf{X}_6)$$

La procédure de conversion de l'expression logique en DNF (minimale) proposée par Schwender and Ickstadt [2008] est basée sur l'algorithme de Quine-McCluskey (Quine [1952] ; McCluskey [1956]) modifié.

L'avantage de la DNF est que les interactions sont directement identifiables car elles sont données par les combinaisons de "AND". L'objectif ici est d'identifier toutes les interactions qui peuvent avoir une influence sur la variable réponse. Dans Schwender [2007a] un algorithme basé sur le calcul matriciel pour générer un tel DNF d'une expression logique est présenté. L'algorithme 9 de la logicFS peut donc à présent être défini.

Algorithm 9 Algorithme logicFS

1. Créer un échantillon bootstrap de taille N à partir des N observations de l'ensemble des variables d'intérêt.
2. Construire un modèle de régression logique sur la base de l'échantillon bootstrap.
3. Convertir chaque expression logique sous la forme DNF.
4. Répétez les étapes 1 à 3, B fois (100 ou 200 fois).

Certaines interactions identifiées par logicFS sont très importantes pour la prédiction. D'autres en revanche ne le sont pas ou pourraient faire obstruction à la bonne prédiction du statut d'une observation. Il est donc nécessaire de quantifier l'importance de chacune de ces interactions potentiellement intéressantes.

3.2.4 Mesure de l'importance des interactions identifiées

Les modèles sont utilisés pour calculer pour chaque variable, chaque paire et chaque triplet de variables la proportion de modèles dans lequel les variables respectives apparaissent conjointement dans le même arbre logique. Les combinaisons de variables les plus fréquentes sont alors supposées être les interactions les plus importantes. Une mesure appropriée quantifie de combien l'interaction améliore la classification (Variable Importance

Measurement (VIM)). Cette amélioration ne doit pas être calculée sur le même jeu de données que celui utilisé pour effectuer les règles de classification, mais sur un jeu de données indépendant contenant de nouvelles observations. L'algorithme logicFS est un modèle de régression logique construit sur la base d'un sous-ensemble de données, les observations Out-Of-Bag (OOB) (observations ne figurant pas dans l'échantillon bootstrap) peuvent être utilisées pour estimer l'importance des interactions.

Dans le cas d'un seul arbre l'importance d'une variable (VIM) ou d'une interaction P est calculée de la manière suivante :

$$VIM_{\text{single}} = \frac{1}{B}\Big(\sum_{b:P\in L_b}(N_b - N_b^-) + \sum_{b:P\notin L_b}(N_b^+ - N_b)\Big), \tag{3.4}$$

avec L_b l'ensemble des premiers impliquants identifiés dans la b-ème itération de logicFS, N_b est le nombre d'observations OOB dans la b-ème itération qui sont correctement classées par le modèle de régression logique et N_b^+/N_b^- est le nombre d'observations OOB correctement classées par le b-ème modèle après que P ait été ajouté/ sorti du modèle.
Dans le cas d'arbres multiples, il n'est pas possible d'ajouter une interaction à l'un des arbres sans ambiguïté car il n'est pas évident de savoir à quelle expression logique l'ajouter. Le premier impliquant P est donc seulement retiré (et non ajouté) aux modèles, et la mesure multi-arbre est déterminée par : (a) calculer le nombre N_b d'observations OOB correctement classées pour chacune des B itérations, (b) enlever P de tous les modèles, (c) recalculer le nombre d'observations OOB correctement classées (maintenant notées N_b^*) pour chacune des B itérations, (d) calcul de :

$$VIM_{multi} = \frac{1}{B}\sum_{b=1}^{B}(N_b - N_b^*) = \frac{1}{B}\sum_{b:P\in L_b}(N_b - N_b^*) \tag{3.5}$$

Une valeur élevée du VIM signifie que l'interaction correspondante est très importante. En revanche, une valeur proche de zéro suggère que l'interaction n'est pas importante pour la classification.

3.3 Exemple d'application

De nombreuses méthodes de construction d'arbres de décision dont les méthodes CART (Breiman et al. [1984]) et CHAID (Kass [1980]) peuvent servir à la détection d'interactions. La principale différence entre ces méthodes réside dans le processus de construction de l'arbre. L'algorithme CART (Classification And Regression Tree) construit des arbres de décision binaires (c'est-à-dire que chaque nœud ne peut avoir que deux branches) et les variables considérées peuvent être binaires, qualitatives ou quantitatives. La première phase de construction de l'arbre est la phase d'expansion réalisée sur l'ensemble d'apprentissage et le critère de segmentation utilisé (mesure de la division) est l'indice de Gini (Gini [1921]). Dans cette phase, on décide si un nœud est terminal, on sélectionne un test à associer à un nœud et on affecte une classe à une feuille. Cela permet de construire l'arbre maximal. Arrive ensuite la phase d'élagage (post-élagage) pour éviter le sur-apprentissage des données, qui consiste à réduire la taille de l'arbre complet. Les sous-arbres sont ordonnés selon une séquence emboîtée suivant la décroissance d'un critère pénalisée de déviance ou de taux de mal-classés et le sous-arbre optimal est sélectionné. L'élagage de l'arbre se fait par estimation de l'erreur réelle, en utilisant un ensemble test. La phase d'élagage peut être modifiée pour le cas d'échantillons de plus petite taille, on utilise alors la validation croisée comme estimation de l'erreur réelle.

La méthode CHi-squared Automatic Interaction Detector (CHAID) considère tout type de variables et peut générer des arbres binaires mais également non binaires contrairement à CART. L'enjeu de la recherche de la taille optimale d'un arbre consiste à stopper (pré-élagage) ou à réduire l'arbre (post-élagage). Contrairement à CART qui réalise du post-élagage, CHAID va éviter une trop grande croissance de l'arbre en fixant une règle d'arrêt qui permet de stopper la construction de l'arbre lors de la phase de construction. Pour évaluer la pertinence de la variable dans la segmentation CHAID utilise le Khi-2 d'écart à l'indépendance. On accepte la segmentation si le Khi-2 calculé sur un nœud est significativement supérieur à un seuil théorique lié à un risque α fixé.

Dans cette section les résultats et la performance de la régression logique sont comparés à ceux des deux méthodes CART et CHAID, ainsi qu'à la régression logistique (Bishop and

Nasrabadi [2006]), l'analyse discriminante linéaire (Bardos [2001]), à la méthode d'agrégation de modèles : randomforest (Breiman [2001]), et au SVM (Support Vecteur Machine, Cortes and Vapnik [1995]) sur un exemple dont les données sont issues d'une enquête cas-témoins (De Micheaux et al. [2011]). Le but est d'évaluer l'existence d'un risque plus élevé de survenue d'un infarctus du myocarde chez les femmes qui utilisent ou ont utilisé des contraceptifs oraux. L'étude a été menée auprès de 149 femmes ayant eu un infarctus du myocarde (cas) et 300 femmes n'en ayant pas eu (témoins). Le facteur d'exposition principal est la prise de contraceptifs oraux (variable binaire oui/non nommée "CO"), les autres facteurs recueillis sont :

1. l'âge (variable continue),
2. le poids (variable continue),
3. la taille (variable continue),
4. la consommation de tabac (variable catégorielle 0 : non, 1 : fumeuse actuelle, 2 : ancienne fumeuse),
5. l'hypertension artérielle (variable binaire oui/non nommée "HTA"),
6. les antécédents familiaux de maladies cardio-vasculaires (variable binaire oui/non nommée "ATCD").

La régression logique cherche des combinaisons booléennes de variables binaires. Chaque variable doit donc être transformée. Les trois variables continues (l'âge, le poids et la taille) ont été discrétisées en 2 classes (découpage par rapport à la médiane). Le tableau de données comporte donc à présent 7 variables catégorielles et la variable "infarct" (codée en oui/non) que l'on souhaite expliquer.

Pour la procédure de régression logique des paramètres doivent être considérés tels que le nombre maximum de branches souhaité (ici fixé à 4 mais les résultats sont les mêmes pour un nombre plus élevé de branches), le nombre d'itérations (fixé à 200) et le nombre d'arbres (ici un seul). Pour la méthode CHAID seul le critère d'arrêt statistique peut être défini. En SVM, l'hyperplan plan optimal dépend des paramètres C (constante du terme de régularisation dans la formule de Lagrange) et gamma (qui contrôle la forme de l'hyperplan de séparation). Aucun paramètre n'est à fixer pour les autres méthodes utilisées.

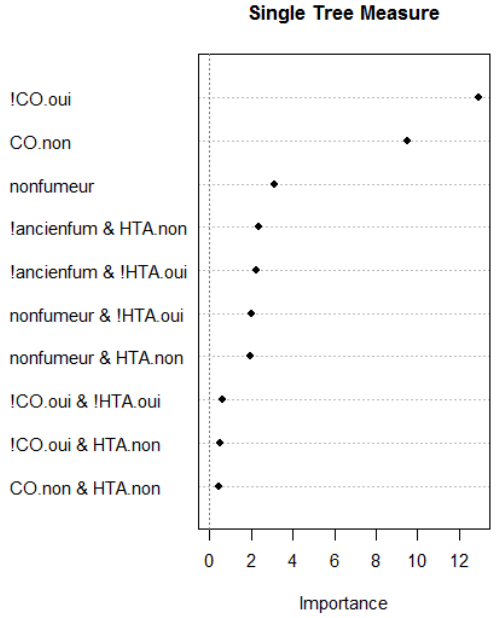

FIGURE 3.5 – Données myocarde : Importance des interactions détectées par régression logique.

La figure 3.5 présente les 10 variables et/ou interactions les plus importantes parmi les 23 détectées par régression logique. La variable la plus importante est la prise de contraceptif oral ou non, comme attendu, la deuxième est la variable tabac. L'interaction la plus importante est celle entre les variables tabac et HTA. Ces règles peuvent s'interpréter de la manière suivante :

SI prise de contraceptif oral **OU** fumeur **OU** (fumeur **ET** antécédents familiaux de maladies cardio-vasculaires)

ALORS risque plus élevé d'infarctus du myocarde.

Les figures 3.7 et 3.8 présentent les arbres de classification obtenus avec les méthodes CART et CHAID. Les variables importantes sont exactement les mêmes que celles obtenues avec la régression logique et l'ordre est conservé : variable CO, puis tabac, puis HTA. Quatre règles sont obtenues avec CART et 7 avec CHAID, dont les principales sont :

1. prise de contraception orale **ET** fumeuse → risque plus élevé d'infarctus

2. prise de contraception orale **ET** non fumeuse **ET** HTA → risque plus élevé d'infarctus

3. prise de contraception orale **ET** non fumeuse **ET** pas HTA → risque d'infarctus moindre

4. pas de contraception orale → risque d'infarctus moindre

La lecture de l'arbre obtenu avec CHAID est simplifiée car cette méthode permet de visualiser les étapes successives au cours desquelles l'algorithme identifie les variables qui permettent de séparer au mieux les différentes catégories de la variable dépendante. Par exemple, C0 ET fumeuse → 60% de chance que l'individu ait un infarctus du myocarde. Ainsi les résultats obtenus pour les trois méthodes se recoupent.

Les performances des différentes méthodes (en termes de prédiction) ont été comparées à l'aide de courbes ROC (Receiver Operating Characteristic, Metz [1978], Zweig and Campbell [1993]) estimées sur l'échantillon test. La courbe ROC est une représentation graphique de la relation existante entre la sensibilité (probabilité que le test soit positif s'il y a infarctus, se mesurant chez les "infarctus" seulement) et la spécificité d'un test (probabilité que le test soit négatif s'il n'y a pas infarctus, se mesurant chez les "non infarctus" seulement) pour toutes les valeurs seuils possibles (le test donne un résultat numérique avec un seuil s tel que la prédiction est positive si $x > s$, et la prédiction est négative si $x < s$). L'ordonnée représente la sensibilité et l'abscisse correspond à la quantité (1 - spécificité). La courbe ROC relie donc le taux de vrais positifs au taux de faux négatifs. La figure 3.6 représente la courbe ROC des 7 différentes méthodes citées précédemment. On remarque que la plupart des courbes se croisent, ce qui signifie qu'il n'y a pas de méthode uniformément meilleure qu'une autre. Afin de comparer les modèles obtenus, les aires sous la courbe ROC (Area Under Curve (AUC)) pour chacune des méthodes peuvent être calculés (indice global de la qualité de la prédiction variant de 0,5 à 1 ; 0,5 dans le cas où le modèle classe au hasard les individus et 1 pour une prédiction idéale), cependant ces indices ne permettraient pas de donner un ordre total pour classer les modèles car les courbes ROC se croisent. Nous allons donc comparer les modèles en fonction du taux de mauvais classement appelé Misclassification Rate (MCR) obtenu sur l'ensemble test. Les résultats sont présentés dans la table 3.1.

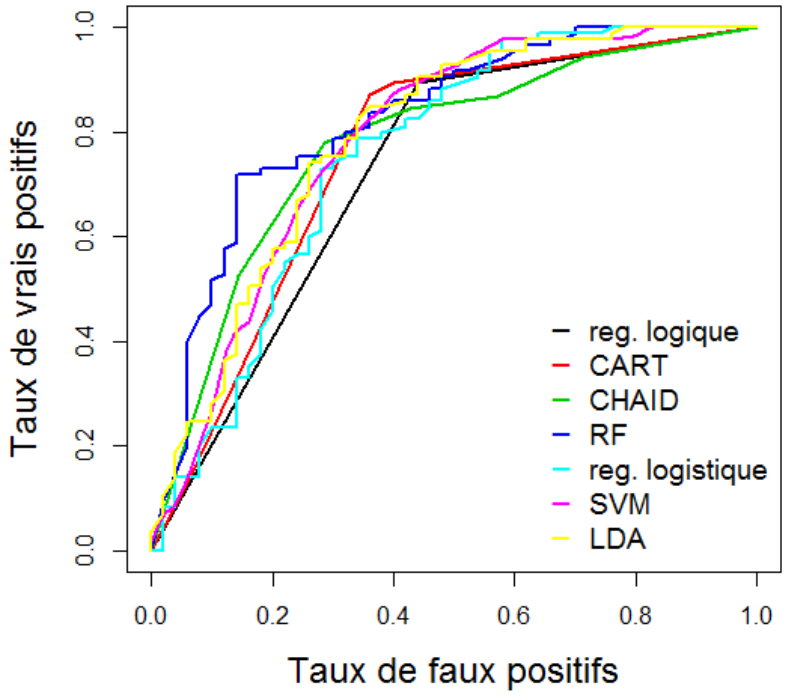

FIGURE 3.6 – Données myocarde : courbes ROC des différentes méthodes testées.

TABLE 3.1 – Données myocarde : taux de mauvais classement (en %) pour les 7 méthodes testées.

Méthodes	Taux de mauvais classement
Régression logique	22,96 %
CART	21,48 %
CHAID	23,70 %
Random Forest	24,44 %
Régression logistique	26,66 %
SVM	22,96 %
Analyse Discriminante	22,96 %

Les taux de mauvais classement obtenus sont proches pour la plupart des méthodes. CART présente le taux le plus faible et ceux obtenus pour la régression logique, SVM, CHAID et l'analyse discriminante sont très proches. La régression logistique présente le taux le plus élevé.

Les résultats obtenus pour ces méthodes sont donc proches sur ce jeu de données, néanmoins pour l'étude de données de SNPs de très grande dimension, la régression logique s'est avérée être plus performante (Schwender et al. [2004],Chen et al. [2011]). Elle sera utilisée dans les analyses présentées au chapitre 4. Les interactions SNP-SNP peuvent néanmoins être détectées par random forest (Winham et al. [2012]) qui calcule un critère d'importance pour chaque interaction.

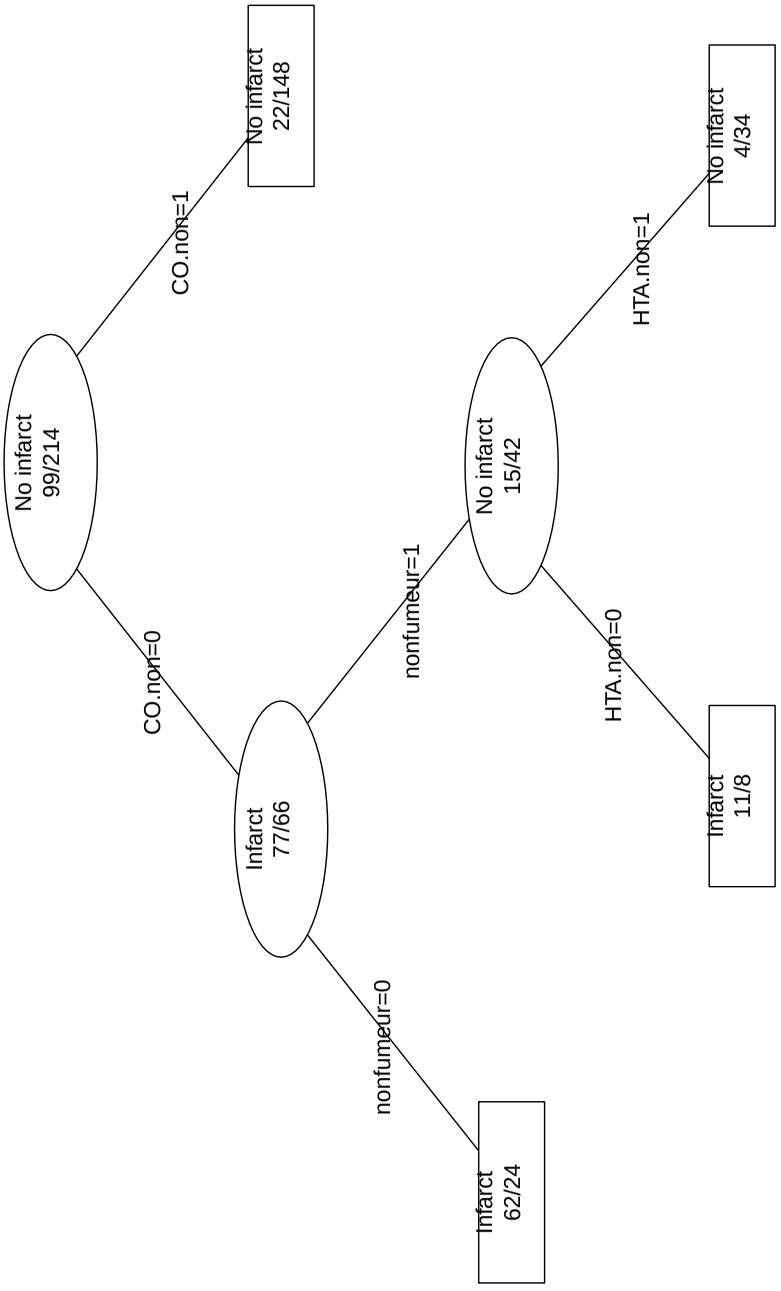

FIGURE 3.7 – Données myocarde : Arbre de classification obtenu par la méthode CART.

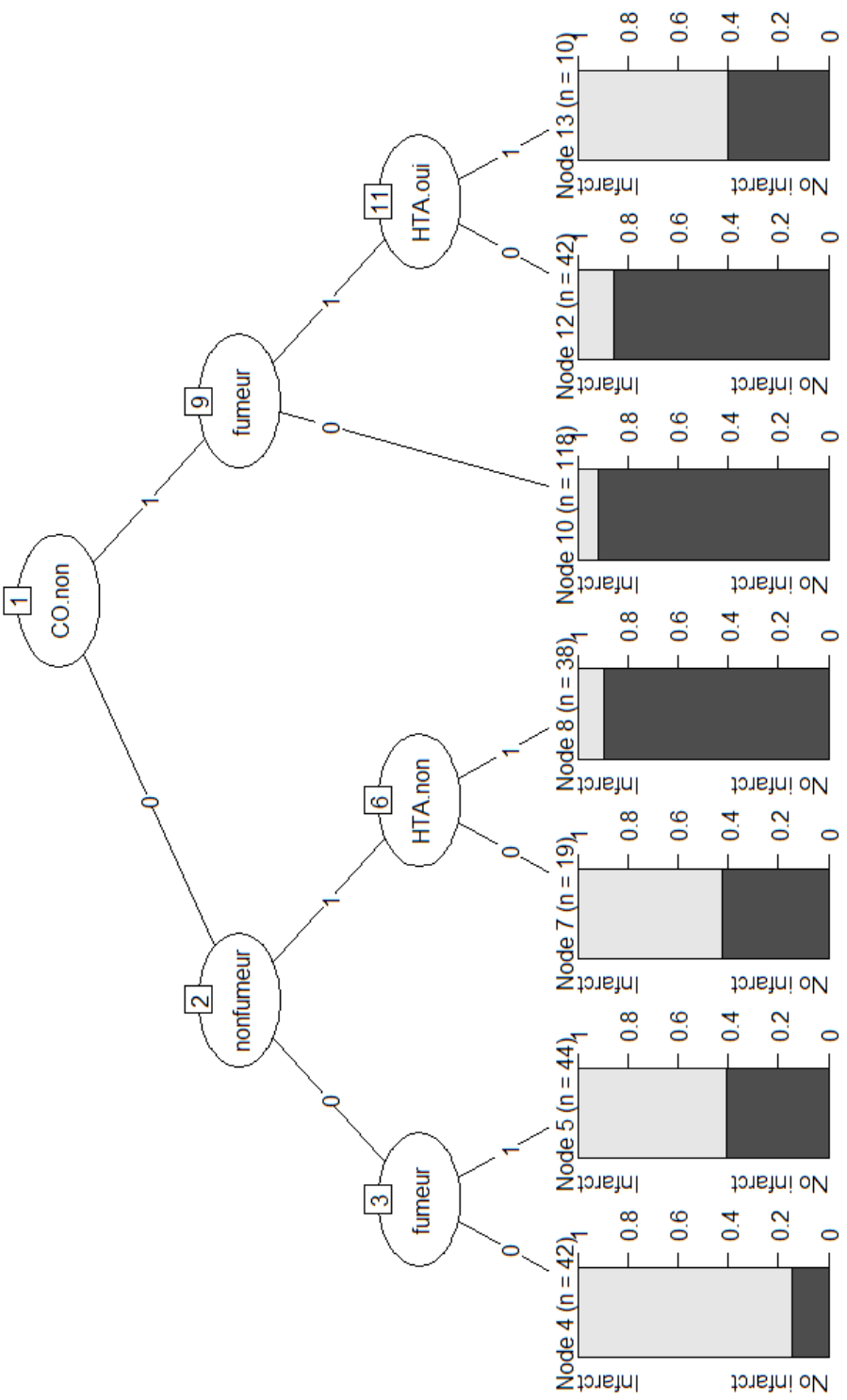

FIGURE 3.8 – Données myocarde : Arbre de classification obtenu par la méthode CHAID.

Chapitre 4

Application sur des données génétiques de biopuces

4.1 Données et pré-traitements

Jusqu'à récemment aucune étude d'association génomique n'avait spécifiquement recherché de liens avec le vieillissement cutané. Après avoir conduit différents travaux pour étudier les déterminants de l'état cutané des volontaires de la cohorte SU.V.I.MAX., le CEntre de Recherche et d'Investigation Epidémiologique et Sensorielle (CE.R.I.E.S.) a mis en place en 2002 une étude transversale en collaboration avec l'Association Projet SU.V.I.MAX. (Hercberg et al. [1998b]) afin d'identifier des gènes impliqués dans l'expression du vieillissement cutané (Elfakir et al. [2009], Latreille et al. [2009], Latreille et al. [2011], Ezzedine et al. [2012], Jdid et al. [2013]). Une approche gène candidat a tout d'abord été adoptée, puis en 2010 un projet de recherche Genome Wide Association Study (GWAS) a été mené en collaboration avec la chaire de Bioinformatique du Conservatoire National des Arts et Métiers (CNAM) dont le responsable scientifique du projet est le Professeur Jean-François Zagury (Le Clerc et al. [2012]). Les données supplémentaires nécessaires à cette étude spécifique ont été collectées sur un sous-échantillon de femmes de la cohorte SU.VI.MAX résidant en Ile de France sur la période automne/hiver 2002-2003. Une présentation de l'étude SU.VI.MAX est réalisée en l'annexe C.1. Ces travaux ont déjà donné lieu à des publications décrivant le matériel et les méthodes. Toutefois pour faciliter la compréhension du manuscrit la description de la population est reprise ci-après.

4.1.1 Population

Cette étude transversale a été conduite sur un échantillon de 570 femmes, âgées de 44 à 70 ans en 2002 vivant en région parisienne et ayant fourni un consentement éclairé (voir annexe C.1 pour plus d'informations). Les critères d'inclusion étaient l'absence de pathologies dermatologiques connues et l'absence de procédures esthétiques anti-âge au niveau du visage. Les participantes ont été invitées à suivre des consignes spécifiques de soins cutanés notamment l'interdiction d'appliquer des produits de nettoyage ou cosmétiques pour le visage pendant les 12 heures précédant la visite pour l'étude. Le jour de la visite, elles ont rempli un questionnaire auto-administré concernant leurs habitudes d'exposition au soleil au cours de la vie, questionnaire qui a permis de construire un score basé sur 5 items pour estimer l'intensité d'exposition au soleil sur la vie de chaque individu (Guinot et al. [2001]). Des informations générales, phénotypiques et médicales sur ces femmes ont été recueillies au cours d'un interrogatoire médical standardisé. Par la suite, trois photographies haute résolution (2008×3032 px) standardisées ont été prises pour chaque participante (une vue de face et deux de profil) avec un appareil photo numérique Kodak DSC 760 et un objectif 105 mm. L'appareil a été monté sur un monopode avec une chaise spécialement conçue pour permettre une normalisation de la position du sujet. Les conditions d'éclairage ont également été normalisées au moyen de deux lampes symétriques fournissant un spectre de lumière diurne continue, placées à 45° de chaque côté du visage. Chaque série de photographie a ensuite été examinée par un dermatologue qui a évalué la sévérité des différents signes de vieillissement au niveau du visage et un échantillon de sang a été prélevé pour les analyses génétiques.

Sur les 570 femmes participant à l'étude, 41 ont été exclues de l'analyse pour les raisons suivantes :

- 18 ont eu une intervention esthétique de rajeunissement (détectées lors de l'examen des photographies par le dermatologue) ;
- 10 étaient d'ascendance non caucasienne ;
- 1 exclue pour cause de quantité insuffisante d'ADN dans le prélèvement ;
- 12 exclues car leur ADN avait été endommagé lors des manipulations.

Les analyses génétiques ont donc été réalisées sur un total de 529 femmes.

4.1.2 Variables à expliquer : phénotypes analysés

Les photographies prises ont été examinées par un dermatologue et le photo-vieillissement a été apprécié cliniquement à l'aide d'une échelle ordinale photographique (figure 4.1, Larnier et al. [1994]). Sur cette échelle à 6 niveaux, chaque niveau est représenté par trois photographies de référence afin d'illustrer la diversité et la variété du photo-vieillissement : troubles pigmentaires, rides et relâchement.

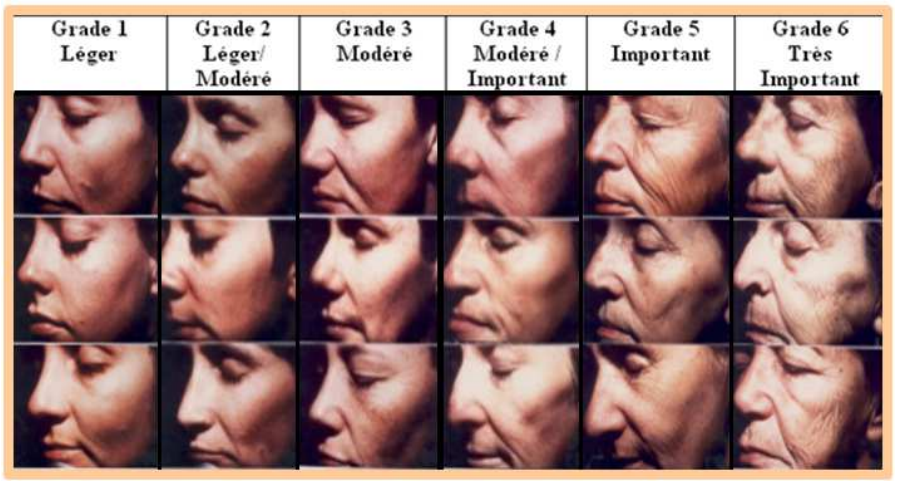

FIGURE 4.1 – Echelle photographique de photo-vieillissement à 6 niveaux (adaptée de Larnier et al. [1994]).

Outre la détermination de la sévérité du photo-vieillissement, une série de douze signes cliniques individuels de vieillissement cutané a été évaluée par le même dermatologue à l'aide d'échelles ordinales photographiques : lentigines [1] sur la joue, lentigines sur le front, poches sous les yeux, affaissement de l'ovale du visage, relâchement des paupières, sillon naso-génien, rides inter-sourcilières, rides de la patte d'oie, rides sous les yeux, rides fines sur les joues, rides d'expression joue et rides du contour de la bouche.

A partir de ces données, trois scores de phénotypes plus ciblés ont été calculés : un score de ride, un score de relâchement et un score de lentigines. Ces trois scores de sévérité

1. Les lentigines correspondent à une hyperpigmentation de la peau avec un contour très bien défini avec une taille variant de quelques millimètres à quelques centimètres de diamètres, de couleur allant de marron clair à marron foncé (Hodgson [1963]). Les lentigines apparaissent le plus souvent après 50 ans et une exposition chronique au soleil (Cario-Andre et al. [2004]).

ont été calculés grâce à une ACP et une régression linéaire (Johnson and Wichern [2002]). Les valeurs de chaque individu pour chacun des scores ont ensuite été transformées pour être compris entre 0 et 10. Ce travail ayant été réalisé en amont de la thèse, le détail du calcul des scores est présenté en annexe C.2.

Afin de valider les scores obtenus avec ACP, plusieurs méthodes ont été comparées durant cette thèse : ACM, procédure transreg du logiciel SAS®(Inc [1990]), procédure prinqual de SAS®, approche PLS. Le tableau 4.1 indique les coefficients de corrélation de Pearson entre les scores obtenus par les 5 méthodes utilisées, pour les scores de rides, relâchement et lentigines.

	ACP	ACM	proc transreg (SAS)	proc prinqual (SAS)	approche PLS
Score de rides					
ACP	**1.000**				
ACM	0.996	**1.000**			
proc transreg	0.981	0.976	**1.000**		
proc prinqual	0.996	0.999	0.976	**1.000**	
approche PLS	0.915	0.904	0.901	0.906	**1.000**
Score de relâchement					
ACP	**1.000**				
ACM	0.993	**1.000**			
proc transreg	0.994	0.996	**1.000**		
proc prinqual	0.993	0.998	0.997	**1.000**	
approche PLS	0.764	0.727	0.768	0.753	**1.000**
Score de lentigines					
ACP	**1.000**				
ACM	0.998	**1.000**			
proc transreg	0.977	0.976	**1.000**		
proc prinqual	1.000	0.998	0.977	**1.000**	
approche PLS	0.696	0.678	0.749	0.696	**1.000**

TABLE 4.1 – Coefficients de corrélation de Pearson entre les scores obtenus à l'aide des cinq méthodes de calcul pour les trois scores de vieillissement : ACP, ACM, proc transreg, proc prinqual, approche PLS.

Les scores calculés avec les différentes méthodes présentent une corrélation très forte (l'approche PLS restant la méthode la moins proche des autres). Les scores de départ calculés avec l'ACP pour les 3 scores de vieillissement seront donc conservés (considérés comme les trois premières variables à expliquer), tout comme celui évalué par le dermatologue pour le photo-vieillissement (considéré comme la quatrième variable à expliquer). Dans la suite des analyses, la variable photo-vieillissement sera transformée en variable dichotomique (léger à modéré, grades 1 à 3, versus modéré/important à très important, grades 4 à 6).

4.1.3 Ajustement des scores sur les covariables

Lorsque l'on teste l'association entre une variable et un phénotype, d'autres variables (covariables) peuvent avoir une influence sur cette association. Les covariables connues doivent être prises en compte lors des analyses.

Le vieillissement cutané est un phénomène complexe, reflet de deux phénomènes principaux qui se superposent et interagissent (voir annexe A) : le vieillissement intrinsèque considéré comme génétiquement déterminé, et le vieillissement extrinsèque lié aux facteurs environnementaux et comportementaux (Yaar and Gilchrest [2007]). Parmi les différents facteurs responsables du vieillissement cutané extrinsèque, l'exposition aux radiations Ultra-Violets (UV) et le tabagisme en sont les principaux. Mais d'autres facteurs comme le statut hormonal ou encore l'indice de masse corporelle peuvent également avoir une influence (Malvy et al. [2000]). Afin de tenir compte de la variabilité des scores liés à ces facteurs, chacun des scores a été régressé sur l'ensemble des covariables suivantes :

- Age : variable continue renseignée en années (entre 44 et 70 ans)
- Score d'exposition au soleil : variable continue définie comme un score compris entre 0 et 10 issu d'une combinaison linéaire de 5 variables (exposition volontaire au soleil, exposition du corps et/ou du visage, exposition durant les heures les plus chaudes de la journée, auto-évaluation de l'intensité de l'exposition au soleil tout au long de la vie et importance accordée au bronzage, Guinot et al. [2001])
- Indice de Masse Corporelle (IMC) : variable continue (entre 14 et 42 kg/m^2)
- Statut hormonal : variable catégorielle en 3 classes (non ménopausée/ménopausée sans Traitement Hormonal de Substitution/ménopausée avec Traitement Hormonal

de Substitution)

- Statut tabagique : variable catégorielle en 3 classes (non-fumeur/ancien fumeur/fumeur actuel)

Les résidus des scores ajustés sur ces facteurs ont ensuite été calculés pour les quatre phénotypes et seront les quatre variables que nous souhaiterons expliquer dans la suite des analyses. Pour faciliter la lecture et la compréhension du manuscrit, nous appellerons "phénotypes" les résidus des scores ajustés. Les scores de vieillissement cutané originaux ainsi que la variable photo-vieillissement en 6 grades ne seront plus considérés.

4.1.4 Variables explicatives : Single Nucleotide Polymorphisms (SNPs)

Les 529 femmes ont été génotypées avec la puce Illumina Infinium HumanOmni1-Quad contenant 1 140 419 marqueurs (SNPs). Les SNPs sont des variations ponctuelles d'un seul nucléotide (figure 4.2). Ils sont répartis sur l'ensemble du génome humain et constituent la forme la plus abondante de variations génétiques. Leur nombre est estimé à environ 10 millions et ils représentent ainsi plus de 90% des différences entre individus (voir Annexe B).

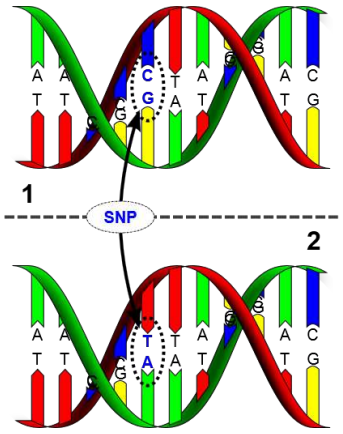

FIGURE 4.2 – Représentation du polymorphisme d'un nucléotide entre 2 individus. La molécule d'ADN de l'individu 1 diffère de celle de l'individu 2 par un seul nucléotide (polymorphisme C/T).

Seuls les SNPs ont été considérés dans la suite des analyses. Après contrôle qualité du génotypage, et des procédures de filtrage en plusieurs étapes réalisée sous PLINK (Purcell et al. [2007]), 795 063 SNPs sont retenus pour un total de 519 sujets. Le détail des procé-

dures de filtrage, du nombre de SNPs éliminés à chaque étape ainsi que l'explication de la suppression de certains individus dans les analyses sont donnés en annexe C.3.

Par ailleurs, pour corriger l'éventuelle stratification de notre échantillon au niveau intercontinental, les génotypes de tous les individus ont été analysés en utilisant le logiciel EIGENSTRAT issu de la suite EIGENSOFT basé sur une analyse en composantes principales qui permet de modéliser les différences ancestrales selon des axes continus de variation (Price et al. [2006]). Cette procédure a permis de mettre en évidence 18 individus atypiques qui ont été retirés pour la suite de l'analyse (le détail de la procédure de stratification est très intéressante et est décrite en annexe C.4). La base de données finale obtenue est donc composée de 501 individus et de 795 063 SNPs.

4.1.5 Codage des SNPs

Les SNPs sont en général bialléliques avec deux des quatre bases A (adénine), C (cytosine), T (thymine) et G (guanine). D'un point de vue statistique, un SNP peut être considéré comme une variable catégorielle avec trois modalités : si les allèles des SNPs sont, par exemple, l'adénine et la cytosine, les génotypes possibles sont "AA", "AC" et "CC". Si C est l'allèle majeur (c'est-à-dire l'allèle le plus fréquent) et A l'allèle mineur, alors "CC" est codé 0 (homozygote de référence), "AC" est codé 1 (hétérozygote) et enfin "AA" est codé 2 (homozygote variant). La base de données contient donc 795 063 variables catégorielles avec trois modalités (ou deux si l'une des modalités n'est pas présente dans la population). Ce codage en 0, 1, 2 est souvent rencontré mais peut être contesté d'un point de vue statistique et biologique lorsque ces variables sont traitées comme des variables quantitatives, c'est-à-dire numériques, dont les valeurs suivent une échelle discrète ou ordinale (Beaton et al. [2013]). En effet, il n'y a pas de notion d'ordre et/ou de proportionnalité entre les différentes modalités, c'est pourquoi nous avons envisagé un autre codage pour l'ensemble des analyses. Le principe général va être de coder une variable qualitative à k modalités (ici 2 ou 3 modalités pour un SNP) dans un modèle de régression par k variables binaires Z_i (dummy variable). La variable Z_i vaut 1 si elle est au niveau i, 0 sinon. Nous allons donc considérer un SNP comme étant l'ensemble de deux ou trois variables binaires (voir figure 4.3).

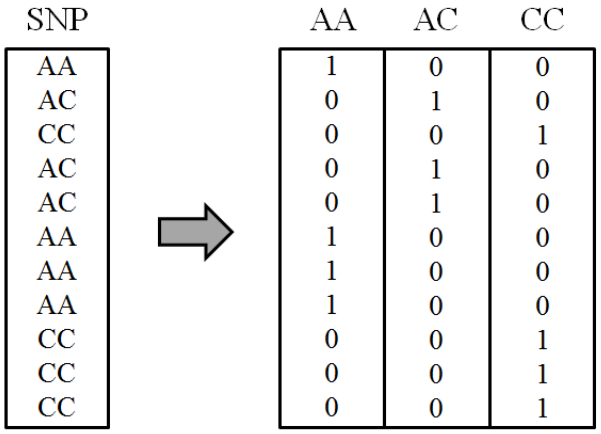

FIGURE 4.3 – Recodage des variables SNPs en variables binaires.

4.1.6 Pré-sélection des SNPs

Les SNPs peuvent se retrouver au sein de régions codantes de gènes (exon), de régions non codantes de gènes (intron), ou de régions intergéniques, entre les gènes. Dans la suite des analyses, nous avons choisi de nous concentrer sur les SNPs intragéniques (situés dans des gènes) afin de pouvoir localiser directement les gènes potentiellement impliqués dans le processus de vieillissement cutané précoce.

Sur les 795 063 SNPs conservés après le contrôle qualité, 362 223 SNPs se sont avérés être localisés dans un total de 15 198 gènes. Nous nous focaliserons sur ces SNPs pour l'ensemble des analyses. Une pré-sélection de SNPs a été effectuée grâce à la méthode de régression elastic net, afin de travailler sur un ensemble de données plus restreint. L'intérêt de l'approche multivariée, contrairement aux analyses qui ont été réalisées SNP par SNP (batterie de tests univariés), est que l'on considère l'ensemble des SNPs dans le modèle ainsi que les corrélations entre ces SNPs.

4.1.6.1 Résultats de la pré-sélection

La régression régularisée elastic net est effectuée en considérant les données génétiques (SNPs) comme variables explicatives et les différents phénotypes décrits section 4.1.3 et pris individuellement comme variable réponse. La figure 4.4 permet de visualiser l'évolution des

coefficients de chaque SNP en fonction du paramètre de régularisation λ pour chacun des quatre phénotypes. La valeur optimale de λ (qui minimise l'erreur quadratique moyenne) n'a pas pu être estimée à l'issue d'une procédure de validation croisée 10-fold habituelle. Nous avons donc choisi de prendre un λ conservant un nombre acceptable de SNPs selon les biologistes. Le paramètre de régularisation a été fixé à 0,014 afin d'obtenir entre 600 et 700

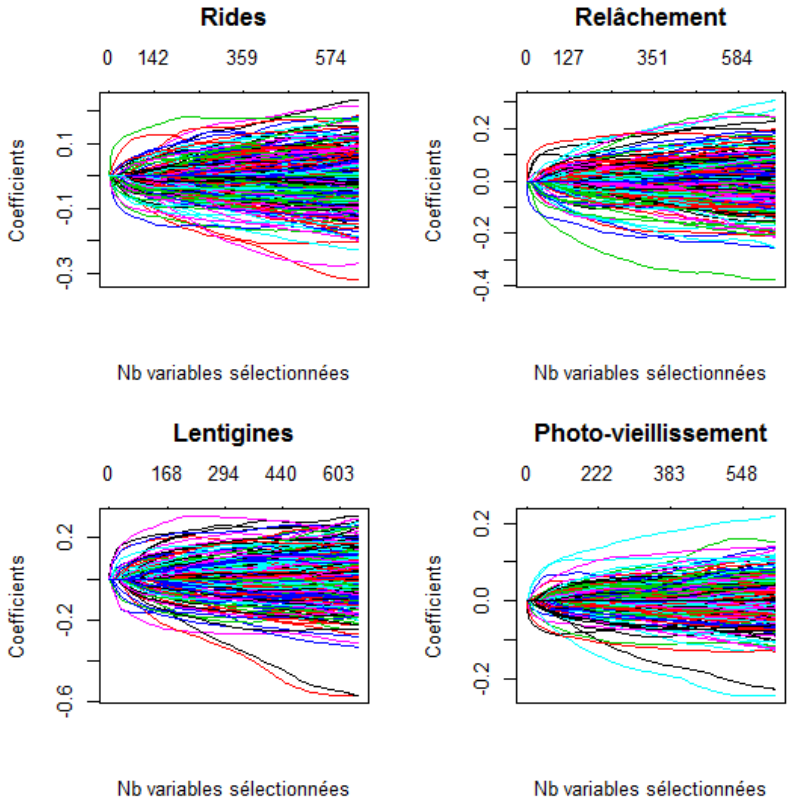

FIGURE 4.4 – Données de SNPs : Coefficients de chaque SNP en fonction du paramètre de régularisation pour chacun des phénotypes dans la régression elastic net.

SNPs en moyenne pour chaque phénotype. Ainsi, 658 SNPs potentiellement liés à des rides plus sévères ont été sélectionnés, 700 SNPs pour le relâchement, 670 SNPs pour les lentigines et enfin 640 SNPs pour le photo-vieillissement global. La sélection de SNPs réalisée par l'approche elastic net a permis de sélectionner des SNPs différents de ceux trouvés statistiquement significatifs avec une approche univariée. Parmi les SNPs sélectionnés, aucun SNP n'est en commun entre les quatre phénotypes, ce qui s'avère normal étant donné que les phénotypes correspondent à des phénomènes biologiques bien différents et n'impliquent

pas forcément les mêmes SNPs ou gènes. Nous étudions donc quatre phénomènes différents et il sera utile de présenter les résultats pour chacun d'entre eux séparément par la suite.

Les SNPs sélectionnés par elastic net ne sont pas forcément ceux dont la valeur p (en anglais, "p-value") dans l'approche univariée était la plus faible. Les informations apportées par les approches multivariées sont plus riches et plus informatives. La figure 4.5 présente la distribution des valeurs p (plus exactement le -$\log_{10}$ des valeurs p par souci de lisibilité) des SNPs dans l'approche univariée (en noir) et celles des SNPs conservés après la pré-sélection par elastic net (en bordeaux) pour le phénotype "rides". Les SNPs sélectionnés par elastic net font partie de ceux ayant une valeur p les plus faibles (moins de SNPs dont la valeur p est supérieure à 10^{-2} sont conservés) mais les SNPs dont la valeur p était la plus petite ($< 10^{-5}$) n'ont pas été sélectionnés. La même chose est observée pour les autres phénotypes (figures non montrées). Le fait de considérer l'ensemble des SNPs dans le modèle permet alors d'obtenir de nouvelles pistes que nous allons explorer tout au long de ce chapitre.

Chaque SNP étant contenu dans un chromosome, il est facile de visualiser les chromosomes potentiellement impliqués dans l'expression de signes de vieillissement plus sévères. Les informations sur les gènes et chromosomes associés à chaque SNP sont fournies par la base de données : Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al. [2001]). Deux approches sont présentées. Dans la première nous nous intéressons aux chromosomes dont les SNPs qu'ils contiennent sont les plus fréquemment sélectionnés peu importe la valeur du coefficient de régression, c'est donc une approche globale. Dans la deuxième, nous considérons les chromosomes contenant les SNPs dont le signal dans la sélection elastic net est le plus fort et dans ce cas c'est une approche centrée sur des SNPs en particulier.

Pour la première approche, nous considérons les figures 4.6, 4.7, 4.8, et 4.9. Elles présentent, à gauche, la fréquence de répartition des SNPs dans les chromosomes correspondants en % dans la base contenant les SNPs intragéniques avant pré-sélection (en noir) et après pré-sélection par elastic net (en bordeaux), et à droite, la différence de fréquence en % entre ceux pré-sélectionnés par elastic net et ceux avant pré-sélection. Cela permet de visualiser très rapidement les chromosomes qui sont plus fréquemment sélectionnés (différence positive des fréquences).

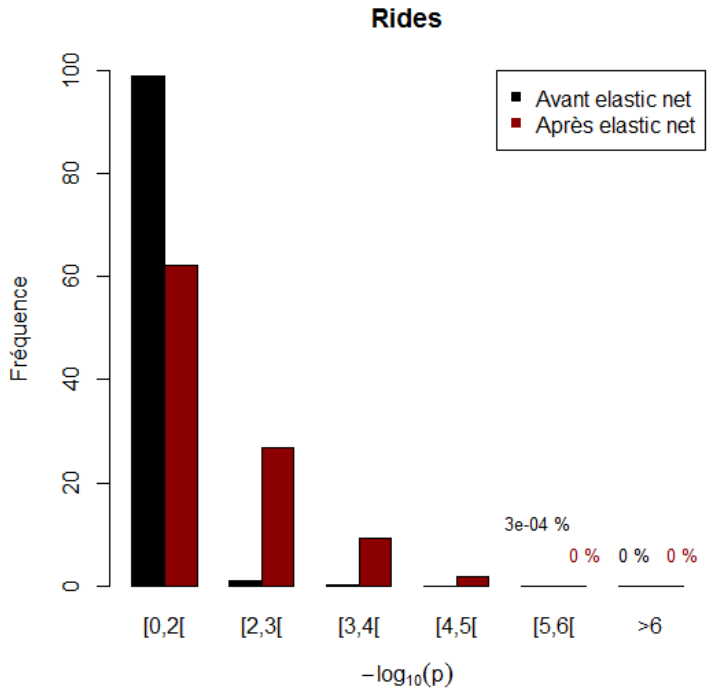

FIGURE 4.5 – Distribution du -$\log_{10}$ des valeurs p des SNPs dans l'approche univariée (en noir) et celles des SNPs conservés après la pré-sélection elastic net (en bordeaux) pour le phénotype "rides".

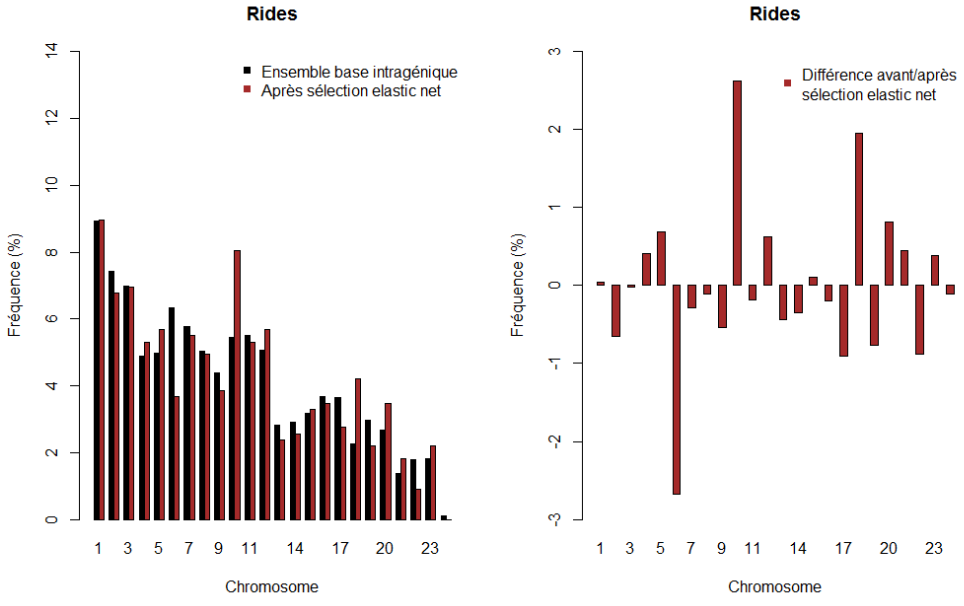

FIGURE 4.6 – Fréquence de répartition des SNPs dans les chromosomes pour le phénotype "rides" avant (en noir) et après sélection (en rouge) elastic net (graphe de gauche) et différence de fréquence de répartition des SNPs avant/après sélection elastic net (graphe de droite).

Sur ces graphiques on voit apparaitre un "chromosome 25" qui correspond en fait à l'ADN mitochondrial. En effet, en plus des 23 paires de chromosomes contenus dans le noyau, les cellules humaines possèdent de l'ADN contenu dans les mitochondries. La mitochondrie est l'unique générateur d'énergie de nos cellules. Situées dans le cytoplasme de chaque cellule, on peut comparer les mitochondries à des "piles" chargées de produire, stocker et distribuer de l'énergie nécessaire à la cellule. L'ADN contenu dans les mitochondries est transmis essentiellement par la mère à 99%, car les mitochondries sont surtout transmises par le cytoplasme de l'ovocyte. Comme cet ADN n'est pas soumis aux lois génétiques de la reproduction sexuée, il n'est pas ou peu soumis aux recombinaisons génétiques. Cependant le taux de mutation reste relativement élevé. C'est d'ailleurs pour ces raisons que cet ADN a été privilégié pour l'étude des grandes migrations humaines. Par ailleurs la présence d'un ADN propre dans les mitochondries a ouvert une nouvelle voie de recherche sur le vieillissement (Wallace [1997]). Il est donc intéressant de conserver ces informations dans nos analyses.

La figure 4.6 montre que la répartition des SNPs sélectionnés par elastic net est semblable à celle avant sélection à l'exception des chromosomes 10 et 18 en particulier dont la fréquence de SNPs sélectionnés est plus élevé que ce à quoi on pouvait s'attendre (voir graphe de droite). Ces chromosomes pourraient contenir plus de SNPs ou de combinaisons de SNPs potentiellement liés à l'expression de rides que les autres.

Pour les phénotypes "relâchement" et lentigines les résultats sont moins homogènes (figures 4.7 et 4.8, respectivement). Il y a de nombreuses différences dans la répartition des SNPs dans les chromosomes après sélection. Le chromosome 3 en particulier pour le phénotype "relâchement" (graphe de droite) et le chromosome 6 pour phénotype "lentigines" possèdent plus de SNPs sélectionnés proportionnellement à avant la sélection. Par ailleurs, il y a une proportion de SNPs plus importante dans le chromosome 11 après la sélection pour le phénotype "photo-vieillissement" (figure 4.9). Il pourra donc être intéressant d'étudier plus en détail les SNPs sélectionnés sur ce chromosome et leur position chromosomique afin d'identifier des effets possibles de groupement de SNPs.

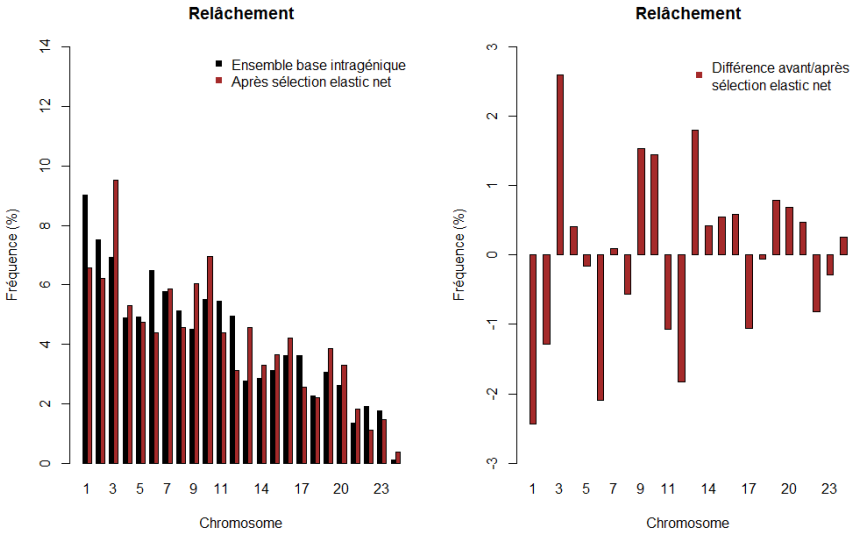

FIGURE 4.7 – Fréquence de répartition des SNPs dans les chromosomes pour le phénotype "relâchement" avant (en noir) et après sélection (en rouge) elastic net (graphe de gauche) et différence de fréquence de répartition des SNPs avant/après sélection elastic net (graphe de droite).

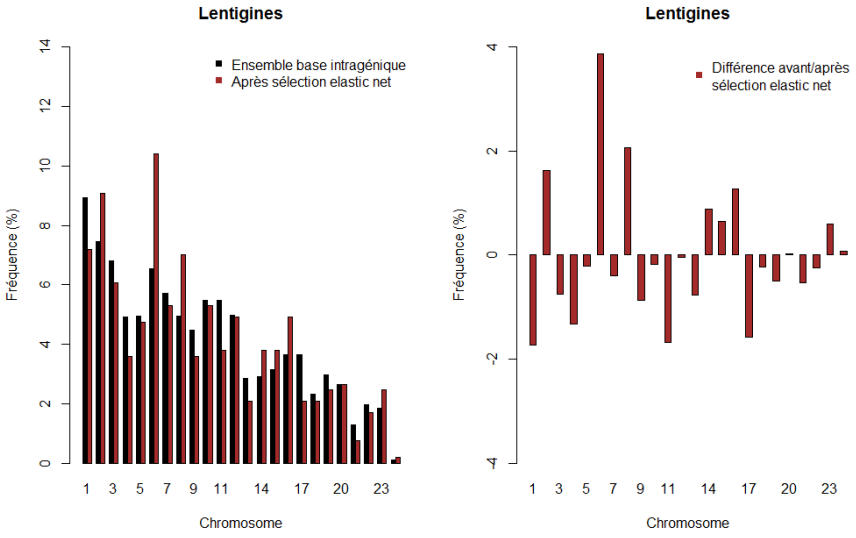

FIGURE 4.8 – Fréquence de répartition des SNPs dans les chromosomes pour le phénotype "lentigines" avant (en noir) et après sélection (en rouge) elastic net (graphe de gauche) et différence de fréquence de répartition des SNPs avant/après sélection elastic net (graphe de droite).

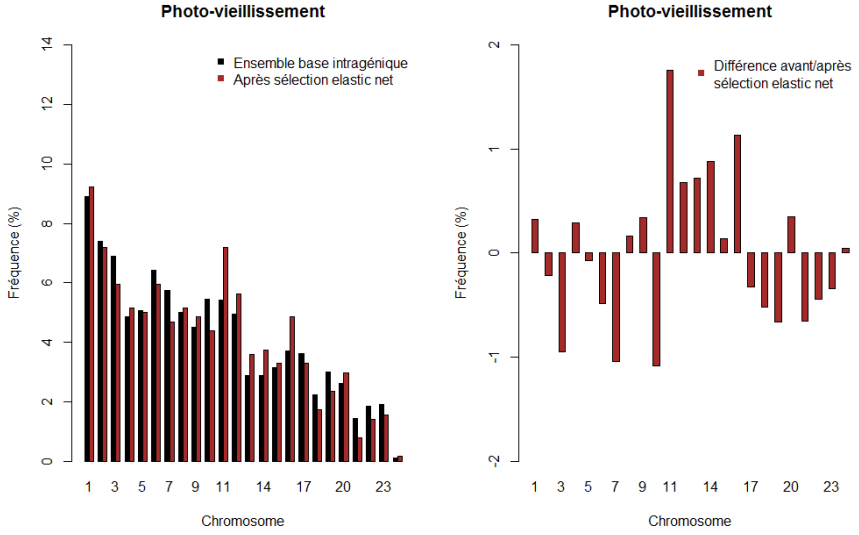

FIGURE 4.9 – Fréquence de répartition des SNPs dans les chromosomes pour le phénotype "photo-vieillissement" avant (en noir) et après sélection (en rouge) elastic net (graphe de gauche) et différence de fréquence de répartition des SNPs avant/après sélection elastic net (graphe de droite).

Cependant, les SNPs ayant les coefficients de régression elastic net les plus élevés ne sont pas forcément situés dans les chromosomes cités précédemment. Pour la deuxième approche, nous nous concentrerons sur la figure 4.10. Elle présente la valeur des coefficients elastic net des SNPs sélectionnés en fonction de leurs coordonnées génomiques pour les quatre phénotypes. Chacun des SNPs est contenu dans un gène lui-même situé sur un chromosome. Les coordonnées génomiques des SNPs dans les chromosomes sont affichées le long de l'axe des abscisses et la valeur du coefficient elastic net est affichée sur l'axe des ordonnées. Chaque point représente un SNP. Des points d'une même couleur et qui sont proches sont situés sur le même chromosome. Cette représentation est équivalente à celle des Manhattan plot souvent utilisés en génomique.

Pour le phénotype "rides" on remarque que les SNPs présentant les coefficients les plus élevés sont situés dans plusieurs chromosomes et plus particulièrement les 9 et 16, pour le phénotype "relâchement" deux chromosomes se détachent des autres : les chromosomes 2 et 7, pour le phénotype "lentigines" on observe le même phénomène de dispersion que pour le phénotype "rides" néanmoins les chromosomes 6 et 12 sont ceux contenant les SNPs

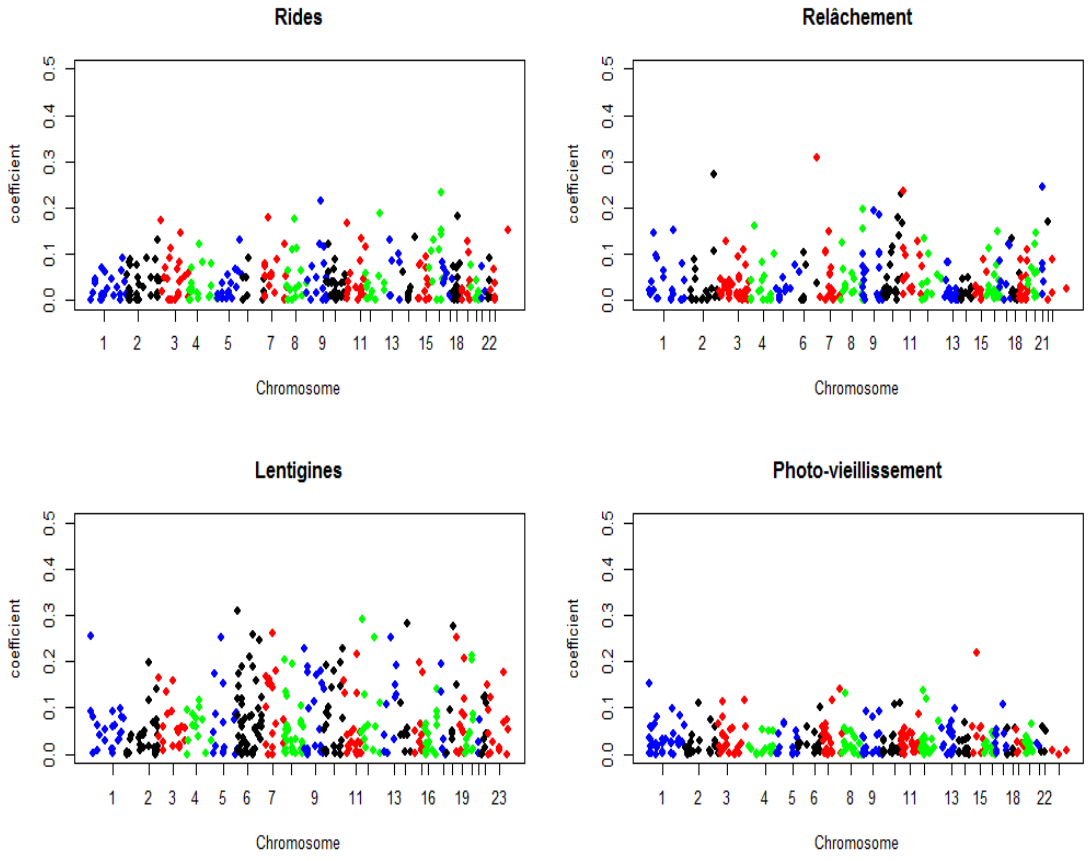

FIGURE 4.10 – Représentation des coefficients des SNPs sélectionnés par elastic net en fonction de leurs coordonnées génomiques pour les quatre phénotypes.

dont les coefficients sont les plus élevés et enfin pour le phénotype "photo-vieillissement" un chromosome en particulier possède un SNP avec un coefficient plus important que les autres : le chromosome 15. Les SNPs présentant un coefficient élevé ne sont donc pas contenus dans les mêmes chromosomes que ceux identifiés dans la première approche.

Les deux approches (la proportion de SNPs sélectionnés dans les chromosomes et la répartition des SNPs en fonction de la valeur de leur coefficient dans la régression elastic net) ouvrent donc deux voies d'investigation différentes. La première est intéressante car permet une approche globale en se focalisant sur des chromosomes dont les SNPs qu'ils contiennent ne correspondent pas aux signaux les plus élevés mais sont pourtant plus souvent sélectionnés que ce à quoi on pouvait s'attendre. Des recherches plus fines sont

actuellement en cours par les biologistes impliqués sur ce projet concernant la région chromosomique (c'est à dire la partie du chromosome) dans laquelle se trouvent ces SNPs pour tenter de comprendre ou d'expliquer des liens potentiels avec les phénotypes considérés. La deuxième voie, quant à elle est basée sur des SNPs en particulier. Elle permet de focaliser l'attention sur des SNPs isolés dont le signal est important. Des recherches de "pathways" à partir de ces SNPs, c'est-à-dire de voies de signalisations biologiques, sont elles aussi en cours d'investigation par les experts du domaine car elles pourraient permettre de signaler des mécanismes moléculaires potentiellement liés au processus de vieillissement cutané du phénotype concerné.

4.1.6.2 Pertinence de la pré-sélection

Afin d'étudier les liens entre chacun des SNPs sélectionnés par la régression elastic net et les liens entre ceux-ci et les variables réponses, une ACM a été effectuée pour chacun des quatre phénotypes. La figure 4.11 représente la répartition des individus sur le premier plan factoriel de l'ACM réalisée sur les SNPs pré-sélectionnés pour le phénotype "rides". Les triangles (en rouge) représentent les individus dont les résidus du score ajusté de rides sont élevés, et les points (en bleu) représentent les individus dont les résidus du score ajusté de rides sont faibles. On remarque que le premier axe discrimine parfaitement les individus en fonction du phénotype "rides". Ce qui signifie que les SNPs sélectionnés par elastic net sont pertinents et permettent de discriminer correctement les individus. On observe exactement la même chose pour les trois autres phénotypes (figures non montrées).

L'étude des SNPs les plus contributifs du premier axe peut nous donner une information sur les SNPs potentiellement liés à un score élevé (l'information discriminante étant situé sur le premier axe) et nous permettre d'identifier les SNPs les plus contributifs à cette discrimination des individus en fonction du score. La figure 4.12 présente les SNPs les plus contributifs du premier axe pour chacune des quatre phénotypes. En rouge apparaissent les gènes dans lesquels sont contenus les SNPs associés. Par souci de confidentialité des résultats, les noms des SNPs et des gènes n'ayant pas encore fait l'objet d'une publication sont recodés. Les trois SNPs les plus contributifs du premier axe pour le phénotype "rides" appartiennent tous au gène GENEri1, le gène GENEre3 apparait à trois reprise parmi

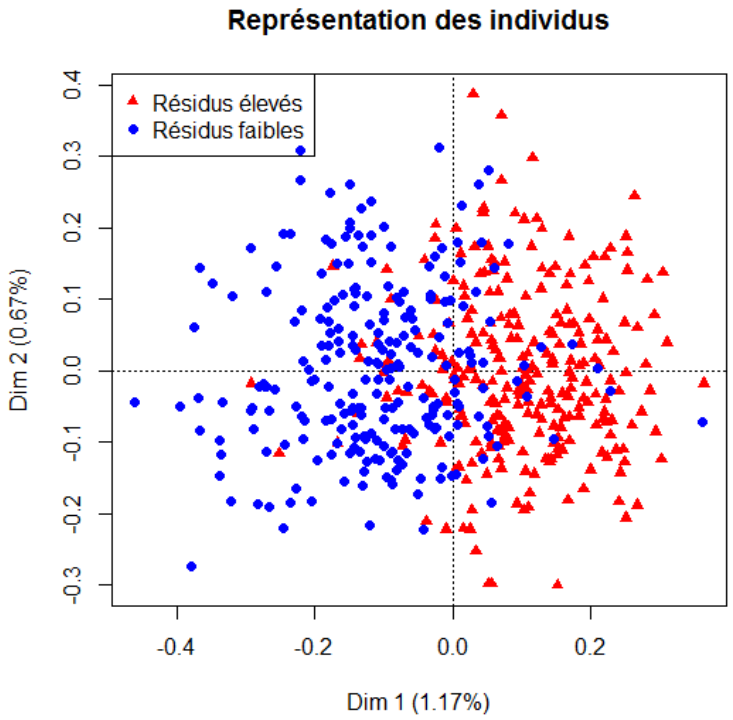

FIGURE 4.11 – Répartition des individus sur le premier plan factoriel de l'ACM des SNPs pré-sélectionnés par elastic net pour le phénotype "rides".

les plus contributif de l'axe 1 pour le phénotype "relâchement", les gènes GENEle1 et GENEle4 apparaissent à plusieurs reprises pour le phénotype "lentigines" et enfin ce qui est très intéressant de remarquer pour le phénotype "photo-vieillissement" est que, parmi les 10 SNPs les plus contributifs de l'axe 1, la moitié est contenue dans le gène GENEpv1 et l'autre dans le gène SynTaxin Binding Protein 5-Like ($STXBP5L$). Ce gène avait passé le seuil de Bonferroni lors de l'analyse univariée réalisée dans Le Clerc et al. [2012] ce qui souligne la pertinence de cette pré-sélection. Des recherches bibliographiques sur ces gènes sont en cours par les biologistes. Ils s'avèrent être impliqués dans la peau et liés à des processus biologiques très intéressant pour notre problématique.

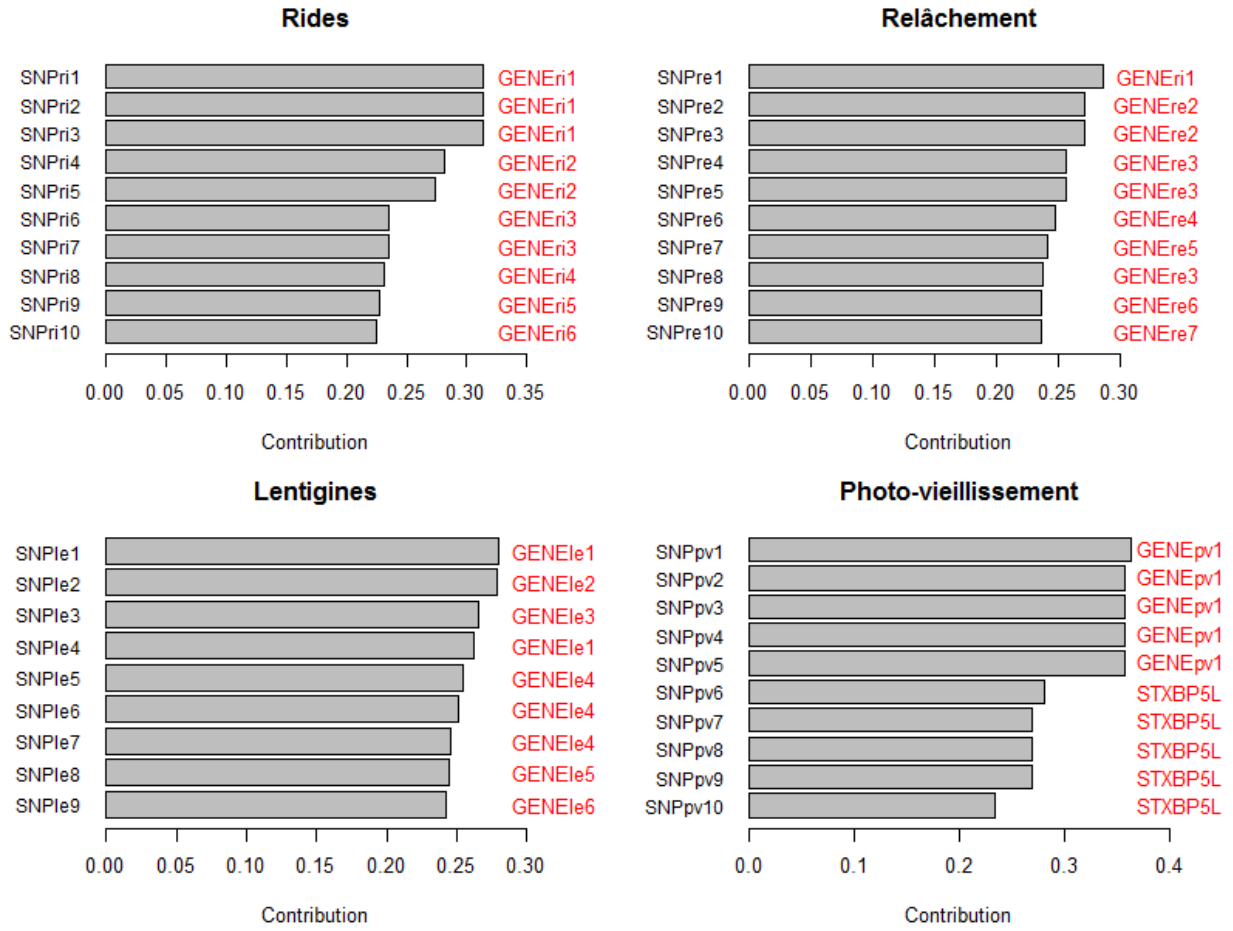

FIGURE 4.12 – SNPs pré-sélectionnés par elastic net les plus contributifs du premier axe de l'ACM pour les quatre phénotypes.

Malgré la pré-sélection le nombre de variables restant est encore trop élevé et les interprétations restent difficiles car de nombreuses voies d'investigations sont possibles. Pour y remédier nous aurons recours à la nouvelle méthode développée au cours de cette thèse : l'ACM sparse. Explicitée dans la section 2.3, elle permet de sélectionner des variables sur chacun des axes de l'ACM et d'interpréter plus facilement les résultats. Les résultats de l'application de cette méthode sont présentés dans la section suivante.

4.2 Approche multiblocs non supervisée : ACM Sparse

Afin de mettre en évidence les relations entre les modalités des différentes variables ainsi qu'entre les individus et faciliter l'interprétation des résultats une ACM sparse a été réalisée pour chacun des phénotypes sur les SNPs pré-sélectionnés par elastic net.

4.2.1 ACM sparse sur données de SNPs

Les résultats de l'ACM sparse sont présentés sur les quatre premiers axes. L'ACM réalisée dans la section 4.1.6.2 a montré sur le premier axe l'existance de deux groupes d'individus bien distincts : un groupe contenant les individus dont la valeur des résidus des scores ajustés de vieillissement cutané est élevé, et un groupe contenant les individus avec une valeur des résidus des scores ajustés de vieillissement cutané faible. Nous nous focaliserons alors sur les SNPs sélectionnés sur ce premier axe dans l'ACM sparse.

L'ACM sparse a été réalisée sur l'ensemble des SNPs pré-sélectionnés par régression elastic net. Dans un premier temps, elle a été effectuée pour plusieurs valeurs de λ comprises entre 0 et 0,01. Une détermination du λ optimal aurait pu être réalisée par validation croisée, cependant les temps de calculs nécessaires étant trop longs, nous avons choisi de déterminer le λ par une approche "ad-hoc". Cette approche consiste à éliminer un maximum de variables sur chacun des axes tout en limitant la perte du pourcentage d'inertie cumulé par rapport à la valeur de départ ($\lambda = 0$). Le but est donc de trouver un compromis entre le nombre de variables sélectionnées et la perte du pourcentage d'inertie par une analyse graphique. Le pourcentage d'inertie est représenté ici uniquement de manière informative étant donné qu'il n'a pas le même sens qu'en ACP. La valeur de λ peut alors être choisie en fonction du nombre de variables que l'utilisateur souhaite conserver (si cela est possible, le nombre de SNPs ou gènes est issu du choix ou de l'objectif du biologiste). Lorsque $\lambda = 0$, les résultats de l'ACM sparse sont les mêmes que ceux de l'ACM (car aucune régularisation effectuée). Dans ce cas, le pourcentage d'inertie cumulé obtenu en ACM sparse correspond à celui de l'ACM et le nombre de variables sélectionnées correspond au nombre de variables au départ.

Pour le phénotype "rides" le tableau de données est donc de dimension 501 individus par 658 SNPs. La figure 4.13 présente l'évolution du nombre de variables sélectionnées sur les quatre premiers axes en fonction de la valeur du paramètre de régularisation λ, et la figure 4.14 présente l'évolution du pourcentage d'inertie cumulé pour ces mêmes valeurs de λ. Les valeurs obtenues par ACM sont signalées en pointillés sur ces deux figures. Si l'on se concentre sur le premier axe, on remarque que pour $\lambda = 0,004$ le % inertie vaut 0,47% et 437 variables (soit 145 SNPs) sont sélectionnées. Pour $\lambda = 0,005$, seules 156 variables (soit 52 SNPs) sont sélectionnées et le % inertie vaut 0,40%. Étant donné la faible différence de pourcentage de variance obtenu entre les deux valeurs de λ, il est préférable de choisir la valeur de λ pour laquelle il y a le moins de variables conservées, soit $\lambda = 0,005$, pour simplifier l'interprétation des résultats.

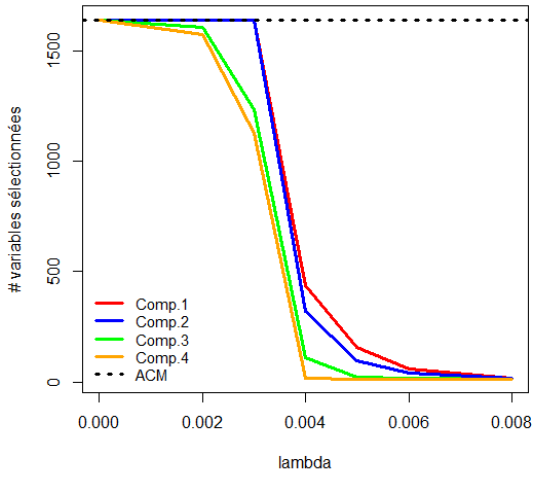

FIGURE 4.13 – Evolution du nombre de variables sélectionnées par ACM sparse en fonction du paramètre λ pour le phénotype "rides".

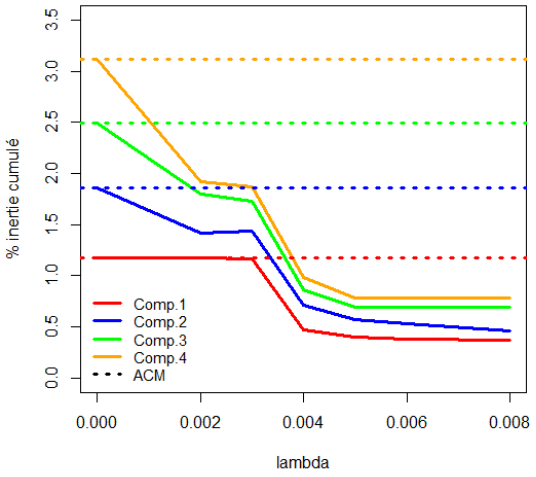

FIGURE 4.14 – Evolution du pourcentage d'inertie cumulé avec ACM sparse en fonction du paramètre λ pour le phénotype "rides".

Les figures pour les autres phénotypes ne sont pas montrées ici car elles sont très similaires à celles obtenues pour le phénotype "rides". Le λ choisi est égal à 0,005 également. Pour le phénotype "relâchement", le tableau de données est de dimension 501 individus par 700 SNPs. Après ACM sparse 225 variables (soit 75 SNPs) sont sélectionnées. Le tableau

de données pour le phénotype "lentigines" est de dimension 501 individus par 670 SNPs. L'ACM sparse permet de sélectionner 174 variables (soit 58 SNPs). Enfin dans le cas du phénotype "photo-vieillissement" le tableau de données est de dimension 501 individus $\times$ 640 SNPs et 96 variables (soit 32 SNPs) sont sélectionnées par ACM sparse.

La robustesse des sélections par l'ACM sparse a ensuite été testée par bootstrap avec un λ choisi par approche "ad-hoc" pour les quatre phénotypes valant 0,005.

4.2.2 Robustesse des sélections : bootstrap

Afin d'obtenir une sélection stable de variables, un bootstrap a été réalisé. Ici encore une validation croisée aurait pu être réalisée afin de déterminer le lambda optimal. Cependant par souci de complexité algorithmique et de temps de calcul, l'approche par bootstrap a été préférée. Le bootstrap a été réalisé pour $B = 100$ itérations pour un lambda égal à 0,005 et ce pour chacun des quatre phénotypes. Lors de chaque itération, les variables sélectionnées sont comptabilisées. On calcule le pourcentage de fois où chaque variable a été sélectionnée au cours des 100 itérations. D'autres méthodes similaires à celle-ci sont également envisageables (Bach [2008], Meinshausen and Bühlmann [2010]). La figure 4.15 présente la distribution des fréquences de sélection des modalités des variables au cours des 100 itérations pour chacun des quatre phénotypes. En rouge apparait le 3ème quartile. Nous avons choisi de sélectionner les variables dont la fréquence des modalités sélectionnées par bootstrap est supérieure à ce quartile. Nous aurions pu également garder les variables sélectionnées dans plus de 50% des cas (médiane) mais le choix du 3ème quartile nous permet d'obtenir un nombre plus restreint de SNPs ce qui simplifie l'interprétation des résultats (environ 150 SNPs conservés en moyenne). Par ailleurs, étant donnée la parfaite discrimination des individus obtenue avec l'ACM sur le premier axe, nous nous concentrerons sur les SNPs conservés par l'ACM sparse sur le premier axe.

Le bootstrap a permis de conserver 138 SNPs sur le premier axe parmi les 658 pré-sélectionnés pour le phénotype "rides", 155 SNPs parmi les 700 pré-sélectionnés pour le phénotype "relâchement", 145 SNPs parmi les 670 pré-sélectionnés pour le phénotype "lentigines" et enfin 165 SNPs sur le premier axe parmi les 640 pré-sélectionnés pour le phénotype "photo-vieillissement". Cette étude de stabilité aurait également pu être réalisée lors

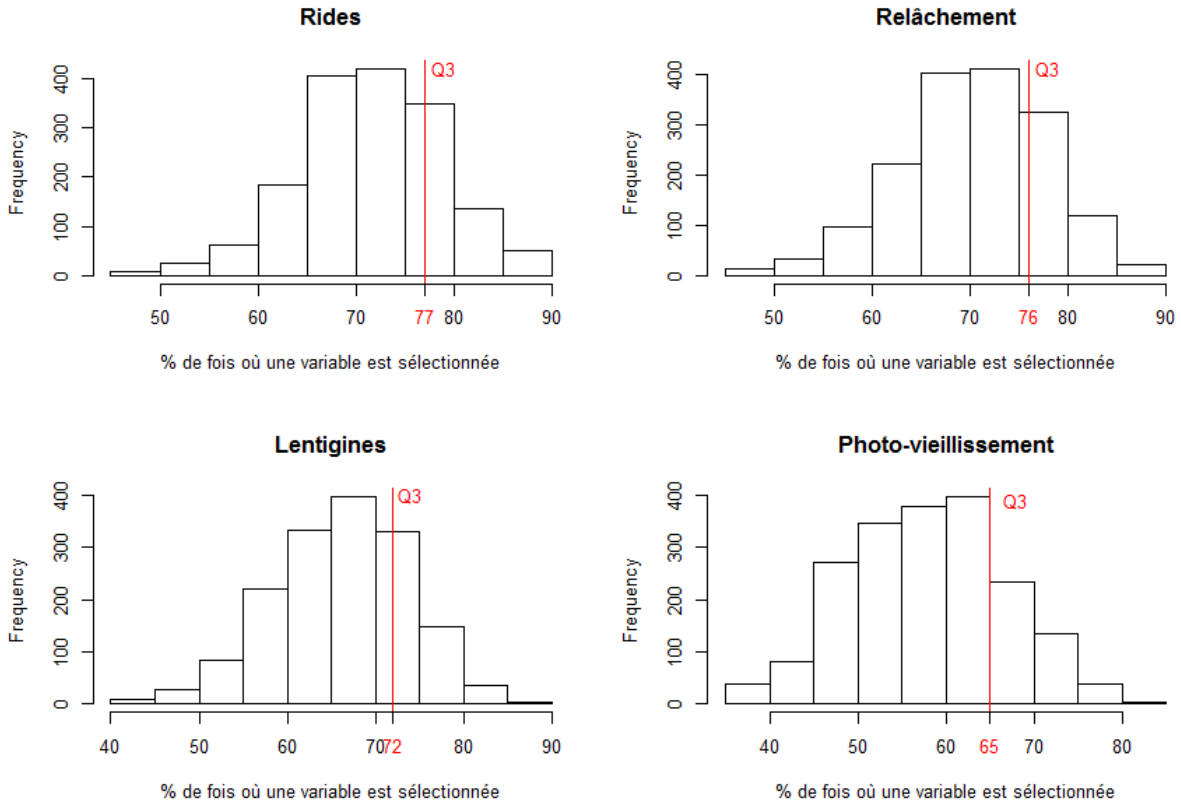

FIGURE 4.15 – Distribution du pourcentage de fois où une variable est sélectionnée par ACM sparse à la suite d'un bootstrap à 100 itérations.

de la pré-sélection par elastic net afin de donner plus de poids à la sélection réalisée.

Le tableau 4.2 est un tableau comparatif des "loadings" des quatre premiers SNPs de l'échantillon pour le phénotype "rides" obtenus avec l'ACM et l'ACM sparse pour un λ égal à 0,005. En ACM, le paramètre de régularisation étant nul ($\lambda = 0$), aucun SNP n'est éliminé. Les "loadings" sont tous non nuls. En revanche en ACM sparse le paramètre λ ($\lambda = 0,005$) permet la sélection de SNPs sur chacun des axes en fixant l'ensemble des "loadings" des modalités d'une variable à zéro, ou non (coefficients en rouge). La sélection se fait donc SNPs par SNPs en supprimant les blocs entiers des modalités correspondantes. La sélection est donc réalisée blocs de modalités par blocs de modalités. Grâce à cette sélection par ACM sparse, l'interprétation des résultats est simplifiée car le nombre de SNPs sélectionnés sur chaque axe est divisé par plus de 10.

TABLE 4.2 – Données SNPs : "Loadings" et inertie obtenus avec ACM et ACM sparse sur les quatre premières composantes pour les quatre premiers SNPs pré-sélectionnés pour le phénotype "rides".

	ACM				**ACM sparse**			
Variable	CP1	CP 2	CP 3	CP 4	CP1	CP 2	CP 3	CP 4
SNPri1.CC	0,277	0,258	-0,159	0,189	0,133	-0,125	**0,000**	**0,000**
SNPri1.CG	-0,108	-0,064	-0,053	-0,097	-0,080	0,007	**0,000**	**0,000**
SNPri1.GG	-0,025	-0,087	0,014	0,021	-0,015	0,012	**0,000**	**0,000**
SNPri2.AA	0,115	-0,025	0,636	-0,731	-0,192	**0,000**	-0,403	0,423
SNPri2.AG	0,072	-0,049	-0,194	0,520	-0,129	**0,000**	-0,124	0,323
SNPri2.GG	-0,351	0,125	-0,368	-0,360	0,235	**0,000**	0,240	-0,240
SNPri3.AA	0,191	-0,151	-0,031	0,016	-0,044	0,019	**0,000**	**0,000**
SNPri3.AG	-0,173	0,204	0,033	0,050	0,032	-0,025	**0,000**	**0,000**
SNPri3.GG	-0,294	-0,133	0,023	-0,381	-0,106	0,009	**0,000**	**0,000**
SNPri4.AA	0,197	0,490	0,499	-0,161	**0,000**	0,280	0,292	**0,000**
SNPri4.AT	0,176	-0,064	0,115	0,010	**0,000**	-0,082	0,102	**0,000**
SNPri4.TT	-0,157	-0,012	-0,148	0,012	**0,000**	0,002	0,121	**0,000**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Nb modalités sélectionnées	1638	1638	1638	1638	156	96	21	12
cumulé	1,18	1,85	2,49	3,12	0,40	0,57	0,69	0,78

4.2.3 Pertinence de la sélection

Une ACM a été réalisée sur les SNPs sélectionnés sur le premier axe par ACM sparse afin de vérifier que la discrimination des individus que l'on observait sur le premier axe avant sélection est toujours présente après sélection par ACM sparse. La représentation des individus sur le premier plan est présentée dans la figure 4.16 pour chacun des quatre phénotypes. On remarque que le premier axe discrimine encore parfaitement les individus en fonction des phénotypes (en marron, un résidu du score ajusté élevé et en orange, un résidu du score ajusté faible). Cela signifie que les SNPs sélectionnés expliquent encore la différence de sévérité des résidus des scores ajustés observée entre les individus. Ceci est

très intéressant car cela signifie que la sélection des SNPs par ACM sparse est pertinente et nous pouvons alors explorer plus en détails les SNPs sélectionnés, leur appartenance aux gènes, ainsi qu'aux chromosomes afin de tenter de trouver une interprétation biologique à ces résultats.

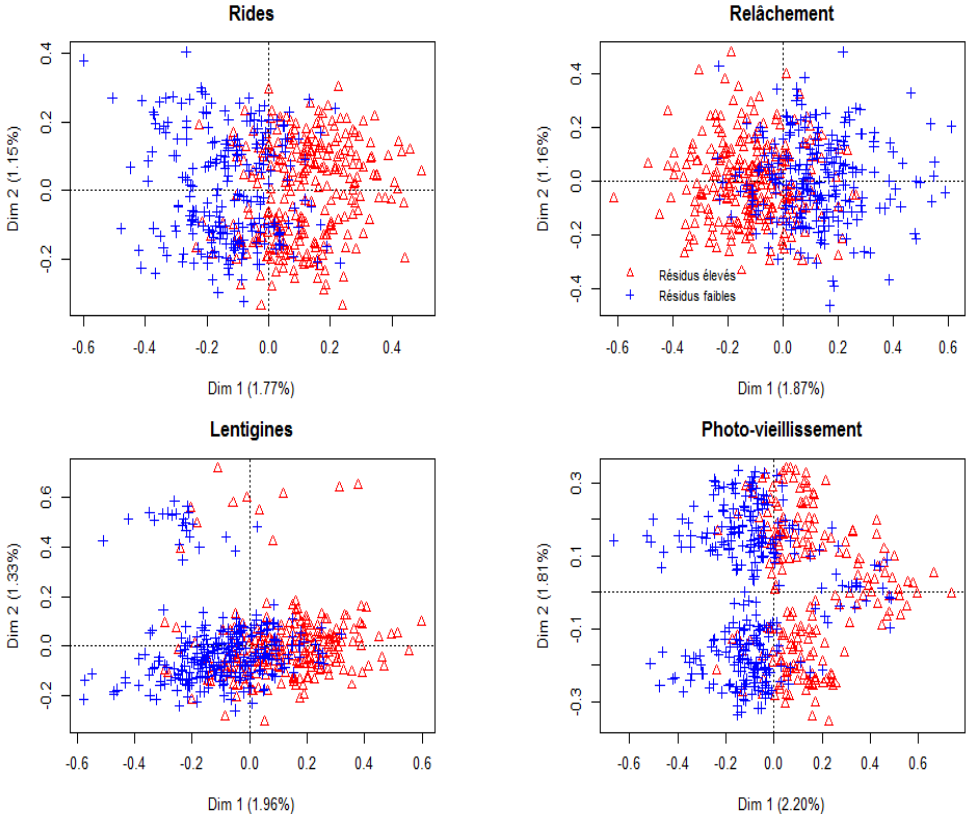

FIGURE 4.16 – Représentation des individus sur le premier plan de l'ACM des SNPs sélectionnés par ACM sparse pour chacun des phénotypes (en rouge, les individus avec un score ajusté élevé, en bleu, ceux avec un score ajusté faible).

Si comme dans la section précédente on s'intéresse aux chromosomes dont les SNPs qu'ils contiennent sont les plus fréquemment sélectionnés on remarque que pour les quatre phénotypes la distribution des SNPs dans les chromosomes est presque la même avant et après sélection par ACM sparse. L'ACM sparse permet donc l'élimination de SNPs tout en conservant l'information principale contenue avant la sélection. Les résultats sont semblables à ceux obtenus après la pré-sélection elastic net et ouvrent des voies de recherchent intéressantes concernant les chromosomes mis en exergue car ils ne contiennent pas les

SNPs les plus contributifs au premier axe, c'est-à-dire contributifs à la distinction des individus en fonction des phénotypes. En effet si l'on considère à présent les chromosomes contenant les SNPs les plus contributifs du premier axe dans l'ACM sparse on remarque que le signal n'est pas contenu dans les mêmes chromosomes que lors de l'approche décrite ci-dessus.

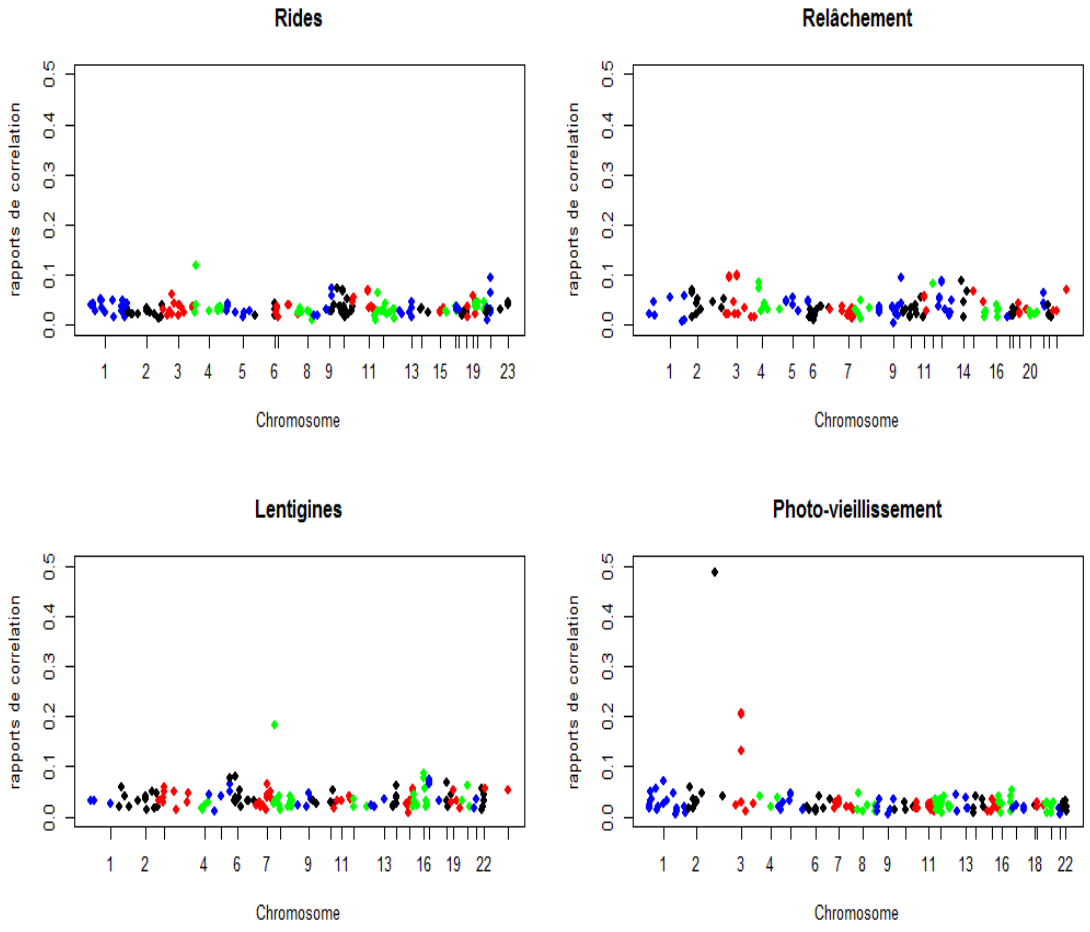

FIGURE 4.17 – Représentation des rapports de corrélation entre les SNPs sélectionnés par ACM sparse et le premier axe de l'ACM sparse en fonction de leurs coordonnées génomiques pour les quatre phénotypes.

La figure 4.17 présente la valeur des rapports de corrélation (variant de 0 à 1) entre les SNPs et le premier axe de l'ACM sparse en fonction de leurs coordonnées génomiques pour les quatre phénotypes. le rapport de corrélation, noté η^2 permet d'étudier la relation entre

une variable qualitative et une variable quantitative (rapport défini comme la variance inter-groupe qui est le carré des écarts entre la moyenne du groupe et la moyenne globale, divisé par la variance totale qui est la somme des carrés des écarts à la moyenne). Les SNPs les plus structurants de l'axe 1 sont ceux avec un rapport de corrélation proche de 1. Les SNPs les plus contributifs au premier axe auraient pu également être représentées (mêmes résultats). Les coordonnées génomiques des SNPs dans les chromosomes sont affichées le long de l'axe des abscisses et la contribution le long de l'axe des ordonnées. Les SNPs les plus contributifs au premier axe de l'ACM sparse pour le phénotype "rides" sont situés sur le chromosome 4, pour le phénotype "relâchement" sur un ensemble assez homogène de chromosomes et pour le phénotype "lentigines" sur le chromosome 8. Les rapports de corrélation les plus élevés pour le phénotype "photo-vieillissement" concernent les SNPs situés sur les chromosomes 2 et 3.

Un zoom sur les SNPs les plus contributifs a été effectué pour chacun des phénotypes pour une approche plus détaillée de ces SNPs. Les SNPs les plus contributifs du premier axe sont présentés figure 4.18 pour chacun des phénotypes. En gras apparaissent les gènes déjà mis en exergue lors de la pré-sélection par elastic net (voir section 4.1.6.2). Le gène GENEri1 est le plus corrélé au premier axe pour le phénotype "rides" et avait déjà été trouvé lors de la pré-sélection elastic net. Ce gène est exprimé dans de nombreux organes y compris dans la peau. Le gène GENEre2 est retrouvé parmi les plus corrélés au premier axe pour le phénotype "relâchement" ainsi que les gènes GENEle1 et GENEle5 pour les "lentigines". Cependant de nouveaux gènes sont mis en exergue (comme les gènes GENEre8 et surtout GENEle7 qui apparait à plusieurs reprises pour le phénotype "lentigines") et les premières recherches bibliographique menées par les biologistes sur ces gènes ouvrent des pistes intéressantes pour la suite car la plupart de ces gènes sont exprimés dans la peau et liés à des mécanismes biologiques bien particuliers. Enfin les gènes les plus contributifs au premier axe pour le "photo-vieillissement" sont les mêmes qu'avant la sélection : GENEpv1 et $STXBP5L$. Le gène $STXBP5L$ (contenu dans le chromosome 3) étant sélectionné par l'ACM sparse, et ce pour plusieurs SNPs, cela donne un poids supplémentaire aux résultats et travaux effectués précédemment sur ce gène et rapporté dans Le Clerc et al. [2012]. Ces SNPs sont retrouvés sur la figure 4.17.

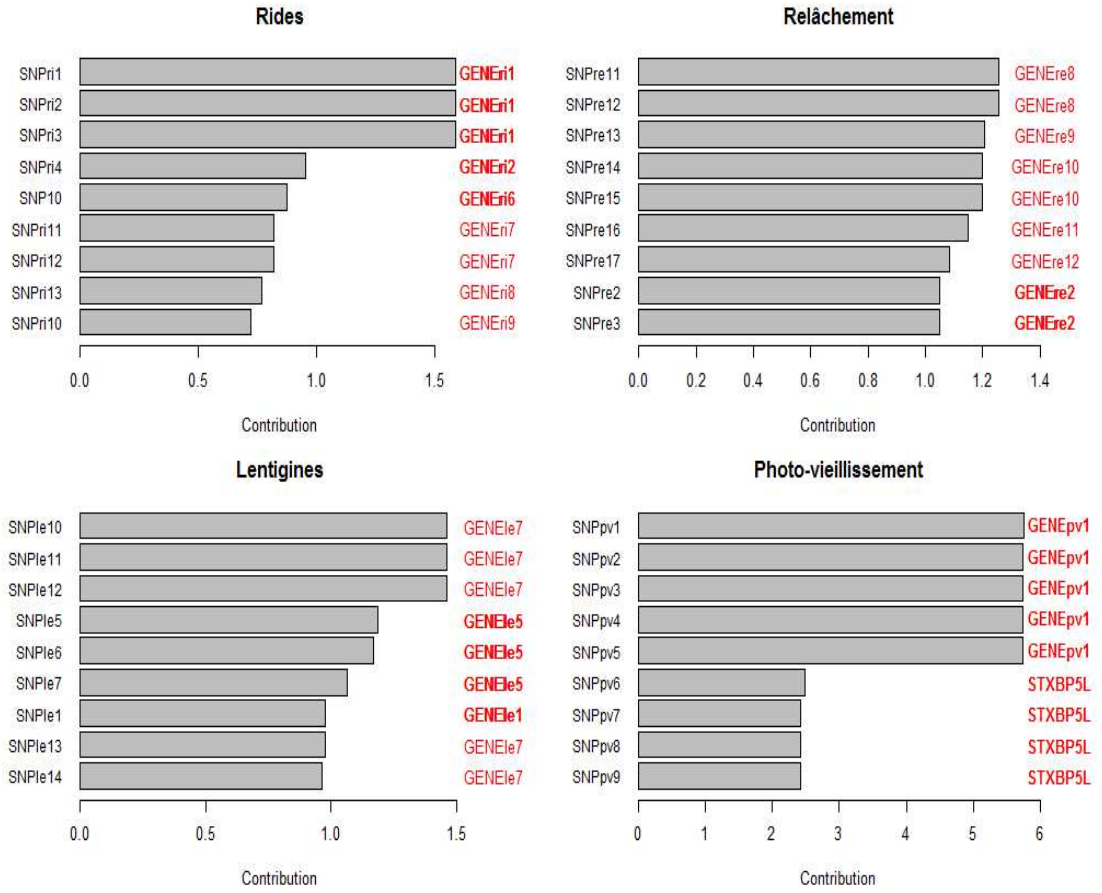

FIGURE 4.18 – Représentation des SNPs les plus contributifs au premier axe de l'ACM sparse pour les quatre phénotypes.

La sélection par ACM sparse s'avère pertinente car elle permet la sélection de variables tout en conservant l'information principale contenue dans la base de départ. Les figures 4.19 et 4.20 présentent les différentes fonctions moléculaires (les fonctions moléculaires étant les activités élémentaires d'un gène au niveau moléculaire) des gènes présents avant (voir figure 4.19) et après sélection par ACM sparse (voir figure 4.20) pour le phénotype "rides". Cette représentation a été réalisée à partir de la base de données en ligne Protein ANalysis THrough Evolutionary Relationships (PANTHER) qui est un système conçu pour classer des protéines (et leurs gènes) afin de faciliter l'analyse à haut débit (Mi et al. [2013] ; "http ://www.pantherdb.org").

Les protéines sont classées en fonction de différentes notions :

- Famille et sous-famille : les familles sont des groupes de protéines évolutives connexes et les sous-familles sont des protéines connexes qui ont la même fonction ;
- Fonction moléculaire : la fonction de la protéine elle-même ou des protéines avec lesquelles elle interagit directement au niveau biochimique, par exemple, une protéine kinase ;
- Procédé biologique : la fonction de la protéine dans le contexte d'un grand réseau de protéines qui interagissent pour accomplir un processus au niveau de la cellule ou l'organisme, par exemple, la mitose.
- Pathway : similaire au processus biologique, mais un "pathway" précise aussi explicitement les relations entre les molécules qui interagissent.

Les SNPs pré-sélectionnés par elastic net (500 en moyenne par phénotype) sont considérés comme la "base de départ". On remarque que malgré la sélection et l'élimination de variables par ACM sparse, l'ensemble des différentes fonctions moléculaires répertoriées reste inchangé (voir figures 4.19 et 4.20). Le même phénomène est observé pour les autres phénotypes (figures non montrées). Le nombre de gènes a diminué après sélection par ACM sparse mais les différentes catégories sont toujours représentées. La sélection ne permet pas d'identifier une fonction moléculaire spécifiquement liée à un phénotype mais cela s'explique par le fait que, le vieillissement étant un processus complexe, il implique plusieurs fonctions moléculaires différentes et ne peut être dû à une seule en particulier.

Pour des raisons de confidentialité des résultats les gènes sélectionnés ne pourront pas être cités ici, mais une étude globale peut être effectuée. Dans l'ensemble des gènes sélectionnés par ACM sparse, des gènes de la famille du collagène sont retrouvés pour le phénotype "relâchement" et "photo-vieillissement". Le collagène est une famille de protéines ayant pour fonction de conférer aux tissus une résistance mécanique à l'étirement. Les protéines ont donc certainement une importance capitale dans le processus de relâchement de la peau. Pour le phénotype "lentigines" des gènes codant des protéines membres de la famille des protéines kinases ont été sélectionnés par ACM sparse. Ce sont des récepteurs pour les membres de la famille du facteur de croissance épidermique. D'autres gènes codant des protéines membres de la famille des protéines tyrosine kinase exprimées dans les tissus adi-

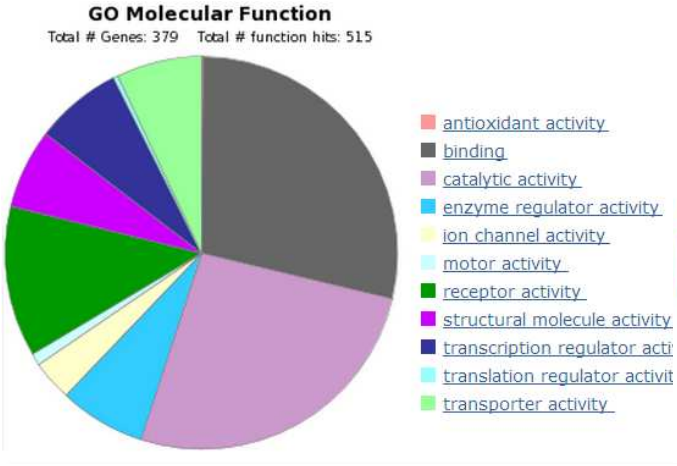

FIGURE 4.19 – Répartition des fonctions moléculaires des gènes avant sélection par ACM sparse.

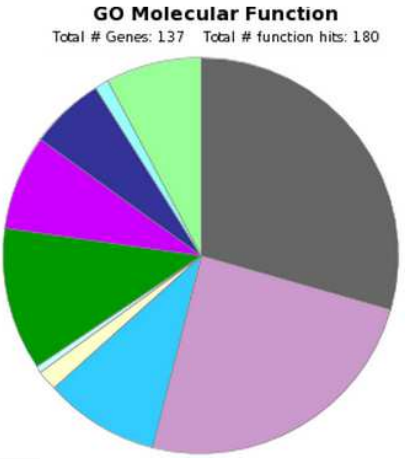

FIGURE 4.20 – Répartition des fonctions moléculaires des gènes après sélection par ACM sparse.

peux et initiant la voie d'activation de récepteurs de surface ont été sélectionnés. Ces gènes pourraient être liés à l'apparition de taches pigmentaires comme les lentigines. Les résultats obtenus (SNPs sélectionnés) semblent avoir un sens biologique ce qui nous conforte dans l'idée que l'ACM sparse permet une sélection pertinente de variables. Les analyses biologiques sont encore en cours car les informations mises en exergue par cette sélection sont nombreuses et ouvrent plusieurs voies d'investigation intéressantes pour les biologistes, notamment des "pathways" liés au métabolisme énergétique (enzymes mitochondriales), des voies de signalisation map-kinases (implication de facteurs de croissance) ainsi qu'une toute nouvelle voie révélée récemment comme étant impliquée dans le renouvellement cellulaire et les cellules souches pour le phénotype "rides", ou encore des mécanismes carbo-hydrates impliqués dans le cycle cellulaire pour le phénotype "photo-vieillissement". L'étude des "pathways est intéressante car elle permet de mettre en évidence des effets plus modérés mais combinés de gènes agissant sur un même mécanisme biologique (le contexte biologique est pris en compte dans l'analyse).

4.3 Approche multiblocs supervisée : RGCCA

Cette section est consacrée à l'utilisation de la RGCCA dans le but d'étudier et de modéliser des liens possibles entre la variable réponse et les SNPs à partir d'une analyse supervisée multiblocs. Chacun des SNPs considérés dans nos analyses sont les SNPs intragéniques pré-sélectionnés par elastic net. Ils peuvent donc être regroupés en fonction de leur appartenance aux gènes. Ainsi, chacun des blocs considérés dans la RGCCA correspond à un gène (contenant des variables SNPs pré-sélectionnés par elastic net), et la variable réponse prise séparément (phénotype "rides", "relâchement", "lentigines" ou "photo-vieillissement") est contenue dans un autre bloc. Il y a donc autant de blocs de SNPs que de gènes, chacun des blocs contenant des SNPs (pour un SNP donné, les variables dichotomiques associées seront considérées). Le bloc contenant la variable à expliquer (un des quatre phénotypes) est relié à chacun des blocs de SNPs mais les liens entre les blocs de SNPs ne sont pas considérés dans cette section. En RGCCA, les conditions de représentation sont similaires à celles de la PLS-PM, comme vu section 1.2.3. Chaque bloc est représenté par une ellipse, et chaque variable du bloc par un rectangle. Chaque variable est connectée à son bloc par une flèche. Deux sous-modèles sont considérés : le modèle externe qui modélise les relations entre le bloc de variables et les composantes du bloc, et le modèle interne qui met en avant les relations entre les éléments au sein d'un bloc.

On rappelle que 658 SNPs ont été pré-sélectionnés par elastic net pour le phénotype "rides". Ces SNPs sont contenus dans 509 gènes au total. Certains blocs (398 précisément) ne contiennent qu'un seul SNP (soit deux ou trois variables binaires selon le nombre de modalités des SNPs). Le modèle considéré présente des liens entre la variable réponse et chacune des variables explicatives. Cela revient à réaliser une régression PLS classique sur la première composante seulement. Le schéma factoriel est utilisé dans les analyses car les résultats ne sont pas très sensibles au choix des schémas centroide ou factoriel (voir Tenenhaus and Tenenhaus [2011]). Le mode Ridge a été choisi afin d'obtenir un bon compromis entre le nouveau mode A et le mode B car il permet d'obtenir des composantes stables et aussi corrélées que possible aux composantes des blocs avec lesquels le bloc est

connecté. Les 510 constantes de régularisation optimales (τ) sont calculées à partir de la formule de Schäfer and Strimmer [2005] mais ne sont pas montrées ici. La matrice de "design" est une matrice de 0 avec des 1 sur la dernière ligne et colonne, sauf le dernier élément de la diagonale ($c_{i,510} = 1$ pour $i = 1, \ldots, 509$ et 0 sinon). La figure 4.21 présente le modèle obtenu pour ce phénotype avec un schéma factoriel et le mode Ridge.

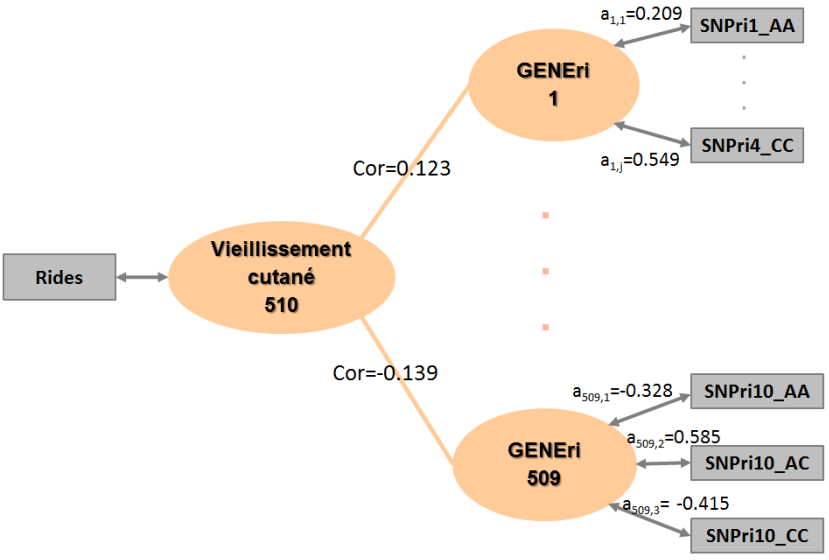

FIGURE 4.21 – RGCCA sur SNPs pré-sélectionnés : schéma factoriel + mode Ridge.

Les poids de chacune des variables d'un bloc sont donnés par la lettre "a" et les corrélations entre les blocs de variables et la variable réponse ont été calculées et testées par bootstrap et sont données pour chacun des liens entre les blocs de SNPs et le phénotype. Les intervalles de confiance calculés par bootstrap pour chacune des corrélations révèlent que 503 des 509 liens entre les blocs de SNPs et la variable réponse sont significatifs. Ceci n'est pas surprenant étant donné que les SNPs ont été pré-sélectionnés de manière supervisée. Ainsi les liens sont également significatifs dans la RGCCA. Les indices de qualité du modèle externe (Average Variance Explained (AVE)) ont été calculés pour chacun des blocs et montrent que chaque composante explique correctement son propre bloc. Cela prouve que le regroupement des SNPs par gène a un sens (en plus du sens biologique).

Les mêmes résultats ont été obtenus pour les autres phénotypes. Pour le "relâchement" la RGCCA a été effectuée sur les 700 SNPs pré-sélectionnés contenus dans 542 gènes au total et la quasi-totalité des liens (537 sur 542) s'est avérée significative par bootstrap. Pour les "lentigines" 523 gènes au total (réunissant les 670 SNPs pré-sélectionnés par elastic-net) ont été considérés dans la RGCCA et 518 liens sur les 523 se sont avérés significatifs. Enfin pour la RGCCA calculée sur les 640 SNPs contenus dans 504 gènes pour le "photo-vieillissement", 490 liens sur 504 se sont avérés significatifs.

Il est intéressant de regarder les gènes les plus corrélés à la variable réponse afin de s'en servir de base dans la comparaison entre la RGCAA sans interaction et celle qui sera réalisée par la suite avec interactions.

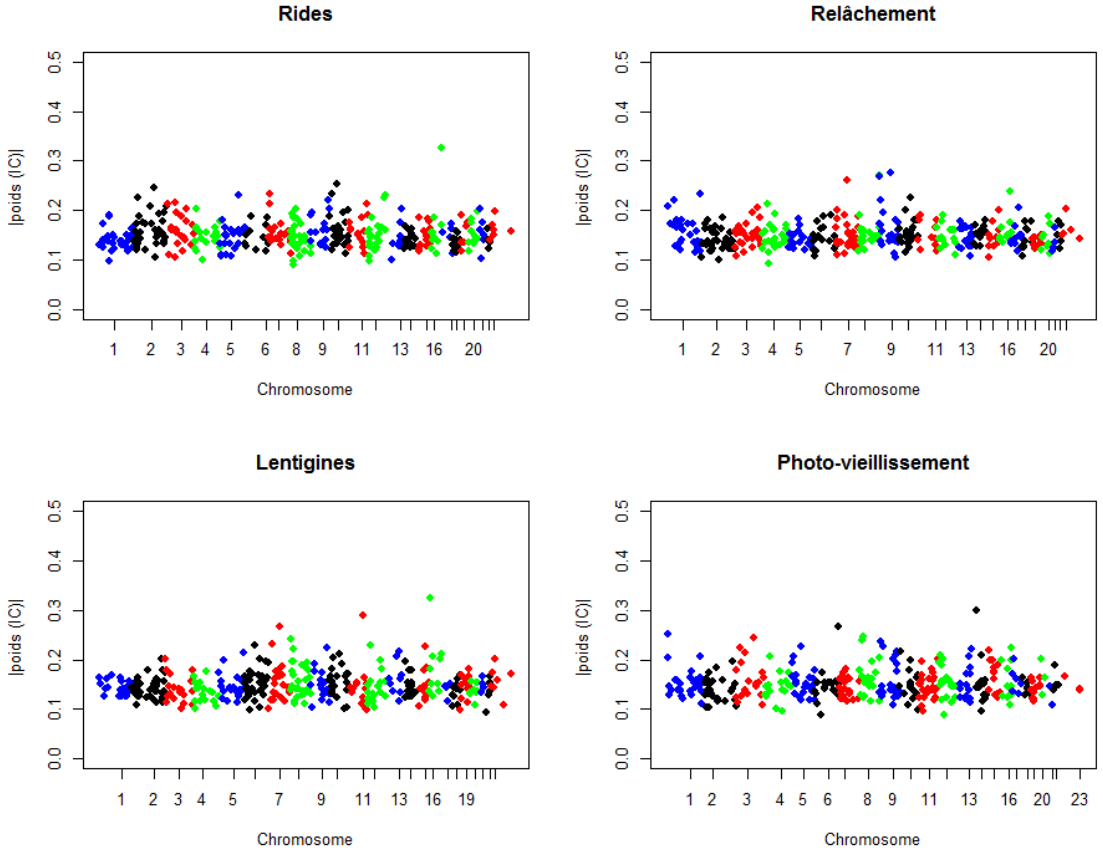

FIGURE 4.22 – Représentation de la valeur absolue des corrélations entre chacun des blocs de SNPs (gènes) et chacun des quatre phénotypes dans la RGCCA en fonction de leurs coordonnées génomiques.

La figure 4.22 présente la valeur absolue des corrélations entre chacun des blocs de SNPs (gènes) et chacun des quatre phénotypes dans la RGCCA en fonction de leurs coordonnées génomiques. On remarque que les corrélations les plus élevées correspondent aux gènes situés dans les mêmes chromosomes que dans la représentation elastic net pour les phénotypes "rides" et relâchement" (voir figure 4.10). Le chromosome 16 principalement pour le phénotype "rides" et le numéro 7 (ainsi que les 8 et 9) pour le phénotype "relâchement". Pour les phénotypes "lentigines" et "photo-vieillissement" ce sont des gènes situés dans les chromosomes 16 et 14, respectivement, qui ont une corrélation plus élevée que les autres.

Cette étude par RGCCA n'a pas apporté de nouveaux résultats étant donné qu'aucun lien entre blocs de SNPs n'est considéré. Cela revient à réaliser une régression PLS des SNPs sur la variable réponse mais cela permet de présenter les données sous une autre forme et de les traiter différemment grâce à une structure par blocs. Les modèles présentés servent d'introduction et peuvent servir au lecteur à mieux comprendre les modélisations réalisées dans la suite des analyses après la prise en compte des interactions.

4.4 Détection d'interactions

Les interactions génétiques correspondent à des modifications dans l'action d'un gène induites par l'expression d'un autre gène. Elles sont mises en évidence par une observation des phénotypes. Par exemple, si le phénotype d'un individu portant un SNP variant sur un premier gène est aggravé ou au contraire sauvé par une "mutation" d'une SNP sur un deuxième gène, alors il y a interaction entre les deux gènes. Nous allons donc ici tenter de détecter de possibles interactions entre SNPs qui modifieraient l'expression du phénotype étudié.

4.4.1 Interactions SNP-SNP par régression logique

Les interactions SNP-SNP sont censées expliquer les différences entre individus à faible et à haut risque de présenter une caractéristique particulière. Dans notre cas, nous cher-

chons à trouver les interactions entre SNPs expliquant la sévérité du vieillissement (pour chacun des scores de rides, relâchement et lentigines, ainsi que le photo-vieillissement global). Il est donc intéressant de construire des règles de classification du type suivant : si le SNP 1 est le génotype hétérozygote ET que le SNP 2 est le génotype homozygote ou que les SNPs 3 et 4 ne sont pas le génotype homozygote de référence, alors l'individu présente un risque plus élevé d'avoir un score de vieillissement important. Pour ce faire, nous utiliserons la régression logique définie dans le chapitre 3 dans le cas de la prédiction du statut d'une "observation" pour une variable à expliquer binaire, méthode très souvent employée pour la détection d'interactions SNP-SNP (Ruczinski et al. [2001], Dinu et al. [2012]).

4.4.1.1 Codage des variables

Codage des SNPs

Les SNPs considérés dans cette analyse sont ceux qui ont été sélectionnés précédemment par la méthode elastic net. Les analyses sont réalisées sous R à l'aide du package logicFS (Schwender [2013a]). Le codage des SNPs diffère quelque peu de celui utilisé dans l'ACM sparse. On considère les trois génotypes possibles :

- homozygote de référence (les 2 allèles sont les variants les plus fréquents)
- hétérozygote (un des 2 allèles est un variant fréquent, l'autre rare)
- homozygote variant (les 2 allèles sont des variants rares ou les moins fréquents)

La régression logique, et donc les fonctions du package R, ne peuvent gérer que des prédicteurs binaires. Les variables catégorielles SNPs doivent donc être transformées en variables binaires. Si le SNP est codé 1 pour "homozygote de référence", 2 pour "hétérozygote" et 3 pour "homozygote variant", alors la fonction "make.dummy.snp" du package LogicFS peut être utilisée pour transformer chaque SNP en 2 variables binaires SNP.1 et SNP.2. Le codage binaire des variables est représenté dans la table 4.3.

TABLE 4.3 – Codage des SNPs en variables binaires pour la régression logique.

	SNP	SNP.1	SNP.2	Génotype supposé
1	1	0	0	Homozygote de référence
2	2	1	0	Hétérozygote
3	3	1	1	Homozygote variant

Les variables nominales binaires peuvent ensuite être utilisées dans l'algorithme logicFS décrit en section 9 afin d'identifier des combinaisons intéressantes de ces variables et de mesurer l'importance de ces interactions.

Codage des variables à expliquer

Les variables résidus des scores ajustés de rides, relâchement, lentigines et photo-vieillissement sont des variables centrées. On considère un codage binaire pour appliquer par la suite une régression logique avec :

- Codage "1" : individus pour lesquels les valeurs des résidus sont supérieures à 0 (vieillissement cutané plus important),
- Codage "0" : individus pour lesquels les valeurs des résidus sont inférieures ou égales à 0 (vieillissement cutané moins important),

On obtient pour les quatre phénotypes :

Rides 266 individus sont codés 1, 235 sont codés 0,

Relâchement 258 individus sont codés 1, 243 sont codés 0,

Lentigines 225 individus sont codés 1, 276 sont codés 0,

Photo-vieillissement 199 individus sont codés 1, 302 sont codés 0.

4.4.1.2 Détection des interactions

L'algorithme de régression logique est réalisé $B = 200$ fois en considérant un seul arbre et un maximum de 2 feuilles dans le modèle (les interactions simples, doubles entre SNPs sont donc considérées). Les interactions triples ou plus peuvent être testées, mais l'intégration par la suite de ces interactions dans un modèle multiblocs par la suite pourrait s'avérer plus coûteux en temps de calcul. De ce fait, nous nous limitons aux interactions doubles.

Phénotype "rides"

L'algorithme logicFS a été utilisé sur les SNPs pré-sélectionnés par elastic net pour le phénotype "rides" et les 5 interactions les plus importantes sont rapportées dans la figure 4.23.

```
Selection of Interactions Using Logic Regression

Number of Iterations:  200
Sampling Method:       Bagging
Logic Regression Type: Classification
Max. Number of Leaves: 2

Importance Measure: Single Tree
Based On: Number of OOB Observation


The 5 Most Important Interactions:

         Importance         Expression
1              9.17       !X376_2 & X725_1
2              8.65        !X30_2 & X725_1
3              8.36       !X376_2 & X1275_1
4              8.27        !X30_2 & X1275_1
5              8.06       !X376_2 & !X856_2
```

FIGURE 4.23 – Sortie du programme de régression logique avec le logiciel R pour les 5 premières interactions obtenues pour le phénotype "rides".

La sortie obtenue peut s'écrire également sous la forme disjonctive normale suivante : "$(\text{X376\_2}^c \wedge \text{X725\_1}) \vee (\text{X30\_2}^c \wedge \text{X725\_1}) \vee (\text{X376\_2}^c \wedge \text{X1275\_1}) \vee (\text{X30\_2}^c \wedge \text{X1275\_1}) \vee (\text{X376\_2}^c \wedge \text{X856\_2}^c) \vee \ldots$".

L'expression "!X376_2" code pour "NOT" X376_2, i.e. X376_2 = 0 et "X725_1" correspond à X725_1 = 1. Ainsi l'expression logique "!X376_2 & X725_1" signifie que si un individu possède un génotype qui n'est pas homozygote variant pour le SNP X376, et qu'il possède un génotype homozygote variant ou hétérozygote pour le SNP X725, alors l'individu a une probabilité plus élevée de présenter une expression de ride sévère.

Lorsque l'on considère un seul arbre 205 SNPs et SNP-SNP interactions potentiellement intéressants de degré deux au maximum sont détectés. L'importance des interactions (calculé sous la forme d'un score nommé VIM présenté au chapitre 3) peut être représentée comme sur la figure 4.24.

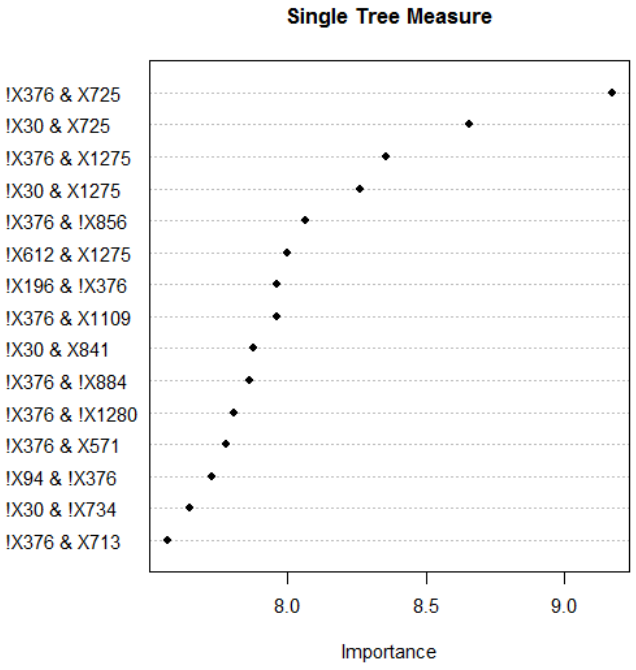

FIGURE 4.24 – Importance des interactions obtenues (VIM) pour le phénotype "rides".

Afin de vérifier la significativité statistique de ces interactions, un test du Chi2 a été réalisé sur l'ensemble de ces interactions. Pour ce faire, chaque interaction a été codée comme une variable binaire. La variable "interaction.1" correspondant à la première interaction observée est codée 1 si un individu présente l'ensemble des génotypes contenus dans l'interaction 1, 0 sinon. On obtient donc un tableau de 501×205 de variables binaires. Parmi les 205 interactions doubles (au maximum), 35 se sont avérées statistiquement significatives. La correspondance SNP/gène nous permet d'obtenir 35 interactions gène-gène (en fonction de l'appartenance des SNPs aux gènes). Un des gènes apparait dans 15 des 35 interactions et le gène GENEri3 dans 8 des 35 interactions. Les interactions génétiques sont des interactions dont on ne connaît pas le mécanisme moléculaire pour la plupart. Pour y remédier, ces interactions sont en cours d'investigation par les biologistes afin de trouver une interprétation biologique et des pistes sur des potentiels réseaux de gènes (dont ces deux derniers qui sont en interaction avec un grand nombre d'autres gènes) liés à l'expression des rides.

Phénotype "relâchement"

Pour le phénotype "relâchement" 220 SNPs et SNPs-interactions potentiellement intéressants de degré deux au maximum ont été détectés (figure non montrée). La significativité statistique de ces interactions a été vérifiée de la même manière que précédemment à l'aide d'un test du Chi2 sur un tableau de 501×220 variables binaires. 23 interactions se sont avérées significatives. Parmi celles-ci, un gène apparait dans 8 des 23 interactions.

Phénotype "lentigines"

La régression logique a permis de détecter 209 SNPs et SNPs-interactions potentiellement intéressants de degré deux au maximum pour le phénotype "lentigines". A l'aide d'un test du Chi2 sur un tableau de 501×209 variables binaires 27 interactions se sont révélées significatives. Un gène apparait dans 8 interactions sur 27.

Phénotype "photo-vieillissement"

Enfin 192 SNPs et SNPs-interactions potentiellement intéressants de degré deux au maximum ont été détectées pour le phénotype "photo-vieillissement". La significativité statistique de ces interactions a été vérifiée de nouveau sur un tableau de 501×192 variables binaires et 111 interactions se sont avérées significatives. Le gène GENEpv1 apparait dans 4 interactions. Ici de nombreux gènes différents sont donc impliqués dans les interactions.

Notre attention se porte sur les gènes impliqués dans de nombreuses interactions car ils pourraient être impliqués dans des réseaux biologiques ("pathways") potentiellement liés à l'expression des rides. Ces gènes ne sont peut-être pas corrélés à la variable réponse lorsqu'ils sont pris séparément dans les analyses mais pourraient s'exprimer en présence d'autres gènes. C'est pourquoi les interactions statistiquement significatives pour chacun des phénotypes seront prises en compte dans la suite des analyses en section 4.5.

4.4.2 Interactions biologiques rapportées dans la littérature

Un autre moyen d'analyser les interactions entre gènes ou entre SNPs est l'utilisation de bases de données réunissant l'ensemble des informations déjà connues sur le sujet telles que le site "http ://string-db.org/" (voir section 3.1). Afin d'avoir une vision globale des interactions déjà connues dans la littérature, l'ensemble des SNPs pré-sélectionnés par elastic net ont été rentrés dans la base de données STRING pour chacun des quatre phénotypes. Pour des raisons de confidentialité des résultats le nom des gènes n'apparait pas sur les figures présentées dans ce document. Ces analyses ont fait l'objet d'un rapport interne dans lequel les noms des gènes apparaissent et des interprétations biologiques sont explicitées.

Phénotype "rides"

La figure 4.25 présente le réseau d'interactions protéiques déjà connues pour l'ensemble des 658 gènes pré-sélectionnés par elastic net pour le phénotype "rides". Seuls les gènes en interaction sont représentés.

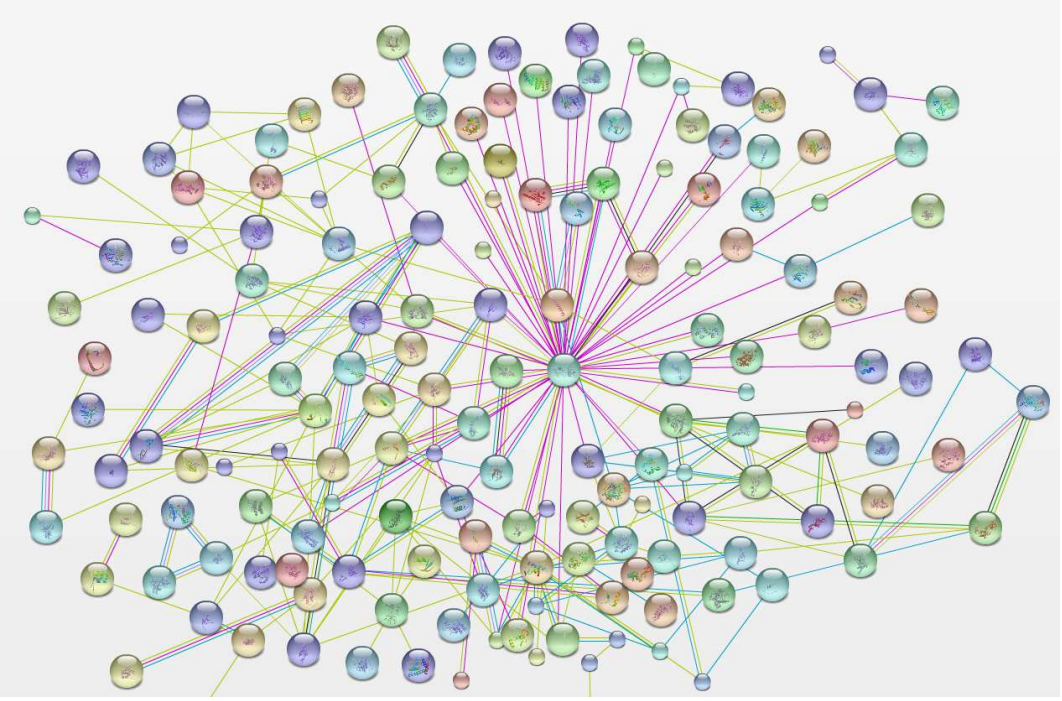

FIGURE 4.25 – Réseau d'interactions protéiques connues entre gènes avant sélection par ACM sparse pour le phénotype "rides" à partir de la base de données STRING.

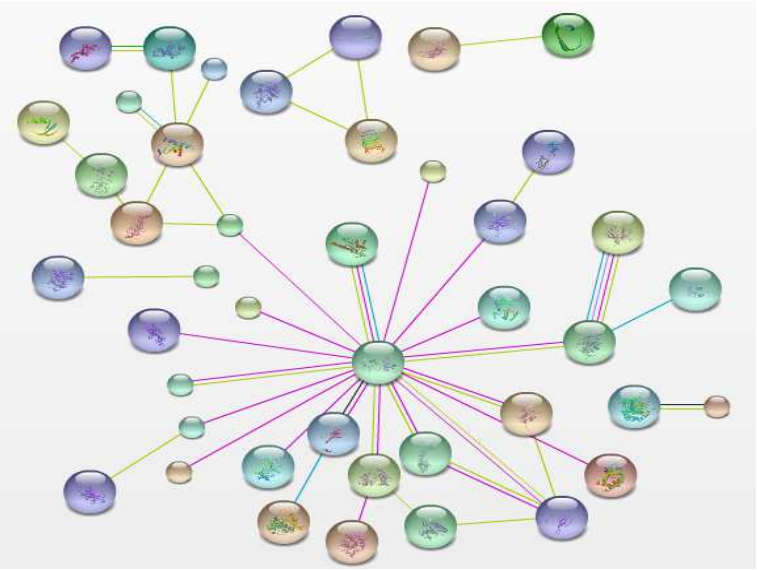

FIGURE 4.26 – Réseau d'interactions protéiques connues entre gènes après sélection par ACM sparse pour le phénotype "rides" à partir de la base de données STRING.

L'analyse et l'interprétation de ces interactions est compliquée en raison du très grand nombre de gènes impliqués. C'est pourquoi la sélection de variables prend tout son sens dans un contexte de grande dimension. La même analyse a donc été réalisée sur l'ensemble des gènes sélectionnés par ACM sparse. Le nombre de gènes étant réduit à 138 nous espérons pouvoir visualiser plus facilement les interactions. La figure 4.26, représentation de même type que la figure précédente, concerne les 138 SNPs sélectionnés par ACM sparse. Seuls les gènes présentant des interactions connues apparaissent. Les autres sont masqués par souci de lisibilité. La visualisation des interactions est simplifiée et les gènes les plus fréquemment impliqués dans les interactions sont plus facilement détectés. Nous remarquons que le noyau central des interactions est conservé comme si un zoom avait été effectué sur la partie centrale des interactions. Cela nous conforte une nouvelle fois dans l'idée que l'ACM sparse réalise une sélection des SNPs tout en conservant l'information principale contenue dans base de départ.

Le même type de représentation d'interactions peut être réalisé pour les phénotypes "relâchement", "lentigines" et "photo-vieillissement". Pour le phénotype "relâchement" certains gènes de la famille du collagène apparaissent dans ces réseaux d'interactions. Ceci

est très intéressant car le collagène est une famille de protéines ayant pour fonction de conférer aux tissus une résistance mécanique à l'étirement, et donc lié au relâchement de la peau. Il s'avère que la plupart des gènes sélectionnés par ACM sparse pour chacun des phénotypes sont impliqués dans la peau et mettent en exergue des réseaux d'interactions entre gènes très intéressants qui ouvrent de nouvelles pistes d'exploration pour les biologistes.

4.4.3 Synthèse des résultats

Les deux approches décrites précédemment nous ont permis d'identifier des interactions potentiellement intéressantes. Elles vont permettre aux biologistes d'explorer des pistes pouvant mener à l'identification de nouveaux complexes biologiques pouvant expliquer l'expression du vieillissement cutané. Parmi les interactions détectées statistiquement grâce à la régression logique, une seule s'est avérée être déjà connue biologiquement (pour le phénotype "rides"). On aurait pu s'attendre à en découvrir plus en commun seulement cela peut s'expliquer par le fait que l'implication des gènes sélectionnés par ACM sparse dans la peau n'a pas encore été investigué et les interactions n'ont pas encore été mises en exergue.

4.5 Intégration des interactions dans les approches multiblocs

Cette section présente les résultats des analyses non supervisées et supervisées effectuées en tenant compte des interactions significatives obtenues par régression logique. Dans un premier temps, une analyse "sparse" non supervisée a été effectuée à l'aide de l'ACM sparse.

4.5.1 Approche multiblocs non supervisée : ACM Sparse

L'étude de détection d'interactions sur l'ensemble de la base de données de SNPs intragéniques a permis de sélectionner des interactions significativement liées aux phénotypes étudiés. Afin de prendre en compte ces interactions dans un modèle général une ACM sparse introduisant les interactions trouvées pour chacun des quatre phénotypes a été réalisée.

La prise en compte des interactions dans l'ACM sparse a été envisagée de deux manières

différentes. La première consiste à considérer l'ensemble des SNPs pré-sélectionnés par elastic net et à rajouter les SNPs en interaction dans des groupes supplémentaires, sachant qu'un groupe de variables SNPs correspond à une interaction donnée. Pour le phénotype "rides" par exemple, le tableau de données comprendrait 658 SNPs + 35 groupes de SNPs correspondant aux 35 interactions significatives trouvées section 4.4.1.2 (voir figure 4.27 pour la visualisation du tableau de données). L'ACM sparse permettrait la sélection de SNPs ainsi que de groupes de SNPs lorsque ceux-ci sont en interaction. Cependant un SNP peut être impliqué dans plusieurs interactions et donc apparaitre dans plusieurs groupes ce qui provoquerait une redondance de plusieurs colonnes dans la table.

	SNP1	. . .	SNP658	Interaction 1	. . .	Interaction 35
1	AA		AA	0		1
	AC		AA	0		0
	CC		AG	0		0
.	AC		AG	1		0
	AC		AA	1		0
.	AA	. . .	AA	0	. . .	1
.	AA		AA	1		1
	AA		AA	1		1
	CC		GG	1		0
	CC		AG	1		0
501	CC		GG	1		0

FIGURE 4.27 – Tableau de données considéré dans l'ACM sparse avec interaction pour le phénotype "rides".

La deuxième possibilité, celle choisie pour la suite des analyses, consiste à traiter chaque interaction comme une variable binaire. Pour une interaction donnée, la variable binaire correspondante sera codée de la manière suivante :

- 1 si l'individu possède les génotypes impliqués dans l'interaction considérée
- 0 sinon.

Phénotype "rides"

Nous considérons le tableau de données de SNPs pré-sélectionnés par elastic net de dimensions 501×658 pour le phénotype "rides". La régression logique et un test du Chi2 ont permis de mettre en évidence 35 interactions doubles significatives recodées en 35 variables

binaires. Le tableau de données final contient alors 501 individus et 658+35 = 683 variables catégorielles, soit un total de 1 824 modalités.

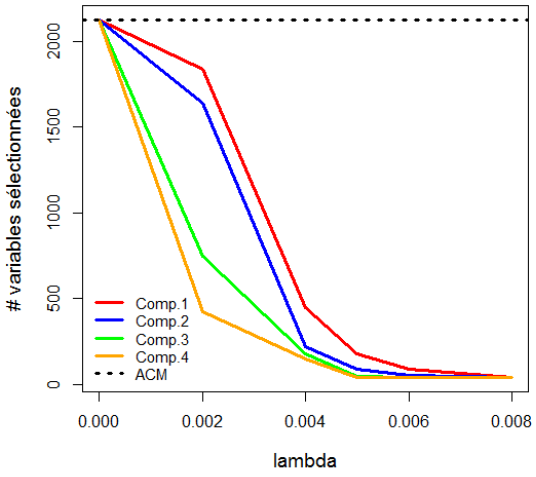

FIGURE 4.28 – Évolution du nombre de variables sélectionnées par ACM sparse (avec interactions) en fonction de λ pour le phénotype "rides".

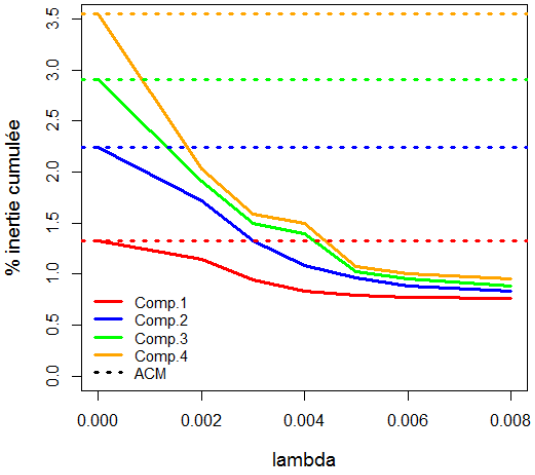

FIGURE 4.29 – Évolution du % d'inertie obtenu avec ACM sparse (avec interactions) en fonction de λ pour le phénotype "rides".

Si l'on se concentre sur le premier axe de l'ACM sparse on remarque que, pour un $\lambda = 0,004$, 437 variables (soit 145 SNPs) sont sélectionnées (figure 4.28) et le % d'inertie vaut 0,47% (figure 4.29). Pour $\lambda = 0,005$, 156 variables (soit 52 SNPs) sont sélectionnées et le % d'inertie est égal à 0,40%. Etant donné la faible différence de pourcentage d'inertie entre les deux valeurs de λ, nous choisissons la valeur de λ pour laquelle la sélection est la plus fine (c'est-à-dire le λ pour lequel il y a le plus de variables écartées), soit $\lambda = 0,005$. La stabilité de la sélection a été testée par un bootstrap à 100 itérations. Tout comme en section 4.2.2, les variables et les interactions dont la fréquence de sélection est supérieure au 3ème quartile sont conservées. Au final 27 SNPs contenus dans 22 gènes et 21 interactions ont été sélectionnés. Les figures présentant le nombre de variables conservées et le pourcentage cumulé de variance expliquée ne seront pas montrées pour les autres phénotypes car les courbes obtenues ont la même allure.

Phénotype "relâchement"

On considère le tableau de données de dimensions 501×700 des SNPs pré-sélectionnés par elastic net pour le phénotype "relâchement" auquel on ajoute les 23 interactions significatives misent en évidence et recodées en 23 variables binaires. Le tableau de données final contient 501 individus et $700 + 23 = 723$ variables catégorielles, soit un total de 2 120 modalités. Pour les même raisons que celles citées dans le paragraphe précédent, le paramètre de pénalisation λ est fixé à 0,005. Le bootstrap a permis de conserver 13 SNPs contenus dans 9 gènes et 10 interactions.

Phénotype "lentigines"

Pour le phénotype "lentigines" les 27 interactions significatives et recodées en variables binaires ont été ajoutées au tableau de données de SNPs de dimensions 501×670. 687 variables catégorielles sont considérées, soit un total de 1 890 modalités. Le bootstrap a été réalisé pour $\lambda = 0,005$: 24 SNPs contenus dans 20 gènes et 24 interactions ont été sélectionnés.

Phénotype "photo-vieillissement"

Nous considérons le tableau de données de SNPs de dimensions 501×640 auquel les 111 interactions doubles significatives, recodées en variables binaires, ont été ajoutées. Le tableau de données final contient 501 individus et $640 + 111 = 751$ variables catégorielles, soit un total de 2 230 modalités. Le bootstrap réalisé avec $\lambda = 0,005$ a retenu 27 SNPs contenus dans 25 gènes et 37 interactions.

Lorsque les interactions sont prises en compte dans l'ACM sparse le nombre de variables sélectionnées est en moyenne 6 fois inférieur à celui obtenu lorsque les interactions ne sont pas intégrées au modèle. Les résultats ne nous permettent plus de discriminer les individus en fonction de la sévérité du vieillissement cutané mais montrent que les variables sélectionnées expliquent la variabilité des individus en fonction des interactions entre SNPs qu'ils présentent (figures non montrées). Ceci pourrait s'expliquer par le fait qu'il puisse y avoir une corrélation forte entre les SNPs et les variables "interaction". Pour y remédier

nous avons envisagé pour de futurs travaux de réaliser l'ACM sparse sur les blocs de SNPs en interaction sans considérer les SNPs de manière individuelle.

4.5.2 Approche multiblocs supervisée : RGCCA

Cette section est consacrée à une autre approche de la prise en compte des interactions : l'approche supervisée. Les résultats de l'analyse supervisée multiblocs RGCCA effectuée en tenant compte des interactions significatives obtenues par régression logique sont présentés. Dans ce contexte chacun des blocs considérés correspond à un gène contenant des SNPs et la variable réponse est contenue dans un autre bloc.

La RGCCA permet de tenir compte des liens possibles entre les différents blocs. Comme dans la section 4.3, chacun des blocs de SNPs sont liés à la variable phénotype. Si K est le nombre de blocs de SNPs, la matrice de design vaut $c_{i,\text{pheno}} = 1$ pour $i = 1, \ldots, K$ et 0 sinon. Par ailleurs, les interactions significatives entre SNPs (et donc entre gènes en fonction des SNPs qu'ils contiennent) mises en exergue par la régression logique et les tests du Chi2 section 4.4.1.2 sont intégrées au modèle. Les gènes en interaction sont reliés par un trait et considérés dans la matrice de design. Si une interaction est détectée entre le SNP 1 contenu dans le gène 1 et le SNP 12 dans le gène 4, alors un lien est établi entre le gène 1 et le gène 4 et est modélisé dans la matrice de design de la manière suivante : $c_{1,4} = c_{4,1} = 1$.

Phénotype "rides"

Sur 509 gènes sélectionnés pour le phénotype "rides" (voir section 4.3), la régression logique et des tests du Chi2 ont permis de détecter 35 interactions significatives qui ont été prises en compte dans le modèle RGCCA. La matrice de design est modifiée en fonction des interactions considérées. Pour une interaction entre le gène 4 et le gène 509 : $c_{4,509} = 1$ et $c_{509,4} = 1$. A l'aide d'un bootstrap à 100 itérations, 496 liens entre les gènes et le phénotype se sont avérés significatifs ainsi que 2 interactions. Les indices de qualité du modèle obtenu nous indiquent que les composantes de chaque bloc expliquent correctement leur propre bloc et que la qualité du modèle interne est supérieure à celle obtenue sans interactions, ce qui était prévisible compte tenu des interactions.

Les mêmes résultats sont obtenus pour les autres phénotypes analysés. Une RGCCA a été réalisée sur les 542 gènes sélectionnés pour le phénotype "relâchement" (voir section 4.3) en tenant compte des 23 interactions significatives détectées par la régression logique et un test du Chi2. Au final 515 liens sur les 542 entre les gènes et le phénotype ainsi que 4 interactions sur les 23 étaient significatifs. Pour le phénotype "lentigines", les 524 gènes sélectionnés par elastic net ont été considérés (voir section 4.3) et les 27 interactions significatives ont été prises en compte. Finalement, 504 liens sur les 524 entre les gènes et le phénotype et 6 interactions sur les 27 se sont avérés significatifs. Enfin, une RGCCA sur 504 gènes sélectionnés pour le phénotype "lentigines" (voir section 4.3) tenant compte des 111 interactions significatives a été réalisée et 486 liens sur les 504 entre les gènes et le phénotype ainsi que 6 interactions sur les 111 seulement étaient significatifs.

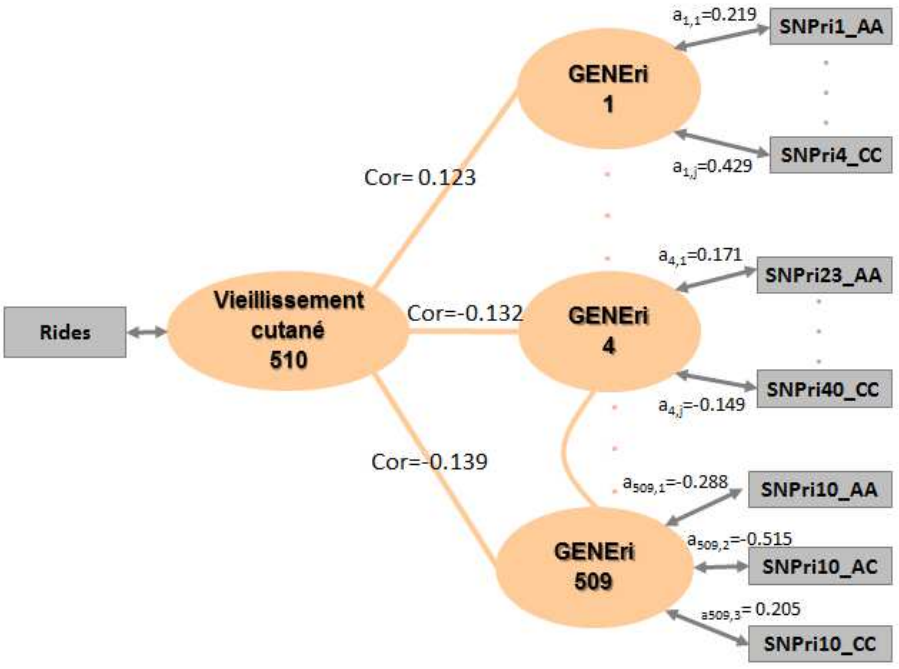

FIGURE 4.30 – RGCCA sur SNPs avec interactions pour le phénotype "rides" : schéma factoriel + mode Ridge.

Les résultats obtenus sont surprenants car nous nous attendions à obtenir plus de liens (interactions) significatifs entre les gènes considérés comme étant en interaction étant donné que la détection d'interactions s'est faite de manière supervisée à l'aide d'une régression logique. Des modifications dans l'ordre des corrélations entre les gènes et le phénotype au-

raient pu être observées, ce qui n'est pas le cas ici. La figure 4.31 représente la valeur absolue des corrélations entre chacun des blocs de SNPs et le phénotype "rides" dans la RGCCA sans (figure de gauche) et avec (figure de droite) la prise en compte des interactions. Étant donné le faible nombre d'interactions significatives dans le modèle, aucune évolution dans l'ordre des corrélations entre gènes et phénotype n'est observée. Il pourrait être intéressant de modifier la structure des blocs en considérant 1 bloc comme 1 chromosome. Le bloc rassemblerait donc l'ensemble des SNPs appartenant à ce chromosome. Cela limiterait le nombre de blocs à 23 (ou 24 en présence d'ADN mitochondrial) et permettrait d'avoir une d'idée plus générale des chromosomes contenant des gènes ayant un lien possible avec le phénotype considéré.

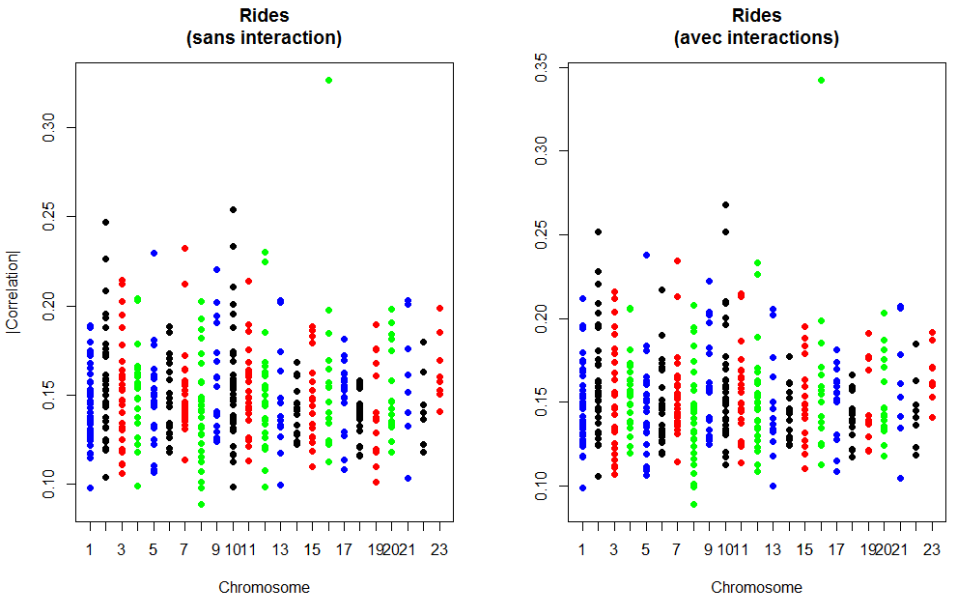

FIGURE 4.31 – Représentation de la valeur absolue des corrélations entre chacun des blocs de SNPs et la variable réponse dans la RGCCA sans (figure de gauche) et avec (figure de droite) prise en compte des interactions pour le phénotype "rides".

Des recherches en collaboration avec les biologistes sont actuellement en cours pour tenter d'apporter une explication biologique aux résultats obtenus avec la RGCCA. La version "sparse" de cette méthode (la SGCCA) sera utilisée dans des analyses futures dans le but de sélectionner des variables au sein des blocs et de, peut-être, faciliter la compréhension des modèles obtenus.

Conclusion et perspectives

Dans le contexte de données de grandes dimensions rencontrées lors du traitement de données génétiques, le problème de l'interprétabilité des résultats se pose pour les biologistes. L'utilisation de méthodes de sélection de variables permet d'obtenir des résultats parcimonieux tout en conservant l'information principale contenue au départ. Une première étape consistait à explorer les données grâce à des analyses non supervisées afin de comprendre et de visualiser les liens entre les variables et entre les individus. Lors de cette approche, la sélection de variables a été importante pour l'exploitation des résultats par les experts du domaine. Des méthodes d'analyse de données dites "sparse" existaient déjà pour sélectionner des variables quantitatives dans ce contexte mais aucune ne permettait la sélection de blocs de variables lorsque les données sont structurées par blocs a priori. Par ailleurs, dans le cas spécifique de l'analyse de données catégorielles de SNPs, ces méthodes ne pouvaient pas être utilisées et aucune version parcimonieuse d'analyse de données catégorielles n'avait encore été proposée à ce jour. Ceci a donc fait l'objet du développement de deux nouvelles méthodes de sélection de variables dans le cas non supervisé.

1. Apports de la thèse

Les deux méthodes développées permettent l'exploration des données de manière sélective. La première, nommée GSPCA, est une adaptation de l'ACP sparse permettant la sélection de variables quantitatives axe par axe dans le cas où des blocs de variables sont définis à priori. La deuxième, l'ACM sparse, est une extension de la précédente dans le cas de données qualitatives. Lorsque le paramètre de régularisation est fixé à 0, la GSPCA et l'ACM sparse reviennent à réaliser une ACP et une ACM, respectivement, ce qui d'après

Zou et al. [2006] est une propriété essentielle d'une "bonne" méthode "sparse". La sélection n'est pas globale mais réalisée axe par axe. Des variables conservées sur un axe peuvent être éliminées sur un autre et vice versa. Ces méthodes créent de la parcimonie au niveau des "loadings" tout en limitant la perte du pourcentage de variance expliquée ce qui facilite l'interprétation des axes obtenus et la visualisation des données. Étant donné la nature catégorielle de nos données (SNPs) seule l'ACM sparse a pu être appliquée et testée. Une ACM a été réalisée dans un premier temps sur notre jeu de données de SNPs et a permis de dissocier 2 groupes d'individus sur le premier axe présentant donc un patrimoine génétique proche. Une étude sur ces individus a été réalisée et a permis de révéler que les deux groupes étaient parfaitement discriminés en fonction de la sévérité de l'expression du vieillissement cutané (ceci étant lié à la pré-sélection supervisée réalisée en amont). Néanmoins, ce qu'il est intéressant de remarquer est qu'après l'application de l'ACM sparse sur ces données, un grand nombre de variables ont été éliminées sur chacun des axes mais la discrimination entre les deux groupes d'individus était toujours parfaitement observée. L'utilité et l'efficacité de l'ACM sparse ont donc été prouvées sur ce jeu de données : l'information principale de départ est conservée, les résultats obtenus sont parcimonieux et l'interprétation est simplifiée. Par ailleurs les gènes contenant les SNPs impliqués dans la discrimination des groupes se sont avérés être impliqués dans la peau et dans des mécanismes biologiques intéressants pour notre étude. L'interprétation des résultats a fait l'objet d'une collaboration étroite avec les biologistes de l'unité. Ils mènent actuellement des recherches plus approfondies sur ces gènes et sur les voies biologiques impliquées afin de réaliser par la suite des tests in-vitro. Ces premiers résultats sont donc très encourageants.

Dans un deuxième temps, une étude de détection d'interaction SNP-SNP a été réalisée sur nos données grâce à une régression logique. Les interactions entre SNPs peuvent être à l'origine du développement de certaines maladies et leur prise en compte dans les analyses est donc importante. La régression logique a permis de mettre en évidence des interactions statistiquement significatives entre SNPs qui n'avaient pas encore été rapportées dans la littérature. Les biologistes de l'équipe travaillent actuellement sur ces résultats afin de trouver une interprétation biologique possible et d'explorer des pistes intéressantes qui ne

seraient pas encore exploitées.

Ces interactions ont été prises en compte par la suite dans un modèle supervisé multiblocs à l'aide de la méthode RGCCA. Les SNPs ont été regroupés dans des blocs en fonction de leur appartenance aux gènes et les interactions ont été modélisées par des liens reliant les blocs concernés. Les analyses n'ont pas révélé de nouvelles voies de recherches potentiellement intéressantes mais elles ont permis d'aborder le problème de la prise en compte des interactions d'une nouvelle manière. Les SNPs auraient également pu être regroupés en fonction de leur appartenance aux chromosomes ce qui aurait réduit le nombre de blocs à 23. Cette approche peut donc s'adapter en fonction de la problématique posée et du but de l'étude.

2. Limitations rencontrées

Au cours de ce travail de recherche nous avons été confrontés à certaines limitations qui ouvrent de nouvelles perspectives de recherches. Une de ces limites concerne le choix du paramètre de pénalisation. Réalisé ici à partir d'une approche "ad-hoc", ce choix peut être discuté. La validation croisée aurait pu être préférée (car reposant sur un critère fixe et objectif) mais le temps de calcul nécessaire s'est avéré trop long. Une des perspectives envisagées consiste à optimiser le critère de validation croisée pour réduire les temps de calculs. D'autre part le paramètre de pénalisation, tel qu'il a été défini dans le chapitre 2, peut prendre des valeurs différentes en fonction des axes. Dans ce travail nous avons choisi de considérer une seule et même valeur pour l'ensemble des axes mais l'intérêt d'en choisir plusieurs en fonction des axes peut être discuté en fonction du contexte. Lorsque l'on a une idée a priori du nombre de variables que l'on souhaite conserver par axe (a priori souvent donné par les biologistes), la deuxième possibilité peut être préférée.

L'ACM sparse a été développée afin d'être utilisée sur l'ensemble de la base de données de SNPs. Cependant le temps de calcul nécessaire à l'obtention de résultats étant trop longs, nous avons été contraints de réaliser une pré-sélection (supervisée) des données et de travailler sur le jeu de données réduit. La réalisation de cette pré-sélection peut cependant

être discutée car elle peut donner lieu à des résultats sensiblement biaisés ("surplus d'optimisme" ou "over-optimism" en anglais) et à des conclusions trop optimistes sur l'efficacité d'une méthode (Jelizarow et al. [2010] ; Mehta et al. [2004]).

3. Perspectives

Extensions possibles

L'application de l'ACM sparse présentée dans le chapitre 4 a été réalisée sur un ensemble réduit du jeu de données initial. Cependant les méthodes "sparse" prennent tout leur sens dans un contexte de données de très grande dimension. L'une des principales perspectives envisagée est donc de réduire les temps de calculs de l'algorithme afin de réaliser l'ACM sparse sur l'ensemble des données (environ 800 000 SNPs).

L'ACM sparse a été utilisée ici comme une méthode de sélection d'une ou plusieurs variables catégorielles, cependant les applications possibles de cette méthode sont nombreuses. En effet, elle peut être généralisée dans le cas où les données sont structurées par blocs a priori. Comme dans le cas de la GSPCA, elle permet la sélection ou l'élimination de blocs entiers de variables lorsqu'ils sont définis a priori. Ceci pourrait s'avérer utile dans le cas où nous souhaiterions regrouper les SNPs en fonction de leur appartenance aux gènes dans la base de départ. Des blocs entiers de SNPs, et donc des gènes, seraient éliminés sur chacun des axes de l'ACM sparse ce qui faciliterait, à plus grande échelle, l'interprétation des résultats. Cependant, dans ce contexte, l'ACM sparse ne permettrait pas la sélection au sein même d'un bloc conservé à cause de la pénalisation group LASSO utilisée dans l'algorithme (voir section 2.3). Ainsi afin de permettre la sélection de variables au sein d'un bloc (lui-même sélectionné), une extension de cette méthode pourra être réalisée en remplaçant la fonction de pénalisation group LASSO par la sparse group LASSO développée par Simon et al. [2013] lorsque le contexte et l'intérêt biologique s'y prête.

Par ailleurs l'approche "sparse" offre des perspectives d'extension à d'autres méthodes d'analyses de données, telle que l'Analyse Factorielle des Correspondances Multiples (AFCM). L'Analyse Factorielle des Correspondances (AFC) est une méthode factorielle qui ne concerne que deux caractères (2 questions) d'une population de I individus. Lorsque celle-ci est caractérisée par plusieurs caractères dans ce cas on utilise une extension de l'AFC que l'on appelle l'AFCM. Dans le cas où le nombre de variables serait trop important pour exploiter facilement les résultats, il pourrait être intéressant d'utiliser une version pénalisée de cette méthode qui permettrait de sélectionner des modalités de variables (ou blocs entiers de modalités) et de faciliter l'interprétation des résultats. Le tableau de départ se présente souvent sous la forme d'un tableau disjonctif complet. Il serait donc intéressant d'adapter les critères sur lesquels sont basées la GSPCA et l'ACM sparse afin d'obtenir une version "sparse" de l'AFCM.

L'étude des interactions SNP-SNP a été réalisée au niveau des gènes durant ce travail de thèse mais cette approche pourrait être étendue au niveau des "pathways" (ou voies de signalisation), dont l'intérêt est de plus en plus évoqué dans des études GWAS (Wang et al. [2009]). Cela pourrait être biologiquement intéressant car cette approche utilise un ensemble de SNPs à l'intérieur de la même voie de signalisation et non à l'intérieur d'un même gène. L'espace logique de recherche des interactions augmenterait sensiblement mais pourrait être restreint grâce à des critères biologiques bien choisis. Cette analyse au niveau des "pathways" pourrait permettre de fournir des indications précieuses sur des interactions génétiques susceptibles de modifier les risques des phénotypes (Dinu et al. [2012]). Par ailleurs, un autre codage des SNPs pourrait être testé afin de comparer les résultats obtenus entre eux.

Une notion importante n'a pas été intégrée à ce travail pour l'instant : celle du déséquilibre de liaison (Reich et al. [2001]). Le déséquilibre de liaison est l'association non aléatoire des allèles de deux ou plusieurs loci polymorphes sur le même chromosome. Il peut être influencé par divers phénomènes : liaison génétique, taux de recombinaison, hétérogénéité spatiale des phénomènes de recombinaison, etc. Les SNPs peuvent être corrélés à cause de ce déséquilibre. Les études d'associations cherchent les associations directes et indirectes

avec le phénotype en question. Dans les analyses de forêts aléatoires (Random Forest), la corrélation entre un SNP "à risque" et un SNP en déséquilibre de liaison peut mener à une diminution de l'importance de la variable pour un SNP "à risque réel". Pour pallier ce problème, Meng et al. [2009] ont proposé une approche consistant à sélectionner les SNPs en équilibre de liaison pour l'analyse. Ils explorent des méthodes alternatives de traitement des données en déséquilibre de liaison modifiant l'algorithme de création des arbres dans les forêts aléatoires. Les résultats obtenus sont meilleurs lorsqu'il existe un déséquilibre de liaison entre les SNPs. Ces techniques sont cependant limitées dans le nombre de SNPs qu'elles peuvent traiter. La considération de ces déséquilibres pourraient donc être pris en compte dans la régression logique afin d'améliorer les performances de cette méthode dans le cas d'analyse de SNPs candidats par exemple (base de données réduite).

Applications

Bien que la GSPCA n'ait pu être appliquée sur nos données (car catégorielles) cette méthode peut avoir une utilité dans le traitement de données d'expression de gènes. Les variables considérées dans ce cas étant quantitatives, elles pourraient être regroupées en fonction de l'appartenance des gènes aux chromosomes par exemple et la GSPCA permettrait une sélection de blocs de gènes et la réalisation de clusters de gènes utiles à la compréhension de certains phénomènes biologiques.

Package R et publications

Un article présentant les deux nouvelles méthodes (GSPCA et ACM sparse) illustrées par des exemples va être soumis au journal *Computational Statistics & Data Analysis (CSDA)* et un package R permettant l'utilisation de ces deux méthodes y sera attaché et soumis au CRAN (article en cours de finalisation présenté en annexe F.2). Par ailleurs les résultats obtenus avec l'ACM sparse sont actuellement en cours d'analyse par les biologistes de l'équipe et feront l'objet d'une publication dans laquelle la méthode sera présentée et des interprétations biologiques plus approfondies seront proposées.

Bibliographie

Abdi, H. (2003). Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage : Thousand Oaks, CA*, pages 792–795.

Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition (gsvd). *Encyclopedia of measurement and statistics*, pages 907–912.

Abdi, H. and Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, pages 651–657.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6) :716–723.

Allen, D. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3) :469–475.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18) :10101–10106.

Bach, F. R. (2008). Bolasso : model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

Bardos, M. (2001). Analyse discriminante, ed. *Dunod, Paris*.

Beaton, D., Filbey, F., and Abdi, H. (2013). Integrating partial least squares correlation and correspondence analysis for nominal data. *In Abdi, H. and Chin, W. W. and Esposito Vinzi, V. et al., New Perspectives in Partial Least Squares and Related Methods*.

Benzécri, J. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult.]. *Cahiers de l'Analyse des Données*, 4(3) :377–378.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2) :123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32.

Breiman, L., Friedman, J., Olshen, R. A., et al. (1984). *Classification and regression trees*. Wadsworth and Brooks, Monterey, CA.

Cario-Andre, M., Lepreux, S., Pain, C., et al. (2004). Perilesional vs. lesional skin changes in senile lentigo. *Journal of cutaneous pathology*, 31(6) :441–447.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. Plenum press New York.

Chen, C. C., Schwender, H., Keith, J., et al. (2011). Methods for identifying snp interactions : a review on variations of logic regression, random forest and bayesian logistic regression. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(6) :1580–1591.

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Collins, F., Lander, E., Rogers, J., et al. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945.

Cooper, D. and Krawczak, M. (1993). *Human gene mutation*. Bios scientific publishers Oxfordshire. UK.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3) :273–297.

De Micheaux, P., Drouilhet, R., and Liquet, B. (2011). *Le logiciel R : Maîtriser le langage Effectuer des analyses statistiques*. Springer.

Dinu, I., Mahasirimongkol, S., Liu, Q., et al. (2012). Snp-snp interactions discovered by logic regression explain crohn's disease genetics. *PloS one*, 7(10) :e43035.

Dray, S. and Dufour, A.-B. (2007). The ade4 package : implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4) :1–20.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2) :407–499.

Elfakir, A., Ezzedine, K., Latreille, J., et al. (2009). Functional mc1r-gene variants are associated with increased risk for severe photoaging of facial skin. *Journal of Investigative Dermatology*, 130(4) :1107–1115.

Ezzedine, K., Mauger, E., Latreille, J., et al. (2012). Freckles and solar lentigines have different risk factors in caucasian women. *Journal of the European Academy of Dermatology and Venereology*.

Falush, D., Stephens, M., and Pritchard, J. (2003). Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics*, 164(4) :1567–1587.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1.

Fritsch, A. and Ickstadt, K. (2007). Comparing logic regression based methods for identifying snp interactions. In *Bioinformatics research and development*, pages 90–103. Springer.

Gauss, C. F. (1855). *Méthode des moindres carrés : mémoires sur la combinaison des observations*. Mallet-Bachelier.

Gemperline, P., Miller, K., West, T., et al. (1992). Principal component analysis, trace elements, and blue crab shell disease. *Analytical Chemistry*, 64(9) :523A–532A.

Gibbs, R., Belmont, J., Hardenbol, P., et al. (2003). The international hapmap project. *Nature*, 426(6968) :789–796.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 31(121) :124–126.

Goldberg, D., Corruble, V., Ganascia, J.-G., et al. (1994). *Algorithmes génétiques : exploration, optimisation et apprentissage automatique*. Addison-Wesley France.

Golub, G. H. and Van Loan, C. F. (1983). Matrix computations. *Johns Hopkins University, Press, Baltimore, MD, USA*.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

González, I., Déjean, S., Martin, P., et al. (2009). Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(02) :173–199.

Goralczyk, R. and Wertz, K. (2009). Skin photoprotection by carotenoids. In *Carotenoids*, pages 335–362. Springer.

Greenacre, M. (2010). *Correspondence analysis in practice*. Chapman and Hall/CRC.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*.

Guinot, C., Malvy, D., Latreille, J., et al. (2001). Sun exposure behaviour of a general adult population in france. *Skin and environment–Perception and protection. 10e congrès de l'EADV, Munich*, pages 10–14.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706) :49–50.

Hercberg, S., Galan, P., Preziosi, P., et al. (1998a). Background and rationale behind the su. vi. max study, a prevention trial using nutritional doses of a combination of antioxidant

vitamins and minerals to reduce cardiovascular diseases and cancers. supplementation en vitamines et mineraux antioxydants study. *International Journal for Vitamin and Nutrition Research*, 68(1) :3–20.

Hercberg, S., Preziosi, P., Briançon, S., et al. (1998b). A primary prevention trial using nutritional doses of antioxidant vitamins and minerals in cardiovascular diseases and cancers in a general population : the su. vi. max study—design, methods, and participant characteristics. *Controlled Clinical Trials*, 19(4) :336–351.

Higham, N. J. (1988). *Matrix nearness problems and applications*. University of Manchester Department of Mathematics.

Hodgson, C. (1963). Senile lentigo. *Archives of Dermatology*, 87(2) :197–207.

Hoerl, A. and Kennard, R. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67.

Hoerl, A. and Kennard, R. (1988). Ridge regression, in 'encyclopedia of statistical sciences', vol. 8.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 42(1) :80–86.

Holter, N. S., Maritan, A., Cieplak, M., et al. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences*, 98(4) :1693–1698.

Holter, N. S., Mitra, M., Maritan, A., et al. (2000). Fundamental patterns underlying gene expression profiles : simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15) :8409–8414.

Husson, F., Josse, J., Le, S., et al. (2013). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.24.

Inc, S. I. (1990). *SAS/STAT user's guide*.

Jdid, R., Ezzedine, K., Latreille, J., et al. (2013). Mc1r major variants are a risk factor of sleep lines in caucasian women. *Journal of the European Academy of Dermatology and Venereology*.

Jelizarow, M., Guillemot, V., Tenenhaus, A., et al. (2010). Over-optimism in bioinformatics : an illustration. *Bioinformatics*, 26(16) :1990–1998.

Jensen, L., Kuhn, M., Stark, M., et al. (2009). String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1) :D412–D416.

Jessup, E. and Sorensen, D. C. (1994). A parallel algorithm for computing the singular value decomposition of a matrix. *Siam Journal on Matrix Analysis and Applications*, 15(2) :530–548.

Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

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.

Jolliffe, I. T. (1995). Rotation of principal components : choice of normalization constraints. *Journal of Applied Statistics*, 22(1) :29–35.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3) :187–200.

Kanehisa, M., Goto, S., Kawashima, S., et al. (2004). The kegg resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1) :D277–D280.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, pages 119–127.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting snps using monte carlo logic regression. *Genetic Epidemiology*, 28(2) :157–170.

Kooperberg, C. and Ruczinski, I. (2012). *LogicReg : Logic Regression*. R package version 1.5.3.

Krämer, N. (2007). *Analysis of high-dimensional data with partial least squares and boosting*. PhD thesis, TU Berlin.

Krishnan, A., Williams, L., McIntosh, A., et al. (2011). Partial least squares (pls) methods for neuroimaging : a tutorial and review. *Neuroimage*, 56(2) :455–475.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 5. California.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., et al. (2008). A sparse pls for variable selection when integrating omics data. *Genetics and Molecular Biology*, 7(1) :35.

Lander, E., Linton, L., Birren, B., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.

Larnier, C., Ortonne, J.-P., Venot, A., et al. (1994). Evaluation of cutaneous photodamage using a photographic scale. *British Journal of Dermatology*, 130(2) :167–173.

Latreille, J., Ezzedine, K., Elfakir, A., et al. (2009). Mc1r gene polymorphism affects skin color and phenotypic features related to sun sensitivity in a population of french adult women. *Photochemistry and photobiology*, 85(6) :1451–1458.

Latreille, J., Ezzedine, K., Elfakir, A., et al. (2011). Polymorphismes du mc1r et photovieillissement du visage. In *Annales de dermatologie et de vénéréologie*, volume 138, pages 385–389. Elsevier.

Le Clerc, S., Limou, S., Coulonges, C., et al. (2009). Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for aids (anrs genomewide association study 03). *Journal of Infectious Diseases*, 200(8) :1194–1201.

Le Clerc, S., Taing, L., Ezzedine, K., et al. (2012). A genome-wide association study in caucasian women points out a putative role of the stxbp5l gene in facial photoaging. *Journal of Investigative Dermatology*, 133(4) :929–935.

Lebart, L., Morineau, A., and Tabard, N. (1977). *Technique de la description statistique : Methodes et logiciels pour l'analyse des grands tableaux*. Dunod.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3) :18–22.

Limou, S., Coulonges, C., Herbeck, J., et al. (2010). Multiple-cohort genetic association study reveals cxcr6 as a new chemokine receptor involved in long-term nonprogression to aids. *Journal of Infectious Diseases*, 202(6) :908–915.

Limou, S., Le Clerc, S., Coulonges, C., et al. (2009). Genomewide association study of an aids-nonprogression cohort emphasizes the role played by hla genes (anrs genomewide association study 02). *Journal of Infectious Diseases*, 199(3) :419–426.

Lohmöller, J. (1989). *Latent variable path modeling with partial least squares*. Physica-Verlag Heidelberg.

Magnus, J. R. and Neudecker, H. (1988). Matrix differential calculus with applications in statistics and econometrics.

Mallows, C. (1973). Some comments on cp. *Technometrics*, 15(4) :661–675.

Malvy, D., Guinot, C., Preziosi, P., et al. (2000). Epidemiologic determinants of skin photoaging : baseline data of the su. vi. max. cohort. *Journal of the American Academy of Dermatology*, 42(1) :47–55.

McCluskey, E. (1956). Minimization of boolean functions. *Bell System Technical Journal*, 35 :1417–1444.

McLachlan, G., Do, K.-A., and Ambroise, C. (2005). *Analyzing microarray gene expression data*, volume 422. Wiley. com.

Mehta, T., Tanik, M., and Allison, D. (2004). Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature genetics*, 36(9) :943–947.

Meier, L. (2013). *grplasso : Fitting user specified models with Group Lasso penalty*. R package version 0.4-3.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(4) :417–473.

Meng, Y. A., Yu, Y., Cupples, L. A., et al. (2009). Performance of random forest when snps are in linkage disequilibrium. *BMC bioinformatics*, 10(1) :78.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Meyer, D., Dimitriadou, E., Hornik, K., et al. (2012). *e1071 : Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-1.

Mi, H., Muruganujan, A., and Thomas, P. (2013). Panther in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1) :D377–D386.

Morgan, J. and Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302) :415–434.

Noah, S., Friedman, J., Trevor, H., et al. (2013). *SGL : Fit a GLM (or cox model) with a combination of lasso and group lasso regularization*. R package version 1.1.

Novembre, J., Johnson, T., Bryc, K., et al. (2008). Genes mirror geography within europe. *Nature*, 456(7218) :98–101.

Petersen, K. B. and Pedersen, M. S. (2006). The matrix cookbook.

Price, A., Patterson, N., Plenge, R., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8) :904–909.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). Plink : a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3) :559–575.

Quine, W. (1952). The problem of simplifying truth functions. *The American Mathematical Monthly*, 59(8) :521–531.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabe, J., Mamelak, A., McElgunn, P., et al. (2006). Photoaging : mechanisms and repair. *Journal of the American Academy of Dermatology*, 55(1) :1–19.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments : application to sporulation time series. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 455. NIH Public Access.

Reich, D. E., Cargill, M., Bolk, S., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834) :199–204.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12(3) :475–511.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004). Exploring interactions in high-dimensional genomic data : an overview of logic regression, with applications. *Journal of Multivariate Analysis*, 90(1) :178–195.

Ruczinski, I., Kooperberg, C., LeBlanc, M. L., et al. (2001). Sequence analysis using logic regression. *Genetic Epidemiology*, 21(1) :S626–S631.

Sage, E., Girard, P.-M., and Francesconi, S. (2012). Unravelling uva-induced mutagenesis. *Photochemical & Photobiological Sciences*, 11(1) :74–80.

Saporta, G. (2006). *Probabilités, analyses des données et statistiques*. Editions Technip.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica : Journal of the Econometric Society*, pages 1273–1291.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1) :32.

Schork, N., Fallin, D., and Lanchbury, J. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics*, 58(4) :250–264.

Schott, J. R. (2005). Matrix analysis for statistics.

Schwender, H. (2007a). Minimization of boolean expressions using matrix algebra. Technical report, SFB 475, Department of Statistics, TU Dortmund University.

Schwender, H. (2007b). *Statistical analysis of genotype and gene expression data*. PhD thesis, Department of Statistics, University of Dortmund.

Schwender, H. (2013a). *logicFS : Identification of SNP Interactions*. R package version 1.28.1.

Schwender, H. (2013b). *logicFS : Identification of SNP Interactions*. R package version 1.28.1.

Schwender, H. and Ickstadt, K. (2008). Identification of snp interactions using logic regression. *Biostatistics*, 9(1) :187–198.

Schwender, H., Zucknick, M., Ickstadt, K., et al. (2004). A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicology Letters*, 151(1) :291–299.

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

Sherry, S., Ward, M.-H., Kholodov, M., et al. (2001). dbsnp : the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1) :308–311.

Simon, N., Friedman, J., Hastie, T., et al. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2) :231–245.

Sing, T., Sander, O., Beerenwinkel, N., et al. (2005). Rocr : visualizing classifier performance in r. *Bioinformatics*, 21(20) :7881.

Stamey, T., Kabalin, J., McNeal, J., et al. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of Urology*, 141(5) :1076–1083.

Stark, C., Breitkreutz, B., Reguly, T., et al. (2006). Biogrid : a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1) :D535–D539.

Strang, G. (2003). *Introduction to linear algebra*. SIAM.

Team, T. F. S. P. (2009). *CHAID : CHi-squared Automated Interaction Detection*. R package version 0.1-1.

Tenenhaus, A. and Guillemot, V. (2013). *RGCCA : RGCCA and Sparse GCCA for multi-block data analysis*. R package version 2.0.

Tenenhaus, A., Philippe, C., Guillemot, V., et al. (2013). Variable selection for generalized canonical correlation analysis. *Biostatistics*.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2) :257–284.

Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Editions Technip.

Tenenhaus, M. (2007). *Statistique : Méthodes pour décrire, expliquer et prévoir*. Dunod.

Tenenhaus, M., Vinzi, V., Chatelin, Y.-M., et al. (2005). Pls path modeling. *Computational statistics & data analysis*, 48(1) :159–205.

Therneau, T., Atkinson, B., and Ripley, B. (2013). *rpart : Recursive Partitioning*. R package version 4.1-3.

Thurstone, L. L. (1947). Multiple factor analysis.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Troyanskaya, O., Cantor, M., Sherlock, G., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6) :520–525.

Tucker, L. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23(2) :111–136.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer, New York, fourth edition. ISBN 0-387-95457-0.

Venter, J., Adams, M., Myers, E., et al. (2001). The sequence of the human genome. *Science*, 291(5507) :1304–1351.

Vines, S. (2000). Simple principal components. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 49(4) :441–451.

Vinzi, V. E. (2010). *Handbook of partial least squares : Concepts, methods and applications.* Springer.

Wallace, D. (1997). Adn mitochondrial, maladies et vieillissement. *Pour la science*, (240) :52–61.

Wang, K., Zhang, H., Kugathasan, S., et al. (2009). Diverse genome-wide association studies associate the il12/il23 pathway with crohn disease. *The American Journal of Human Genetics*, 84(3) :399–405.

Weinberg, W. (1908). On the demonstration of heredity in man. *Naturkunde*, 64 :368–382.

Wigginton, J., Cutler, D., and Abecasis, G. (2005). A note on exact tests of hardy-weinberg equilibrium. *The American Journal of Human Genetics*, 76(5) :887–893.

Winham, S., Colby, C., Freimuth, R., et al. (2012). Snp interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, 13(1) :164.

Witte, J. and Fijal, B. (2001). Introduction : analysis of sequence data and population structure. *Genetic Epidemiology*, 21 :S600.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 1 :391–420.

Wold, H. (1985). Partial least squares. *in S. Kotz and N. L. Johnson (Eds.), Encyclopedia of Statistical Sciences*, 6 :581–591.

Wold, S., Josefson, M., Gottfries, J., et al. (2004). The utility of multivariate design in pls modeling. *Journal of chemometrics*, 18(3-4) :156–165.

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method. *Matrix pencils*, pages 286–293.

Wold, S., Sjöström, M., and Eriksson, L. (2001). Pls-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2) :109–130.

Yaar, M. and Gilchrest, B. (2007). Photoageing : mechanism, prevention and therapy. *British Journal of Dermatology*, 157(5) :874–887.

Yanai, H., Takeuchi, K., and Takane, Y. (2011). *Projection Matrices*. Springer.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9) :763–774.

Yeung, M. S., Tegnér, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9) :6163–6168.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67.

Zou, H. (2005). *Some perspectives of sparse statistical modeling*. PhD thesis, Citeseer.

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.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2) :265–286.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots : a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4) :561–577.

Liste des publications

Parue (annexe F.1)

Le Clerc, S., Taing, L., Ezzedine, K., Latreille, J., Labib, T., Coulonges, C., Bernard, A., Melak, S., Carpentier, W., Malvy, D., Jdid, R., Galan, P., Hercberg, S., Morizot, F., Guinot, G., Tschachler, E., Zagury J.-F. (2013). A genome-wide association study in Caucasian women points out for a role of the STXBP5L gene in facial photoageing. *Journal of Investigative Dermatology*, 133, p.929-935.

A soumettre (annexe F.2)

Bernard, A., Abdi, H., Tenenhaus, A., Guinot, C., Saporta, G. Sparse Principal Component Analysis for multiblocks data and its extension to Sparse Multiple Correspondence. *Computational Statistics & Data Analysis.*

A soumettre

Bernard, A., Gardinier, S., Mallet de Chauny, E., Latreille, J., Staner, L., Cornette, F., Pross, N., Metzger, D., Tenenhaus, M., Tschachler, E., Guinot, C., Morizot, F. Effects of a 6-day 4-hour sleep restriction protocol on biophysical skin properties in healthy women. *Skin Research and Technology.*

Liste des communications

Communications orales

Congrès internationaux

Bernard, A., Latreille, J., Gardinier, S., Cornette, F. , Mallet de Chauny, E. , Pross, N., Staner, L., Metzger, D., Tenenhaus, M., Morizot, F., Tschachler, E., Guinot, C. Effect of chronic sleep deprivation on skin status in healthy young women. 14th Applied Stochastic Models and Data Analysis International Conference (ASMDA), 7-10 juin 2011, Rome, Italie. Résumé : CDROM.

Leclerc, S., Tiang, L., Bernard, A., Latreille, J., Ezzedine, K., Malvy, D., Jdid, R., Galan, P., Hercberg, S., Zagury, J.-F., Tschachler, E., Guinot, C. A genome-wide association study on 520 adult Caucasian women identifies a gene associated with facial photoageing. 41st Annual Meeting of the European Society for Dermatological Research (ESDR), 7-10 septembre 2011, Barcelone, Espagne. Résumé : Journal of Investigative Dermatology 2011, 131 :S52, 307.

Morizot, F., Latreille, J., Gardinier, S., Staner, S., Guinot, C., Bernard, A., Porcheron, A., Tschachler, E. Effects of partial sleep deprivation on face appearance and on skin properties. 41st Annual Meeting of the European Society for Dermatological Research (ESDR), 7-10 septembre 2011, Barcelone, Espagne. Résumé : Journal of Investigative Dermatology 2011, 131 :S61, 365.

Bernard, A., Saporta, G., Guinot, C. Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. 20th International Conference on COMPUTATIONAL STATISTICS (Compstat 2012), 27-31 août 2012, Limassol, Chypre. Abstract : Proceedings, p.99-106.

Bernard, A., Guinot, C., Saporta, G. A generalisation of sparse PCA to multiple correspondence analysis. 5th International Conference of the ERCIM WG on Computing & Statistics, 1-3 décembre 2012, Oviedo, Espagne. Résumé : Program and Abstracts, p.33.

Leclerc, S., Tiang, L., Ezzedine, K., Bernard, A., Latreille, J., Malvy, D., Jdid, R., Galan, P., Hercberg, S., Morizot, F., Guinot, C., Tschachler, E., Zagury, J.-F. A genome-wide approach to skin ageing. International Investigative Dermatology (IDD 2013), 8-11 mai 2013, Edimbourg, Ecosse.

Congrès nationaux

Bernard, A. Développement de méthodes statistiques pour le traitement de données génomiques. 4ème Rencontres des Jeunes Statisticiens, 5-9 septembre 2011, Aussois, France.

Bernard, A., Tenenhaus, A., Zagury, J.-F., Saporta, G., Guinot, C. Méthodes multiblocs pour l'identification de gènes associés au vieillissement cutané chez 502 femmes Caucasiennes adultes. XXXXIVe Journées de Statistique, 21-25 mai 2012, Bruxelles, Belgique. Résumé : Programme des journées, p.53.

Morizot, F., Bernard, A., Latreille, J., Gardinier, S., Porcheron, A., Staner, L., Guinot, C., Tschachler, E. Effets de la privation partielle de sommeil sur l'apparence du visage et sur les propriétés biophysiques cutanées. Journées Dermatologiques de Paris, 11-15 décembre 2012, Paris, France. Résumé : Annales de Dermatologie et de Vénéréologie 2012, 139 Hors serie 3 : B247, p.287.

Leclerc, S., Taing, L., Bernard, A., Latreille, J., Ezzedine, K., Malvy, D., Jdid, R., Galan, P., Hercberg, S., Zagury, J.-F., Tschachler, E., Guinot, C. Identification d'un gène associé au photovieillissement cutané par analyse génomique (GWAS) chez 520 femmes Caucasiennes adultes. Journées Dermatologiques de Paris, 11-15 décembre 2012, Paris, France. Résumé : Annales de Dermatologie et de Vénéréologie 2012, 139 Hors serie 3 : B181, p.141.

Bernard, A., Guinot, C., Saporta, G. Analyse en Composantes Principales Sparse pour données multiblocs et extension à l'Analyse des Correspondances Multiples Sparse. XXXXVe Journées de Statistique, 27-31 mai 2013, Toulouse, France. Résumé : Programme des journées, p.62.

Posters

Bernard, A., Tenenhaus, A., Zagury, J.-F., Saporta, G., Guinot, C. Méthodes multiblocs pour l'identification de gènes associés au vieillissement cutané chez 502 femmes Caucasiennes adultes. 12ème Journées Francophones d'Extraction et Gestion des Connaissances (EGC'2012), 31 janvier au 3 février 2012, Bordeaux, France. Dans : Revue des Nouvelles Technologies de l'Information (RNTI E.23), (Guy Melançon, Bruno Pinaud, Yves Lechevallier, editors), ISBN 978 2 7056 8310 8, Edition Herman, Paris, p.555-556.

Annexes

Annexe A

Peau et vieillissement cutané

A.1 Description

La peau est une enveloppe qui protège l'individu. Cet organe se compose de trois couches : l'épiderme (couche supérieure), le derme et l'hypoderme. La représentation schématique de la coupe de peau présentée figure A.1 est tirée de "http ://www.esthetique.qc.ca/services_fr/peau/schema_peau.html".

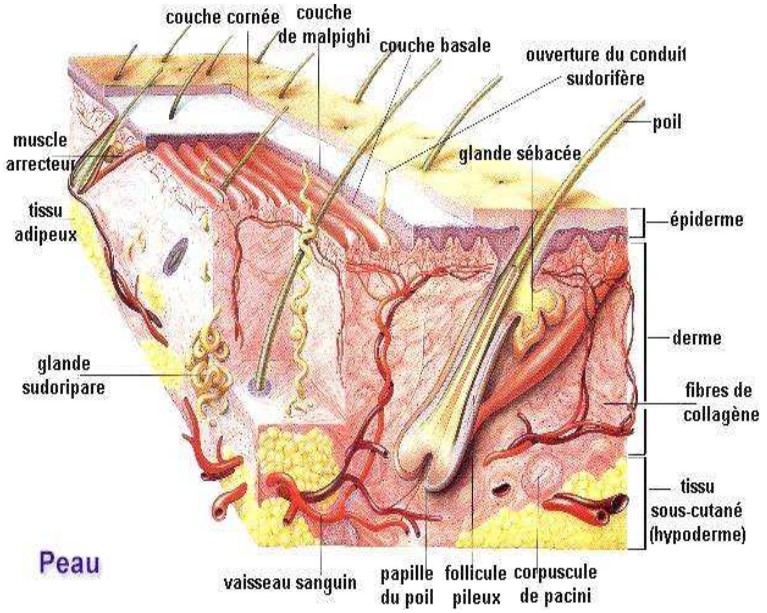

FIGURE A.1 – Représentation schématique d'une coupe de peau

L'épiderme est constitué de kératinocytes, de mélanocytes et de cellules de Langerhans.

Les kératinocytes sont présents dans toutes les couches : de la couche basale, zone de multiplication de ces cellules, aux couches les plus superficielles. Les kératinocytes migrent progressivement vers la surface, ce phénomène étant accompagné d'une modification de leur composition pour aboutir à la couche cornée. La couche cornée est constituée de cellules entièrement kératinisées sans noyaux, séparées les unes des autres par des lipides. En surface, les cellules de la couche cornée perdent leur adhérence puis tombent (phénomène de desquamation). Les mélanocytes sont uniquement situés dans la couche basale. Ces cellules ont une forme étoilée, leurs dendrites étant en contact avec les kératinocytes de leur voisinage. Elles fabriquent de la mélanine sous la forme de petits grains (les mélanosomes) qui migrent dans les dendrites et pénètrent dans les kératinocytes. Les cellules de Langerhans, qui sont présentes dans toutes les couches de l'épiderme, forment un réseau de cellules étoilées reliées entre elles ayant la capacité de capter toute substance étrangère. Ces cellules jouent un rôle dans l'inflammation cutanée et dans la fonction immunitaire. Le derme contient des fibres de collagène et des fibres élastiques, différentes cellules dont les fibroblastes, des vaisseaux de petit diamètre, et des nerfs "libres" et des nerfs reliés à différents corpuscules sensoriels. L'hypoderme est une couche graisseuse où se trouvent vaisseaux et nerfs.

La peau assure différentes fonctions :

- Une fonction d'échanges en intervenant dans la régulation de la température corporelle.
- Une fonction sensorielle par le biais de divers récepteurs (chaleur, froid, toucher, douleur et prurit).
- Une fonction métabolique, dont la synthèse de la vitamine D2 sous l'action des Ultra-Violets B (UVB) dans la partie profonde de l'épiderme.
- une fonction d'auto-réparation (cicatrisation) et de régulation de la teneur en eau : elle limite les risques de déshydratation et constitue une barrière efficace face aux agressions externes.
- Une fonction de souplesse et d'élasticité : l'épiderme est peu élastique, son rôle étant de protéger les couches profondes. La couche basale qui sépare l'épiderme du derme est ondulée ce qui permet la transmission des déformations de la surface de la peau jusqu'au

derme. Cette couche s'aplatit progressivement avec l'âge : elle devient plate vers 60-70 ans accompagnant une diminution de l'épaisseur de l'épiderme et une perte d'élasticité et de souplesse.

- Une fonction de protection contre les agressions extérieures : aux agressions mécaniques, chimiques et microbiennes, au rayonnement solaire et à la chaleur. La fonction de photoprotection et de bronzage peut être résumée de la façon suivante : sous l'action des rayons solaires l'épiderme s'épaissit, et les mélanocytes fabriquent les mélanosomes (phénomène de bronzage) - la qualité et la quantité de mélanine dépendant des individus (facteur génétique). Les dendrites des mélanocytes jouent un rôle de parasol en protégeant les noyaux des kératinocytes de la couche basale.

A.2 Vieillissement cutané et facteurs influençant l'aspect cutané

Le vieillissement cutané est un processus plurifactoriel complexe, qui découle de deux processus : le vieillissement chronologique encore appelé intrinsèque pouvant être considéré comme programmé génétiquement qui touche l'ensemble du revêtement cutané, et le vieillissement actinique lié à l'action néfaste des rayons ultraviolets au niveau des zones cutanées exposées. La ménopause chez la femme accentue le vieillissement en raison des modifications du statut hormonal.

A.2.1 Vieillissement intrinsèque

Le vieillissement intrinsèque est responsable de nombreuses modifications épidermiques parmi lesquelles : une diminution de l'épaisseur de la peau liée à un aplatissement de la jonction dermoépidermique et une perte des expansions dermiques, et une diminution de la teneur en lipides, du nombre de mélanocytes et de cellules de Langerhans. Le derme s'atrophie aussi avec l'âge : le nombre et la taille des fibroblastes dermiques diminuent, les fibres élastiques sont altérées et on constate entre autres choses une diminution de la microvascularisation dermique. Les manifestations cliniques majeures du vieillissement cutané chronologique concernent la formation de rides et la perte d'élasticité. Les manifestations cliniques du photovieillissement cutané font apparaître une peau épaissie, rugueuse,

jaunâtre et hyperlaxe. On note l'apparition de ridules puis de rides profondes, des télangiectasies (atteintes vasculaires au niveau du derme), des taches pigmentaires témoins des altérations des mélanocytes, et une hyperplasie sébacée constituée de multiples papules jaunes, molles, ombiliquées en leur centre. Le vieillissement hormonal est lui aussi responsable de modifications cutanées : diminution de l'épaisseur de la peau, du contenu en collagène dermique, de la prolifération des kératinocytes et des fibroblastes, de la vascularisation cutanée, de l'hydratation cutanée et de la secrétion sébacée. Ces modifications sont donc proches des conséquences du vieillissement chronologique et différentes de celles du photovieillissement.

A.2.2 Photo-vieillissement

Modifications histologique et clinique de la peau lors du photo-vieillissement

Le photo-vieillissement se superpose au vieillissement intrinsèque et engendre également des modifications de l'épiderme et du derme. L'épiderme devient irrégulier, parfois atrophié, parfois hyperplasique. Le nombre de cellules de langerhans diminue tandis que le nombre de mélanocytes hyperplasique augmente. Le tissu conjonctif dermique est altéré. La microvascularisation est déteriorée (perte des plexus papillaires avec aplatissement des crêtes papillaires mais également vaisseaux dilatés et élargis dans le derme papillaire et le derme moyen). Le collagène diminue et le tissu élastique dystrophique s'accumule. Le photo-vieillissement cutané se caractérise alors par une peau plus épaisse (élastose solaire), rugueuse, jaunâtre et hyperlaxe. Des ridules, puis des rides profondes apparaissent, des télangiectasies (reflet des altérations vasculaires au niveau du derme), de taches pigmentaires encore appelées lentigines (témoins des altérations des mélanocytes), et une hyperplasie sébacée constituée de multiples papules jaunes, molles, ombiliquées en leur centre et de kératoses actiniques, considérées comme des lésions précancéreuses.

Mécanismes biologiques liés au photo-vieillissement

L'exposition aux radiations UV est considérée comme le principal facteur responsable du vieillissement extrinsèque. Elle va être à l'origine d'une forte production d'**ERO!** qui mettra à mal les défenses antioxydantes de la peau (Rabe et al. [2006] ; Sage et al. [2012] ; Yaar

and Gilchrest [2007]). Seulement 5 à 10% du rayonnement UVB atteint la surface terrestre par rapport à 90% des Ultra-Violets A (UVA) mais il est le plus énergétique. Il pénètre principalement au niveau de l'épiderme où il induit directement des dommages ADN (dimères cyclobutane et les photoproduits 6-4) dans les kératinocytes et les mélanocytes. Le rayonnement UVA, bien que moins énergétique, est reconnu pour être également fortement impliqué dans le photo-vieillissement. Il pénètre plus profondément que les UVB dans la peau et peut atteindre le derme profond et le tissu sous cutané (Figure A.2).

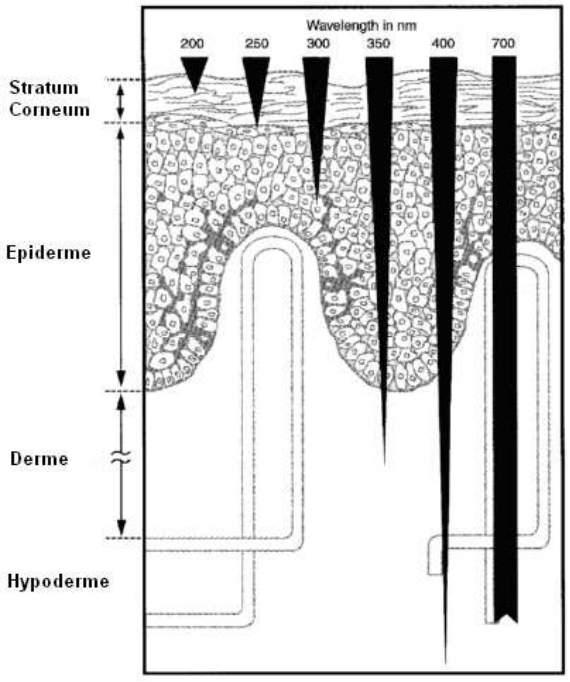

FIGURE A.2 – Couches de la peau atteintes par les UV en fonction de leur longueur d'onde (tirée de Goralczyk and Wertz [2009])

Annexe B

Bases de la génétique

Le vivant : protéines et ADN

Les protéines sont des molécules complexes qui interviennent dans tous les mécanismes moléculaires de développement et de fonctionnement du vivant. Elles sont synthétisées par les cellules et peuvent avoir des fonctions très diverses : enzymes pour catalyser les réactions biochimiques, hormones pour servir de messagers dans l'organisme, anticorps pour la défense immunitaire, et bien d'autres formes qu'illustrent les bases de données actuelles où plusieurs centaines de milliers de protéines sont déjà recensées. Elles jouent donc un rôle dans notre aspect physique, nos risques de contracter des maladies et la réponse de notre corps aux stimuli rencontrés dans l'environnement. Une protéine est une chaîne composée de molécules plus petites, appelées acides aminés, dont l'ordre d'agencement va déterminer la structure et la fonction au sein de l'organisme. Ce "collier de perles" va de quelques acides aminés dans le cas de petits peptides à plusieurs milliers pour les plus grosses protéines. Chez l'homme et les mammifères, il y a vingt acides aminés de base. Pour fabriquer une protéine, la cellule a besoin d'une part des acides aminés et d'autre part d'un schéma d'assemblage. Les acides aminés proviennent de l'alimentation ou sont néosynthétisés. Leur enchaînement au sein de la protéine est quant à lui dicté par l'information génétique. Cette information génétique est constituée d'une molécule bien particulière, l'Acide Désoxyribo-Nucléique (ADN).

Structure moléculaire de L'ADN

L'ADN est une très longue séquence de perles appelées nucléotides, qui sont de quatre natures : Adénine, Thymine, Cytosine et Guanine (ou A, T, G et C). Au début des années 50, deux découvertes majeures ont permis de mieux comprendre sa structure moléculaire. La première, menée par Chargaff, a montré que quel que soit l'espèce dont on extrait l'ADN, les quantités de nucléotides A et C sont respectivement égales à celles de T et G. Seul le rapport (A+T)/(C+G) est susceptible de changer d'une espèce à l'autre. La seconde menée par Franklin a montré par diffraction de rayons X que l'ADN devait avoir une structure hélicoïdale. Ceci permit finalement à Watson et Crick d'entrevoir que l'ADN était composée de deux séquences complémentaires de nucléotides (A <=> T et C <=> G) entrelacées au sein d'une double hélice (figure B.1). On a donc commencé à mieux comprendre ses propriétés : d'un côté, elle stocke le patrimoine génétique, et de l'autre, elle en permet des copies rapides et robustes durant les phases de réplication cellulaire (mitose). Sa structure stable permet de minimiser les erreurs de copies tout en laissant une place à l'évolution.

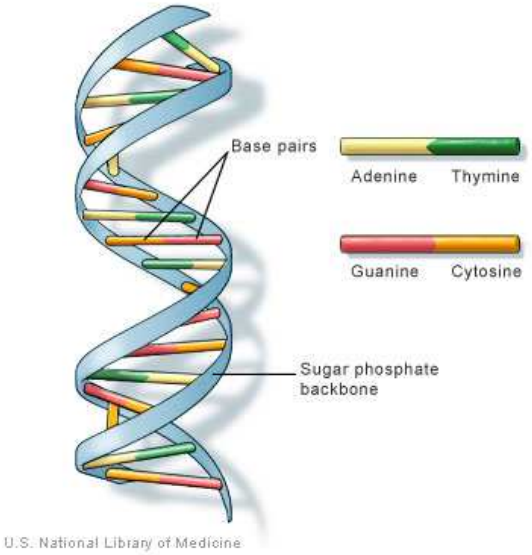

FIGURE B.1 – Représentation d'une portion de la molécule d'ADN. Les nucléotides sont appariés suivant leur complémentarité. Les deux séquences complémentaires s'entrelacent pour former une double hélice.

L'ADN, support de l'information génétique

En informatique, chaque nombre entier est codé en mémoire sur 8 bits de valeurs binaires 0 ou 1. De façon analogue, en génétique, les 20 acides aminés différents sont codés dans l'ADN sur 3 nucléotides de valeurs A, T, G ou C. Cette correspondance s'appelle le code génétique, et permet d'entrevoir le premier rôle de l'ADN : être la mémoire du vivant. L'ADN correspond donc à une très longue séquence de triplets de nucléotides dont la lecture dans un sens particulier de certaines régions code l'assemblage d'acides aminés en protéines. Ces régions codantes sont appelées les gènes (figure B.2). Chez l'homme, on en compte de l'ordre de 20 000 à 25 000. Mais paradoxalement, ils ne représentent que 1.5% des 3 milliards de nucléotides de l'ADN humain (Lander et al. [2001] ; Venter et al. [2001] ; Collins et al. [2004]).

Pour le moment, on ne comprend que partiellement le rôle des vastes régions intergéniques. Les plus connues restent les promoteurs ; des régions régulatrices que l'on trouve en amont des gènes et sur lesquelles des protéines dites "régulatrices" peuvent se fixer. Elles permettent à la cellule de contrôler l'expression des gènes. Ce niveau d'expression dépend généralement de conditions environnementales particulières, comme par exemple, les stimuli reçus par la cellule, le tissu dans lequel se trouve la cellule, ou le stade de développement de l'organisme.

Organisation de l'ADN

L'ensemble des gènes d'un individu est réparti sur plusieurs molécules d'ADN dans la cellule, chacune sous la forme compacte de chromosome. La position d'un gène est donc spécifiée par le chromosome auquel il appartient et par la distance en nucléotides qui sépare le début du chromosome du début du gène (figure B.3). Cette distance est exprimée en paires de bases (base pairs (bp)) : 1b pour 1 nucléotide, 1 Kilo base (Kb) pour 1000 nucléotides et 1 Million base (Mb) pour 106 nucléotides. L'ensemble des chromosomes d'un individu, donc l'ensemble de ses gènes, forme son génome. Dans l'exemple, Le gène débute à peu près à 50 101 Mb et fini à 50 105 Mb : il fait donc 4 Kb soit 4 000 nucléotides.

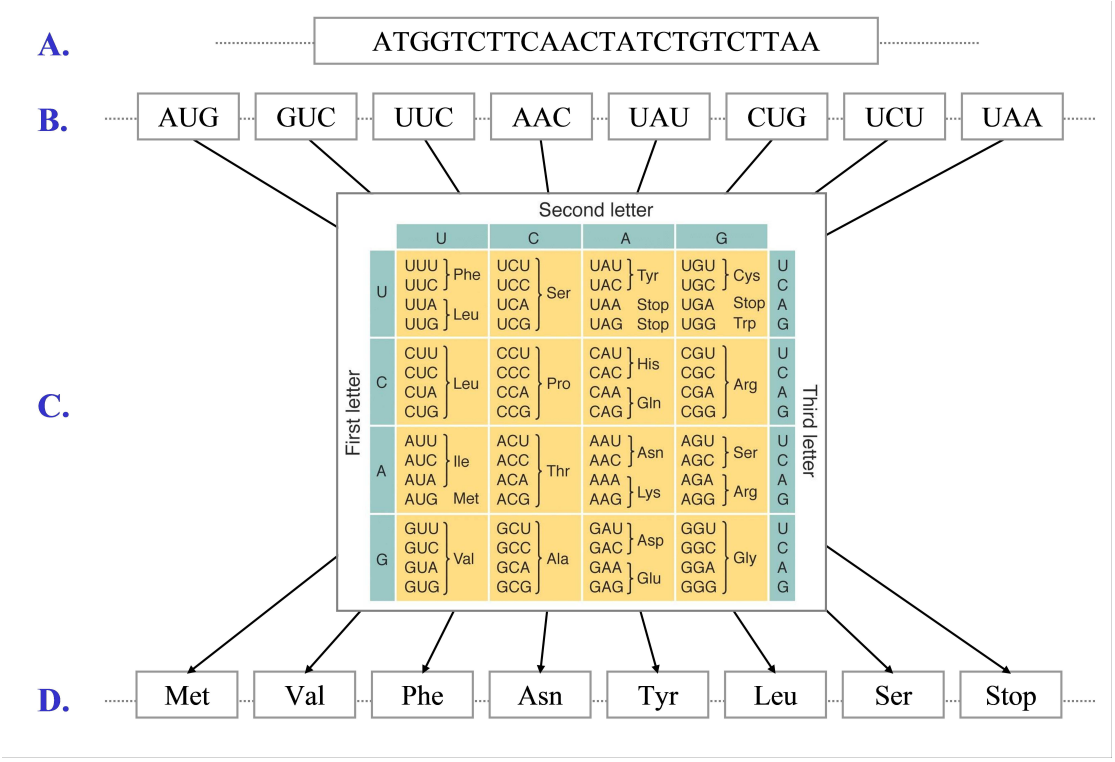

FIGURE B.2 – Passage des triplets de nucléotides à la protéine. A. La séquence d'ADN initiale. B. La séquence sous forme d'ARN : elle commence par un triplet d'initialisation (AUG) et fini par un triplet stop (UAA). C. Le code génétique qui permet le passage de l'ARN en acides aminés. D. Les acides aminés sont assemblés en protéine suivant l'ordre codé par l'ADN.

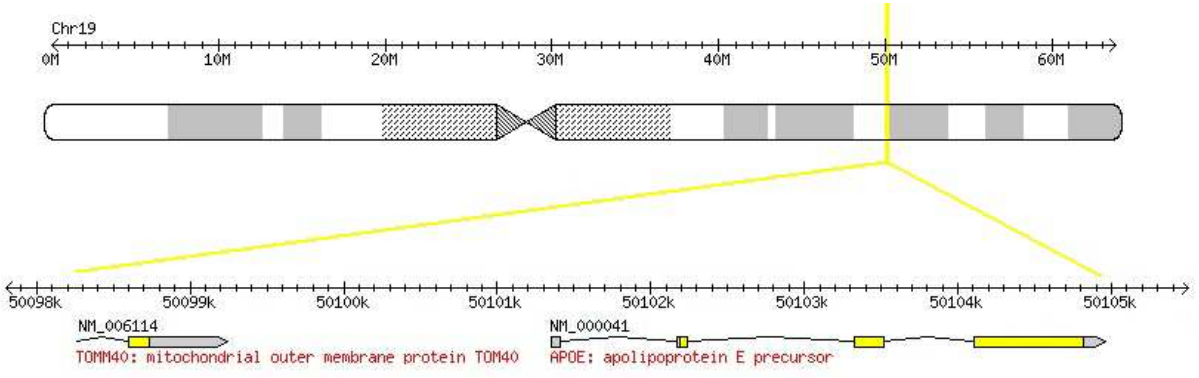

FIGURE B.3 – Carte génétique centrée sur le gène **ApoE!** du chromosome 19

.

L'ADN, support de l'hérédité

Le génome humain est diploïde car il est composé de 22 paires de chromosomes autosomiques et d'une paire de chromosomes sexuels (XX pour les femmes et XY pour les hommes). Lors de la reproduction, chaque parent transmet à ses enfants une version haploïde de son génome ; c'est-à-dire un chromosome de chacune de ses 23 paires. Le génome de l'enfant est ensuite formé par l'union des chromosomes parentaux transmis (figure B.4).

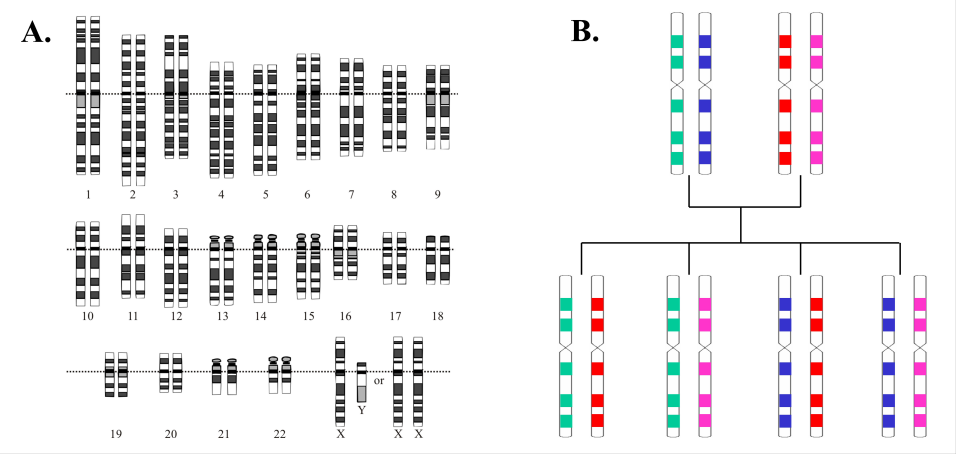

FIGURE B.4 – Les chromosomes chez l'homme. A. Représentation des 23 paires de chromosomes du génome humain ; numérotés par taille décroissante. B. En haut, les chromosomes parentaux, et en bas, les 4 combinaisons possibles pour former les enfants.

ADN : Support de l'évolution

Durant la production des gamètes, deux types de modifications de l'ADN peuvent survenir, créant ainsi une variabilité dans les gènes transmis à la génération suivante :

- Les mutations ; des événements très rares d'erreur dans le processus de copie de l'ADN. Par exemple, un nucléotide A d'une séquence peut être substitué par un nucléotide C (figure B.5.A).

- Les recombinaisons ; un brassage des chromosomes parentaux regroupés en paires au moment de la méiose. Dans la pratique, un parent ne transmet pas à sa descendance un chromosome entier de chaque paire, mais plutôt un chromosome hybride (figure B.5.B).

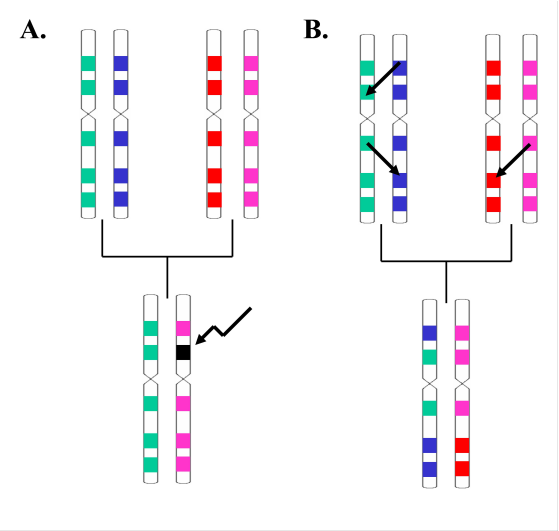

FIGURE B.5 – A. Exemple de mutation (en noir) intervenant sur un chromosome parental transmis. B. Exemple de trois recombinaisons chromosomiques.

Le polymorphisme génétique

Les mutations qui peuvent apparaître sur un chromosome vont peu à peu se diffuser dans la population avec les générations successives, surtout celles correspondant à un "avantage sélectif". Certains variants deviennent partagés par de nombreux individus alors que d'autres disparaissent. Les sites du génome où se concentrent les différences génétiques entre individus d'une même population sont les polymorphismes. On estime à l'heure actuelle qu'ils représentent 1% du génome humain. Les différentes formes que les séquences d'ADN prennent au niveau de ces polymorphismes sont les allèles. Un polymorphisme est biallèlique lorsqu'il possède deux allèles distincts dans la population. Lorsque ce nombre est plus grand, le polymorphisme est multiallèlique.

L'homme, possédant deux copies de chaque chromosome, a donc deux allèles au niveau de chaque polymorphisme : c'est son génotype. Si les deux allèles sont identiques, il est homozygote, dans le cas contraire, il est hétérozygote. Prenons un exemple : imaginons dans une population, les allèles A et a observés pour un polymorphisme biallèlique donné. Si un individu de cette population possède deux allèles identiques, son génotype est alors AA ou aa, et il donc est homozygote. Mais si l'individu possède deux allèles différents, son génotype est Aa et il est donc hétérozygote.

Dans la pratique, il existe différents types de polymorphismes, correspondant à différents types d'altérations héritables de l'ADN. On peut citer par exemple les microsatellites ; des séquences d'ADN formées par la répétition continue d'un motif de quelques nucléotides. On en trouve environ un tous les 3 à 10 Kb (Cooper and Krawczak [1993]). Ce type de polymorphisme est généralement multiallèlique car le nombre de répétitions peut beaucoup varier d'un individu à un autre. Dans la suite, nous nous intéressons plutôt à un type particulier de polymorphismes : les Single Nucléotide Polymorphisme (SNP). Ce sont les variations du génome les plus fréquemment rencontrées et les mieux réparties : on en retrouve dans tous les gènes, promoteurs, régions intergéniques, etc. Ils correspondent à de simples mutations où un nucléotide a été substitué par un autre (ex : A => T). On estime de nos jours qu'il existe 10 millions de SNP à travers le génome humain (Gibbs et al. [2003]), c'est-à-dire un tous les 300 nucléotides en moyenne. Ces 10 millions suffisent à capturer 90% de la diversité génétique observée chez l'homme. Les 10% qui restent correspondent aux microsatellites ou à des variations rares et diverses qui complexifient leur étude. Ces SNP, très nombreux et uniformément répartis sur tout le génome, constituent de nos jours le marqueur génétique de prédilection dans les études de populations humaines. Pour une revue sur les SNP, on peut se référer à l'article de Schork et al. [2000].

Séquençage de l'ADN

Toute la génétique moderne repose avant tout sur notre capacité à observer les données génétiques. On obtient les séquences d'ADN d'un individu par séquençage. Il a été inventé en 1975 indépendamment par l'équipe de Gilbert, aux États-Unis, et celle de Sanger, en Grande-Bretagne : toutes deux gratifiées du prix Nobel en 1980. Avec les années, les progrès réalisés dans les techniques de biologie moléculaire ont permis d'augmenter considérablement les débits tout en réduisant les coûts. Les volumes de données séquencés ont ainsi été démultipliés (figure B.6). Pour preuve, en 2003, le génome humain a été complètement séquencé par le consortium public international Human Genome Project et la société privée Celera genomics. Les 3 milliards de nucléotides ont été rendu disponibles à la communauté scientifique à travers de nombreuses bases de données, dont GenBank ("http ://www.ncbi.nlm.nih.gov/").

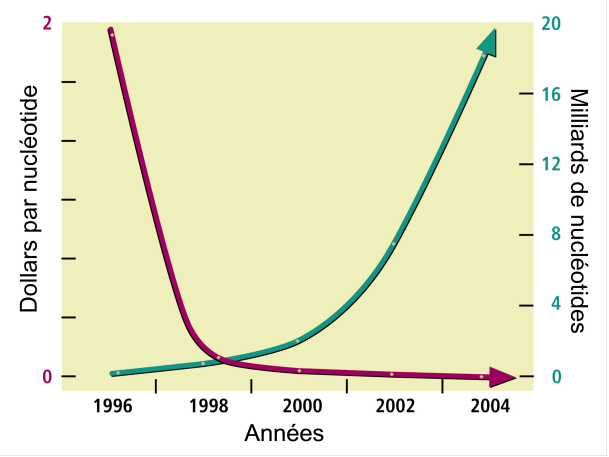

FIGURE B.6 – Représentation entre 1996 et 2004 de l'augmentation du débit de séquençage et de la baisse des coûts

Les SNP étant des marqueurs privilégiés, beaucoup d'attention est portée au génotypage, dont le but est d'obtenir les deux allèles d'un individu au niveau de certains SNP connus, c'est-à-dire le génotype (figure B.7). A l'heure actuelle, des sociétés comme Illumina et Affymetrix proposent des puces permettant d'obtenir facilement les génotypes de plusieurs dizaines d'individus sur plusieurs centaines de milliers de SNP à travers tout le génome, sans pour autant avoir besoin de séquencer systématiquement les 3 milliards de nucléotides.

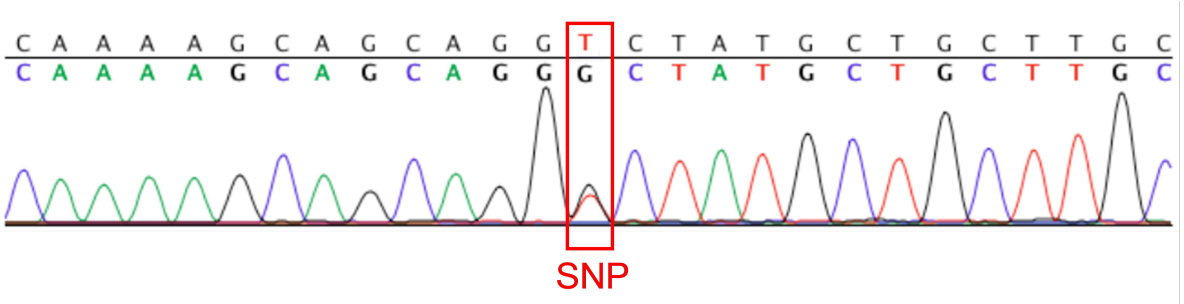

FIGURE B.7 – Chromatogramme de séquençage sur 27 nucléotides, dont un SNP hétérozygote (T/G).

Annexe C

Présentation de l'étude SU.VI.MAX et pré-traitements des données réalisés en amont de la thèse

C.1 Présentation de l'étude SU.VI.MAX

L'étude SU.VI.MAX (acronyme pour "SUpplémentation en VItamines et Minéraux AntioXydants") est une étude épidémiologique d'intervention nutritionnelle qui s'est intéressée aux grandes pathologies chroniques caractéristiques des pays industrialisés, réalisée à l'échelon national (Hercberg et al. [1998b] ; Hercberg et al. [1998a]). Les sujets ont été recrutés sur une base de volontariat en fonction du sexe, et de certaines variables (catégorie socio-professionnelle, comportement vis à vis du tabac, lieu d'habitation) permettant d'approcher la constitution de la cohorte de celle de la population générale française. L'objectif principal de l'étude SU.VI.MAX (1994-2002) était de mesurer l'impact d'une supplémentation en vitamines et minéraux antioxydants sur l'incidence des cancers et des maladies cardio-vasculaires. L'étude réalisée était randomisée en double aveugle, testant l'effet d'une supplémentation journalière en minéraux et vitamines antioxydants à des doses nutritionnelles – 1 à 3 fois les apports nutritionnels recommandés versus placebo. L'attribution du traitement était stratifiée sur le sexe, l'âge, le tabagisme et le lieu de résidence. Le nombre de sujets nécessaire a été estimé entre 12 500 et 15 000 selon les différentes hypothèses envisagées. Les trois principaux critères de jugement étaient la mortalité générale, l'incidence de cancers tous sites confondus et l'incidence des maladies cardiovasculaires is-

chémiques. L'objectif de l'étude SU.VI.MAX était également de préciser les relations entre alimentation et santé. L'étude a été approuvée par le comité d'éthique pour les études sur l'homme (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale (CCPPRB) n°706) de Paris-Cochin, et la Commision Nationale de l'Informatique et des Libertés (CNIL) (n°334641) qui implique que toutes les informations médicales sont confidentielles et anonymes.

L'inclusion des sujets adultes volontaires dans l'étude a commencé en octobre 1994. Suite à une campagne multimédia menée de mars à juin 1994 79 976 candidats se sont déclarés volontaires. Un mailing contenant une information détaillé sur l'étude SU.VI.MAX, 15 capsules du traitement, un questionnaire et une demande de consentement éclairé ont été envoyés à chacun des candidats. Seulement 1/4 des dossiers (n=21 481) ont été correctement renseignés et rentournés. Pour être définitivement éligibles, les sujets devaient être dans la tranche d'âge définie, soit [35-60] pour les femmes et [45-60] pour les hommes, se déclarer ne pas être atteint de pathologie sévère qui puisse restreindre leur participation pendant la durée de l'étude, ne prendre aucun supplément contenant des vitamines ou minéraux étudiés, ne manifester aucune inquiétude ou réticence à se conformer aux contraintes du protocole, en particulier à recevoir un placebo, et n'exprimer aucune motivation ambiguë ou comportement obsessionnel concernant l'alimentation et la santé. Après vérification des critères d'inclusion 14 412 ont été retenus. En réalité, seuls 13 017 sujets ont été inclus entre octobre 1994 et avril 1995. Parmi eux, 270 sujets (2% ; dont 115 dans le groupe antioxydant et 155 dans le groupe placebo) ont rompu leur consentement le jour de l'enrôlement au sein de la cohorte ou dans les 3 jours qui ont suivi. Tout au long de l'étude, des données cliniques et biologiques ont été collectées afin de pouvoir vérifier la compliance des sujets à la supplémentation et d'évaluer leur état de santé. Une fois par an, un examen clinique ou biologique (en alternance) a été réalisé par les équipes médicales de SU.VI.MAX. Chaque participant devait choisir parmi une des 65 principales villes françaises pour réaliser sa visite annuelle. En fonction de la ville, la visite avait lieu soit dans un centre de médecine préventive soit dans une des deux unités mobiles SU.VI.MAX qui sillonnaient la France. Lors du bilan clinique, un certain nombre de tests de dépistage des cancers étaient réalisés (recherche de sang dans les selles, mammographie pour les femmes

de plus de 50 ans, frottis cervical pour les femmes n'en ayant pas eu dans l'année...), ainsi qu'un électrocardiogramme, une mesure de la pression artérielle, un examen clinique et des mesures anthropométriques. En cas d'anomalie aux tests de dépistage, un contact se faisait avec les structures de soins en charge des sujets afin d'assurer le suivi des investigations complémentaires et documenter les diagnostics. Lors du bilan biologique, un prélèvement le matin à jeun de 35 ml de sang était réalisé afin d'effectuer différents dosages (bêta-carotène, rétinol, vitamine C, vitamine E, zinc et sélénium sériques, hémoglobine, glycémie, iodurie, cholestérol total, triglycérides, apolipoprotéines A1 et B). De plus, tous les événements liés à la santé étaient recueillis chaque mois. Toutes les consultations et hospitalisations étaient ainsi analysées et ont fait l'objet d'investigations détaillées par les médecins de l'équipe SU.VI.MAX afin de documenter ces événements médicaux.

C.2 Calcul des scores de vieillissement

Le score de rides a été calculé comme indiqué figure C.1.

Score de rides			Coefficient
Constante			-0.64
Rides inter-sourcillières	(grades 0/1,2,3,4,5)	X	0.44
Rides de la patte d'oie	(grades 0/1,2,3,4,5)	X	0.54
Rides sous les yeux	(grades 1/2,3,4,5)	X	0.64
Rides fines sur les joues	(grades 0,1,2)	X	0.70
Rides d'expression joues	(absent/peu marquées, très marquées)	X	1.06
Rides contour de la bouche	(grades 0,1,2,3,4)	X	0.42
Score		**=**	**Somme (Σ)**

FIGURE C.1 – Calcul du score de rides (entre 0 et 10)

Le score de relâchement a été calculé comme suit (voir figure C.2) :

Score de relâchement			Coefficient
Poches sous les yeux	(absence, présence)	X	0.87
Relâchement de l'ovale du visage	(grades <3, 3, >3)	X	0.93
Relâchement des paupières	(grades <3, 3, >3)	X	1.07
Sillon naso-génien	(grades <3, 3, 4, >4)	X	0.78
Score		=	**Somme (Σ)**

FIGURE C.2 – Calcul du score de relâchement (entre 0 et 10)

Le score de lentigines a été calculé comme suit (voir figure C.3) :

Score de lentigines			Coefficient
Constante			0.00
Lentigines sur la joue	(grades 0,1,2,3,4)	X	1.25
Lentigines sur le front	(grades 0,1,2,3,4)	X	1.25
Score		=	**Somme (Σ)**

FIGURE C.3 – Calcul du score de lentigines (entre 0 et 10)

C.3 Génotypage et contrôle qualité

C.3.1 Génotypage

Les individus ont été génotypées avec la puce Illumina Infinium HumanOmni1-Quad contenant 1 140 419 marqueurs (SNPs). L'ADN génomique (250 ng) a été amplifié, fragmenté, dénaturé et hybridé sur la puce pendant au moins 16 heures à 48 ° C. Les fragments hybridés de manière non spécifique ont été éliminés après lavage et les SNPs restants ont été labellisés par fluorescence par extension d'une simple base et ensuite lus avec un scanner IScan (Illumina). Les intensités de fluorescence ont ensuite été normalisées et l'inférence des SNPs a été effectuée à l'aide du logiciel GenomeStudio (v 1.6.3 ; Illumina).

C.3.2 Contrôle de qualité du génotypage

Seuls les SNPs ont été considérés dans la suite des analyses les 91 706 Copy Number Variations (CNVs) ayant été écartés. De plus, 2 182 SNPs se situant sur le chromosome Y ont été exclus, la population étant composée uniquement de femmes. Par la suite, un filtrage des données en trois étapes a été réalisé :

- **Analyse BeadStudio :** Les données brutes issues du génotypage ont été analysées à l'aide du logiciel Illumina BeadStudio v3.1 et filtrées selon plusieurs paramètres. Dans un premier temps les génotypes sont attribués d'après une classification fournie par Illumina générée sur une population caucasienne. Cette étape assure la robustesse des génotypes attribués. Ensuite, les individus avec un "call rate" (pourcentage de SNPs génotypés par individu) inférieur à 95% sont éliminés. D'autre part, les SNPs avec un "call frequency" (pourcentage d'individus génotypés par SNP) inférieur à 99% ont été re-classifiés. Après la re-classification, les individus avec un "call rate" inférieur à 98% ont été supprimés. Ainsi neuf participantes ont dû être écartées de la suite des analyses, réduisant la taille de l'échantillon à 520 femmes. Les étapes de classification peuvent induire des erreurs dans l'attribution des génotypes évitables en suivant la procédure mise en place par Illumina. Cette procédure permet d'évaluer la qualité de la re-classification selon différents critères qui peuvent être corrigés manuellement si nécessaire. Enfin, les SNPs avec un "call frequency" inférieur à 98% (i.e. un taux de données manquantes $>2\%$) ont été exclus, soit 56 479 SNPs. L'ensemble de ces étapes assure des données de génotypage fiables avec peu de données manquantes. Cette méthode a été utilisée sur nos données et dans de nombreuses études (Le Clerc et al. [2009], Limou et al. [2009], Limou et al. [2010]).

- **Equilibre d'Hardy-Weinberg :** La loi de Hardy-Weinberg postule que la diversité génétique de la population se maintient et tend vers un équilibre stable des fréquences des allèles et des génotypes au cours des générations (Weinberg [1908], Hardy [1908]). Cet équilibre est observé si les hypothèses suivantes sont respectées : la population étudiée est de taille infinie ; absence de migration, mutation et sélection ; la panmixie (rencontre aléatoire des individus) et la pangamie (rencontre aléatoire des gamètes).

Dans ce cas d'équilibre, pour un SNP bi-allélique a_1/a_2 où fa_1 est la fréquence de l'allèle a_1 et $fa_2 = (1 - fa_1)$ est la fréquence de l'allèle a_2, alors :
- la fréquence du génotype homozygote a_1/a_1 est de fa_1^2,
- la fréquence du génotype hétérozygote a_1/a_2 est de $2fa_1fa_2$,
- la fréquence du génotype homozygote a_2/a_2 est de fa_2^2.

En pratique, cette loi est bien respectée, et un écart à l'équilibre d'Hardy-Weinberg dans un groupe de patients pour un SNP donné suggère un effet biologique, tandis qu'une déviation dans un groupe contrôle suggère généralement une erreur de génotypage. La déviation de l'équilibre de Hardy-Weinberg a été évaluée pour chaque SNP dans chaque groupe en comparant la répartition des fréquences génotypiques observées et des fréquences génotypiques théoriques en utilisant un test statistique exact (Wigginton et al. [2005]). Les SNPs déviant de cet équilibre dans la population contrôle ($p<10^-3$) ont été exclus, soit 3 866 SNPs.

- **SNPs de faible fréquence :** L'élimination des SNPs de faible fréquence est une étape classique du contrôle qualité assurant la fiabilité des données de génotypage et facilitant l'analyse statistique postérieure. Les SNPs dont la fréquence de l'allèle mineur est inférieure à 1% dans la population globale ont donc été éliminés, soit un total de 191 123 SNPs.

En conclusion, nous avons appliqué les seuils standards de contrôle qualité à savoir :

- données manquantes par individu : inférieur à 2% (suppression de 9 individus) ;
- données manquantes par marqueur : inférieur à 2% (suppression de 56 479 SNPs) ;
- seuil au test d'équilibre d'Hardy-Weinberg par marqueur : pvalue $< 5.10^{-3}$ (suppression de 3 866 SNPs) ;
- fréquence allélique mineure par marqueur : inférieur à 1% (suppression de 191 123 SNPs).

Par ailleurs, une des 520 femmes restantes s'est présentée avec un léger fond de teint unifiant lors de la prise photographique. La couleur de sa peau étant masquée, les lentigines n'ont pu être appréciées par le dermatologue. Il a été décidé de l'écarter de toutes les analyses. Les données génétiques dont nous disposons concernent alors n = 501 femmes

caucasiennes génotypées sur 795 063 SNPs (tableau final de dimension 501 × 795 063). Cette procédure de filtrage en plusieurs étapes, réalisée sous PLINK (Purcell et al. [2007]) amène finalement à retenir 795 063 SNPs sur 519 sujets.

C.4 Stratification

En génétique la stratification désigne un phénomène de "différenciation" dû à des phénomènes de migrations ancestrales. Ces différences peuvent être observées à l'échelle continentale mais aussi à celle de pays voisins comme en Europe (figure C.4, Novembre et al. [2008]). La figure C.4 est un exemple de carte génétique avec les profils génétiques des Européens. Chaque population est représentée par une couleur spécifique, par exemple orange pâle pour les français et rose pâle pour les anglais). La répartition des couleurs sur cette carte (les différentes populations) correspond de façon spectaculaire à la carte géographique de l'Europe. Ces profils illustrent l'histoire de déplacement de l'Homo sapiens sur la planète. Les différences dans les aspects à la fois physiques et culturelles résultent de l'adaptation de ces populations au climat et à la géographie.

Pour corriger l'éventuelle stratification de notre échantillon au niveau intercontinental, les génotypes de tous les individus ont été analysés en utilisant le logiciel EIGENSTRAT issu de la suite EIGENSOFT basé sur une analyse en composantes principales qui permet de modéliser les différences ancestrales selon des axes continus de variation (Price et al. [2006]). Pour ce faire, un jeu de 328 SNPs informatifs de l'origine ancestrale (index de fixation **FST!** $>$0,2) d'après les données Perlegen et distants de plus de 5Mb (afin d'éviter le déséquilibre de liaison) est sélectionné pour l'ensemble des 519 femmes de l'échantillon. Les génotypes des individus non apparentés issus des populations du projet HapMap sont également inclus dans l'analyse afin de séparer au mieux les individus selon leur origine continentale et exclure ceux d'origine non européenne. Les deux premières passes d'EIGENSTRAT ont permis de mettre en évidence 18 individus atypiques qui ont été retirés pour la suite de l'analyse (figure C.5).

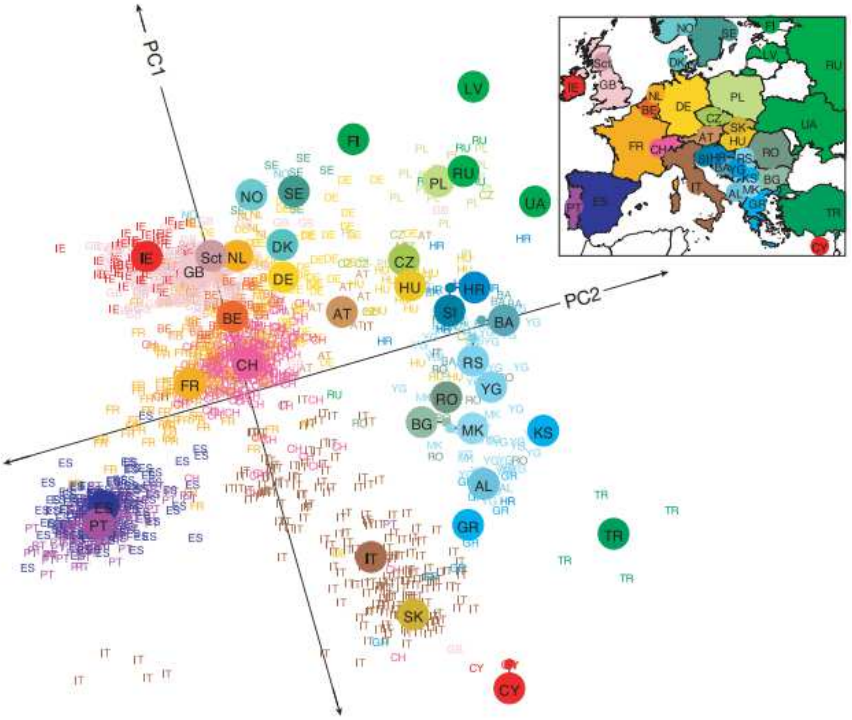

FIGURE C.4 – Carte génétique à partir de biopuces

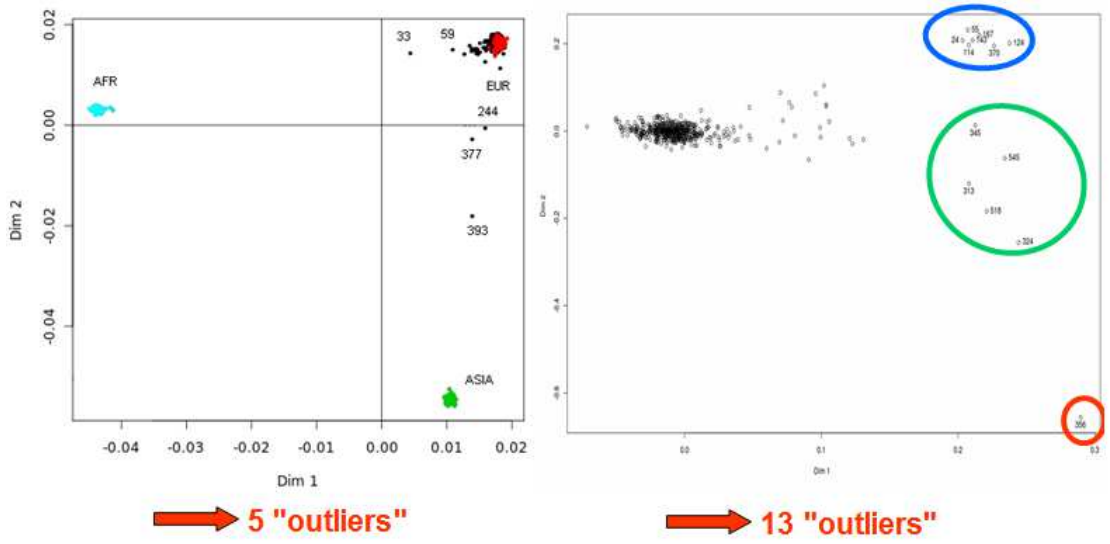

FIGURE C.5 – Stratification : Identifications successives des individus "atypiques" à l'aide de deux ACPs

De plus, le logiciel STRUCTURE v2.2 basé sur la phylogénie a également été utilisé et a permis d'identifier 4 individus atypiques déjà inclus dans les 18 individus précédemment écartés. (Pritchard et al. [2000], Falush et al. [2003]).

Annexe D

Packages R

L'ensemble des analyses présentées dans le manuscrit a été réalisé à l'aide du logiciel R. Les programmes des deux nouvelles méthodes développées : GSPCA et ACM sparse (voir chapitre 2) ont été écris et mis sous la forme d'un package R qui sera soumis très prochainement au CRAN (The Comprehensive R Archive Network). Les deux fonctions correspondantes se nomment "GSPCA" et "SMCA".

La comparaison des méthodes régression logique, CART et CHAID dans le chapitre 3 a été réalisée à l'aide de trois packages R : le package "logicFS" (Schwender [2013a]) pour la régression logique, la package "rpart" (Therneau et al. [2013]) pour la méthode CART et le package "CHAID" (Team [2009]), contenant la fonction "chaid", pour la méthode CHAID. Certaines fonctions du package "logicFS" utlisées dans nos analyses sont détaillées par la suite. La méthode de régression logistique a été appliquée à l'aide de la fonction "glm" du package "stats" de R, la random forest à l'aide de la fonction "randomForest" du package portant le même nom (Liaw and Wiener [2002]), la méthode SVM avec la fonction "svm" de la librairie "e1071" (Meyer et al. [2012]), et l'analyse discriminante linéaire à partir de la fonction "lda" de la librairie "MASS" (Venables and Ripley [2002]). Les courbes ROC ont été tracée à partir des fonctions contenues dans le package "ROCR" (Sing et al. [2005]).

Dans le chapitre 4, la régression elastic net a été réalisée à l'aide du package "glmnet" (Friedman et al. [2010]) et de la fonction portant le même nom. Le package "FactoMineR" comprenant la fonction "MCA" a permis de calculer les Analyses des Correspondances Multiples (Husson et al. [2013]), et le package "ade4" les rapports de corrélations grâce

à la fonction "dudi.acm" (Dray and Dufour [2007]). La RGCCA dont les résultats sont présentés section 4.3 a été réalisée grâce au package "RGCCA" et à la fonction du package portant le même nom (Tenenhaus and Guillemot [2013]). Les intervalles de confiance ont été calculés à l'aide d'une fonction n'étant pas encore incorporée dans le package.

La détection des interactions (voir section refdetectinter) a été possible grâce au package "logicFS" (Schwender [2013a]) qui s'installe à partir des commandes suivantes :

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("logicFS")
```

Le codage des SNPs décrit dans le paragraphe 4.4.1.1 du chapitre 4 a été réalisé grâce à la fonction "make.snp.dummy" du package "logicFS" (Schwender [2013b]). Elle permet de transformer les SNPs en variables binaires. Par la suite, pour accélérer la recherche des meilleurs modèles de régression logiques, seulement un petit nombre d'itérations est utilisé dans le recuit simulé grâce à la fonction "logreg.anneal.control" du package "LogicReg" (Kooperberg and Ruczinski [2012]). Enfin la sélection par régression logique est effectuée grâce à la fonction "logicFS" suivante :

```
> log.out<-logicFS(bin.snps,B=20, nleaves=2, anneal.control=my.anneal)
```

avec bin.snps la table contenant les variables binaires calculées à l'aide de la fonction "make.snp.dummy" (chaque colonne correspond à une variable binaire, chaque ligne à une observation), B le nombre d'itérations, nleaves le nombre de branches que l'on considère (2 branches dans notre cas), anneal.control la liste contenant les paramètres pour le recuit simulé (liste obtenue grâce à la fonction précédente "logreg.anneal.control"). L'importance des interactions représentée dans la figure 4.24 est obtenue grâce à la commande :

```
> plot(log.out)
```

Les autres représentations graphiques, telles que les histogrammes ou les figure de type "manhattan plot" ont été réalisées grâce à des fonctions R de base (comme la fonction "hist") et modifiées selon le type de données considérées et la représentation voulue.

Annexe E

Démonstrations

E.1 Démonstration de l'équation (1.28) page 55

Démonstration. Le problème des moindres carrés permet de trouver une matrice d'approximation de rang inférieur $\mathbf{X}^{(1)}$ d'une matrice donnée $\mathbf{X}$. On peut le formuler de la manière suivante :

$$\underset{\mathbf{X}^{(1)}\in\mathbf{R}^{I\times J},\mathbf{q}\in\mathbf{R}^{J}}{\arg\min} \|\mathbf{X}-\mathbf{X}^{(1)}\|_2^2 \qquad \text{tel que} \qquad \mathbf{p}^T\mathbf{p}=\mathbf{1} \text{ et } \mathbf{q}^T\mathbf{q}=\mathbf{1}. \tag{E.1}$$

La matrice $\mathbf{X}^{(1)}$ est de la forme $\mathbf{p}\delta\mathbf{q}^T = \mathbf{f}\mathbf{q}^T$ et l'on veut déterminer le $\mathbf{p}$ et $\mathbf{q}$ optimal solution du problème (E.1). L'expression $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_2^2$ est égale à :

$$\begin{aligned}
\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_2^2 &= \operatorname{tr}\left((\mathbf{X}-\mathbf{X}^{(1)})^T(\mathbf{X}-\mathbf{X}^{(1)})\right) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X}-\mathbf{X}^T\mathbf{X}^{(1)}-\mathbf{X}^{(1)^T}\mathbf{X}+\mathbf{X}^{(1)^T}\mathbf{X}^{(1)}) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X})-2\operatorname{tr}(\mathbf{X}^{(1)^T}\mathbf{X})+\operatorname{tr}(\mathbf{X}^{(1)^T}\mathbf{X}^{(1)}) \\
&= \|\mathbf{X}\|_2^2-2\operatorname{tr}(\mathbf{q}\mathbf{f}^T\mathbf{X})+\operatorname{tr}(\mathbf{q}\mathbf{f}^T\mathbf{f}\mathbf{q}^T) \\
&= \|\mathbf{X}\|_2^2-2\delta\operatorname{tr}(\mathbf{q}\mathbf{p}^T\mathbf{X})+\delta^2\operatorname{tr}(\mathbf{q}\mathbf{p}^T\mathbf{p}\mathbf{q}^T) \\
&= \|\mathbf{X}\|_2^2-2\delta\operatorname{tr}(\mathbf{q}\mathbf{p}^T\mathbf{X})+\delta^2
\end{aligned} \tag{E.2}$$

car $\mathbf{q}^T\mathbf{q}=\mathbf{1}$ et $\mathbf{p}^T\mathbf{p}=\mathbf{1}$ donc $\mathbf{f}^T\mathbf{f}=\delta^2\mathbf{p}^T\mathbf{p}=\delta^2$. Supposons α et $\gamma\in\mathbf{R}$ les multiplicateurs de Lagrange ; le lagrangien peut alors être écrit de la manière suivante :

$$\begin{aligned}
\mathcal{L}(\mathbf{p},\mathbf{q},\alpha,\gamma) &= \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_2^2+\alpha(\mathbf{p}^T\mathbf{p}-1)+\gamma(\mathbf{q}^T\mathbf{q}-1) \\
&= \|\mathbf{X}\|_2^2-2\delta\operatorname{tr}(\mathbf{q}\mathbf{p}^T\mathbf{X})+\delta^2+\alpha(\mathbf{p}^T\mathbf{p}-1)+\gamma(\mathbf{q}^T\mathbf{q}-1)
\end{aligned} \tag{E.3}$$

En fixant chaque dérivée à zéro, on obtient :

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{p}} &= -2\delta \mathbf{X}\mathbf{q} + 2\alpha \mathbf{p} = 0 \quad &\Rightarrow \quad \delta \mathbf{X}\mathbf{q} = \alpha \mathbf{p} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{q}} &= -2\delta \mathbf{X}^T\mathbf{p} + 2\gamma \mathbf{q} = 0 \quad &\Rightarrow \quad \delta \mathbf{X}^T\mathbf{p} = \gamma \mathbf{q} \end{aligned} \tag{E.4}$$

Cependant, quand $\mathbf{X}$ est multipliée à droite et à gauche par un vecteur, la matrice devient une matrice de rang 1.

$$\begin{aligned} \delta \mathbf{X}\mathbf{q} = \alpha \mathbf{p} \quad &\Leftrightarrow \quad \delta \mathbf{p}\delta \mathbf{q}^T\mathbf{q} = \alpha \mathbf{p} \\ &\Leftrightarrow \quad \delta^2 \mathbf{p} = \alpha \mathbf{p} \\ &\Leftrightarrow \quad \delta^2 = \alpha \end{aligned} \tag{E.5}$$

et

$$\begin{aligned} \delta \mathbf{X}^T\mathbf{p} = \gamma \mathbf{q} \quad &\Leftrightarrow \quad \delta \mathbf{q}\delta \mathbf{p}^T\mathbf{p} = \gamma \mathbf{q} \\ &\Leftrightarrow \quad \delta^2 \mathbf{q} = \gamma \mathbf{q} \\ &\Leftrightarrow \quad \delta^2 = \gamma \end{aligned} \tag{E.6}$$

Ainsi, on peut réécrire (E.4) :

$$\begin{aligned} \mathbf{X}\mathbf{q} = \delta \mathbf{p} \quad &\text{avec} \quad \mathbf{p}^T\mathbf{p} = \mathbf{1} \\ \mathbf{X}^T\mathbf{p} = \delta \mathbf{q} \quad &\text{avec} \quad \mathbf{q}^T\mathbf{q} = \mathbf{1} \end{aligned} \tag{E.7}$$

ce qui implique que $(\delta, \mathbf{p}, \mathbf{q})$ doit être un triplet singulier de $\mathbf{X}$ avec δ une valeur singulière, $\mathbf{p}$ un vecteur singulier gauche et $\mathbf{q}$ un vecteur singulier droit. Pour obtenir le minimum de $\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_2^2$, δ doit être la valeur singulière la plus grande. Le triplet singulier sera celui correspondant à la plus grande valeur singulière : $\mathbf{p} = \mathbf{p}_1$ et $\mathbf{q} = \mathbf{q}_1$, avec $\mathbf{X}^{(1)} = \delta_1 \mathbf{p}_1 \mathbf{q}_1^T$ la meilleure matrice d' approximation de rang 1 de la matrice $\mathbf{X}$. □

E.2 Démonstration de l'équation (2.8) page 83

Démonstration. Le problème des moindres carrés permettant de trouver la meilleure matrice de rang 1 $\mathbf{X}^{(1)}$ peut être réécrit sous la forme suivante :

$$\underset{\mathbf{X}^{(1)}\in\mathbb{R}^{I\times J},\mathbf{q}\in\mathbb{R}^{J}}{\arg\min} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^{2} \qquad \text{avec} \qquad \mathbf{p}^T\mathbf{M}\mathbf{p}=\mathbf{1} \text{ et } \mathbf{q}^T\mathbf{W}\mathbf{q}=\mathbf{1}. \tag{E.8}$$

La matrice $\mathbf{X}^{(1)}$ est de la forme $\mathbf{f}\mathbf{q}^T$ et l'on souhaite obtenir le $\mathbf{f}$ et le $\mathbf{q}$ optimal, solutions du problème (E.8). L'expression $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^{2}$ vaut :

$$\begin{aligned}
\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^{2} &= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}-\mathbf{X}^{(1)})\mathbf{W}(\mathbf{X}-\mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right) \\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}-\mathbf{X}^{(1)}\mathbf{W})(\mathbf{X}-\mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right) \\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}\mathbf{X}^T-\mathbf{X}\mathbf{W}\mathbf{X}^{(1)^T}-\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T+\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^{(1)^T})\mathbf{M}^{\frac{1}{2}}\right) \\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}\mathbf{X}^T-2\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T+\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^{(1)^T})\mathbf{M}^{\frac{1}{2}}\right) \\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}-2\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}+\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^{(1)^T}\mathbf{M}^{\frac{1}{2}}\right) \\
&= \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^{(1)^T}\mathbf{M}^{\frac{1}{2}}) \\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\delta\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\delta\mathbf{q}^T\mathbf{W}\mathbf{q}\delta\mathbf{p}^T\mathbf{M}^{\frac{1}{2}}) \\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\delta\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\delta^2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{p}^T\mathbf{M}^{\frac{1}{2}}) \\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\delta\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\delta^2.
\end{aligned} \tag{E.9}$$

car $\mathbf{p}^T\mathbf{M}\mathbf{p}=\mathbf{1}$ et $\mathbf{q}^T\mathbf{W}\mathbf{q}=\mathbf{1}$. Supposons α et $\gamma\in\mathbf{R}$ les multiplicateurs de Lagrange ; alors le lagrangien s'écrit :

$$\begin{aligned}
\mathcal{L}(\mathbf{p},\mathbf{q},\alpha,\gamma) &= \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^{2}+\alpha(\mathbf{p}^T\mathbf{M}\mathbf{p}-1)+\gamma(\mathbf{q}^T\mathbf{W}\mathbf{q}-1) \\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\delta\operatorname{tr}(\mathbf{M}^{1/2}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{1/2})+\delta^2 \\
&\quad +\alpha(\mathbf{p}^T\mathbf{p}-1)+\gamma(\mathbf{q}^T\mathbf{q}-1).
\end{aligned} \tag{E.10}$$

Si l'on fixe les dérivées à zéro :

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{p}} &= -2\delta\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q}+2\alpha\mathbf{M}\mathbf{p}=0 \qquad &\Rightarrow \qquad \delta\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q}=\alpha\mathbf{M}\mathbf{p} \\
\frac{\partial\mathcal{L}}{\partial\mathbf{q}} &= -2\delta\mathbf{W}\mathbf{X}^T\mathbf{M}\mathbf{p}+2\gamma\mathbf{W}\mathbf{q}=0 \qquad &\Rightarrow \qquad \delta\mathbf{W}\mathbf{X}^T\mathbf{M}\mathbf{p}=\gamma\mathbf{W}\mathbf{q}
\end{aligned} \tag{E.11}$$

Cependant,

$$\begin{aligned} \delta\mathbf{MXWq} = \alpha\mathbf{Mp} \quad &\Leftrightarrow \quad \delta\mathbf{Mp}\delta\mathbf{q}^T\mathbf{Wq} = \alpha\mathbf{Mp} \\ &\Leftrightarrow \quad \delta^2\mathbf{Mp} = \alpha\mathbf{Mp} \\ &\Leftrightarrow \quad \delta^2 = \alpha \end{aligned} \tag{E.12}$$

et

$$\begin{aligned} \delta\mathbf{WX}^T\mathbf{Mp} = \gamma\mathbf{Wq} \quad &\Leftrightarrow \quad \delta\mathbf{Wq}\delta\mathbf{p}^T\mathbf{Mp} = \gamma\mathbf{Wq} \\ &\Leftrightarrow \quad \delta^2\mathbf{Wq} = \gamma\mathbf{Wq} \\ &\Leftrightarrow \quad \delta^2 = \gamma \end{aligned} \tag{E.13}$$

On a alors :

$$\begin{aligned} \mathbf{Xq} = \delta\mathbf{p} \quad &\text{avec} \quad \mathbf{p}^T\mathbf{p} = \mathbf{1} \\ \mathbf{X}^T\mathbf{p} = \delta\mathbf{q} \quad &\text{avec} \quad \mathbf{q}^T\mathbf{q} = \mathbf{1} \end{aligned} \tag{E.14}$$

ce qui implique que $(\delta, \mathbf{p}, \mathbf{q})$ doit être le triplet singulier de $\mathbf{X}$ et δ la plus grande valeur propre. De la même manière que dans la section 1.1.2.3, on conclut que $\mathbf{p} = \mathbf{p}_1$ et $\mathbf{q} = \mathbf{q}_1$, avec $\mathbf{X}^{(1)} = \delta_1\mathbf{p}_1\mathbf{q}_1 = \mathbf{f}_1\mathbf{q}_1$ la meilleure approximation de la matrice $\mathbf{X}$. □

Annexe F

Publications

F.1 Publication parue

Le Clerc, S., Taing, L., Ezzedine, K., Latreille, J., Labib, T., Coulonges, C., Bernard, A., Melak, S., Carpentier, W., Malvy, D., Jdid, R., Galan, P., Hercberg, S., Morizot, F., Guinot, G., Tschachler, E., Zagury J.-F. (2013). A genome-wide association study in Caucasian women points out for a role of the STXBP5L gene in facial photoageing. *Journal of Investigative Dermatology*, 133, p.929-935.

A Genome-Wide Association Study in Caucasian Women Points Out a Putative Role of the *STXBP5L* Gene in Facial Photoaging

Sigrid Le Clerc[1,11], Lieng Taing[1,11], Khaled Ezzedine[2,3], Julie Latreille[4,12], Olivier Delaneau[1,5], Toufik Labib[1], Cédric Coulonges[1], Anne Bernard[4,12], Safa Melak[1], Wassila Carpentier[6], Denis Malvy[2,7], Randa Jdid[4,12], Pilar Galan[2], Serge Hercberg[2,8], Frederique Morizot[4,12], Christiane Guinot[4,9,12], Erwin Tschachler[4,10,11,12] and Jean F. Zagury[1,11]

A genome-wide association study (GWAS) was conducted on 502 French middle-aged Caucasian women to identify genetic factors that may affect skin aging severity. A high-throughput Illumina Human Omni1-Quad beadchip was used. After single-nucleotide polymorphism (SNP) quality controls, 795,063 SNPs remained for analysis purposes. Possible stratification was first examined using the Eigenstrat method, and then the relationships between genotypes and four skin aging indicators (global photoaging, lentigines, wrinkles, and sagging) were investigated separately by linear regressions adjusted on age, smoking habits, lifetime sun exposure, hormonal status, and the two main Eigen vectors. One signal passed the Bonferroni threshold ($P = 1.53 \times 10^{-8}$) and was significantly associated with global photoaging. It was also correlated with the wrinkling score and the sagging score. According to HapMap, this SNP, rs322458, was in linkage disequilibrium (LD) with intronic SNPs of the *STXBP5L* gene, which is expressed in the skin. In addition, it was also in LD with another SNP that increases the expression of the *FBXO40* gene in the skin. These two genes, which were not previously described in the context of aging, may constitute good candidates for the investigation of molecular mechanisms of skin photoaging.

Journal of Investigative Dermatology advance online publication, 6 December 2012; doi:10.1038/jid.2012.458

INTRODUCTION

Similar to other organs, skin ages owing to passage of time. Skin aging is influenced both by inherited intrinsic factors and by extrinsic or environmental factors, such as chronic UV exposure and smoking (Malvy *et al.*, 2000; Yaar and Gilchrest, 2007). Intrinsic aging is an ineluctable process and is due to the genetically determined natural degeneration of the cell functioning and loss of extracellular matrix with age (Yaar and Gilchrest, 1990). Its clinical phenotype on the skin is mainly characterized by fine wrinkles and dry, thin, and pale skin (Fisher *et al.*, 2002; Makrantonaki and Zouboulis, 2007).

The main factor responsible for extrinsic aging of the skin is UVR. UV-induced skin aging or photoaging is defined as the premature occurrence of signs of aging on the skin, and presents with characteristic morphological changes of both the epidermal and dermal compartments (Rabe *et al.*, 2006; Yaar and Gilchrest, 2007). A number of hereditary phenotypic features influence the severity of photoaging, most notably skin color (Kligman and Kligman, 1999; Malvy *et al.*, 2000), and skin phototype (Fitzpatrick, 1988). Individuals with dark phototypes (III–IV) commonly exhibit more ''hypertrophic responses'' such as deep wrinkling, coarseness, and lentigines, whereas fair phototype individuals (I–II) generally show fewer wrinkles with epidermal atrophy, focal depigmentation, as well as dysplastic changes, such as actinic keratosis, nonmelanoma, and melanoma skin cancers (Rabe *et al.*, 2006; Yaar and Gilchrest, 2007; Puizina-Ivié, 2008).

Up to now, the exploration of the genes affecting skin aging has remained limited to *MC1R* gene (Elfakir *et al.*, 2010; Suppa *et al.*, 2011), or to genes involved in genetic pathologies with accelerated skin aging (Rooryck *et al.*,

[1]*Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, Paris, France;* [2]*UMR U557, INSERM/U1125 INRA/CNAM, University Paris 13/Centre de Recherche en Nutrition Humaine Ile-de-France, Bobigny, France;* [3]*Department of Dermatology, Hôpital Saint-André, Bordeaux, France;* [4]*CE.R.I.E.S., Neuilly-sur-Seine, France;* [5]*Department of Statistics, University of Oxford, Oxford, UK;* [6]*Plateforme Post-Génomique P3S, Hôpital Pitié-Salpêtrière, Paris, France;* [7]*Department of Internal Medicine and Tropical Diseases, Hôpital Saint-André, Bordeaux, France;* [8]*Department of Public Health, Hôpital Avicenne, Bobigny, France;* [9]*Computer Science Laboratory, University François Rabelais, Tours, France and* [10]*Department of Dermatology, University of Vienna Medical School, Vienna, Austria*

[11]*These authors contributed equally to this work.*

[12]*CE.R.I.E.S. is a research center on human skin founded by Chanel.*

Correspondence: Erwin Tschachler, Department of Dermatology, University of Vienna Medical School, Währinger Gürtel 18–20, A-1090 Vienna, Austria. E-mail: erwin.tschachler@meduniwien.ac.at or Jean-François Zagury, Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, 292 Rue Saint Martin, 75003 Paris, France. E-mail: zagury@cnam.fr

Abbreviations: BMI, body mass index; GWAS, genome-wide association study; LD, linkage disequilibrium; SNP, single-nucleotide polymorphism; SU.VI.MAX, SUpplémentation en VItamines et Minéraux AntioXydants

Received 4 July 2012; revised 28 September 2012; accepted 9 October 2012

2008; Soufir *et al.*, 2010). A candidate gene approach has previously established associations between *MC1R* gene variants, particularly loss-of-function variants, with an increased risk of severe photoaging (Elfakir *et al.*, 2010). In addition, a few studies conducted in twin cohorts have explored the associations between environmental factors, skin aging, and gene expression (Plomin *et al.*, 1994; Shekar *et al.*, 2005, 2006; Christensen *et al.*, 2009).

To unravel new genetic associations with skin aging in a systematic way, we have undertaken a genome-wide study on a well-defined sample of Caucasian women from the SU.VI.MAX (*SUpplémentation en VItamines et Minéraux Anti-oXydants* (Antioxidant Vitamin and Mineral Supplementation)) cohort (Hercberg *et al.*, 2004). To the best of our knowledge, no genome-wide association study (GWAS) targeting skin aging in middle-aged women of European-derived ancestry has been previously reported.

RESULTS

Using the Illumina HumanOmni1-Quad BeadChips, we conducted a GWAS by testing associations between single-nucleotide polymorphisms (SNPs) and global skin photoaging on a large sample of French middle-aged women from the SU.VI.MAX cohort. After the various quality-control tests (see Materials and Methods), 795,063 genotyped SNPs were available for 502 women.

Table 1 describes the sample of women according to the severity of photoaging. We also computed the correlations between the age and the outcome variables (Table 2). We found that the correlations with age were all statistically significant ($P<0.0001$): 0.56 for the grade of photoaging, 0.61 for the score of wrinkling and the score of sagging, and 0.27 for the score of lentigines. Similarly, the correlations between the grade of photoaging and the other outcome variables were also statistically significant ($P<0.0001$): 0.78 for the score of wrinkling, 0.66 for the score of sagging, and 0.31 for the score of lentigines; the correlation between the score of wrinkling and the score of sagging reached 0.71 ($P<0.0001$; Table 2).

Our core association analysis focused on genotypic associations obtained using linear regressions, after correction for stratification and nongenetic skin aging factors. Figure 1 presents the distribution of the *P*-values obtained for each SNP along the chromosomes (Manhattan plot). One SNP located on the chromosome 3 (locus 3q13.33), rs322458, passed the Bonferroni threshold (6.28×10^{-8}) with $P=1.53\times10^{-8}$. According to HapMap, this SNP is in linkage disequilibrium (LD) with five SNPs positioned in intronic regions of the *STXBP5L* gene (rs470647, rs612545, rs617332, rs645045, and rs1795413), and with two intergenics SNPs (rs377374 and rs450614; Figure 2). A more refined analysis suggested that the effect was likely recessive. Indeed, when regrouping the individuals according to their grade of skin photoaging, the frequency of the homozygous rs3224587-AA genotype was clearly inversely proportional with photoaging severity (Figure 3): from 28% of homozygous subjects among grade 1 to 4% among grade 5. To further investigate the rs322458 SNP, we assessed its putative impact on each phenotype: lentigines, wrinkling, and sagging. No relationship was found with the lentigines score ($P=0.63$), whereas significant links were found with wrinkling and sagging scores (respectively, $P=5.6\times10^{-5}$ and $P=1.76\times10^{-4}$).

Moreover, bioinformatics databases were investigated for possible associations between SNPs and mRNA expression, regulation (splicing, polyadenylation, and miRNA), and also for putative transcription binding sites. According to Genevar (Nica *et al.*, 2011), the genotype rs470647-AA (rs470647 is in LD with the rs322458; see Figure 2) increases the expression in skin of a neighboring gene, *FBXO40* ($P=6\times10^{-4}$; Figure 4). The rs470647 SNP and *FBXO40* are at a distance of 683 kb.

To further investigate other possible associations, we also computed all the haplotypes based on two SNPs derived from both *STXBP5L* and *FBXO40* genes. Only three haplotypes were strongly associated with photoaging (Figure 4) and they implicated the rs322458 SNP. These haplotypes involved one exonic SNP and one 3′-untranslated region of the *STXBP5L* gene (respectively, rs17740066, $P=6.27\times10^{-9}$ and rs6782033, $P=3.96\times10^{-9}$), and one intronic SNP of the *FBXO40* gene (rs6775899, $P=9.52\times10^{-10}$). The rs17740066 and rs6782033 SNPs were in partial LD with rs322458 ($D'=1$); in other words, the G allele frequency of rs322458 SNP was identical with that of the haplotypes GG (rs322458–rs17740066) and GA (rs322458–rs6782033). Interestingly, the rs17740066 SNP corresponds to the Val855Ile protein variation and rs6782033 SNP corresponds to a putative binding site for a miRNA (hsa-mir-892b; Figure 4). There was no LD between the two SNPs, rs6775899 and rs322458 ($r^2=0.014$ and $D'=0.2$). However, the GA haplotype (rs322458–rs6775899) also exhibited a significant *P*-value ($P=9.52\times10^{-10}$), suggesting it might also be a haplotype of interest.

DISCUSSION

We have described here a GWAS investigating possible associations between SNPs and global skin photoaging. This research yielded an association for the rs322458 SNP connected to the *STXBP5L* gene with severity of skin photoaging, the rs322458-AA genotype being inversely linked with the severity of skin aging. This SNP was also associated with the wrinkle and sagging scores that are defined independently from the grade of photoaging, but it was not associated with the lentigines score, suggesting that: (1) its role in photoaging does not include pigmentary disorders; and (2) molecular mechanisms might be shared by sagging and wrinkling. According to the HapMap database, this SNP is also polymorphic in the Asian and African populations, and thus it would also be worth investigating these populations. As for any GWAS, additional genetic studies will be needed to affirm this association.

Another alias for *STXBP5L* is *LLGL4*, as it is homologous to the Lethal giant larvae (*Lgl*) drosophila gene (Katoh and Katoh, 2004). The protein coded by *STXBP5L* contains five WD40 repeats (or β-transducin repeats) and a C-terminal syntaxin-binding (STXB) domain. *Lgl* regulates epithelial polarity and, when mutated, may lead to tumor-like

Table 1. Description of the population according to photoaging severity

	Photoaging severity						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5/6 [1]		
	N = 43	*N* = 86	*N* = 174	*N* = 150	*N* = 49	Total, *N* = 502	*P*-value of test
Age (years)	50.1 ± 4.2[2]	54.1 ± 5.0	56.8 ± 5.5	60.9 ± 5.6	62.6 ± 5.2	57.6 ± 6.4	< 0.0001[3]
Lifetime sun exposure (score)	5.3 ± 3.4	5.1 ± 3.5	5.2 ± 3.6	5.5 ± 3.5	5.5 ± 3.5	5.3 ± 3.5	0.84[3]
BMI classification							0.49[4]
Normal	28 (8.4)[5]	57 (17.0)	121 (36.1)	94 (28.1)	35 (10.4)	335 (66.7)	
Overweight	9 (7.4)	19 (15.6)	37 (30.3)	45 (36.9)	12 (9.8)	122 (24.3)	
Obese	6 (13.3)	10 (22.2)	16 (35.6)	11 (24.5)	2 (4.4)	45 (9.0)	
Hormonal status							< 0.0001[4]
Nonmenopausal	27 (28.7)	24 (25.5)	32 (34.0)	9 (9.6)	2 (2.2)	94 (18.7)	
Menopausal with HRT	9 (3.4)	40 (15.3)	98 (37.4)	92 (35.1)	23 (8.8)	262 (52.2)	
Menopausal without HRT	7 (4.8)	22 (15.1)	44 (30.1)	49 (33.6)	24 (16.4)	146 (29.1)	
Smoking habits							0.61[4]
Never	23 (8.0)	45 (15.7)	100 (35.0)	86 (30.1)	32 (11.2)	286 (57.0)	
Former smoker	15 (9.3)	34 (21.3)	50 (31.2)	47 (29.5)	14 (8.7)	160 (31.9)	
Current smoker	5 (8.9)	7 (12.5)	24 (42.8)	17 (30.4)	3 (5.4)	56 (11.1)	
Eye color							0.21[4]
Blue/gray	14 (10.3)	18 (13.2)	50 (36.8)	36 (26.5)	18 (13.2)	136 (27.2)	
Green/hazel/brown/black	28 (7.8)	68 (18.7)	122 (33.6)	114 (31.4)	31 (8.5)	363 (72.8)	
Hair color at 20 years							0.08[4]
Blond/red	4 (3.7)	20 (18.6)	40 (37.0)	28 (25.9)	16 (14.8)	108 (21.6)	
Light and dark brown/black	38 (9.7)	66 (16.9)	132 (33.8)	122 (31.2)	33 (8.4)	391 (78.4)	
Skin color without tanning							0.78[4]
Fair	35 (9.0)	65 (16.8)	136 (35.0)	113 (29.2)	39 (10.0)	388 (77.8)	
Dark	7 (6.3)	21 (18.9)	36 (32.4)	37 (33.3)	10 (9.0)	111 (22.2)	
History of facial freckles							0.40[4]
No	25 (8.5)	55 (18.7)	104 (35.4)	87 (29.6)	23 (7.8)	294 (58.9)	
Yes	17 (8.3)	31 (15.1)	68 (33.2)	63 (30.7)	26 (12.7)	205 (41.1)	
Suntan intensity							0.71[4]
None/slight/light	23 (7.7)	49 (16.3)	101 (33.7)	96 (32.0)	31 (10.3)	300 (60.1)	
Dark/very dark	19 (9.5)	37 (18.6)	71 (35.7)	54 (27.2)	18 (9.0)	199 (39.9)	
Sunburn event frequency							0.74[4]
None/rare	28 (7.8)	61 (17.0)	123 (34.4)	113 (31.6)	33 (9.2)	358 (71.7)	
Frequent/constant	14 (9.9)	25 (17.7)	49 (34.8)	37 (26.2)	16 (11.4)	141 (28.3)	

Abbreviations: BMI, body mass index; HRT, hormonal replacement therapy.
[1]As a single woman had grade 6, she had been grouped with grade 5 individuals.
[2]Mean ± SD.
[3]Analysis of variance (ANOVA) test.
[4]The χ^2 test.
[5]Frequency and (%): because of possible missing values, the sum of the cell frequencies can be smaller than the total indicated in the top of the columns.

phenotype development. According to bioinformatics analysis (UniProt, 2011), *STXBP5L* seems to be implicated in vesicle trafficking and could have a role in exocytosis ((Katoh and Katoh, 2004; UniProt, 2011). Interestingly, *STXBP5L* has previously been associated with liver fibrosis risk in Caucasians and with chronic hepatitis C infection (Li *et al.*, 2009). *STXBP5L* is expressed in several tissues, including the skin (Safran *et al.*, 2010), and is also expressed in lung carcinoid and germ cell tumors (Katoh and Katoh, 2004).

Bioinformatics database exploration pointed out the possible role of the SNP rs322458 in the skin expression of a neighboring gene, *FBXO40*. Haplotype analysis of the

Table 2. Correlation coefficients between age and outcome variables

	Age	Score of wrinkling	Score of sagging	Score of lentigines	Grade of photoaging
Age	1	0.61	0.61	0.27	0.56
Score of wrinkling		1	0.71	0.31	0.78
Score of sagging			1	0.26	0.66
Score of lentigines				1	0.31
Grade of photoaging					1

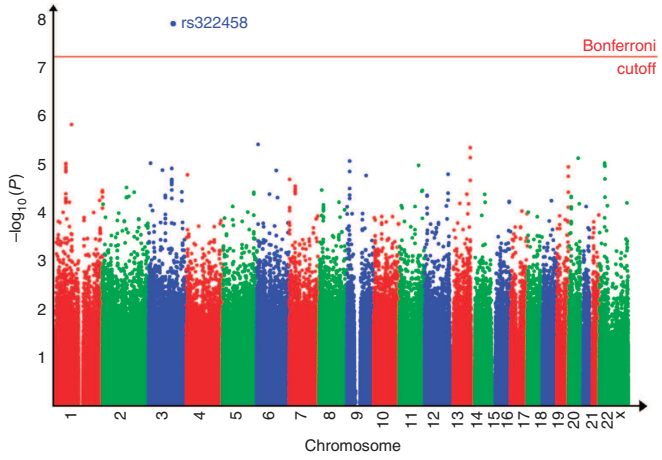

Figure 1. Manhattan plot of the association study with the photoaging score. Distribution of $-\log10(P)$ obtained for the associations tested between the genotypes and skin photoaging, according to Larnier's scale, along the human chromosomes (Manhattan plot).

STXBP5L and *FBXO40* gene region also revealed positive signals ($P \sim 10^{-9}$), bringing up a second hypothesis in which rs3322458 G allele in the dominant mode, possibly in combination with other alleles, might be implicated in the phenotype.

FBXO40 encodes a protein characterized by a 40 amino-acid F-box motif. This gene is expressed specifically in the muscle (Ye *et al.*, 2007), may function as a regulator involved in the postnatal myogenesis (Ye *et al.*, 2007), and has a role in muscle hypertrophy (Shi *et al.*, 2011). F-box proteins are involved in the SCF (Skp, Cullin, F-box containing) complex, known to act as protein–ubiquitin ligases (Skowyra *et al.*, 1997), and a recent study demonstrated that the SCF–F-box40 complex prevented skeletal muscle hypertrophy by limiting the IGF1 pathway in the muscle (Shi *et al.*, 2011).

Both *STXBP5L* and *FBXO40* were not known before for any skin function. How could they affect skin aging? *FBXO40* is linked with the IGF1 pathway known for its role in inflammation, and its direct link with myogenesis could also explain its impact on wrinkling and sagging severity. Knowing that

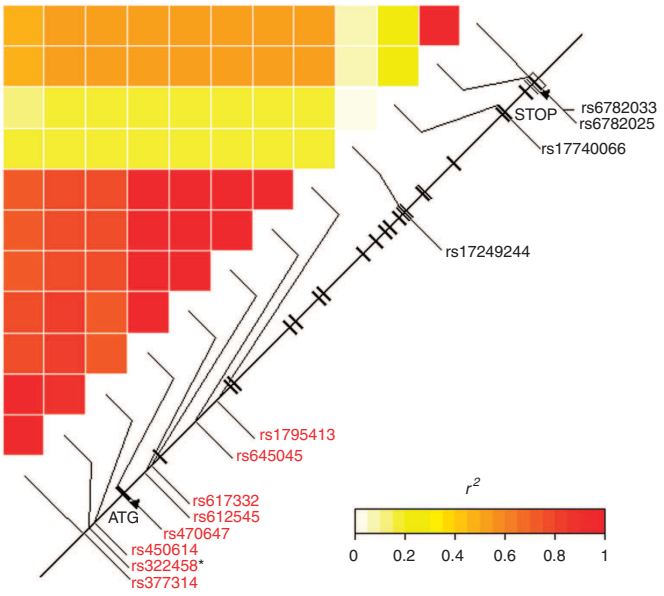

Figure 2. Genetic map of the *STXBP5L* gene. The single-nucleotide polymorphisms (SNPs) in high linkage disequilibrium (LD) with the SNP rs322458 are in red, and the exonic SNPs genotyped in the study are in black. The LD map (providing the r^2 coefficient between SNPs) is given below the genetic map. The SNP rs322458 is flagged with an asterisk (*).

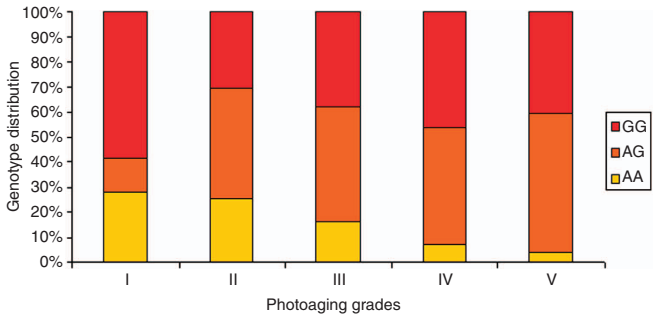

Figure 3. Distribution of the rs322458 genotypes in function of the photoaging severity.

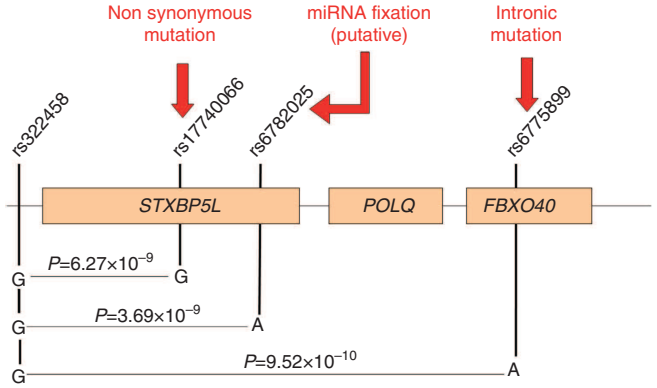

Figure 4. Haplotype analysis. All the two-single-nucleotide polymorphism (SNP) haplotypes of the region were computed using the Shape-IT software. Associations were computed, and the figure presents the three best signals obtained, which all involve the SNP rs322458.

photoaging is intimately associated with the occurrence of dysplastic skin changes, such as actinic keratosis as well as nonmelanoma and melanoma skin cancer, it is striking to see that *STXBP5L* has been linked to cancer (Katoh and Katoh, 2004; Li *et al.*, 2009). Therefore, the search for gene polymorphisms involved in photoaging may also help to identify risk factors for skin carcinogenesis.

MATERIALS AND METHODS

Study design and population

A cross-sectional study was conducted to investigate skin aging in the context of the SU.VI.MAX cohort, a longitudinal cohort study, conducted in French middle-aged adults (Hercberg *et al.*, 1998). The protocol was approved by the Hospital Medicals Ethics Committee of Paris-Cochin (CCPPRB no. 706) and the "*Commission Nationale de l'Informatique et des Libertés*" (CNIL no. 334641). The study was conducted according to the Declaration of Helsinski Principles. All participants gave their written, informed consent. The SU.VI.MAX cohort included 13,017 volunteers who were representative of the French adult middle-aged population for most sociodemographic features (Hercberg *et al.*, 2004).

This study was conducted in the autumn/winter of 2002–2003. All women living in the Paris area were requested to participate in this research. Among them ($n=2{,}257$), 570 women, aged 44–70 years, agreed to take part in this study and provided informed consent. The participants were asked to follow specific skin care instructions; notably, application of detergents or cosmetics to the face was not authorized for at least 12 hours before the study visit. On the day of the visit, they were first asked to complete a self-administered questionnaire related to lifetime sun exposure behavior. Subsequently, three standardized, high-resolution digital images ($2{,}008 \times 3{,}032$ pixels) of the face were taken for each participant (one frontal view of the face and one of each profile), using a Kodak DCS 760 digital camera with a 105 mm camera lens (Kodak, Paris, France). The camera was mounted on a monopod and a specifically developed chair was used to allow standardized positions of the camera with respect to the face. Lighting conditions were standardized by means of two symmetrical lamps, which provided a continuous daylight spectrum, placed at 45° to each side of the face. Finally, a blood sample was collected for genetic analysis.

Assessment of skin aging features

The facial photographs was examined for each woman by a dermatologist, and the severity of global skin photoaging was rated using a six-grade ordinal scale (Larnier *et al.*, 1994), each grade being depicted by three reference photographs that illustrate the diversity and range of pigmentation disorders, wrinkling, and sagging. In addition, the severity of 12 age-related skin features was also assessed on forehead and on cheeks using specific ordinal photographic scales (Morizot *et al.*, 2002).

Outcome variables: phenotypes analyzed

The primary outcome variable is the global photoaging grade (1–6) and the secondary outcomes variables are the three independent scores: wrinkling, sagging, and lentigines scores. On the basis of the 12 age-related skin features, the global severity of wrinkling, sagging, and solar lentigines was estimated by three scores built using principal component analysis and linear regression methods (Jobson, 1992). Then, each individual's score values were transformed to fit a range between 0 and 10.

The solar lentigines score is computed as follows: $1.25\times$ severity on cheeks $+1.25\times$ severity on forehead (with grade $0=0$, grade $1=1$, grade $2=2$, grade $3=3$, and grade $>3=4$ for each skin area). The sagging score is based on four features: 0.87 when presence of bags under the eyes $+0.78\times$ severity of nasolabial fold (with grade $<3=1$, grade $3=2$, grade $4=3$, and grade $>4=4$) $+0.93\times$ severity of tissue slackening $+1.07\times$ severity of drooping eyelids (with grade $<3=1$, grade $3=2$, and grade $>3=3$ for the two preceding features). Finally, the wrinkling score is computed using the six remaining features: $-0.64+0.42\times$ severity of wrinkles above the upper lip (with grade $0=0$, grade $1=1$, grade $2=2$, grade $3=3$, and grade $4=4$) $+0.64\times$ severity of wrinkles under the eyes (with grade $<3=0$, grade $3=1$, grade $4=2$, and grade $5=3$) $+0.70\times$ severity of fine lines on cheek (with grade $0=0$, grade $1=1$, and grade $2=2$) $+0.44\times$ severity of furrows between eyebrows $+0.54\times$ severity of crow's feet $+1.06\times$ severity of coarse wrinkles on cheek (with grade $<2=0$, grade $2=1$, grade $3=2$, grade $4=3$, and grade $5=4$ for the three preceding features).

Covariables used for the statistical analysis: general and phenotypic data

To focus more specifically on the genetic factors affecting skin aging, several characteristics known to affect aging had to be taken into account: age (in years), body mass index (BMI; in $kg\,m^{-2}$), smoking habits (never, former, and current), and hormonal status (nonmenopausal, menopausal with hormone replacement therapy, and menopausal without hormone replacement therapy). BMI was categorized as underweight/normal (BMI $<25\,kg\,m^{-2}$), overweight ($25\leqslant BMI<30\,kg\,m^{-2}$), or obese (BMI $\geqslant 30\,kg\,m^{-2}$) according to the World Health Organization (WHO) recommendations (WHO, 1995). In addition, phenotypic data such as natural hair color at the age of 20 years, eye color, skin color in winter, sunburn event frequency, suntan intensity, and history of facial freckles were also collected. Moreover, lifetime sun exposure intensity was estimated by a score based on data collected by a self-reported questionnaire. This score is a linear combination of five items weighted according to their relative contribution to the score: voluntary sun exposure, exposure of the body and the facial skin, exposure during the hottest hours of the day, intensity of self-reported lifetime sun exposure, and consideration for sunbathing. The design, validation, and description of this score have been described previously (Guinot *et al.*, 2001).

Genotyping method

The 529 women were genotyped using Illumina Infinium HumanOmni1-Quad BeadChips (Illumina, San Diego, CA) that contain 1,140,419 markers. Genomic DNA (250 ng) was whole-genome amplified, fragmented, denatured, and hybridized on prepared HumanOmni1-Quad BeadChips for a minimum of 16 hours at 48 °C. Nonspecifically hybridized fragments were removed by washing, and the remaining specifically hybridized DNA was fluorescently labeled by a single base extension reaction and detected using a IScan scanner (Illumina). Normalized bead-intensity data obtained for each sample were loaded into GenomeStudio software (version 1.6.3; Illumina), which converted fluorescence intensities into SNP genotypes. For the analysis, we considered only SNPs, consequently

excluding the copy-number variations that represented 91,706 markers on the HumanOmni1-Quad BeadChips. Moreover, 2,182 SNPs were removed because they were located on the Y chromosome and they could not be analyzed as the population was composed of women.

Quality control

Using the GenomeStudio software (version 1.6.3; Illumina), we analyzed the crude genotyping data, and SNPs were filtered according to the following parameters. First, nine samples with a call rate (percentage of SNPs genotyped by sample) of <95% in the Illumina clusters were removed. Second, the SNPs with a call frequency (percentage of samples genotyped by SNP) of <99% were reclustered. Third, after reclustering, samples with a call rate <98% were deleted. This method has been already used in several studies (Le Clerc *et al.*, 2009; Limou *et al.*, 2009, 2010). The clustering step can create SNP genotyping errors, which can be prevented by following the Illumina procedure (http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf).

This method evaluates the quality of the newly created clusters according to several criteria, which can be manually checked and corrected as necessary. In total, after all the quality control steps were carried out, 56,479 SNPs with a call frequency of <98% (2% of missing data) were excluded. This procedure ensures reliable genotyping data with little missing data. Hardy–Weinberg equilibrium analysis was performed for each SNP in each group by using an exact statistical test implemented in PLINK software (Purcell *et al.*, 2007). Deviation from Hardy–Weinberg equilibrium in a group of patients suggests an error in genotyping. Thus, 3,866 SNPs, which were not in the Hardy–Weinberg equilibrium ($P<1.0\times10^{-3}$), were rejected in this way. We removed 191,123 SNPs with minor allele frequency <1% to avoid error of genotyping, leaving a total of 795,063 SNPs.

Identification of population stratification

To correct for possible population stratification, genotypes were analyzed using EIGENSTRAT utility of the EIGENSOFT package version 2.0 (Price *et al.*, 2006). The two first pass with the Eigenstrat software pointed out 18 outliers, which were removed from further analyses. Then, a third pass without outliers was performed to determine the Eigen vectors. In the statistical analysis, we used the top two Eigen vectors as covariables to correct for population substructure in the association analyses (Price *et al.*, 2006).

Statistical analysis

Of the 570 women who participated in the study, 68 were excluded from the analysis: 18 had a history of recent antiaging invasive procedures and 10 were observably non-Caucasian. In addition, one sample was removed because of insufficient DNA concentration, 12 samples were removed because the DNA was damaged, and nine samples were removed after quality control. Furthermore, 18 outliers appeared during the stratification analysis. Thus, the population investigated for our genome-wide association study was composed of 502 individuals.

The population was first described according to the severity of photoaging, using a series of analyses of variance for quantitative variables and using χ^2 tests for qualitative variables. In addition, Kendall rank correlation coefficients were calculated between age and each outcome variable, and between each pair of outcome variable (Armitage, 1971). Then, for the remaining 795,063 SNPs and 502 women, the associations between the genotypes and skin photoaging were tested. The statistical analysis was performed by a multivariate linear regression (PLINK software; Purcell *et al.*, 2007) in the genotypic mode, taking as covariables the two first Eigenstrat principal components and the potential confounding factors (smoking habits, BMI, hormonal status, lifetime sun exposure intensity, and age). The *P*-values were adjusted by the Bonferroni correction (statistical threshold $=6.28\times10^{-8}$). Finally, for the secondary outcome variables, additional analyses were performed using the same methodology.

Haplotype inference and LD

Haplotype inference was obtained using the rapid and accurate Shape-IT algorithm (Delaneau *et al.*, 2008, 2012). Then, for each SNP exhibiting a significant association, we looked for other SNPs in LD ($r^2>0.8$) in the HapMap population of Western European ancestry (CEU, HapMap data Release 24/phase II November 2008, on NCBI B36 assembly, dbSNP126; available at: http://www.hapmap.org) to identify the genes possibly involved with the associations. A SNP was assigned to a gene if it was located in the gene or in the 2-kb flanking regions (potential regulatory sequence); otherwise, it was considered intergenic. It is important to note that LD in HapMap population of Western European ancestry is very similar in our group of patients.

Bioinformatics exploration

To further explore the signals observed by the GWAS by using bioinformatics exploration we tried to look for modifications in mRNA expression levels (Yang *et al.*, 2010; Nica *et al.*, 2011; Dixon *et al.*, 2007; Zeller *et al.*, 2010), splicing (NetGene2, http://www.cbs.dtu.dk/services/NetGene2/), polyadenylation regions (polyAH, http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter and polyApred, http://www.imtech.res.in/raghava/polyapred/submission.html), transcription factor binding sites (SignalScan, http://www-bimas.cit.nih.gov/molbio/signal/, TESS, http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME, and TFSearch, http://www.cbrc.jp/research/db/TFSEARCH.html, derived from TRANSFAC database), and miRNA genes or miRNA targets (miRBAse, http://www.mirbase.org/, miRTarBase, http://mirtarbase.mbc.nctu.edu.tw/, MicroCosm Targets, http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/).

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

We gratefully acknowledge the dedicated efforts of all the SU.VI.MAX volunteers, the investigators, and the staff members involved in this study, especially Dr Sandrine Bertrais, and Ms Nathalie Arnault and Mr Gwenael Monot who coordinated the data management.

REFERENCES

Armitage P (1971) *Statistical Methods in Medical Research*. Blackwell Scientific: Oxford, 504 pp

Christensen K, Doblhammer G, Rau R *et al.* (2009) Ageing populations: the challenges ahead. *Lancet* 374:1196–208

Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 9:540

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9:179–81

Dixon AL, Liang L, Moffatt MF *et al.* (2007) A genome-wide association study of global gene expression. *Nat Genet* 39:1202–7

Elfakir A, Ezzedine K, Latreille J *et al.* (2010) Functional MC1R-gene variants are associated with increased risk for severe photoaging of facial skin. *J Invest Dermatol* 130:1107–15

Fisher GJ, Kang S, Varani J *et al.* (2002) Mechanisms of photoaging and chronological skin aging. *Arch Dermatol* 138:1462–70

Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol* 124:869–71

Guinot C, Malvy D, Latreille J *et al.* (2001) Sun exposure behaviour of a general adult population in France. *In:* Ring J, Weidinger S, Darsow U eds *Skin and Environment - Perception and Protection*. Monduzzi editore S.p.A: Bologna, 1099–106

Hercberg S, Galan P, Preziosi P *et al.* (2004) The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch Intern Med* 164:2335–42

Hercberg S, Galan P, Preziosi P *et al.* (1998) Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Mineraux AntioXydants Study. *Int J Vitam Nutr Res* 68:3–20

Jobson JD (1992) *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*. Springer Verlag: New York, 731 pp

Katoh M, Katoh M (2004) Identification and characterization of human LLGL4 gene and mouse Llgl4 gene in silico. *Int J Oncol* 24:737–42

Kligman A, Kligman L (1999) Photoageing. In: Freeberg IM, Eisen AZ, Wolff K *et al.*, (eds) *Fitzpatrick's Dermatology in General Medicine*. McGraw-Hill: New York, 1717–23

Larnier C, Ortonne JP, Venot A *et al.* (1994) Evaluation of cutaneous photodamage using a photographic scale. *Br J Dermatol* 130:167–73

Le Clerc S, Limou S, Coulonges C *et al.* (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J Infect Dis* 200:1194–201

Li Y, Chang M, Abar O *et al.* (2009) Multiple variants in toll-like receptor 4 gene modulate risk of liver fibrosis in Caucasians with chronic hepatitis C infection. *J Hepatol* 51:750–7

Limou S, Coulonges C, Herbeck JT *et al.* (2010) Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J Infect Dis* 202:908–15

Limou S, Le Clerc S, Coulonges C *et al.* (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 199: 419–26

Makrantonaki E, Zouboulis CC (2007) Molecular mechanisms of skin aging: state of the art. *Ann NY Acad Sci* 1119:40–50

Malvy J, Guinot C, Preziosi P *et al.* (2000) Epidemiologic determinants of skin photoaging: baseline data of the SU.VI.MAX. cohort. *J Am Acad Dermatol* 42:47–55

Morizot F, Lopez S, Guinot C *et al.* (2002) Development of photographic scales documenting features of skin ageing based on digital images. *Ann Dermatol Venereol* 129(Suppl 1 Part 2):1s402

Nica AC, Parts L, Glass D *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 7:e1002003

Plomin R, Owen MJ, McGuffin P (1994) The genetic basis of complex human behaviors. *Science* 264:1733–9

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–9

Puizina-Ivié N (2008) Skin aging. *Acta Dermatovenerol Alp Panonica Adriat* 17:47–54

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–75

Rabe JH, Mamelak AJ, McElgunn PJ *et al.* (2006) Photoaging: mechanisms and repair. *J Am Acad Dermatol* 55:1–19

Rooryck C, Morice-Picard F, Elcioglu NH *et al.* (2008) Molecular diagnosis of oculocutaneous albinism: new mutations in the OCA1-4 genes and practical aspects. *Pigment Cell Melanoma Res* 21:583–7

Safran M, Dalah I, Alexander J *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010:baq020

Shekar SN, Duffy DL, Montgomery GW *et al.* (2006) A genome scan for epidermal skin pattern in adolescent twins reveals suggestive linkage on 12p13.31. *J Invest Dermatol* 126:277–82

Shekar SN, Luciano M, Duffy DL *et al.* (2005) Genetic and environmental influences on skin pattern deterioration. *J Invest Dermatol* 125:1119–29

Shi J, Luo L, Eash J *et al.* (2011) The SCF-Fbxo40 complex induces IRS1 ubiquitination in skeletal muscle, limiting IGF1 signaling. *Dev Cell* 21:835–47

Skowyra D, Craig KL, Tyers M *et al.* (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* 91:209–19

Soufir N, Ged C, Bourillon A *et al.* (2010) A prevalent mutation with founder effect in xeroderma pigmentosum group C from north Africa. *J Invest Dermatol* 130:1537–42

Suppa M, Elliott F, Mikeljevic JS *et al.* (2011) The determinants of periorbital skin ageing in participants of a melanoma case-control study in the U.K. *Br J Dermatol* 165:1011–21

UniProt (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39:D214–9

WHO (eds) (1995) *Physical Status: The Use and Interpretation of Anthropometry*. Report of a WHO Expert Committee. WHO Technical Report Series 854Geneva: World Health Organization

Yaar M, Gilchrest BA (1990) Cellular and molecular mechanisms of cutaneous aging. *J Dermatol Surg Oncol* 16:915–22

Yaar M, Gilchrest BA (2007) Photoageing: mechanism, prevention and therapy. *Br J Dermatol* 157:874–87

Yang TP, Beazley C, Montgomery SB *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* 26:2474–6

Ye J, Zhang Y, Xu J *et al.* (2007) FBXO40, a gene encoding a novel muscle-specific F-box protein, is upregulated in denervation-related muscle atrophy. *Gene* 404:53–60

Zeller T, Wild P, Szymczak S *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5:e10693

F.2 Publication à soumettre

Bernard, A., Abdi, H., Tenenhaus, A., Guinot, C., Saporta, G. Sparse Principal Component Analysis for multiblocks data and its extension to Sparse Multiple Correspondence. *Computational Statistics & Data Analysis.*

Sparse Principal Component Analysis for Multiblock Data and its Extension to Sparse Multiple Correspondence Analysis

A. Bernard[a,b,*], C. Guinot[c], A. Tenenhaus[d], H. Abdi[e], G. Saporta[a]

[a] *CNAM, laboratoire CEDRIC, 292 rue Saint-Martin, Paris, France*
[b] *CE.R.I.E.S., 20 rue Victor Noir, Neuilly sur Seine, France*
[c] *Université Francois Rabelais, département d'informatique, 64 avenue Jean Portalis, Tours, France*
[d] *SUPELEC, 3, rue Joliot-Curie, Gif-sur-Yvette, France*
[e] *Department of Brain and Behavioral Sciences, The University of Texas at Dallas, Richardson, TX, USA*

Abstract

Principal Component Analysis (PCA) for quantitative data, and Multiple Correspondence Analysis (MCA) for qualitative data are well-known dimension reduction methods. However, the principal components obtained are combinations of all the original variables, making interpretation of results difficult for high dimensional data. To overcome these difficulties, we propose two new methods for selecting groups of quantitative and qualitative variables: Group Sparse Principal Component Analysis (GSPCA) and Sparse Multiple Correspondence Analysis (SMCA), respectively. GSPCA is an extension of SPCA-RSVD of Shen and Huang for data structured by blocks. It uses the connection between PCA and singular value decomposition of a data matrix to extract components through solving a low rank matrix approximation problem. A regularization penalty "group Lasso" is introduced to the corresponding minimization problem to obtain components that are combinations of a few number of groups of vari-

*Corresponding author
Email addresses: anne.bernard@ceries-lab.com (A. Bernard), christiane.guinot@univ-tours.fr (C. Guinot), arthur.tenenhaus@supelec.fr (A. Tenenhaus), herve@utdallas.edu (H. Abdi), gilbert.saporta@cnam.fr (G. Saporta)

ables. All loadings of a block of variables are set to zero to reduce the number of selected variables. Since MCA is a special case of PCA for blocks of dummy variables, SMCA is defined as an extension of GSPCA. An application of this method will be presented on a real genetic data set of SNPs.
Keywords: Dimension reduction, Sparse Principal Component Analysis, Multiple Correspondence Analysis, singular value decomposition, multiblocks methods.

1. Introduction

In all areas of science and engineering, such as machine learning, statistics, finance and more particularly genetics, the challenge is to analyze and interpret high dimensional data. In recent years, the search for the association between complex diseases and single nucleotide polymorphisms (SNPs) has received great attention. Modern scientific technology, led by the microarray, has produced data dramatically above the conventional scale, with a number of variables that can reach more than 1 million, but most of them are non informative or noisy. Therefore, the challenge to develop appropriate sparse methodologies is very important, to reduce the number of variables and infer reliable and interpretable results. Principal component analysis (PCA) [10] is a popular tool for quantitative data analysis and dimensionality reduction. PCA aims at finding linear combinations of all the input variables, which makes principal components difficult to interpret and explain. Accordingly, components that are linear combinations of a small number of variables are easier to interpret. Therefore, in the recent years many versions of sparse principal component analysis (sparse PCA) has been developed to find a reasonable trade-off between explaining as much variability in the data as possible and achieving representation sparsity simultaneously (using components constructed from as few variables as possible), such as SCoTLASS [6], Sparse PCA [5] and Sparse PCA-rSVD [2].

In the field of genetic, data are naturally structured by blocks (chromosomes, genes, mitochondria, etc ...) and can be qualitative kind (SNPs). Analysis of multiblocks quantitative data such as MFA, and analysis of qualitative data such as MCA are well-known dimension reduction methods, but did not exist as sparse. In this paper we propose two new sparse data analysis techniques: Group Sparse PCA for high dimensional multiblocks quantitative data, and Sparse MCA for high dimensional qualitative data. This paper is organized as follows. The extension of sparse PCA via regularized SVD of Shen and Huang for data structured by blocks (GSPCA) is presented in section 2. The link between this new method and MCA is done and the adaptation of this method for blocks of categorical variables is developed in section 3 with the sparse extension of multiple correspondence analysis (sparse MCA). Properties of these new methods are explained in section 4 and the usefulness of SMCA is demonstrated on an application in genomic data analysis in section 5. Section 6 discusses some extension that can be done on this new method.

2. Notations and preliminary

Matrices are denoted by boldface uppercase letters (e.g., $\mathbf{X}$), vectors by boldface lowercase letters (e.g., $\mathbf{q}$), elements of vectors and matrices are denoted by italic lower case letters with appropriate indices if needed (e.g., $x_{i,j}$ is an element of $\mathbf{X}$). The identity matrix is denoted by $\mathbf{I}$, a column vector of ones is denoted by $\mathbf{1}$. The rank of a matrix is denoted rank(). The transpose of a matrix is denoted ^{T} and the inverse $^{-1}$. When applied to a square matrix, the diag operator denoted diag() takes the diagonal elements of this matrix and stores them into a column vector; when applied to a vector, the diag operator stores the elements of this vector on the diagonal elements of a diagonal matrix. The trace operator denoted tr() computes the sum of the diagonal elements of

a square matrix.

The L_2 norm is denoted by $||.||$ and is defined as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \left(\sum_{i=1}^{I} |x_i|^2\right)^{1/2}. \tag{1}$$

The L_2 norm $\mathbf{W}$-generalized is denoted by $\|.\|_{\mathbf{W}}$ and is defined as follows

$$\|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}^T\mathbf{W}\mathbf{x}}. \tag{2}$$

with $\mathbf{W}$ a positive definite matrix.

For all squares matrices $\mathbf{A}$ and $\mathbf{B}$, the following properties are verified:

Property 1.

$$\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \tag{3}$$

Property 2.

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}((\mathbf{AB})^T) = \mathrm{tr}(\mathbf{B}^T\mathbf{A}^T) \tag{4}$$

Property 3.

$$\mathrm{tr}(\mathbf{A}\mathbf{A}^T) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) \tag{5}$$

The standard product between matrices is implicitly denoted by simple juxtaposition or by $\times$ when it needs to be explicitly stated (e.g., $\mathbf{XY} = \mathbf{X} \times \mathbf{Y}$ is the product of matrices $\mathbf{X}$ and $\mathbf{Y}$).

The rank-one matrices are denoted $\mathbf{X}^{(1)}$ and the rank-L matrices $\mathbf{X}^{(L)}$. If the data consist of K data sets collected on the same observation, each data set is called a sub-table (or also a block). The data for each table are stored in an $I \times J_{[k]}$ rectangular data matrix denoted $\mathbf{X}_{[k]}$, where I is the number of

observations and $J_{[k]}$ the number of variables collected on the observations for the k-th table. The total number of variables is denoted J (*i.e.*, $J = \sum_{k=1}^{K} J_{[k]}$). Each data matrix is, in general, preprocessed (e.g., centered, normalized) and the preprocessed data matrices actually used in the analysis are denoted $\mathbf{X}_{[k]}$. Blocks of variables are considered as sub-matrices of larger matrices and are represented in brackets separated by vertical bars. For example, the K data matrices $\mathbf{X}_{[k]}$, each of dimensions I rows by $J_{[k]}$ columns, are concatenated into the complete I by J data matrix denoted $\mathbf{X}$:

$$\mathbf{X} = \left[\mathbf{X}_{[1]}|\ldots|\mathbf{X}_{[k]}|\ldots|\mathbf{X}_{[K]}\right]. \tag{6}$$

The operator diag applied on a block matrix $\mathbf{X}$ produces a square matrix block diagonal with each block of $\mathbf{X}$being a block of the diagonal.

Matrix derivative properties

Let us consider a matrix $\mathbf{X}$ and, $\mathbf{A}$ and $\mathbf{B}$, matrices of constants which does not depend on $\mathbf{X}$. We consider the following properties

Property 4.

$$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} = \frac{\operatorname{tr}(\partial \mathbf{X})}{\partial \mathbf{X}} \tag{7}$$

Property 5.

$$\frac{\partial \operatorname{tr}(\mathbf{X}^T\mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{X} \tag{8}$$

Property 6.

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial \operatorname{tr}(\mathbf{X}\mathbf{A})}{\partial \mathbf{X}} = \mathbf{A}^T \tag{9}$$

Property 7.

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{B}^T \tag{10}$$

Property 8. The sub-gradient of the L_2 norm $\|\mathbf{w}\|_2$, with $\mathbf{w}$ a vector $\in \mathbb{R}^J$ is equal to:

$$\partial \|\mathbf{w}\|_2 = \begin{cases} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} & \text{if} \quad \mathbf{w} \neq 0 \\ \\ \in \{\mathbf{z}|\ \|\mathbf{z}\|_2 \leq 1\} & \text{if} \quad \mathbf{w} = 0 \end{cases} \tag{11}$$

with $\mathbf{z}$ a vector $\in \mathbb{R}^J$.

2.1. SVD

2.1.1. Definition

The singular value decomposition (SVD) is a factorization of a rectangular matrix, with many useful applications in multivariate statistics. The SVD of an $I \times J$ matrix $\mathbf{X}$ of rank L is equal to

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \text{ with } \mathbf{P}^T\mathbf{P} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}. \tag{12}$$

where $\mathbf{P}$ is an $I \times L$ orthonormal matrix of the left singular vectors, $\mathbf{Q}$ the $J \times L$ orthonormal matrix of the right singular vectors, and $\mathbf{\Delta}$ the $L \times L$ diagonal matrix of the singular values with δ_ℓ the diagonal elements of $\mathbf{\Delta}$, $\ell = 1, ..., L$. $\mathbf{P}$ is also the matrix of the normalized eigenvectors of $\mathbf{X}\mathbf{X}^T$, $\mathbf{Q}$ the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and the singular values are the square root of the eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$. We define $\mathbf{F} = \mathbf{P}\mathbf{\Delta}$ the factor scores with $\mathbf{F}^T\mathbf{F} = \mathbf{\Delta}^2$ and $\mathbf{Q}$ the loadings.

2.1.2. Properties of the SVD

The SVD provides the best reconstitution (in a least squares sense) of the original matrix by a matrix with a lower rank. The best rank-one matrix ap-

proximation $\mathbf{X}^{(1)}$ of $\mathbf{X}$ is the solution of the following optimization problem:

$$\arg\min_{\mathbf{X}^{(1)}} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 = \arg\min_{\mathbf{X}^{(1)}} \left(\operatorname{tr}\left((\mathbf{X}-\mathbf{X}^{(1)})^T(\mathbf{X}-\mathbf{X}^{(1)})\right)\right). \tag{13}$$

where

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T. \tag{14}$$

Proof. The least squares problem reduces to finding a rank-deficient matrix approximation $\mathbf{X}^{(1)}$ to a given matrix $\mathbf{X}$. This can be formulated as

$$\arg\min_{\mathbf{X}^{(1)}\in\mathbb{R}^{I\times J},\mathbf{q}\in\mathbb{R}^J} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 \qquad \text{subject to} \qquad \mathbf{p}^T\mathbf{p}=1 \text{ and } \mathbf{q}^T\mathbf{q}=1. \tag{15}$$

The matrix $\mathbf{X}^{(1)}$ has the form $\mathbf{f}\mathbf{q}^T$ and we want to find the optimal $\mathbf{f}$ and $\mathbf{q}$ which are solutions to the problem (15). The expression $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2$ is equal to

$$\begin{aligned}
\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 &= \operatorname{tr}\left((\mathbf{X}-\mathbf{X}^{(1)})^T(\mathbf{X}-\mathbf{X}^{(1)})\right) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}^{(1)} - {\mathbf{X}^{(1)}}^T\mathbf{X} + {\mathbf{X}^{(1)}}^T\mathbf{X}^{(1)}) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X}) - 2\operatorname{tr}({\mathbf{X}^{(1)}}^T\mathbf{X}) + \operatorname{tr}({\mathbf{X}^{(1)}}^T\mathbf{X}^{(1)}) \\
&= \|\mathbf{X}\|^2 - 2\operatorname{tr}(\mathbf{q}\mathbf{f}^T\mathbf{X}) + \operatorname{tr}(\mathbf{q}\mathbf{f}^T\mathbf{f}\mathbf{q}^T) \\
&= \|\mathbf{X}\|^2 - 2\operatorname{tr}(\mathbf{q}\mathbf{f}^T\mathbf{X}) + \delta^2\operatorname{tr}(\mathbf{q}\mathbf{q}^T) \\
&= \|\mathbf{X}\|^2 - 2\delta\operatorname{tr}(\mathbf{q}\mathbf{p}^T\mathbf{X}) + \delta^2.
\end{aligned} \tag{16}$$

because $\mathbf{f}^T\mathbf{f} = \delta^2\mathbf{p}^T\mathbf{p} = \mathbf{1}$ and $\mathbf{q}^T\mathbf{q} = \mathbf{1}$. We are looking for the min of $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2$. Let α and $\gamma \in \mathbb{R}$ be Lagrange multipliers; then the Lagrangian

can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{p},\mathbf{q},\boldsymbol{\theta},\gamma) &= \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 + \alpha(\mathbf{p}^T\mathbf{p}-1) + \gamma(\mathbf{q}^T\mathbf{q}-1) \\ &= \|\mathbf{X}\|^2 - 2\delta\,\mathrm{tr}(\mathbf{q}\mathbf{p}^T\mathbf{X}) + \delta^2 + \alpha(\mathbf{p}^T\mathbf{p}-1) + \gamma(\mathbf{q}^T\mathbf{q}-1).\end{aligned} \tag{17}$$

The min on $\mathbf{p}$ and $\mathbf{q}$ is obtained when all derivatives are set to zero:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{p}} &= -2\delta\mathbf{X}\mathbf{q} + 2\alpha\mathbf{p} = 0 \qquad &\Rightarrow \qquad \delta\mathbf{X}\mathbf{q} &= \alpha\mathbf{p} \\ \frac{\partial\mathcal{L}}{\partial\mathbf{q}} &= -2\delta\mathbf{X}^T\mathbf{p} + 2\gamma\mathbf{q} = 0 \qquad &\Rightarrow \qquad \delta\mathbf{X}^T\mathbf{p} &= \gamma\mathbf{q}\end{aligned} \tag{18}$$

However, when $\mathbf{X}$ is multiplied by a vector on the left and the right, the matrix becomes a rank-one matrix.

$$\begin{aligned}\delta\mathbf{X}\mathbf{q} = \alpha\mathbf{p} \qquad &\Leftrightarrow \qquad \delta\mathbf{p}\delta\mathbf{q}^T\mathbf{q} = \alpha\mathbf{p} \\ &\Leftrightarrow \qquad \delta^2\mathbf{p} = \alpha\mathbf{p} \\ &\Leftrightarrow \qquad \delta^2 = \alpha\end{aligned} \tag{19}$$

and

$$\begin{aligned}\delta\mathbf{X}^T\mathbf{p} = \gamma\mathbf{q} \qquad &\Leftrightarrow \qquad \delta\mathbf{q}\delta\mathbf{p}^T\mathbf{p} = \gamma\mathbf{q} \\ &\Leftrightarrow \qquad \delta^2\mathbf{q} = \gamma\mathbf{q} \\ &\Leftrightarrow \qquad \delta^2 = \gamma\end{aligned} \tag{20}$$

Hence we can rewrite (18) as

$$\begin{aligned}\mathbf{X}\mathbf{q} = \delta\mathbf{p} \qquad &\text{with} \qquad \mathbf{p}^T\mathbf{p} = \mathbf{1} \\ \mathbf{X}^T\mathbf{p} = \delta\mathbf{q} \qquad &\text{with} \qquad \mathbf{q}^T\mathbf{q} = \mathbf{1}\end{aligned} \tag{21}$$

which implies that $(\delta, \mathbf{p}, \mathbf{q})$ must be the singular triplet of $\mathbf{X}$ with δ a singular value, $\mathbf{p}$ a left singular vector and $\mathbf{q}$ a right singular vector. To obtain the min of $\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|^2$, we need δ to be the largest singular value according to the problem (15) and (16). So the singular triplet corresponds to the largest singular value. We conclude that $\mathbf{p} = \mathbf{p}_1$ and $\mathbf{q} = \mathbf{q}_1$, with $\mathbf{X}^{(1)} = \delta_1 \mathbf{p}_1 \mathbf{q}_1^T$ being the best rank-one approximation matrix of $\mathbf{X}$. □

Knowing that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ (which implies that $\mathbf{f}_1^T\mathbf{f}_1 = \delta_1^2\mathbf{p}_1^T\mathbf{p}_1 = \delta_1^2$) and taking into account Properties 1 and 2 of section 2, we can rewrite

$$\begin{aligned}
\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|^2 &= \operatorname{tr}\left((\mathbf{X} - \mathbf{f}_1\mathbf{q}_1^T)^T(\mathbf{X} - \mathbf{f}_1\mathbf{q}_1^T)\right) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{f}_1\mathbf{q}_1^T - \mathbf{q}_1\mathbf{f}_1^T\mathbf{X} + \mathbf{q}_1\mathbf{f}_1^T\mathbf{f}_1\mathbf{q}_1^T) \\
&= \operatorname{tr}(\mathbf{X}^T\mathbf{X}) - \operatorname{tr}(\mathbf{X}^T\mathbf{f}_1\mathbf{q}_1^T) - \operatorname{tr}(\mathbf{q}_1\mathbf{f}_1^T\mathbf{X}) + \operatorname{tr}(\mathbf{q}_1\mathbf{f}_1^T\mathbf{f}_1\mathbf{q}_1^T) \\
&= \|\mathbf{X}\|^2 - 2\operatorname{tr}(\mathbf{X}^T\mathbf{f}_1\mathbf{q}_1^T) + \operatorname{tr}(\mathbf{q}_1\mathbf{f}_1^T\mathbf{f}_1\mathbf{q}_1^T) \\
&= \|\mathbf{X}\|^2 - 2\operatorname{tr}(\mathbf{X}^T\mathbf{f}_1\mathbf{q}_1^T) + \delta_1^2\operatorname{tr}(\mathbf{q}_1\mathbf{q}_1^T) \\
&= \|\mathbf{X}\|^2 - 2\operatorname{tr}(\mathbf{X}^T\mathbf{f}_1\mathbf{q}_1^T) + \delta_1^2.
\end{aligned} \tag{22}$$

The subsequent best rank-one approximations can be obtained sequentially via rank-one approximation of residual matrices.
The pairs $(\mathbf{f}_k, \mathbf{q}_k)$, $k > 1$, provides the best rank one approximation of the corresponding residual matrix. For example, $\mathbf{f}_2\mathbf{q}_2^T$ is the best rank one approximation of the first deflated matrix:

$$\mathbf{X}^{\perp_1} = \mathbf{X} - \mathbf{f}_1\mathbf{q}_1^T. \tag{23}$$

$\mathbf{X}^{\perp_1}$ is the orthogonal complement of $\mathbf{X}$ and if the same procedure is repeated on $\mathbf{X}^{\perp_1}$, the first singular vector of $\mathbf{X}^{\perp_1}$ will be the second one of the matrix

$\mathbf{X}$. The low rank approximation matrices can be generalized to rank-L.

In this case, the optimization problem becomes

$$\underset{\mathbf{X}^{(L)}}{\arg\min} \left\| \mathbf{X} - \mathbf{X}^{(L)} \right\|^2 = \underset{\mathbf{X}^{(L)}}{\arg\min} \left(\mathrm{tr} \left((\mathbf{X} - \mathbf{X}^{(L)})^T (\mathbf{X} - \mathbf{X}^{(L)}) \right) \right) \tag{24}$$

and the solution is

$$\mathbf{X}^{(L)} = \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^T = \sum_{\ell=1}^{L} \mathbf{f}_\ell \mathbf{q}_\ell^T \tag{25}$$

where $\mathbf{X}^{(L)}$ is the closest rank-L matrix approximation to $\mathbf{X}$ ([1]).

2.1.3. SVD on a block matrix

When $\mathbf{X}$ is made of K sub-matrices (see (6)), the SVD can be accommodate to the block structure:

$$\begin{aligned} \mathbf{X} &= \left[\mathbf{X}_{[1]} | \dots | \mathbf{X}_{[k]} | \dots | \mathbf{X}_{[K]} \right] \\ &= \left[\mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[1]}^T | \dots | \mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[k]}^T | \dots | \mathbf{P}\mathbf{\Delta}\mathbf{Q}_{[K]}^T \right] \\ &= \mathbf{P}\mathbf{\Delta} \left[\mathbf{Q}_{[1]}^T | \dots | \mathbf{Q}_{[k]}^T | \dots | \mathbf{Q}_{[K]}^T \right] \end{aligned} \tag{26}$$

where $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T | \dots | \mathbf{Q}_{[k]}^T | \dots | \mathbf{Q}_{[K]}^T \right]^T$ is the matrix of the singular vectors with $\mathbf{Q}_{[k]}^T$ the sub-matrices, each of dimension $J_{[k]} \times L$ (see Figure 2.1.3).

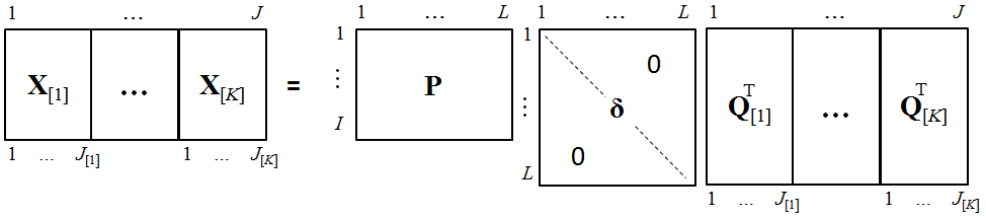

Figure 1: SVD on a block matrix $\mathbf{X}$

When the blocks $\left[\mathbf{X}_{[1]} | \dots | \mathbf{X}_{[k]} | \dots | \mathbf{X}_{[K]} \right]$ are reduced to one variable (col-

umn), SVD on a block matrix reduces to the standard SVD.

2.1.4. Properties of the SVD on a block matrix

In this framework, the SVD also provides the best approximation of the original matrix structured by blocks to a lower rank matrix. The best rank-one matrix approximation $\mathbf{X}^{(1)}$ of $\mathbf{X}$ is the solution of the following optimization problem

$$\arg\min_{\mathbf{X}^{(1)}} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 = \arg\min_{\mathbf{X}^{(1)}} \left(\operatorname{tr}\left((\mathbf{X}-\mathbf{X}^{(1)})^T(\mathbf{X}-\mathbf{X}^{(1)})\right)\right) \tag{27}$$

where

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T. \tag{28}$$

According to Property 2 and Equation (16), the expression 27 can be reconsidered as

$$\begin{aligned} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|^2 &= \operatorname{tr}\left((\mathbf{X}-\mathbf{f}_1\mathbf{q}_1^T)^T(\mathbf{X}-\mathbf{f}_1\mathbf{q}_1^T)\right) \\ &= \|\mathbf{X}\|^2 - 2\operatorname{tr}\left(\mathbf{q}_1\mathbf{f}_1^T\mathbf{X}\right) + \delta_1^2 \\ &= \|\mathbf{X}\|^2 - 2\sum_{k=1}^{K}\operatorname{tr}\left(\mathbf{q}_{1,[k]}\mathbf{f}_1^T\mathbf{X}_{[k]}\right) + \delta_1^2. \end{aligned} \tag{29}$$

where $\mathbf{q}_1^T=(\mathbf{q}_{1,[1]},\ldots,\mathbf{q}_{1,[k]},\ldots,\mathbf{q}_{1,[K]})^T$ is the first row of $\mathbf{Q}^T$ with $\mathbf{q}_{1,[k]}$ a vector of dimension $J_{[k]} \times 1$ (see Figure 2.1.4). Note that the matrices $\mathbf{q}_{1,[k]}\mathbf{q}_{1,[k]}^T$ and $\mathbf{q}_{1,[k]}\mathbf{f}_1^T\mathbf{X}_{[k]}$ are $J_{[k]} \times J_{[k]}$ matrices.

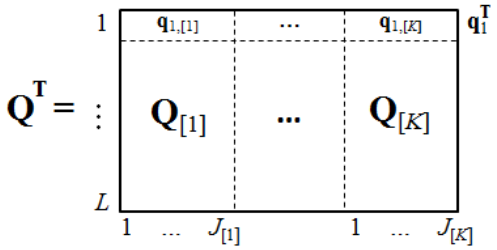

Figure 2: Explanation of the **Q** matrix

2.1.5. PCA as SVD

Principal component analysis (PCA) ([6]) is a popular tool for data analysis and dimensionality reduction. PCA can be defined from the SVD with **F**=**PΔ** the matrix of factor scores and **Q** the matrix of the loadings.

2.1.6. PCA as SVD on a block matrix

When **X** is made of K sub-matrices (see (6)), PCA of **X** can be defined from the SVD on a block matrix (cf. section 2.1.3) with :

$$\begin{aligned}\mathbf{X} &= \left[\mathbf{X}_{[1]}|\dots|\mathbf{X}_{[k]}|\dots|\mathbf{X}_{[K]}\right] \\ &= \mathbf{P\Delta}\left[\mathbf{Q}_{[1]}^T|\dots|\mathbf{Q}_{[k]}^T|\dots|\mathbf{Q}_{[K]}^T\right] \\ &= \mathbf{F}\left[\mathbf{Q}_{[1]}^T|\dots|\mathbf{Q}_{[k]}^T|\dots|\mathbf{Q}_{[K]}^T\right]\end{aligned} \tag{30}$$

In this framework, the matrix of the loadings is $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T|\dots|\mathbf{Q}_{[k]}^T|\dots|\mathbf{Q}_{[K]}^T\right]^T$ with $\mathbf{Q}_{[k]}^T$ the sub-matrices, each of dimension $J_{[k]} \times L$, and the matrix of factor scores is **F**=**PΔ**.

2.1.7. SVD as a regression-type problem

Consider the SVD of $\mathbf{X} = \mathbf{P\Delta Q}$ with $\mathbf{F} = \mathbf{P\Delta}$ being the matrix of factor scores. Knowing that the factor scores $\mathbf{f}_1$ are a linear combination of the

variables, we want to recover the loadings. The Ordinary Least Squares (OLS) estimates $\hat{\boldsymbol{\beta}}$ are obtained by minimizing this expression

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{f}_1 - \mathbf{X}\boldsymbol{\beta} \right\|^2 . \tag{31}$$

The OLS is solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{f}_1. \tag{32}$$

In the case of high-dimensional data, $\operatorname{rank}(\mathbf{X}) \leq L < J$, so $\mathbf{X}$ is not invertible. To circumvent this problem, we consider the Moore-Penrose pseudoinverse matrix $\mathbf{X}^+$ defined as

$$\mathbf{X}^+ = \mathbf{Q}\boldsymbol{\Delta}^+\mathbf{P}^T. \tag{33}$$

with $\boldsymbol{\Delta}^+$ the pseudoinverse matrix of the diagonal matrix $\boldsymbol{\Delta}$, obtained by taking the reciprocal of the non-zero diagonal elements, and transposing the resulting matrix. Expression (32) becomes

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^+\mathbf{f}_1. \tag{34}$$

Because $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, $\mathbf{Q}^T\mathbf{q}_1$ is the first column of the $L \times L$ identity matrix, and so $\mathbf{Q}^T\mathbf{q}_1 = [1, 0, \ldots, 0]^T$ of dimension $L \times 1$. From the SVD of $\mathbf{X}$

$$\begin{aligned} \mathbf{X}\mathbf{q}_1 &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{q}_1 \\ &= \mathbf{F}\mathbf{Q}^T\mathbf{q}_1 \\ &= \mathbf{F}\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{f}_1. \end{aligned} \tag{35}$$

So the OLS solution becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{X}^{+}\mathbf{f}_1 \\ &= \mathbf{X}^{+}\mathbf{X}\mathbf{q}_1 \\ &= \mathbf{q}_1.\end{aligned} \tag{36}$$

In conclusion, the loadings $\mathbf{q}_1$ can be recovered by regressing the factor scores $\mathbf{f}_1$ on the J variables because each column of $\mathbf{F}$ is a linear combination of the J variables. Thereby the SVD and so PCA, can be seen as a regression-type optimization problem.

2.2. GSVD

2.2.1. Definition

The Generalized SVD (GSVD) generalizes SVD by incorporating the following constraints

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \text{ with } \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}. \tag{37}$$

The matrix $\mathbf{M}$ is an $I \times I$ positive definite matrix representing the constraints imposed on the rows. If $\mathbf{M}$ is diagonal, the elements of $\mathbf{M}$ are called masses. The matrix $\mathbf{W}$ is a $J \times J$ positive definite matrix representing the constraints imposed on the columns of $\mathbf{X}$ (metric on the variables). If $\mathbf{W}$ is diagonal, the elements of $\mathbf{W}$ are called the weights.

2.2.2. Properties of the GSVD

The GSVD also provide the best reconstitution of the original matrix by a matrix with a lower rank. The best rank-one matrix approximation $\mathbf{X}^{(1)}$ of $\mathbf{X}$ is the solution of the minimization of the norm $\mathbf{W}$-generalized weighted by the

masses $\mathbf{M}$

$$\arg\min_{\mathbf{X}^{(1)}} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 = \arg\min_{\mathbf{X}^{(1)}} \left(\operatorname{tr}\left(\mathbf{M}^{1/2}(\mathbf{X}-\mathbf{X}^{(1)})\mathbf{W}(\mathbf{X}-\mathbf{X}^{(1)})^T\mathbf{M}^{1/2}\right)\right) \quad (38)$$

where

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T \quad (39)$$

Proof. The least squares problem reduces to finding a rank-deficient matrix approximation $\mathbf{X}^{(1)}$ of a given matrix $\mathbf{X}$. This can be formulated as

$$\arg\min_{\mathbf{X}^{(1)}\in\mathbb{R}^{I\times J},\mathbf{q}\in\mathbb{R}^J} \left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 \qquad \text{subject to} \qquad \mathbf{p}^T\mathbf{M}\mathbf{p}=1 \text{ and } \mathbf{q}^T\mathbf{W}\mathbf{q}=1 \quad (40)$$

The matrix $\mathbf{X}^{(1)}$ has the form $\mathbf{f}\mathbf{q}^T$ and we want to find the optimal $\mathbf{f}$ and $\mathbf{q}$ which are solutions to the problem (40). The expression $\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2$ is equal to

$$\begin{aligned}
\left\|\mathbf{X}-\mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 &= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}-\mathbf{X}^{(1)})\mathbf{W}(\mathbf{X}-\mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right) \quad (41)\\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}-\mathbf{X}^{(1)}\mathbf{W})(\mathbf{X}-\mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right)\\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}\mathbf{X}^T-\mathbf{X}\mathbf{W}{\mathbf{X}^{(1)}}^T-\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T+\mathbf{X}^{(1)}\mathbf{W}{\mathbf{X}^{(1)}}^T)\mathbf{M}^{\frac{1}{2}}\right)\\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X}\mathbf{W}\mathbf{X}^T-2\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T+\mathbf{X}^{(1)}\mathbf{W}{\mathbf{X}^{(1)}}^T)\mathbf{M}^{\frac{1}{2}}\right)\\
&= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}-2\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}+\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}{\mathbf{X}^{(1)}}^T\mathbf{M}^{\frac{1}{2}}\right)\\
&= \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{X}^{(1)}\mathbf{W}{\mathbf{X}^{(1)}}^T\mathbf{M}^{\frac{1}{2}})\\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{f}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{f}\mathbf{q}^T\mathbf{W}\mathbf{q}\mathbf{f}^T\mathbf{M}^{\frac{1}{2}})\\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{f}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{f}\mathbf{f}^T\mathbf{M}^{\frac{1}{2}})\\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{f}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\delta^2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{p}^T\mathbf{M}^{\frac{1}{2}})\\
&= \|\mathbf{X}\|_{\mathbf{W}}^2-2\delta\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}})+\delta^2.
\end{aligned}$$

because $\mathbf{p}^T\mathbf{M}\mathbf{p} = 1$ and $\mathbf{q}^T\mathbf{W}\mathbf{q} = 1$. Let α and $\gamma \in \mathbb{R}$ be Lagrange multipliers; then the Lagrangian can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{p}, \mathbf{q}, \boldsymbol{\theta}, \gamma) &= \left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 + \alpha(\mathbf{p}^T\mathbf{M}\mathbf{p} - 1) + \gamma(\mathbf{q}^T\mathbf{W}\mathbf{q} - 1) \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\delta \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2 \\ &+ \alpha(\mathbf{p}^T\mathbf{p} - 1) + \gamma(\mathbf{q}^T\mathbf{q} - 1).\end{aligned} \tag{42}$$

Setting all derivatives equal to zero we obtain

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{p}} &= -2\delta\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q} + 2\alpha\mathbf{M}\mathbf{p} = 0 \quad &\Rightarrow \quad \delta\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q} = \alpha\mathbf{M}\mathbf{p} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{q}} &= -2\delta\mathbf{W}\mathbf{X}^T\mathbf{M}\mathbf{p} + 2\gamma\mathbf{W}\mathbf{q} = 0 \quad &\Rightarrow \quad \delta\mathbf{W}\mathbf{X}^T\mathbf{M}\mathbf{p} = \gamma\mathbf{W}\mathbf{q}.\end{aligned} \tag{43}$$

However,

$$\begin{aligned}\delta\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q} = \alpha\mathbf{M}\mathbf{p} \quad &\Leftrightarrow \quad \delta\mathbf{M}\mathbf{p}\delta\mathbf{q}^T\mathbf{W}\mathbf{q} = \alpha\mathbf{M}\mathbf{p} \\ &\Leftrightarrow \quad \delta^2\mathbf{M}\mathbf{p} = \alpha\mathbf{M}\mathbf{p} \\ &\Leftrightarrow \quad \delta^2 = \alpha\end{aligned} \tag{44}$$

and

$$\begin{aligned}\delta\mathbf{W}\mathbf{X}^T\mathbf{M}\mathbf{p} = \gamma\mathbf{W}\mathbf{q} \quad &\Leftrightarrow \quad \delta\mathbf{W}\mathbf{q}\delta\mathbf{p}^T\mathbf{M}\mathbf{p} = \gamma\mathbf{W}\mathbf{q} \\ &\Leftrightarrow \quad \delta^2\mathbf{W}\mathbf{q} = \gamma\mathbf{W}\mathbf{q} \\ &\Leftrightarrow \quad \delta^2 = \gamma.\end{aligned} \tag{45}$$

Hence we can rewrite (43) as

$$\begin{aligned} \mathbf{X}\mathbf{q} &= \delta\mathbf{p} \qquad \text{with} \qquad \mathbf{p}^T\mathbf{p} = 1 \\ \mathbf{X}^T\mathbf{p} &= \delta\mathbf{q} \qquad \text{with} \qquad \mathbf{q}^T\mathbf{q} = 1. \end{aligned} \tag{46}$$

which implies that $(\delta, \mathbf{p}, \mathbf{q})$ must be the singular triplet of $\mathbf{X}$ and δ has to be the largest singular value according to the problem (40) and (41). So we need the singular triplet corresponding to the largest singular value. We conclude that $\mathbf{p} = \mathbf{p}_1$ and $\mathbf{q} = \mathbf{q}_1$, with $\mathbf{X}^{(1)} = \delta_1\mathbf{p}_1\mathbf{q}_1$ being the best rank-one approximation matrix of $\mathbf{X}$. □

The low rank approximation matrices can be generalized for the rank-L. The optimization problem becomes

$$\arg\min_{\mathbf{X}^{(L)}} \left\|\mathbf{X} - \mathbf{X}^{(L)}\right\|_{\mathbf{W}}^2 = \arg\min_{\mathbf{X}^{(L)}} \left(\mathrm{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \mathbf{X}^{(L)})\mathbf{W}(\mathbf{X} - \mathbf{X}^{(L)})^T\mathbf{M}^{\frac{1}{2}}\right)\right) \tag{47}$$

and the solution is

$$\mathbf{X}^{(L)} \equiv \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^T \equiv \sum_{\ell=1}^{L} \mathbf{f}_\ell \mathbf{q}_\ell^T. \tag{48}$$

where $\mathbf{X}^{(L)}$ is the closest rank-L matrix approximation to $\mathbf{X}$ ([1]). The term "closest" means that $\mathbf{X}^{(L)}$ minimizes the squared norm between $\mathbf{X}$ and an arbitrary rank-L matrix.

2.2.3. GSVD on a block matrix

The GSVD of a matrix $\mathbf{X}$ which is made of K sub-matrices is

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \text{ with } \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I} \tag{49}$$

with $\mathbf{M}$ a $I \times I$ matrix, $\mathbf{Q} = \left[\mathbf{Q}_{[1]}^T | \dots | \mathbf{Q}_{[k]}^T | \dots | \mathbf{Q}_{[K]}^T\right]^T$ and $\mathbf{W}$ a $J \times J$ positive definite matrix which can be decomposed in K positive definite sub-matrices

$$\mathbf{W} = \operatorname{diag}\left(\left[\mathbf{W}_{[1]}^T | \dots | \mathbf{W}_{[k]}^T | \dots | \mathbf{W}_{[K]}^T\right]\right) \tag{50}$$

with $\mathbf{W}_{[k]}$ of dimensions $J_{[k]} \times J_{[k]}$ and $\mathbf{W}$ a block diagonal matrix which blocks are in line with those of the matrix $\mathbf{Q}$.

2.2.4. Properties of the GSVD on a block matrix

In this framework, the GSVD also provides the best reconstitution of the original matrix structured by blocks with a lower rank. The best rank-one matrix approximation $\mathbf{X}^{(1)}$ of $\mathbf{X}$ is the solution of the following optimization problem

$$\arg\min_{\mathbf{X}^{(1)}} \left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 = \arg\min_{\mathbf{X}^{(1)}} \left(\operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \mathbf{X}^{(1)})\mathbf{W}(\mathbf{X} - \mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right)\right) \tag{51}$$

where

$$\mathbf{X}^{(1)} \equiv \delta_1 \mathbf{p}_1 \mathbf{q}_1^T = \mathbf{f}_1 \mathbf{q}_1^T. \tag{52}$$

Fromquation (41), the expression $\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2$ can be re-expressed as

$$\begin{aligned}\left\|\mathbf{X} - \mathbf{X}^{(1)}\right\|_{\mathbf{W}}^2 &= \operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \mathbf{X}^{(1)})\mathbf{W}(\mathbf{X} - \mathbf{X}^{(1)})^T\mathbf{M}^{\frac{1}{2}}\right) \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\delta \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}\mathbf{q}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2 \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\delta \sum_{k=1}^{K} \operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\mathbf{p}_1\mathbf{q}_{1,[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2.\end{aligned} \tag{53}$$

where $\mathbf{q}_1^T = (\mathbf{q}_{1,[1]}, \dots, \mathbf{q}_{1,[k]}, \dots, \mathbf{q}_{1,[K]})^T$ is the first row of $\mathbf{Q}^T$ with $\mathbf{q}_{1,[k]}$ a vector of dimension $J_{[k]} \times 1$ (see Figure 2.1.4). The best rank approximation property can be generalized for the rank-L as shown in section 51.

2.2.5. Correspondence Analysis as GSVD

Correspondence analysis is a statistical visualization method usual to picture the associations between an $I \times J$ two-way contingency table denoted $\mathbf{N}$. $\mathbf{N}$ is an $I \times J$ matrix of non-negative numbers (and we assume that no row and no column is full of zeros). We define $\mathbf{X}$ as the stochastic matrix derived from $\mathbf{N}$ as

$$\mathbf{X} = \mathbf{N}(\mathbf{1}^T\mathbf{N}\mathbf{1})^{-1}. \tag{54}$$

We define:

- $\mathbf{r} = \mathbf{X}\mathbf{1}$ the vector of row marginal proportions
- $\mathbf{c} = \mathbf{X}^T\mathbf{1}$ the vector of column marginal proportions
- $\mathbf{D_r} = \mathrm{diag}(\mathbf{r})$
- $\mathbf{D_c} = \mathrm{diag}(\mathbf{c})$.

For the GSVD of $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{r}\mathbf{c}^T$ under the constraints of $\mathbf{M}$ and $\mathbf{W}$ we set $\mathbf{M} = {\mathbf{D_r}}^{-1}$ and $\mathbf{W} = {\mathbf{D_c}}^{-1}$. In conclusion the GSVD of $\overline{\mathbf{X}}$ is

$$\overline{\mathbf{X}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \qquad \text{with} \qquad \mathbf{P}^T{\mathbf{D_r}}^{-1}\mathbf{P} = \mathbf{Q}^T{\mathbf{D_c}}^{-1}\mathbf{Q} = \mathbf{I}. \tag{55}$$

The matrices of the coordinates of the row profiles $\mathbf{F}$ and of the columns profiles $\mathbf{G}$ are, respectively

$$\begin{aligned} \mathbf{F} &= {\mathbf{D_r}}^{-1}\mathbf{P}\boldsymbol{\Delta} \\ \mathbf{G} &= {\mathbf{D_c}}^{-1}\mathbf{Q}\boldsymbol{\Delta}. \end{aligned} \tag{56}$$

2.2.6. From Correspondence Analysis to Multiple Correspondence Analysis

Suppose the original matrix of categorical data is $I \times J$, *i.e.*, I cases and J variables. MCA converts the cases-by-variables data to an indicator matrix

where the categorical data have been recoded as dummy variables. If the k-th variable has $J_{[k]}$ categories, this indicator matrix will have $J = \sum_{k=1}^{K} J_{[k]}$ columns. Then MCA is obtained from a standard correspondence analysis on blocks of indicator matrices, resulting in coordinates for the I cases and the J categories. Therefore, sPCA-rSVD can be extended to blocks of indicator variables to produce sparsity in MCA analysis.

2.2.7. GSVD as a regression-type problem

Consider the GSVD of $\mathbf{X}$ with $\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}$ the matrix of factor scores. Knowing the factor scores $\mathbf{f}_1$ are a linear combination of the variables, we want to recover the loadings. The Ordinary Least Squares (OLS) estimates $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the weighted norm under the constraints of the masses $\mathbf{M}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{f}_1 - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{W}} . \tag{57}$$

The OLS is solution is given by

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^{+}_{\mathbf{W}}\mathbf{M}\mathbf{f}_1, \tag{58}$$

with $\mathbf{X}^{+}_{\mathbf{W}} = \mathbf{Q}\boldsymbol{\Delta}^{+}_{\mathbf{W}}\mathbf{P}^T = (\mathbf{X}^T\mathbf{M}\mathbf{X})^{-1}\mathbf{W}\mathbf{X}^T$ Assuming that $\mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}$ we can deduce that the vector $\mathbf{Q}^T\mathbf{W}\mathbf{q}_1$ is the first column of the identity matrix of dimension $L \times L$. So $\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 = [1, 0, \ldots, 0]^T$ is a vector of dimension $L \times 1$.

Knowing that $\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T$ we find that

$$\begin{aligned} \mathbf{X}\mathbf{W}\mathbf{q}_1 &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{F}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{F}\begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{f}_1. \end{aligned} \tag{59}$$

So the OLS solution becomes

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{X}_{\mathbf{W}}^{+}\mathbf{M}\mathbf{f}_1 \\ &= \mathbf{X}_{\mathbf{W}}^{+}\mathbf{M}\mathbf{X}\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{Q}\boldsymbol{\Delta}_{\mathbf{W}}^{+}\mathbf{P}^T\mathbf{M}\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{Q}\boldsymbol{\Delta}_{\mathbf{W}}^{+}\boldsymbol{\Delta}\mathbf{Q}^T\mathbf{W}\mathbf{q}_1 \\ &= \mathbf{q}_1. \end{aligned} \tag{60}$$

Hence, we can conclude that the loadings $\mathbf{q}_1$ can be recovered by regressing the factor scores $\mathbf{f}_1$ on the J variables because each column of $\mathbf{F}$ is a linear combination of the J variables. Thereby, the SVD and so PCA, can be seen as a regression-type optimization problem.

3. Regularization of the decomposition

3.1. Regularized Singular Value Decomposition

Shen and Huang ([2]) have adapted the SVD to compute low-rank matrix approximations of the data matrix under various sparsity-inducing penalties (sparse PCA via regularized Singular Value Decomposition). For a given $\mathbf{f}$, the elements of $\mathbf{q}$ are obtained by regressing the columns of $\mathbf{X}$ on $\mathbf{f}$ (see sec-

tion 2.1.7). But in case of high-dimensional data, the number of non-zero loadings is very important and the interpretation of the results very difficult. Hence, the solution is to create sparsity eliminating some loadings. The principle is similar to a kind of rotation but a large number of loadings will be set exactly to zero.

According to section 2.1.2, $\mathbf{fq}$ with $\|\mathbf{q}\| = \mathbf{1}$ is the best rank-one approximation of the data matrix $\mathbf{X}$, with $\mathbf{f}$ the first factor score and $\mathbf{q}$ the first loading vector. SVD can be seen as a regression type problem, hence to achieve sparseness on $\mathbf{q}$, a shrinkage is performed on the regression coefficients through a penalty function in the optimization problem (27). However, the loading vector $\mathbf{q}$ is typically constrained to have unit length to make the representation unique. This constraint makes direct application of a penalty on $\mathbf{q}$ inappropriate. To overcome this difficulty, we rewrite $\mathbf{fq}^T = \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T$, where $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{q}}$ are re-scaled versions of $\mathbf{f}$ and $\mathbf{q}$ such that $\tilde{\mathbf{f}} = \delta\tilde{\mathbf{p}}$ with $\tilde{\mathbf{p}}$ having unit length and $\tilde{\mathbf{q}}$ free of any scale constraint, and then perform shrinkage on $\tilde{\mathbf{q}}$ through some regularization penalty. After a sparse $\tilde{\mathbf{q}}$ is obtained, we define the corresponding sparse loading vector as $\mathbf{q} = \tilde{\mathbf{q}}/\|\tilde{\mathbf{q}}\|$. The optimization problem becomes

$$\underset{\tilde{\mathbf{f}},\tilde{\mathbf{q}}}{\arg\min} \|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\|^2 + P_\lambda(\tilde{\mathbf{q}}) \tag{61}$$

where $P_\lambda(\tilde{\mathbf{q}})$ is a penalty function and λ a tuning parameter. Here, the expression can be written

$$\begin{aligned} \left\|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\right\|^2 + P_\lambda(\tilde{\mathbf{q}}) &= \|\mathbf{X}\|^2 - 2\,\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T) + \mathrm{tr}(\tilde{\mathbf{q}}\tilde{\mathbf{f}}^T\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T) + P_\lambda(\tilde{\mathbf{q}}) \\ &= \|\mathbf{X}\|^2 - 2\,\mathrm{tr}(\mathbf{X}^T\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T) + \delta^2\,\mathrm{tr}(\tilde{\mathbf{q}}^T\tilde{\mathbf{q}}) + P_\lambda(\tilde{\mathbf{q}}) \end{aligned} \tag{62}$$

The penalty function sets some elements of $\tilde{\mathbf{q}}$ exactly to zero, hence it produces sparse loadings and variables are removed. This penalty can be for example the

lasso penalty, the soft thresholding penalty, the hard thresholding penalty or any others (see [2]).

3.2. Regularized Generalized Singular Value Decomposition

According to section 2.2.7, GSVD can be seen as a regression type problem, hence to achieve sparseness on $\mathbf{q}$, a shrinkage is performed on $\mathbf{q}$ through a penalty function in the optimization problem (51). As seen previously, we will consider $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{q}}$. In a generalized case, the optimization problem (61) can be written as

$$\underset{\mathbf{f},\mathbf{q}}{\arg\min} \left\|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) \tag{63}$$

with $\mathbf{W}$ a positive definite matrix, $P_\lambda(\tilde{\mathbf{q}})$ a penalty function and λ a tuning parameter. Here, the expression can be written

$$\begin{aligned}\left\|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\,\mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) \\ &\quad + \mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}\tilde{\mathbf{f}}^T\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}) \\ &= \|\mathbf{X}\|_{\mathbf{W}}^2 - 2\,\mathrm{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) \\ &\quad + \delta^2\,\mathrm{tr}(\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}) + P_\lambda(\tilde{\mathbf{q}})\end{aligned} \tag{64}$$

When data are divided into K groups each of cardinal $J_{[k]}$, [4] introduced the group lasso penalty to control sparsity at the group level. Consider the general regression problem with K groups:

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_{[k]}\boldsymbol{\beta}_{[k]} + \boldsymbol{\varepsilon} \tag{65}$$

where $\mathbf{y}$ is a $I \times 1$ vector, $\boldsymbol{\varepsilon}$ the error, $\mathbf{X}_{[k]}$ is an $I \times J_{[k]}$ matrix corresponding to the kth block, and $\boldsymbol{\beta}_{[k]}$ is a coefficient vector of size $J_{[k]}$, $k = 1, \ldots, K$. Given positive definite matrices $\mathbf{W}_{[1]}, \ldots, \mathbf{W}_{[K]}$, the group lasso estimate is defined as

the solution to

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]} \boldsymbol{\beta}_{[k]} \right\|_{\mathbf{W}}^2 + \lambda \sum_{k=1}^{K} \left\| \boldsymbol{\beta}_{[k]} \right\|_{\mathbf{W}_{[k]}} \tag{66}$$

where $\lambda \geq 0$ is a tuning parameter. In our case, the expression become

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_{[k]} \mathbf{q}_{[k]} \right\|_{\mathbf{W}}^2 + \lambda \sum_{k=1}^{K} \left\| \mathbf{q}_{[k]} \right\|_{\mathbf{W}_{[k]}} \tag{67}$$

where $\left\| \mathbf{q}_{[k]} \right\|_{\mathbf{W}_{[k]}} = \sqrt{\mathbf{q}_{[k]}^T \mathbf{W}_{[k]} \mathbf{q}_{[k]}}$ (see (50)).

4. Group Sparse PCA via regularized generalized SVD

This section presents the new method developed to select quantitative variables when data are structured by blocks. Let $\mathbf{X}$ be an $I \times J$ matrix of quantitative variables made of K sub-matrices $\mathbf{X}_{[k]}$, $k = 1, ..., K$ as defined in (6). The SVD of $\mathbf{X}$ is defined as in (12) and we will consider the problem as describe in section 3.1. In this context, the problem (61) can be re-expressed as

$$\underset{(\tilde{\mathbf{f}}, \tilde{\mathbf{q}})}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}} \tilde{\mathbf{q}}^T \right\|^2 + P_\lambda(\tilde{\mathbf{q}}) = \underset{(\tilde{\mathbf{f}}, \tilde{\mathbf{q}})}{\arg\min} \left(\operatorname{tr} \left((\mathbf{X} - \tilde{\mathbf{f}} \tilde{\mathbf{q}}^T)^T (\mathbf{X} - \tilde{\mathbf{f}} \tilde{\mathbf{q}}^T) \right) + P_\lambda(\tilde{\mathbf{q}}) \right) \tag{68}$$

$P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]})$ because the penalty function is additive, so according to (62), the problem becomes

$$\underset{\tilde{\mathbf{f}}, \tilde{\mathbf{q}}}{\arg\min} \left(\|\mathbf{X}\|^2 - 2 \sum_{k=1}^{K} \operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \sum_{k=1}^{K} \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right) \tag{69}$$

To find the solution of the minimization problem we use an iterative algorithm with respect to $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{q}}$ under the constraint $\|\tilde{\mathbf{f}}\| = \|\delta \tilde{\mathbf{p}}\| = \delta$ because $\|\tilde{\mathbf{p}}\| = 1$.

First consider the problem of optimizing over $\tilde{\mathbf{f}}$ for a fixed $\tilde{\mathbf{q}}$. The minimizing $\tilde{\mathbf{f}}$ can be obtained by

$$\tilde{\mathbf{f}} = \mathbf{X}\tilde{\mathbf{q}} / \|\mathbf{X}\tilde{\mathbf{q}}\| \tag{70}$$

Then, we will optimize over $\tilde{\mathbf{q}}$ for a fixed $\tilde{\mathbf{p}}$, we will find $\tilde{\mathbf{q}}$ that minimizes the problem. The minimization problem (69) can be written

$$\arg\min_{\tilde{\mathbf{q}}} \left(\|\mathbf{X}\|^2 - 2\sum_{k=1}^{K} \operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \sum_{k=1}^{K} \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) \right) \quad (71)$$

The term $\|\mathbf{X}\|^2$ does not depend on $\tilde{\mathbf{q}}$ and we can optimize over individual components of $\tilde{\mathbf{q}}$ separately. Hence, the optimal $\tilde{\mathbf{q}}_{[k]}$ minimizes

$$R(\tilde{\mathbf{q}}_{[k]}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + P_\lambda(\tilde{\mathbf{q}}_{[k]}) \quad (72)$$

and depends on the form of $P_\lambda(.)$.

Data are structured by blocks, thus the penalty function P_λ chosen is the Group Lasso penalty describe in (67). In conclusion, we have to find the optimal $\mathbf{q}_{[k]}$ that minimizes

$$R(\tilde{\mathbf{q}}_{[k]}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \lambda \|\tilde{\mathbf{q}}_{[k]}\| \quad (73)$$

Minimizing a sum is equivalent to finding the value at which the gradient of this sum is equal to zero. The gradient of a sum is the sum of the gradient of each element of the sum. So we have

$$\begin{aligned} \frac{\partial R}{\partial \tilde{\mathbf{q}}_{[k]}} &= \frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}} \left(\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) + \lambda \|\tilde{\mathbf{q}}_{[k]}\| \right) \quad (74) \\ &= \frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}} \left(\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) \right) - 2\frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}} \left(\operatorname{tr}(\mathbf{X}_{[k]}^T \tilde{\mathbf{f}} \tilde{\mathbf{q}}_{[k]}^T) \right) + \frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}} \left(\lambda \|\tilde{\mathbf{q}}_{[k]}\| \right) \end{aligned}$$

On the one hand, according to property 5

$$\frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}} \left(\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T \tilde{\mathbf{q}}_{[k]}) \right) = 2\delta^2 \tilde{\mathbf{q}}_{[k]} \quad (75)$$

On the other hand, according to property 6

$$-2\frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}}\left(\mathrm{tr}(\mathbf{X}_{[k]}^T\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T)\right) = -2\mathbf{X}_{[k]}^T\tilde{\mathbf{f}} \tag{76}$$

The equation (75) and (76) are the results of the SVD. So we can affirm that the optimal solutions are $\mathbf{q} = \mathbf{q}_1$ et $\mathbf{f} = \mathbf{f}_1$. In conclusion, according to property 8

if $\mathbf{q}_{1,[k]} \neq \mathbf{0}$

$$\frac{\partial R}{\partial \mathbf{q}_{[1,k]}} = 2\delta_1^2\tilde{\mathbf{q}}_{1,[k]} - 2\mathbf{X}_{[k]}^T\tilde{\mathbf{f}}_1 + \lambda\frac{\mathbf{q}_{1,[k]}}{\left\|\mathbf{q}_{1,[k]}\right\|} = 0 \tag{77}$$

if $\mathbf{q}_{1,[k]} = \mathbf{0}$

$$\left\|2\mathbf{X}_{[k]}^T\mathbf{f}_1\right\| \leq \lambda \tag{78}$$

Expressions (77) and (78) can be rewritten as

$$\mathbf{X}_{[k]}^T\tilde{\mathbf{f}}_1 = \delta_1^2\tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2}\frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|} \tag{79}$$

$$\left\|\mathbf{X}_{[k]}^T\tilde{\mathbf{f}}_1\right\| \leq \frac{\lambda}{2} \tag{80}$$

After combining (99) and (100) the minimizer has the form

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2\delta_1^2}\frac{1}{\left\|\mathbf{X}_{[k]}^T\tilde{\mathbf{f}}_1\right\|}\right)_+ \mathbf{X}_{[k]}^T\tilde{\mathbf{f}}_1 \tag{81}$$

where

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{82}$$

A thresholding rule h_λ can be defined as

$$h_\lambda(y) = \left(1 - \frac{\lambda}{2\delta_1^2}\frac{1}{\|y\|}\right)_+ y \tag{83}$$

with $y = \mathbf{X}_{[k]}^T \tilde{\mathbf{f}}_1$.

Algorithm 1 GSPCA-rSVD algorithm

Initialize: Apply the standard SVD to $\mathbf{X}$ and obtain the best rank-one approximation of $\mathbf{X}$ as $\delta \mathbf{pq} = \mathbf{fq}$ where $\mathbf{p}$ and $\mathbf{q}$ are unit vectors.
Set $\tilde{\mathbf{q}}^s = \mathbf{q}_1$ and $\tilde{\mathbf{f}}^s = \delta_1 \mathbf{p}_1$.
Update:

a) $\tilde{\mathbf{q}}^{s+1} = [\tilde{\mathbf{q}}_{1,[1]}^{s+1}, \ldots, \tilde{\mathbf{q}}_{1,[K]}^{s+1}] = [h_\lambda(\mathbf{X}_{[1]}^T \tilde{\mathbf{f}}^s), \ldots, h_\lambda(\mathbf{X}_{[K]}^T \tilde{\mathbf{f}}^s)]$;

b) $\tilde{\mathbf{f}}^{s+1} = \mathbf{X}_{[k]}^T \tilde{\mathbf{q}}^{s+1} / \left\| \mathbf{X}_{[k]}^T \tilde{\mathbf{q}}^{s+1} \right\|$

Repeat Step 2 replacing $\tilde{\mathbf{f}}^s$ and $\tilde{\mathbf{q}}^s$ by $\tilde{\mathbf{f}}^{s+1}$ and $\tilde{\mathbf{q}}^{s+1}$ until convergence
Standardize the final $\tilde{\mathbf{q}}^{s+1}$ as $\mathbf{q}$=$\tilde{\mathbf{q}}^{s+1} / \left\| \tilde{\mathbf{q}}^{s+1} \right\|$, the desired sparse loading.

Setting $\lambda = 0$ in the above algorithm, Step 2a reduces to $\mathbf{q}^{s+1} = \mathbf{X}^T \mathbf{f}^s$ and the algorithm becomes the well-known alternating least-squares algorithm for calculating SVD. The iterative procedure of the GSPCA-rSVD algorithm is defined for one-dimensional vectors, and can be used to obtain the first sparse loading vector $\mathbf{q}_1$. Subsequent sparse loading vectors $\mathbf{q}_i$ $(i > 1)$ can be obtained sequentially via rank-one approximation of residual matrices.

MCA is a particular case of PCA for blocks of indicator variables, thereby sparse MCA introduced in the following part is defined as an extension of the Group sparse PCA. Hence the problem (68) can be generalized for the MCA sparse as

$$\underset{(\tilde{\mathbf{f}}, \tilde{\mathbf{q}})}{\arg\min} \left\| \mathbf{X} - \tilde{\mathbf{f}} \tilde{\mathbf{q}}^T \right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) \tag{84}$$

with a norm $\mathbf{W}$-generalized under the constraints of the masses $\mathbf{M}$.

5. Sparse MCA via regularized SVD

Suppose an $I \times J$ matrix of qualitative variables. The corresponding complete disjunctive table $\mathbf{N}$ is made of K sub-matrices of indicator variables $\mathbf{N}_{[k]}$, $k = 1, ..., K$, each of dimension $I \times J_{[k]}$. The total number of modalities is

denoted by $J = \sum_{k=1}^{K} J_{[k]}$. To select one column in the original table (categorical variable) is equivalent to select a block of indicator variables in the complete disjunctive table. So we can define a new method called Sparse MCA which is a straightforward extension of the GSPCA for blocks of indicator variables. We will consider the GSVD for blocks of indicator variables as in (37). Let $\mathbf{X}$ be the stochastic matrix defined in section 2.2.5. In this context, the problem (63) can be re-expressed as

$$\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left\|\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\right\|_{\mathbf{W}}^2 + P_\lambda(\tilde{\mathbf{q}}) \qquad \text{with} \qquad \tilde{\mathbf{f}}^T\mathbf{M}\tilde{\mathbf{f}} = \tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \tag{85}$$

and according to the constraints of the masses $\mathbf{M}$, the problem can be rewritten

$$\begin{aligned} &\underset{(\tilde{\mathbf{f}},\tilde{\mathbf{q}})}{\arg\min} \left(\operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}(\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T)\mathbf{W}(\mathbf{X} - \tilde{\mathbf{f}}\tilde{\mathbf{q}}^T)^T\mathbf{M}^{\frac{1}{2}}\right) + P_\lambda(\tilde{\mathbf{q}})\right) = \\ &\underset{\mathbf{f},\mathbf{q}}{\arg\min} \left(\|\mathbf{X}\|_{\mathbf{W}}^2 - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T\mathbf{W}\mathbf{X}^T\mathbf{M}^{\frac{1}{2}}) + \delta^2\operatorname{tr}(\tilde{\mathbf{q}}^T\mathbf{W}\tilde{\mathbf{q}}) + P_\lambda(\tilde{\mathbf{q}})\right) \end{aligned} \tag{86}$$

The penalty function is additive, $P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]})$, so according to (64), the problem becomes

$$\begin{aligned} \underset{\tilde{\mathbf{f}},\tilde{\mathbf{q}}}{\arg\min}(\|\mathbf{X}\|_{\mathbf{W}}^2 - 2\sum_{k=1}^{K}\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\mathbf{X}_{[k]}^T\mathbf{M}^{\frac{1}{2}}) + \\ \sum_{k=1}^{K}\delta^2\operatorname{tr}(\tilde{\mathbf{q}}_{[k]}^T\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) + \sum_{k=1}^{K}P_\lambda(\tilde{\mathbf{q}}_{[k]})) \end{aligned} \tag{87}$$

To find the solution of the minimization problem we use an iterative algorithm with respect to $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$ under the constraint $\tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = 1$. First consider the problem of optimizing over $\tilde{\mathbf{p}}$ for a fixed $\tilde{\mathbf{q}}$. The minimizing $\tilde{\mathbf{p}}$ can be obtained by

$$\tilde{\mathbf{p}} = \mathbf{X}\tilde{\mathbf{q}} / \|\mathbf{X}\tilde{\mathbf{q}}\|_{\mathbf{W}}\,. \tag{88}$$

Now we will optimize over $\tilde{\mathbf{q}}$ for a fixed $\tilde{\mathbf{p}}$. The minimization problem (87) can be written

$$\arg\min_{\tilde{\mathbf{q}}}(\|\mathbf{X}\|^2_{\mathbf{W}} + \sum_{k=1}^{K}(\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) \\ - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\mathbf{X}^T_{[k]}\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}_{[k]})) \quad (89)$$

The term $\|\mathbf{X}\|^2_{\mathbf{W}}$ does not depend on $\tilde{\mathbf{q}}$ and we can optimize over individual components of $\tilde{\mathbf{q}}$ separately. Hence, the optimal $\tilde{\mathbf{q}}_{[k]}$ minimizes

$$R(\tilde{\mathbf{q}}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\mathbf{X}^T_{[k]}\mathbf{M}^{\frac{1}{2}}) + P_\lambda(\tilde{\mathbf{q}}_{[k]}) \quad (90)$$

and depends on the form of $P_\lambda(.)$. Data are structured by blocks, thus the penalty function P_λ chosen is the Group Lasso penalty describe in (67)

$$P_\lambda(\tilde{\mathbf{q}}) = \sum_{k=1}^{K} P_\lambda(\tilde{\mathbf{q}}_{[k]}) = \sum_{k=1}^{K} \lambda\|\tilde{\mathbf{q}}_{[k]}\|_{\mathbf{W}_{[k]}} \quad (91)$$

In conclusion, we have to find the optimal $\mathbf{q}_{[k]}$ that minimizes

$$R(\mathbf{q}) = \delta^2 \operatorname{tr}(\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) - 2\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\mathbf{X}^T_{[k]}\mathbf{M}^{\frac{1}{2}}) + \lambda\|\tilde{\mathbf{q}}_{[k]}\|_{\mathbf{W}_{[k]}} \quad (92)$$

Minimizing a sum is equivalent to finding the value at which the gradient of this sum is equal to zero. On thee one hand,

$$\frac{\partial}{\partial \mathbf{q}_{[k]}}\delta^2 \operatorname{tr}(\tilde{\mathbf{q}}^T_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}) = 2\delta^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]} \quad (93)$$

According to property 2 and property 6

$$-2\frac{\partial}{\partial \mathbf{q}_{[k]}}\operatorname{tr}(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\tilde{\mathbf{q}}_{[k]}^{T}\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}^{\frac{1}{2}}) = -2\frac{\partial}{\partial \tilde{\mathbf{q}}_{[k]}}\left(\operatorname{tr}\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}_{[k]}\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{[k]}\tilde{\mathbf{f}}^{T}\mathbf{M}^{\frac{1}{2}}\right)\right) \tag{94}$$

$$= -2\left(\mathbf{M}^{\frac{1}{2}}\mathbf{X}_{[k]}\mathbf{W}_{[k]}\right)^{T}\left(\tilde{\mathbf{f}}^{T}\mathbf{M}^{\frac{1}{2}}\right)^{T}$$

$$= -2\left(\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}^{\frac{1}{2}}\right)\left(\mathbf{M}^{\frac{1}{2}}\tilde{\mathbf{f}}\right)$$

$$= -2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}$$

The equation (93) and (94) are the results of the GSVD presented in section 2.2. Hence the optimal solutions are $\mathbf{q} = \mathbf{q}_1$ et $\mathbf{f} = \mathbf{f}_1$. In conclusion,
if $\mathbf{q}_{1,[k]} \neq \mathbf{0}$

$$\frac{\partial R}{\partial \tilde{\mathbf{q}}_{[1,k]}} = -2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1 + 2\delta_1^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]} + \lambda\mathbf{W}_{[k]}\frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_{\mathbf{W}_{[k]}}} = 0 \tag{95}$$

if $\mathbf{q}_{1,[k]} = \mathbf{0}$

$$\left\|-2\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \lambda\mathbf{W}_{[k]} \tag{96}$$

Expressions (95) and (96) can be rewritten as

$$\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1 = \delta_1^2\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2}\frac{\mathbf{W}_{[k]}\tilde{\mathbf{q}}_{1,[k]}}{\|\tilde{\mathbf{q}}_{1,[k]}\|_{\mathbf{W}_{[k]}}} \tag{97}$$

$$\left\|\mathbf{W}_{[k]}\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \frac{\lambda}{2}\mathbf{W}_{[k]} \tag{98}$$

So we obtain

$$\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1 = \delta_1^2\tilde{\mathbf{q}}_{1,[k]} + \frac{\lambda}{2}\frac{\tilde{\mathbf{q}}_{1,[k]}}{\left\|\tilde{\mathbf{q}}_{1,[k]}\right\|_{\mathbf{W}_{[k]}}} \tag{99}$$

$$\left\|\mathbf{X}_{[k]}^{T}\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}} \leq \frac{\lambda}{2} \tag{100}$$

After combining (99) and (100) the minimizer has the form

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2}\frac{1}{\left\|\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1\right\|_{\mathbf{W}_{[k]}}}\right)_+ \frac{1}{\delta_1^2}\mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1 \tag{101}$$

where

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{102}$$

A thresholding rule h_λ can be defined as

$$h_\lambda(y) = \left(1 - \frac{\lambda}{2}\frac{1}{\|y\|_{\mathbf{W}_{[k]}}}\right)_+ \frac{1}{\delta^2}y \tag{103}$$

with $y = \mathbf{X}_{[k]}^T\mathbf{M}\tilde{\mathbf{f}}_1$.

In the case of the MCA:

$$\mathbf{M} = {\mathbf{D_r}}^{-1} \qquad \mathbf{W} = {\mathbf{D_c}}^{-1} \tag{104}$$

So the norm $\mathbf{W}$-generalized can be written as

$$\begin{aligned} \|\mathbf{X}\|_{\mathbf{W}} &= \sqrt{\mathbf{MXWX}^T} \\ &= \sqrt{{\mathbf{D_r}}^{-1}\mathbf{X}{\mathbf{D_c}}^{-1}\mathbf{X}^T} \end{aligned} \tag{105}$$

The norm $\left\|\mathbf{X}_{[k]}^T\mathbf{M}\mathbf{f}_1\right\|_{\mathbf{W}_{[k]}}$ is equal to $\left\|\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\mathbf{f}_1\right\|_{\mathbf{D_c}_{[k]}^{-1}}$. Hence, the solution becomes

$$\tilde{\mathbf{q}}_{1,[k]} = \left(1 - \frac{\lambda}{2}\frac{1}{\left\|\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\tilde{\mathbf{f}}_1\right\|_{\mathbf{D_c}_{[k]}^{-1}}}\right)_+ \frac{1}{\delta_1^2}\mathbf{X}_{[k]}^T{\mathbf{D_r}}^{-1}\tilde{\mathbf{f}}_1 \tag{106}$$

Algorithm 2 SMCA algorithm

Initialize: Apply the GSVD to $\mathbf{X}$. Calculate the best rank-one approximation of $\mathbf{X}$ as $\delta\mathbf{pq} = \mathbf{fq}$ where $\mathbf{p}$ and $\mathbf{q}$ are unit vectors.
Set $\mathbf{q}^s = \mathbf{q}_1$ and $\mathbf{f}^s = \delta_1\mathbf{p}_1$.
Update:

a) $\mathbf{q}^{s+1} = [h_\lambda(\mathbf{X}_{[1]}^T\mathbf{f}^s), \ldots, h_\lambda(\mathbf{X}_{[J]}^T\mathbf{f}^s)]$;

b) $\mathbf{f}^{s+1} = \mathbf{X}_{[k]}^T\tilde{\mathbf{q}}^{s+1} / \left\|\mathbf{X}_{[k]}^T\tilde{\mathbf{q}}^{s+1}\right\|_{\mathbf{W}_{[k]}}$

Repeat Step 2 replacing $\mathbf{f}^s$ and $\mathbf{q}^s$ by $\mathbf{f}^{s+1}$ and $\mathbf{q}^{s+1}$ until convergence
Standardize the final $\mathbf{q}^{s+1}$ as $\mathbf{q}$=$\mathbf{q}^{s+1}$ /$\left\|\mathbf{q}^{s+1}\right\|_{\mathbf{W}}$, the desired sparse loading.

Subsequent sparse loading vectors $\mathbf{q}_i$ $(i > 1)$ can be obtained sequentially via rank-one approximation of residual matrices.

6. Properties of Group Sparse PCA and Sparse MCA

These two new methods have important properties to realize sparsity.

- Without any sparsity constraint, Group Sparse PCA and Sparse MCA reduces to PCA and MCA.
- They are computationally efficient for both small p and big p data.
- They avoid misidentifying the important variables.

Barycentric properties of MCA are retained in Sparse MCA. However, in PCA and MCA, the PCs are uncorrelated and their loadings are orthogonal. These properties are lost in Sparse PCA and Sparse MCA. The orthogonality among the loadings is lost, a nice property enjoyed by standard PCA and MCA. Several other sparse PCA procedures lose this property as well, which is the price one pays for easy interpretation of the results.

Furthermore, explained variance can not be defined in the same way as PCA and MCA in these two new methods. In Group Sparse PCA the variance explained is defined as in Sparse PCA of [2]. We consider the projection of X

onto the k-dimensional subspace spanned by the k loading vectors as

$$\mathbf{X}_{[k]} = \mathbf{X}\mathbf{Q}_{[k]}(\mathbf{Q}_{[k]}^T\mathbf{Q}_{[k]})^{-1}\mathbf{Q}_{[k]}^T \tag{107}$$

where $\mathbf{Q}_{[k]}$ is the matrix of the first k sparse loadings. We generally define the total variance explained by the first k PCs as $\mathrm{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]})$. The adjusted variance of the kth PC by $\mathrm{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]}) - \mathrm{tr}(\mathbf{X}_{[k-1]}^T\mathbf{X}_{[k-1]})$ and the cumulative percentage of variance (CPEV) explained by the first k PCs by $\mathrm{tr}(\mathbf{X}_{[k]}^T\mathbf{X}_{[k]})/\mathrm{tr}(\mathbf{X}^T\mathbf{X})$.

For Sparse MCA, there is a little modification. MCA codes data by creating several binary columns for each variable with the constraint that one and only one of the columns gets the value 1. This coding schema creates artificial additional dimensions because one categorical variable is coded with several columns. As a consequence, the inertia (i.e., variance) of the solution space is artificially inflated and therefore the percentage of inertia explained by the first dimension is severely underestimated. Two corrections formulas are often used, the first one is due to [7], the second one to [8]. These formulas take into account that the eigenvalues smaller than $1/H$ are coding for the extra dimensions (H is the number of catgeorical variables). J is total number of modalities, i.e. the number of binary variables. If we denote by λ_ℓ the eigenvalues obtained from MCA, then the corrected eigenvalues, denoted λ^c are obtained as

$$\lambda_\ell^c = \begin{cases} \left[\left(\frac{J}{J-1}\right)\left(\lambda_\ell - \frac{1}{J}\right)\right]^2 & \text{si } \lambda_\ell > \frac{1}{J} \\ 0 & \text{si } \lambda_\ell \leq \frac{1}{J} \end{cases} \tag{108}$$

Using this formula gives a better estimate of the inertia, extracted by each eigenvalue. Traditionally, the percentages of inertia are computed by dividing each eigenvalue by the sum of the eigenvalues, and this approach could be used here also. However, it will give an optimistic estimation of the percentage

of inertia. A better estimation of the inertia has been proposed by [8] who suggested instead to evaluate the percentage of inertia relative to the average inertia of the off-diagonal blocks of the Burt matrix. We recall that the Burt matrix is the $J \times J$ matrix $\mathbf{X}^T\mathbf{X}$ associated to $\mathbf{X}$. This average inertia, denoted $\bar{\sigma}$ can be computed as

$$\bar{\sigma} = \frac{H}{H-1} \times \left(\sum_{\ell} \lambda_{\ell}^2 - \frac{J-H}{H} \right)^2 . \tag{109}$$

According to this approach, the percentage of inertia would be obtained by the ratio

$$\frac{\lambda^c}{\bar{\sigma}} \qquad \text{instead of } \frac{\lambda^c}{\sum_{\ell} \lambda_{\ell}^c}. \tag{110}$$

7. Applications in genomic data analysis

To identify genetic factors that may affect skin aging severity, a study have been conducted on a well-defined sample of 501 French middle-aged Caucasian women from the SU.VI.MAX (SUpplémentation en VItamines et Minéraux AntioXydants (Antioxidant Vitamin and Mineral Supplementation) cohort ([9]). Using the Illumina HumanOmni1-Quad BeadChips and after quality control, 795 063 genotyped Single Nucleotide Polymorphisms (SNPs) were available for 501 women. Although all humans share far more than 99% of their DNA, there are still millions of differences between the DNA of 2 individuals. The most common, and so far the best investigated, genetic variations are single nucleotide polymorphisms (SNPs). A SNP occurs when a single nucleotide (A, T, C or G) is altered, that is, when different sequence alternatives exist at a single base-pair position. Since the human genome is diploid, that is, consists of pairs of chromosomes, each SNP is explained by 2 bases. Therefore, each SNP can take

one of the following 3 forms:

- "Homozygous reference genotype": both bases explaining the SNP are the more frequent variant,
- "Heterozygous variant genotype": one of the bases is the more frequent and the other the less frequent variant,
- "Homozygous variant genotype": both bases are the less frequent variant.

Table 1 presents a view of the first six SNPs of the database for the first five women.

Table 1: Excerpt of the first six SNPs of the database for the first five women

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
ind 1	AA	AA	AC	AA	GG	AA
ind 2	AA	AG	CC	AT	GG	AG
ind 3	AT	AG	CC	TT	AA	AA
ind 4	AT	AA	AC	TT	AA	AG
ind 5	AT	GG	AA	AA	AG	AG

A SNP with K levels is encoded by K dummy variables $\mathbf{Z}_i$. $\mathbf{Z}_i$ variable is equal to 1 at level i, 0 otherwise. We will consider a SNP as a set of two or three dummy variables. Table 2 displays the table of the corresponding first 6 dummy variables for the first five women.

Table 2: Excerpt of the first six corresponding dummy variables of the table for the first five women

	SNP1		SNP2			SNP3			SNP4			SNP5			SNP6	
	AA	AT	AA	AG	GG	AA	AC	CC	AA	AT	TT	AA	AG	GG	AA	AG
ind 1	1	0	1	0	0	0	1	0	1	0	0	0	0	1	1	0
ind 2	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1
ind 3	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0
ind 4	0	1	1	0	0	0	1	0	0	0	1	1	0	0	0	1
ind 5	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	1

After several stages of pre-selection, 640 SNPs were retained. Thus, the processed data have $I = 501$ women and $J = 640$ SNPs. We use sparse MCA on these data as a gene selection method to try to unravel new genetic associations with skin aging. The first step is to select he optimal tuning parameter λ by an "ad-hoc" approach. Figure 3 and figure 4 displays the evolution of the number of selected modalities and the evolution of the cumulative percentage variance explained depending on λ for the first four axes, respectively.

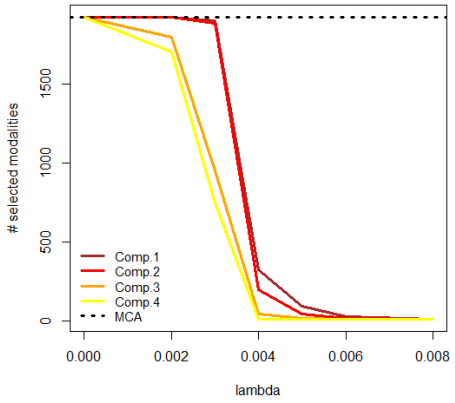

Figure 3: Evolution of the number of selected modalities depending on the tuning parameter for the first four axes of the sparse MCA

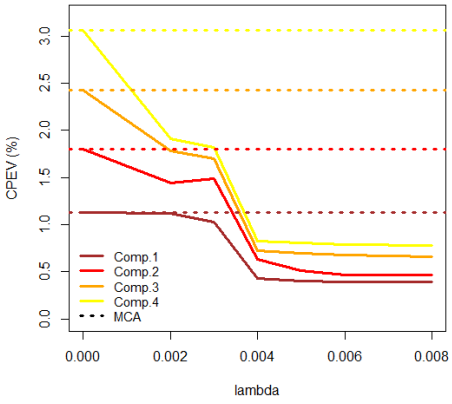

Figure 4: Evolution of the cumulative percentage variance explained (CPEV) depending on the tuning parameter for the first four axes of the sparse MCA

If we focus on the first axis, for $\lambda = 0.004$, CPEV= 0.43% and 324 modalities (i.e. 108 SNPs) are selected. Moreover, for $\lambda = 0.005$, CPEV= 0.40% and 96 modalities (i.e. 32 SNPs) are selected. Thus, given the low difference of the percentage of variance between the two values of λ, we choose the value of λ which conserves the least number of variables, so we set $\lambda = 0.005$.

The stability of the selection has been tested with a bootstrap approach. We consider 100 bootstrap replications of the given sample of 501 women. Then sparse MCA has been applied on these new 100 samples and the most often

selected variables have been conserved (the highest 25% of them). Finally 165 SNPs have been selected on the first axis among the 640 SNPs. Table 3 displays the comparison between the loadings obtained with MCA and with sparse MCA forλ set to 0.005 and for the first five SNPs of the sample. In MCA, no SNP is eliminated because λ is equal to 0. In contrast, in sparse MCA, the non zero tuning parameter allows selection of SNPs on each axis by setting all the loadings of the modalities of a variable to zero.

Table 3: Loadings and variance obtained for the first five SNPs of the database with MCA and sparse MCA on the first four axes

	MCA				**Sparse MCA**			
Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.1	Comp.2	Comp.3	Comp.4
SNPri1.CC	0.277	0.258	-0.159	0.189	0.033	-0.025	**0.000**	**0.000**
SNPri1.CG	-0.108	-0.064	-0.053	-0.097	-0.010	0.007	**0.000**	**0.000**
SNPri1.GG	-0.025	-0.087	0.214	0.021	-0.015	0.012	**0.000**	**0.000**
SNPri2.AA	-0.285	-0.652	-0.179	-0.287	**0.000**	**0.000**	**0.000**	**0.000**
SNPri2.AG	0.113	0.095	-0.133	-0.208	**0.000**	**0.000**	**0.000**	**0.000**
SNPri2.GG	0.119	0.636	0.528	0.835	**0.000**	**0.000**	**0.000**	**0.000**
SNPri3.AA	0.115	-0.025	0.636	-0.731	-0.192	0.193	-0.203	0.203
SNPri3.AG	0.072	-0.449	-0.194	0.520	-0.129	0.129	-0.124	0.123
SNPri3.GG	-0.351	0.125	-0.368	-0.360	0.435	-0.437	0.440	-0.440
SNPri4.AA	0.191	-0.151	-0.031	0.016	-0.024	0.019	**0.000**	**0.000**
SNPri4.AG	-0.173	0.204	0.033	0.050	0.032	-0.025	**0.000**	**0.000**
SNPri4.GG	-0.294	-0.133	0.023	-0.381	-0.016	0.009	**0.000**	**0.000**
SNPri5.AA	0.197	0.490	0.499	-0.161	**0.000**	**0.000**	**0.000**	**0.000**
SNPri5.AT	0.176	-0.064	0.115	0.010	**0.000**	**0.000**	**0.000**	**0.000**
SNPri5.TT	-0.157	-0.012	-0.148	0.012	**0.000**	**0.000**	**0.000**	**0.000**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Nb of selected modalities	1638	1638	1638	1638	165	96	21	12
Adjusted variance (%)	1.18	0.67	0.64	0.63	0.40	0.17	0.12	0.11
Cumulated variance (%)	1.18	1.85	2.49	3.12	0.40	0.57	0.69	0.78

8. Discussion

Sparse MCA is an extension of the GSPCA for qualitative data. It produces sparsity at loadings level (with a minimum loss of percentage of variance explained) which facilitates the interpretation and understanding of the different

axes. When the regularization parameter λ is set to 0, GSPCA and sparse MCA are identical to PCA and MCA, respectively, which is an essential property of a "good" sparse method according to [5]. The application presented in Section 4 was performed on a small data set, but these methods make sense in the context of very large dimension. However it does not produce sparsity within a group. In case you would like to select variables within a block, an extension of these methods could be achieved by replacing the penalty function "group Lasso" by "sparse group Lasso" developed by [3].

References

[1] Eckart, C. et Young, G., 1936. *The approximation of one matrix by another of lower rank.* Psychometrika. **1**. 211–218.

[2] Shen, H. and Huang, J.Z, 2008. *Sparse principal component analysis via regularized low rank matrix approximation.* Journal of Multivariate Analysis. **99**, 1015–1034.

[3] Simon, N., Friedman, J., Hastie, T. and others, 2013. *A Sparse-Group Lasso.*Journal of Computational and Graphical Statistics. **22**, 231–245.

[4] Yuan, M. et Lin, Y., 2006. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B. **68**, 49–67.

[5] Zou, H., Hastie, T. et Tibshirani, R., 2006. *Sparse Principal Component Analysis.* Journal of Computational and Graphical Statistics. **15**, 265–286.

[6] Jolliffe, I.T and Trendafilov, N.T and Uddin, M, 2003. *A modified principal component technique based on the LASSO.* Journal of Computational and Graphical Statistics. **12**, 531–547.

[7] Benzécri, J.P., 1979. *Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.].* Cahiers de l'Analyse des Données. **4**, 377–378.

[8] Greenacre, M., 2010. *Correspondence analysis in practice.* Chapman and Hall/CRC. **4**, 377–378.

[9] Hercberg, S and Galan, P and Preziosi, P and others, 2004. *The SU. VI. MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals.* Archives of Internal Medicine. **164**, 2335.

[10] Jolliffe, IT, 1986. *Principal component analysis.* Spring-verlag, New York.

le cnam

Anne BERNARD

Développement de méthodes statistiques nécessaires à l'analyse de données génomiques

le cnam

Résumé :
Les nouvelles technologies développées ces dernières années dans le domaine de la génétique ont permis de générer des bases de données de très grande dimension, en particulier de SNPs, ces bases étant souvent caractérisées par un nombre de variables largement supérieur au nombre d'individus. L'objectif de ce travail a été de développer des méthodes statistiques adaptées à ces jeux de données de grande dimension et permettant de sélectionner les variables les plus pertinentes au regard du problème biologique considéré. Dans la première partie de ce travail, un état de l'art présente différentes méthodes de sélection de variables non supervisées et supervisées pour 2 blocs de variables et plus. Dans la deuxième partie, deux nouvelles méthodes de sélection de variables non supervisées de type "sparse" sont proposées : la GSPCA et l'Analyse des Correspondances Multiples sparse (ACM sparse). Vues comme des problèmes de régression avec une pénalisation group LASSO elles conduisent à la sélection de blocs de variables quantitatives et qualitatives, respectivement. La troisième partie est consacrée aux interactions entre SNPs et dans ce cadre, une méthode spécifique de détection d'interactions, la régression logique, est présentée. Enfin, la quatrième partie présente une application de ces méthodes sur un jeu de données réelles de SNPs afin d'étudier l'influence possible du polymorphisme génétique sur l'expression du vieillissement cutané au niveau du visage chez des femmes adultes. Les méthodes développées ont donné des résultats prometteurs répondant aux attentes des biologistes, et qui offrent de nouvelles perspectives de recherches intéressantes.
Mots clés :
sélection de variables, ACP sparse, ACM, SNP-SNP interactions, régression logique, méthodes multiblocs, méthodes sparse non supervisées.

Abstract:
New technologies developed recently in the field of genetic have generated high-dimensional databases. These databases are often characterized by a number of variables much larger than the number of individuals. The goal of this dissertation was to develop appropriate statistical methods to analyse high-dimensional data, and to select the most biologically relevant variables. In the first part, I present the state of the art that describes unsupervised and supervised variables selection methods for two or more blocks of variables. In the second part, I present two new unsupervised "sparse" methods: Group Sparse Principal Component Analysis (GSPCA) and Sparse Multiple Correspondence Analysis (Sparse MCA). Considered as regression problems with a group LASSO penalization, these methods lead to select blocks of quantitative and qualitative variables, respectively. The third part is devoted to interactions between SNPs. A method employed to identify these interactions is presented: the logic regression. Finally, the last part presents an application of these methods on a real SNPs dataset to study the possible influence of genetic polymorphism on facial skin aging in adult women. The methods developed gave relevant results that confirmed the biologist's expectations and that offered new research perspectives.
Keywords:
feature selection, sparse PCA, MCA, SNP-SNP interactions, logic regression, multiblocks methods, unsupervised sparse methods.