Modèles de régression pour données fonctionnelles hétérogènes : application à la modélisation de données de spectrométrie dans le moyen infrarouge

Marie Morvan

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions*

Par

Marie MORVAN

Modèles de régression pour données fonctionnelles hétérogènes. Application à la modélisation de données de spectrométrie dans le moyen infrarouge.

Thèse présentée et soutenue à l'université Rennes 1, le 5/12/2019
Unité de recherche : IRMAR - UMR CNRS 6625

Rapporteurs avant soutenance :

Émilie Lebarbier Professeur, Université Paris Nanterre
Nicolas Molinari Professeur, Université de Montpellier

Composition du Jury :

Président :	David Causeur	Professeur, Agrocampus Ouest, Rennes
Examinateurs :	Vincent Brault	Maître de conférence, Université Grenoble Alpes
Dir. de thèse :	Valérie Monbet	Professeur, Université Rennes 1
Co-dir. de thèse :	Joyce Madison Giacofci	Maître de conférence, Université Rennes 2

Invité :

Émilie Devijver Chargée de recherche CNRS, Université Grenoble Alpes

Remerciements

Comme beaucoup de doctorants, je me rends compte en faisant le bilan de ces trois années de thèse que celles-ci ont été intenses tant professionnellement que personnellement. Beaucoup de personnes qui ont croisé mon chemin et ont contribué ou non à ma thèse, avant ou pendant ont rendu cette aventure unique.

Tout d'abord, j'adresse un immense merci à mes directrices de thèse Valérie Monbet et Madison Giacofci. Il est difficile de résumer ces trois années en quelques mots, mais merci pour votre écoute, votre patience, votre disponibilité, votre soutien et votre confiance. Merci de m'avoir encouragée, laissé expérimenter, et guidée. J'ai eu beaucoup de chance de pouvoir faire cette thèse avec vous et dans de si bonnes conditions. J'ai énormément progressé et appris grâce à vous, et j'espère vivement pouvoir continuer de travailler avec vous.

Je souhaite remercier tout particulièrement Émilie Lebarbier et Nicolas Molinari d'avoir eu la gentillesse de rapporter cette thèse. Merci pour votre relecture attentive et les remarques constructives qui m'ont aidé à préparer ma soutenance et à prendre du recul sur mon travail. Un grand merci également à David Causeur et à Vincent Brault d'avoir accepté de faire partie de mon jury.

Ensuite, je souhaite remercier chaleureusement Émilie Devijver pour le temps, la patience et la confiance accordée. Travailler avec toi est un plaisir et tes explications toujours claires m'ont aidée à progresser et à aboutir à cette thèse. Un grand merci à Alessia Pini de m'avoir permis d'aborder un nouveau point de vue sur cette thèse, j'espère que notre travail ensemble se prolongera aussi bien qu'il a débuté.

Merci à l'IRMAR et à toute l'équipe de statistique pour son accueil. Merci à Jean-Louis Marchand pour ses précieux conseils arrivés au bon moments. Un grand merci à Pierre Navaro pour son aide et sa disponibilité dès que des questions de programmation et d'optimisation me venaient. Merci à l'équipe d'Agrocampus qui a été à l'origine de mon engouement pour les statistiques, et que j'ai recroisée avec plaisir tout au long de cette thèse.

Merci à toute l'équipe de Diafir, à Maëna, Laure, Fred, Alexis, Hugues, Jean-Marc et Jérôme. Merci pour votre accueil et vos conseils tout au long de ma thèse. Merci pour toutes les opportunités et les rencontres que vous m'avez offertes et continuez de m'offrir. En particulier, un grand merci à Maëna qui m'a permis de m'approprier pleinement les problématiques en biologie de ma thèse, et avec qui c'est un plaisir de travailler. Merci pour tes conseils et ton écoute.

Je tiens aussi à remercier vivement Olivier Loréal et Olivier Sire, pour leur disponibilité et leurs précieux conseils en hépatologie et en biochimie. Merci de m'avoir suivie depuis le début et de m'avoir permis de mieux comprendre ces domaines et d'approfondir mon travail. Merci pour votre soutien. Merci à Rodolphe Anty pour son aide sur l'étude des données NASH.

Merci à toutes les personnes que j'ai eu la chance de rencontrer dans le cadre de ma thèse, pour les discussions et les conseils reçus. Merci aux doctorants que j'ai pu rencontrer pendant ces trois années ; un grand à merci notamment à Florian pour toutes

nos discussions (toujours optimistes !), qui continueront je l'espère. Merci à Audrey d'avoir été présente au bon moment, merci à Gabriel, Othman, Bilel, Paul, Fabrice.

Ensuite, je souhaite remercier de tout cœur ma (grande) famille, et en particulier mes parents pour leur soutien et leur confiance depuis toujours. Merci maman, pour tout. Merci à Caroline, Ewen, Lise et Tangi de former notre formidable adelphie. Merci à mes grands-parents extraordinaires, d'être toujours là et de me soutenir, et merci aux tartes qui tiennent dans la main. Je ne citerai pas tout le monde un par un, mais je me sens extrêmement chanceuse d'être entourée d'une famille aussi merveilleuse.

Enfin, merci à mes amis, qui par leur soutien ont un peu contribué à cette thèse. Merci à Marjo et Yanne, vous êtes toujours mes préférences, depuis des années et pour des années encore je l'espère. Merci à toute la bande d'être toujours là : Noëmie, Laura, Marianne, Cassin, Malo, Marion, Emma, Pol, Olivier, Milan, Vincent, Gaël, Coco, et tous les autres que je ne cite pas. Merci à Raph et à son merveilleux service d'accueil. Merci à Noëmie ma partenaire de course et de potins. Merci à Alex et Antoine, pour votre bonne humeur permanente et contagieuse. Merci à Valou d'être toujours là quand il faut. Enfin, malgré tout, merci à Brendan d'avoir été mon pilier, et pour tout le bonheur que tu m'as apporté pendant des années.

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte applicatif Les changements de mode de vie de ces dernières décennies entraînent au niveau mondial une augmentation importante de la prévalence des pathologies liées à des dérèglements métaboliques. Ces pathologies peuvent être à l'origine de complications graves et nécessitent une prise en charge de façon précoce. L'inconvénient de ce type de maladies est la difficulté de l'établissement du diagnostic aux stades précoces et intermédiaires d'évolution (Younossi et al., 2018). Dans cette thèse, nous nous intéressons à la Stéatohépatite Non Alcoolique (NASH), une maladie métabolique liée à l'accumulation de lipides dans le foie (EASL–EASD–EASO, 2016). La méthode actuelle de diagnostic de cette maladie étant basée sur une technique invasive, risquée et coûteuse, il n'est pas envisageable de l'utiliser à large échelle. Pour faire face à cette situation, des technologies non invasives permettant d'évaluer la composition moléculaire de biofluides prélevés sur des patients à risques sont en développement (Younossi et al., 2018). Ces technologies pourraient faciliter l'étude des variations de composés associés aux dérèglements métaboliques, et être utilisées pour l'établissement du diagnostic. La spectrométrie infrarouge permet d'étudier la composition d'un échantillon mesuré (Baker et al., 2016), et son potentiel dans le contexte du diagnostic médical a été démontré dans de nombreuses études (Thumanu et al., 2014; Zhang et al., 2013; Le Corvec et al., 2012). Les données obtenues par cette technologie sont des courbes constituées de plusieurs centaines de variables, correspondant chacune à une valeur d'absorbance associée aux nombres d'ondes mesurés. On se trouve alors dans une situation où le nombre de variables est du même ordre de grandeur que le nombre d'observations, voire supérieur. Dans le cadre de la modélisation, cette situation entraîne des difficultés d'estimation des paramètres, et des méthodes spécifiques doivent être utilisées pour la gestion d'un nombre important de variables (Hastie et al., 2001). Tout au long de ce travail, nous considérons des données de spectrométrie mesurées sur une cohorte constituée de patients atteints de NASH et de patients témoins (Anty et al., 2010, 2019). Plusieurs techniques d'analyse sont possibles pour modéliser ces données.

D'une part, il est possible de considérer une approche multivariée en tenant compte des données mesurées aux points de discrétisation. Dans ce cas, les méthodes statistiques utilisées sont des méthodes permettant la modélisation de données multivariées en dimension modérée, incluant une étape de réduction de la dimension. Les avantages de cette approche sont nombreux puisqu'elle permet notamment d'éviter le surapprentissage et de faciliter l'estimation des paramètres (James et al., 2013). En chimiométrie, il est

classique d'utiliser des méthodes de projection sur des composantes combinaisons linéaires des covariables pour construire des modèles de prédiction (Frank and Friedman, 1993). Les méthodes classiques utilisées sont basées sur l'analyse en composantes principales (ACP) et les moindres carrés partiels (PLS). Même si ces méthodes sont efficaces en pratique, la projection des données de départ sur des composantes principales rend difficile l'interprétation des modèles obtenus. Cet aspect est crucial dans notre situation, puisque nous étudions une maladie dont les processus moléculaires impliqués dans l'évolution sont mal connus, nous avons donc un double objectif de compréhension et de prédiction. En spectrométrie, l'information représente une empreinte moléculaire globale d'un biofluide complexe, reflétant le métabolisme des patients mesurés. L'information spectrale est très large, et n'est pas spécifiquement liée au problème étudié. Nous souhaitons alors déterminer les zones du spectre qui reflètent les différences entre les patients malades et les patients témoins. Cela permet d'une part de faciliter la construction d'une règle de diagnostic, et d'autre part de mieux comprendre les processus moléculaires liés à la maladie. Il apparaît alors pertinent d'établir des modèles à partir des variables observées, ce qui pourrait permettre de les relier à des caractéristiques biologiques par la suite. Nous considérons donc une approche de réduction de la dimension différente dans ce travail. Notre choix porte d'abord sur les méthodes de sélection de variables pour réduire la dimension, ce qui permet de travailler sur un sous-ensemble des variables mesurées.

Le chapitre 1 présente en détail le contexte applicatif de cette thèse, ainsi que les premières analyses menées sur les données de spectrométrie mesurées sur la cohorte de patients atteints de NASH. Ces analyses permettent de comparer les performances obtenues avec les approches statistiques évoquées précédemment et mettent en avant les difficultés rencontrées lors du traitement de ce type de données. Ce chapitre permet de poser les problématiques et les objectifs de ce travail.

Finalement, l'objectif de la thèse est d'établir un modèle de diagnostic sur des données complexes, tout en sélectionnant l'information pertinente pour obtenir un modèle interprétable. Le cadre de modélisation utilisé dans cette thèse est d'abord présenté, ainsi qu'une adaptation de ce type de modèles à la prédiction d'une variable réponse suivant une loi usuelle. L'extension de ce modèle au contexte d'étude est ensuite détaillée, avec la construction d'un modèle de prédiction d'une variable binaire de diagnostic grâce une méthode intégrant la sélection de variables. Ensuite, l'adaptation de ce modèle à l'analyse de données fonctionnelles est présentée. Dans ce cadre, une méthode de sélection de portions de courbes est détaillée. Pour finir, l'interaction entre ces portions de spectres est étudiée dans le dernier chapitre.

Classification non supervisée avec des modèles génératifs Un des aspects de la pathologie étudiée est lié à sa prévalence élevée, qui entraîne une répartition de la maladie parmi des types de patients aux caractéristiques très diverses. Nous nous trouvons face à une maladie complexe avec, potentiellement, des trajectoires de maladie différentes, en plus des groupes de diagnostic (Ross and Dy, 2013). En effet, indépendamment du diagnostic, des caractéristiques, morphologiques ou génétiques par exemple, peuvent définir des profils

de patients qui seront marqués par une évolution différente de la maladie, ce qui rend plus difficile le diagnostic. Les méthodes statistiques habituelles consistant à construire un modèle de prédiction unique sur une cohorte de patients n'apparaissent pas adaptées. On suppose que des groupes de patients structurent les observations. Dans le cas de groupes connus *a priori*, les modèles linéaires généralisés mixtes (Breslow and Clayton, 1993) permettent d'établir une règle de prédiction prenant en compte l'hétérogénéité des observations grâce à la combinaison d'effets fixes et aléatoires dans les prédicteurs. Les effets aléatoires permettent notamment de modéliser la non-indépendance observée dans les groupes existant dans les données structurées. Dans notre contexte, les médecins ne sont pas capables d'établir *a priori* les profils de patients structurant les données, qui doivent donc être estimés. Dans ce cas, nous faisons le choix de modéliser de façon discrète l'hétérogénéité de la population étudiée, en prenant en compte une structure latente dans les données, c'est-à-dire une structure en groupes non observés, mais qui a une influence sur les variables caractérisant les patients et sur la règle de diagnostic.

Une des problématiques importantes de cette thèse est donc l'identification de groupes d'observations reflétant des profils de patients particuliers. Dans le chapitre 2, nous introduisons le cadre de classification utilisé comme base de ce travail. En classification, on distingue la classification supervisée de la classification non supervisée, dont les objectifs sont différents (Hastie et al., 2001). En classification supervisée, l'objectif est de construire un modèle de prédiction d'une variable de label, à partir d'observations réparties dans des groupes connus. Dans le cas de la classification non supervisée, aussi appelée partitionnement, on considère des données structurées en groupes latents, c'est-à-dire non observés. La structure de groupes est supposée, et pas forcément réelle. L'objectif est de construire des groupes dans lesquels les observations sont plus similaires les unes des autres que par rapport aux autres groupes, selon une mesure de distance choisie. Cela permet de mieux comprendre les données et notamment d'adapter des méthodes statistiques à la structure des données, ce qui ouvre la voie à de potentielles améliorations des performances de certaines méthodes de prédiction. Il existe plusieurs méthodes de partitionnement : d'une part les méthodes heuristiques de type Classification Ascendante Hiérarchique (CAH) ou K-means (Macqueen, 1967), et d'autre part les méthodes à base de modèles génératifs, basées sur des modèles probabilistes. C'est ce type de modèles qui est considéré dans cette thèse, et détaillé dans ce chapitre.

Les modèles génératifs utilisés dans cette thèse, appelés modèles de mélange, ont été introduits par Newcomb (1886) et Pearson (1894) et permettent d'obtenir une partition des données à partir d'une modélisation probabiliste. Les modèles de mélange tels que décrits historiquement sont basés sur la modélisation de la distribution d'un ensemble de covariables comme une somme pondérée de distributions de probabilités de la même famille, mais dépendant de paramètres différents. McLachlan and Peel (2000) et Grün and Leisch (2007) décrivent l'extension de cette formulation à la modélisation de la dépendance entre une variable réponse et des covariables, grâce aux mélanges de régressions qui peuvent être appliqués aux familles de distributions conditionnelles classiques. Cela permet de considérer divers types de variables réponses, on parle alors de mélange de modèles linéaires généralisés. Ces modèles sont définis avec un design fixe, et ont pour objectif

l'estimation de la structure des données, et non la prédiction. En effet, le mélange est estimé grâce à la distribution conditionnelle d'une variable réponse Y sachant l'ensemble des p covariables $\mathbf{x} \in \mathbb{R}^p$, qui n'est pas observée pour une nouvelle observation, tout comme l'information latente d'appartenance au groupe. Ces modèles ne sont donc pas adaptés à la prédiction d'une variable réponse, et ne peuvent pas être utilisés pour construire une règle de diagnostic dans notre situation. Les mélanges d'experts, qui sont une généralisation des mélanges de régressions, sont définis avec pour objectif la prédiction d'une variable réponse (Jacobs et al., 1991). Dans ce cas, les probabilités *a priori* sont modélisées comme des fonctions des covariables et ainsi le modèle d'experts n'entre pas dans la catégorie des modèles de mélange. Tous ces modèles peuvent être estimés grâce à des approches basées sur le maximum de vraisemblance. Cependant, les expressions à optimiser n'ont pas de solution explicite, et l'estimation est réalisée grâce à un algorithme itératif alternant entre une étape d'estimation (étape E) et une étape de maximisation (étape M), appelé algorithme EM (Dempster et al., 1977). Lors de l'étape E, les probabilités *a posteriori* d'appartenance aux groupes latents sont calculées pour chaque observation. Cela permet le calcul de l'espérance conditionnelle de la log-vraisemblance des données complétées, sachant les données observées. L'étape M permet de mettre à jour les paramètres du modèle maximisant la vraisemblance calculée à l'étape E.

La structure des données étant inconnue *a priori*, le nombre de groupes à considérer dans le mélange est généralement choisi *a posteriori*. Dans le cas d'un modèle simple avec peu de groupes, les paramètres estimés auront tendance à avoir une variance faible, alors que les paramètres des modèles complexes auront un biais faible mais une variance élevée. On se place alors dans le contexte de la sélection de modèles, où l'objectif est de choisir un modèle faisant un meilleur compromis entre biais et variance (Akaike, 1978). Dans ce cadre, il est classique d'estimer une collection de modèles - dans le cas des modèles de mélange, il s'agit de modèles estimés avec différents nombres de groupes - et de sélectionner ensuite le meilleur modèle parmi cette collection (Bozdogan, 1993). Pour cela, la sélection est guidée par des critères de sélection de modèles comme le BIC (Schwarz, 1978) et l'AIC (Akaike, 1973), mais aussi par l'ICL (Biernacki et al., 2000) qui est adapté au cadre des mélanges grâce à la considération d'un terme d'entropie prenant en compte la concentration des groupes estimés.

Les modèles de mélanges sont relativement peu étudiés dans le cadre du design aléatoire dans la littérature. Nous présentons en détails ce cadre dans le cas de mélanges de régressions, ce qui permet de répondre à notre objectif de prédiction d'une variable réponse. Une des contributions de cette thèse est d'adapter les modèles décrits précédemment à la prédiction d'une variable suivant une loi usuelle avec l'utilisation des modèles linéaires généralisés. Un design aléatoire est considéré, et le modèle est estimé sur les covariables et la variable réponse ; les groupes sont donc définis grâce à toutes ces informations. L'avantage de cette approche est de conduire à une règle naturelle de prédiction basée sur les covariables observées.

Dans ce chapitre, nous proposons une famille de modèles de mélange à design aléatoire. Un des points forts du modèle est qu'il permet de définir une règle de prédiction de la réponse. La règle s'appuie notamment sur les probabilités *a posteriori* d'appartenance à la classe latente.

Mélange de régressions pénalisées à design aléatoire Dans le chapitre 3, l'objectif est d'adapter le modèle de mélange de régressions avec design aléatoire au traitement des données de spectrométrie étudiées dans le cadre de cette thèse. Une des difficultés dans l'analyse des données de spectrométrie est liée au risque de surapprentissage important. En effet, la dimension des spectres discrétisés est élevée en comparaison du nombre d'observations, généralement assez faible dans le cas des données médicales. De plus, une grande partie du spectre n'apporte pas d'information utile pour le diagnostic et génère donc du bruit. Dans ce cas, une réduction de la dimension est nécessaire pour l'estimation des paramètres du modèle. L'approche de la sélection de variables est favorisée car elle permet de conserver les absorbances associées aux longueurs d'ondes importantes dans le spectre, et donc de localiser l'information pertinente pour la prédiction, pour éventuellement la relier à des molécules du métabolisme. La stratégie considérée ici est de combiner la sélection de variables avec les modèles génératifs, pour sélectionner l'information pertinente et identifier les classes latentes de façon simultanée.

Il existe de nombreuses méthodes de sélection de variables (Guyon and Elisseeff, 2006), mais l'approche utilisée ici permet de sélectionner les variables de façon simultanée à l'estimation des paramètres du modèle. L'estimation se base alors sur la vraisemblance pénalisée de type Lasso (pour Least Absolute Shrinkage and Selection Operator), introduite par Tibshirani (1994), pour estimer les paramètres de régression. La vraisemblance est pénalisée par une pénalité en norme ℓ_1, ce qui permet de contraindre les régresseurs associés aux variables peu discriminantes à zéro, et donc d'effectuer une sélection de variables. On retrouve cette approche dans le cadre des modèles de mélange notamment dans Lloyd-Jones et al. (2018). De la même façon, les matrices de précision associées à chacun des groupes sont estimées avec parcimonie grâce l'algorithme Graphical Lasso introduit par Friedman et al. (2008). Cet algorithme permet d'estimer les dépendances conditionnelles entre un ensemble de variables grâce à leur matrice de covariance, via une pénalisation de type Lasso appliquée à la matrice précision. La matrice de précision parcimonieuse estimée peut ensuite être utilisée pour la construction de modèles graphiques, reflétant les dépendances conditionnelles entre variables. Dans le cadre d'un modèle de mélange, cette approche permet de sélectionner les paramètres par groupe latent, puisque la régularisation est appliquée aux régressions du mélange par groupe. Les coefficients contraints à zéro peuvent donc être différents selon la composante du mélange.

La pénalisation dépend du choix d'hyper-paramètres permettant de régler le niveau de parcimonie des régresseurs et des matrices de précision. Le modèle estimé dépend donc des valeurs de ces paramètres de régularisation, qui doivent être fixées. Différentes approches sont possibles pour la sélection de ces paramètres. On note par exemple l'utilisation de la validation croisée permettant de choisir les paramètres de régularisation parmi un

ensemble de valeurs possibles, en minimisant une fonction de perte (Khalili and Chen, 2007). Cependant, les méthodes de validation croisée sont coûteuses numériquement. L'approche considérée dans ce chapitre suit donc la même démarche que la sélection du nombre de groupes. La procédure de sélection détaillée est basée sur le choix du meilleur modèle parmi une collection de modèles, et repose sur la minimisation du critère BIC, classiquement utilisé dans ce type de situation (Wang et al., 2007; Jiang et al., 2018; Lloyd-Jones et al., 2018; Khalili and Lin, 2013).

Une étude de simulation permet d'étudier les performances d'estimation et de prédiction du modèle présenté, et de comparer la méthode présentée, appelée PMLR, avec des méthodes classiques de prédiction d'une variable binaire. Cette étude permet aussi d'évaluer et de comparer les critères de sélection de modèles BIC, AIC et ICL.

Nous utilisons ensuite la méthode développée pour l'établissement d'un modèle de diagnostic sur des données issues de la cohorte de patients NASH. De la même manière, cette application permet la comparaison des performances de prédiction obtenues par la méthode PMLR avec les performances obtenues par des méthodes alternatives. Dans ce chapitre, l'objectif est aussi de mettre l'accent sur l'interprétation des modèles et d'utiliser des outils pour faciliter la compréhension du problème médical étudié. Les matrices de précision parcimonieuses estimées pour chaque groupe latent grâce au Graphical Lasso permettent la construction de modèles graphiques (Jordan, 2004), modélisant les dépendances conditionnelles entre les variables retenues dans le modèle. Ces modèles graphiques permettent aussi de mettre en avant les différences entre la sélection de variables selon les groupes, et les liens entre variables estimés au sein des groupes.

Pour résumer, cette partie est basée sur la construction d'une méthode, appelée PMLR, permettant l'estimation d'un modèle de diagnostic, sur une population structurée en groupes latents, tout en réalisant de la sélection de variables dans le modèle mais aussi dans la matrice de précision, pour faire ressortir des dépendances conditionnelles entre variables, avec un but de prédiction d'une variable binaire.

Notre contribution :

- développement d'un modèle de mélange de régressions logistiques à design aléatoire (PMLR)
- sélection de variables par maximum de vraisemblance pénalisée pour le modèle PMLR
- amélioration du diagnostic de la NASH par la mise en œuvre du modèle PMLR
- utilisation des modèles graphiques comme outil de visualisation des dépendances conditionnelles entre variables estimées par les matrices de précision, pour une meilleure interprétation des résultats

Ce chapitre fait l'objet de deux articles.

Mélange de régressions sur données fonctionnelles Jusqu'à présent, les données ont été analysées comme des données multivariées en dimension modérée. Dans le chapitre 4, nous souhaitons prendre en compte la spécificité fonctionnelle des données de spectrométrie, selon les principes introduits par Ramsay and Silverman (1997). En effet, les données étant des courbes mesurées sur une grille de discrétisation fine et régulière, nous pouvons considérer chaque observation comme une fonction. On s'intéresse ainsi à la prise en compte de l'information globale portée par la mesure, c'est-à-dire à l'intensité du spectre mais aussi à sa forme particulière dans les zones considérées. Nous nous plaçons alors dans le cadre de l'analyse de données fonctionnelles, développée en détail dans Ramsay and Silverman (1997) et dans Ferraty and Vieu (2006). Nous cherchons à effectuer du partitionnement de courbes, et de la prédiction de manière simultanée. Les méthodes de classification non supervisée de données fonctionnelles ont notamment été résumées dans Jacques and Preda (2014) et peuvent être décrites en trois grands types. Il existe des méthodes non paramétriques qui utilisent des techniques de classification automatique classiques en considérant des distances ou dissimilarités spécifiques aux données fonctionnelles. D'autre part, il est possible de procéder en deux étapes, avec dans un premier temps une réduction de la dimension et dans un second temps l'utilisation de méthodes de classification usuelles sur les nouvelles données considérées. En outre, il existe des méthodes de partitionnement de données fonctionnelles utilisant les approches s'appuyant sur les modèles génératifs, correspondant à notre cadre d'étude. Ces deux dernières approches sont basées sur l'approximation des courbes observées par leur projection sur une base de fonctions. Dans ce travail, nous utilisons aussi cette méthode. La projection des données fonctionnelles permet d'approcher les fonctions observées dans un espace de dimension infinie, par des combinaisons linéaires de fonctions et de se replacer dans un espace de dimension finie. Les coefficients de projection dans la base de fonctions permettent de résumer l'information fonctionnelle. Nous choisissons une base de fonctions adaptée à notre problématique, constituée de splines, qui permettent de conserver la localisation des zones spectrales, et d'interpréter les modèles estimés. La première étape est donc de projeter les données spectrales sur une base de B-splines. La dimension de la base est déterminée grâce à une procédure permettant de faire un compromis entre une bonne approximation des courbes, via la minimisation de l'erreur de projection, et un nombre minimal de fonctions dans la base. Dans ce travail, nous nous limitons à des nœuds equirépartis.

Notre objectif étant de coupler la classification non supervisée à la prédiction, nous souhaitons utiliser les outils de modélisation développés précédemment et les adapter au cadre fonctionnel en considérant les méthodes de régression fonctionnelle. Nous nous basons notamment sur les modèles linéaires généralisés pour données fonctionnelles, qui sont détaillées par James (2002). Dans ce travail, les prédicteurs sont approchés par des splines cubiques, et les coefficients de projection pour chaque observation sont ensuite utilisés dans la fonction de lien pour lier les prédicteurs et la réponse. Nous considérons une approche similaire en construisant notre modèle sur les coefficients de projection résumant l'information fonctionnelle. Les coefficients de régression estimés sont alors des régresseurs fonctionnels, représentant le lien entre la réponse scalaire et les prédicteurs fonctionnels

(James et al., 2009). Dans ce cas, les régresseurs estimés peuvent être représentés sur la même dimension que les prédicteurs, et dans notre cas sur la fenêtre spectrale.

Les premières analyses de sélection de variables par régression Lasso menées sur les données de spectrométrie NASH mettent en avant la présence de zones spectrales regroupant un grand nombre de variables sélectionnées pour prédire la maladie étudiée. À l'inverse, des zones du spectre ne semblent pas informatives pour le diagnostic. Cependant, dans le cas de dépendance forte entre les prédicteurs comme en spectrométrie où les variables proches sont très corrélées, les propriétés oraculaires de la méthode Lasso sont affectées, notamment sa capacité à déterminer le support du vecteur des régresseurs. On peut faire face à un manque de stabilité dans la sélection des variables, c'est-à-dire qu'une légère modification de l'échantillon des observations étudié entraîne une sélection de variables différente. Nous souhaitons pré-sélectionner les portions de courbes regroupant l'information d'intérêt pour la prédiction sous forme de bandes larges dans le spectre, pouvant être reliées à des molécules. Cela passe par la prise en compte du caractère fonctionnel des spectres avec l'ordre des variables et la forme des courbes dans les zones sélectionnées. La projection des données fonctionnelles sur une base de splines permet d'associer chaque coefficient de projection à un ensemble de variables voisines du spectre. L'information portée par une bande spectrale constituée de plusieurs variables voisines est alors résumée par un coefficient, ce qui permet de prendre en compte les corrélations entre variables consécutives. Cet aspect permet une meilleure stabilité du Lasso. Cette étape de pré-sélection a finalement l'avantage de faciliter l'estimation du modèle PMLR, mais aussi d'améliorer la compréhension du problème étudié.

Dans le cadre fonctionnel, il est plus cohérent de vouloir sélectionner des zones discriminantes du spectre plutôt que des variables. La sélection de domaine, c'est-à-dire la sélection d'intervalles de variables consécutives, est étudiée dans la littérature sur les données fonctionnelles, par exemple par Picheny et al. (2019), Fauvel et al. (2015) et Fraiman et al. (2016), utilisant notamment les outils de pénalisation dans le cadre linéaire. Dans ce travail, nous considérons une méthode différente, basée sur les tests d'hypothèse pour la pré-sélection de variables, ce qui permet de contrôler le risque de variables sélectionnées à tort. Dans cette approche, un score basé sur la p-valeur obtenue à la suite d'un test statistique est calculé pour chaque variable, ce qui permet un classement de l'ensemble des variables et la sélection des variables à meilleur score. Nous considérons les tests de permutation qui permettent d'approcher la loi de la statistique de test par une estimation empirique, ce qui permet l'estimation de la p-valeur (Pesarin and Salmaso, 2010). Ces tests sont basés sur le ré-échantillonnage et peuvent être utilisés pour tester la différence entre deux groupes d'observations. Un des avantages de cette approche est le contrôle du risque d'erreur de première espèce, c'est-à-dire du risque de rejeter à tort l'hypothèse nulle, ce qui n'est pas le cas avec les méthodes de régression pénalisée. La méthode de sélection présentée s'appuie sur l'approche développée par Pini and Vantini (2017), permettant de tester par intervalles la différence entre deux populations, sur des données fonctionnelles. Cette méthode de sélection de portions de courbes repose sur trois étapes. Tout d'abord, les données sont projetées sur une base de fonctions permettant de conserver la localisation. La prise en compte du caractère fonctionnel des données est

basée sur la considération des coefficients de projection sur la base de fonctions pour effectuer les tests. Ensuite, les tests sont réalisés sur un ensemble d'intervalles consécutifs de coefficients de la base, ce qui permet l'obtention d'une p-valeur associée à chaque coefficient. Dans un dernier temps, une correction est appliquée à cette p-valeur, prenant en compte la multiplicité des tests réalisés. Nous proposons une adaptation de cette méthode en modifiant la deuxième étape de la procédure. Les indices des intervalles testés sont modifiés et basés sur les groupes de coefficients construits par classification ascendante hiérarchique (CAH). La matrice de distance sur laquelle la CAH est construite permet de prendre en compte l'ordre des coefficients et donc leur localisation sur la fenêtre spectrale.

Notre contribution :

- extension du mélange de régressions pénalisées à design aléatoire à la prise en compte du caractère fonctionnel des données pour augmenter les performances de prédiction
- méthode de sélection de domaine basée sur les tests d'hypothèse dans notre cadre fonctionnel, permettant d'améliorer les performances et l'interprétabilité des résultats de régression.

Tests par blocs au sein de matrices de corrélation *Ce travail est réalisé en collaboration avec Alessia Pini (Université Catholique du Sacré Coeur, Milan).*

Le dernier travail de cette thèse a notamment pour objectif de permettre une meilleure compréhension des processus moléculaires liés à l'évolution de la maladie. Il existe encore des questionnements concernant les interactions entre les molécules impliquées dans le métabolisme et l'évolution de la NASH. La composition des échantillons prélevés sur des patients atteints de NASH et des patients témoins étant reflétée par les données de spectrométrie, il serait intéressant d'étudier les interactions entre les zones spectrales. Ces zones pouvant être reliées à des types de molécules, leurs interactions sont susceptibles de refléter des processus moléculaires mis en jeu dans la maladie. Ces interactions peuvent être considérées grâce aux matrices de covariance, qui reflètent les dépendances entre variables. De plus, les analyses menées sur les données NASH ont permis l'estimation de groupes de patients caractérisés par des profils métaboliques différents. Nous considérons donc les dépendances entre variables au sein de ces groupes de patients, ce qui pourrait permettre de mieux comprendre les mécanismes métaboliques en action selon le type de patient.

Le chapitre 5 porte donc sur l'étude de la structure des dépendances entre variables, caractérisées par les matrices de covariance. Dans les chapitres précédents, les matrices de covariance sont supposées non structurées. Cependant, si l'on considère la matrice de covariance de données de spectrométrie complètes, la dimension des matrices de covariance est trop grande pour permettre une estimation de bonne qualité. Dans cette situation,

les méthodes usuelles reposent sur des estimateurs pénalisés comme le Graphical Lasso (Friedman et al., 2008). Ici, nous proposons une autre approche qui consiste à identifier les coefficients non nuls de la matrice de covariance par des tests statistiques. On considère la prise en compte d'une structure en blocs au sein de la matrice de covariance, ce qui permet de faciliter son estimation si on est capable d'identifier des blocs de coefficients nuls. Cette approche est notamment suggérée par la présence de structures qui semblent se répartir en blocs au sein des matrices de covariance empiriques. De plus, nous supposons que les bandes spectrales représentant des groupements moléculaires similaires correspondraient à des blocs. Dans notre cadre, nous faisons l'hypothèse que cette structure de blocs affectant la matrice de covariance est liée aux bandes spectrales discriminantes sélectionnées grâce à la méthode de sélection de portions de courbes détaillée dans le chapitre précédent.

Dans ce travail, nous faisons le choix de considérer les matrices de corrélation pour l'étude de la significativité des dépendances entre variables. En effet, la corrélation correspond à la version standardisée de la matrice de covariance, et reflète aussi la structure des dépendances entre variables caractérisées par la matrice de covariance. Dans ce chapitre, nous exposons donc une méthode permettant d'étudier les dépendances entre des zones du spectre pré-sélectionnées, grâce à des tests effectués sur les matrices de corrélation. Là encore, nous adoptons une approche fonctionnelle en projetant les spectres observés sur une base de fonctions et en travaillant ensuite sur les coefficients splines associés. La procédure se déroule en plusieurs étapes. Les spectres observés sont dans un premier temps projetés sur une base de splines. Nous considérons ensuite les coefficients de projection dans la base de fonctions pour les analyses. Dans un deuxième temps, les blocs sont définis en appliquant la procédure de sélection d'intervalles basée sur les tests de permutation détaillée au chapitre 4. Les matrices de corrélation sont ensuite estimées sur les variables restreintes aux blocs discriminants sélectionnés, associés à des zones spectrales. Finalement, les tests d'hypothèse sont réalisés sur des sous-matrices de matrices de corrélation associées aux couples d'intervalles de variables discriminants, dans le but d'évaluer la significativité de la dépendance entre ces zones spectrales. Nous considérons là encore une approche basée sur les tests de permutation pour tester les différences entre blocs de variables. L'aspect fonctionnel est pris en compte grâce à la réalisation de tests des blocs successifs ensembles. La p-valeur ajustée est ensuite calculée et permet la prise en compte de la multiplicité des tests. Cette procédure peut être adaptée à l'étude des dépendances conditionnelles en considérant les matrices de précision. Pour ce faire, une attention particulière doit être portée sur l'estimation des matrices de précision, qui peut se faire selon la méthode proposée par Xia et al. (2018). Le type de permutation effectué est aussi différent dans le cas des matrices de précision, pour garantir un test exact asymptotiquement.

Les performances de la procédure de tests sur les matrices de corrélation sont évaluées grâce à une étude de simulation. Les matrices de covariance utilisées pour simuler les données doivent avoir une forme particulière, avec la présence de blocs de covariance nulle. Pour cela, nous proposons un modèle original de génération de matrices de covariance parcimonieuses par bloc qui permet de garantir la définie positivité et produit des matrices faciles à inverser. Cette approche est basée sur la décomposition en éléments propres,

en simulant des matrices de passage composées de petites rotations aléatoires, avec des valeurs propres fixées. Pour finir, la procédure développée est appliquée à l'analyse des profils métaboliques de patients estimés lors des analyses précédentes sur les données NASH, avec l'étude des dépendances entre zones spectrales par groupe latent de patients. L'objectif ici est de mieux comprendre les interactions entre variables cliniques impliquées dans le développement de la maladie.

Notre contribution :

- développement d'une procédure permettant l'étude de la structure des dépendances entre variables grâce à des tests d'hypothèse par blocs sur les matrices de corrélation estimées sur des données fonctionnelles
- méthode de génération de matrices de covariance parcimonieuses par blocs à l'aide de la décomposition en éléments propres, à partir de matrices de rotation

Chapitre 1

Introduction générale et contexte applicatif

Contents

Les travaux de cette thèse s'appuient sur un contexte particulier, et les méthodes développées sont adaptées à une application médicale utilisant les données de spectrométrie. Dans ce chapitre, le contexte général de la thèse est présenté. Tout d'abord, le contexte médical est développé, avec une description du problème médical étudié. Le principe

de la spectrométrie et la technologie permettant d'obtenir les données étudiées sont ensuite présentés, puis les caractéristiques des données obtenues sont détaillées ainsi que les méthodes statistiques couramment utilisées pour analyser ce type de données. Les premières analyses réalisées sont ensuite présentées sur des données utilisées comme exemple pour toute la suite de la thèse, avec une comparaison des méthodes usuelles utilisées en chimiométrie pour analyser les données spectrales. Ces premières analyses permettent de poser les problématiques de la thèse et de justifier les méthodes développées par la suite.

1.1 Stéatohépatite non alcoolique (NASH)

1.1.1 Définitions

La stéatose non alcoolique du foie (ou NAFLD pour Non-alcoholic fatty liver disease) est caractérisée par une accumulation excessive de graisse dans le foie, sous forme de triglycérides. La NAFLD est associée à une insulinorésistance et définie par la présence de stéatose pour au moins 5% des cellules hépatiques, survenant en l'absence de consommation excessive d'alcool (EASL–EASD–EASO, 2016).

La stéatose peut s'accompagner de lésions des cellules hépatiques (notamment sous forme de ballonisation caractérisée par la déformation des cellules et de nécrose) et d'une inflammation du tissu hépatique. Dans ce cas, on parle de stéatohépatite non alcoolique (ou NASH pour Non-alcoholic steatohepatitis) qui peut progresser vers la fibrose, caractérisée par l'apparition de cicatrices résultant de l'accumulation de constituants nouveaux de la matrice extracellulaire (notamment du collagène), et la cirrhose, qui correspond au stade le plus avancé de la fibrose. Un cancer du foie peut survenir durant cette séquence (EASL–EASD–EASO, 2016).

La NASH est causée par deux mécanismes : l'accumulation hépatique de triglycérides due à une insulinorésistance dans un premier temps, entraînant un stress oxydatif conduisant à une inflammation dans un second temps (McCullough, 2006). Cependant, cette hypothèse apparaît réductrice, et l'évolution de la NAFLD serait plus complexe. Elle résulterait de multiples agressions agissant en association avec des facteurs génétiques, ce que l'on peut nommer l'hypothèse "multiples coups" (Buzzetti et al., 2016). De manière générale, la NASH est liée à des troubles métaboliques importants, comme par exemple l'obésité et le diabète de type 2 (Younossi et al., 2018).

1.1.2 Épidémiologie et facteurs de risque

La NAFLD et la NASH font partie des maladies chroniques du foie les plus répandues dans les pays occidentaux. Ces maladies suivent l'évolution du nombre de patients atteints d'obésité, de diabète de type 2 et de syndrome métabolique, qui est en augmentation mondiale, due aux changements de mode de vie des dernières décennies (Younossi et al., 2018).

Actuellement, la prévalence globale de la NAFLD est estimée à 25% (Younossi et al., 2016). Dans les pays à plus fortes prévalences de NAFLD, en Amérique du sud et au

Moyen-Orient, elle est estimée entre 14% et 32%, ces chiffres très variables étant dus aux différentes méthodes de diagnostic utilisées, ainsi qu'aux différences entre les populations étudiées, caractérisées notamment par l'âge ou l'ethnie (Younossi et al., 2018).

La NAFLD a une prévalence en constante augmentation (passant de 15% en 2005 à 25% en 2010), contrairement aux autres pathologies du foie dont la prévalence stagne, voire diminue. De manière similaire, le taux de NASH parmi les cas de NAFLD a presque doublé dans cet intervalle de temps (59.1% versus 33%)(Younossi et al., 2018). La prévalence globale de NASH est difficile à estimer dans la population générale ; cependant, elle est estimée dans la population de NAFLD ayant subi une biopsie entre 7 et 30%, ce qui permet une estimation de prévalence dans la population générale entre 1.5% et 6.45% selon les études (Younossi et al., 2016).

La NAFLD est associée à une mortalité plus élevée par rapport à la population générale. Cependant, le risque de mortalité lié aux atteintes au foie dans le cas de NAFLD semble être majoritairement associé à l'âge, à la résistance à l'insuline, à l'inflammation hépatique et à la fibrose (Adams et al., 2005). Il apparaît aussi que la majorité des patients atteints de NAFLD et de NASH ne décèdent pas directement de leurs problèmes hépatiques, mais d'événements cardiovasculaires (voir par exemple Pisto et al., 2014). Il a été montré que seule la fibrose est associée à une mortalité globale à long terme, à la transplantation hépatique, et aux autres événements liés au foie (White et al., 2012).

Des facteurs environnementaux, comme un régime trop riche en lipides et en sucre, la consommation en excès et la sédentarité menant à un gain de poids et à l'obésité sont liés à la NASH (Vernon et al., 2011). D'autres facteurs, comme la prise de certains médicaments, augmenteraient le risque de NASH (Pessayre et al., 2002). Le microbiote intestinal aurait un rôle dans le développement de l'obésité et de la NASH (voir par exemple Henao-Mejia et al., 2012), mais il reste à définir si le changement de microbiote est une cause ou une conséquence de la maladie. D'autre part, des facteurs génétiques jouent un rôle dans l'apparition de la maladie (Seko et al., 2018).

Finalement, les processus actuels selon lesquels un patient passe du stade de NAFLD au stade de NASH sont encore mal connus, et résultent de multiples facteurs.

Remarque : il y a une proportion non négligeable de patients minces parmi la population atteinte de NAFLD. Les cas de NAFLD chez les patients minces concerneraient 10 à 20% des américains et européens (Younossi et al., 2018). Ces cas englobent des populations hétérogènes et seraient liés à la sédentarité, à une augmentation de la graisse viscérale, à des régimes riches en fructose et en lipides, ainsi qu'à des facteurs génétiques ayant notamment un effet sur le métabolisme.

1.1.3 Diagnostic

Données cliniques et sériques

Dans la majorité des cas, la NAFLD est asymptomatique, et lorsque des symptômes apparaissent, ils sont peu spécifiques.

Les dysfonctionnements métaboliques liés à la NAFLD et à la NASH, pourraient se

refléter sur un profil sérique de patient qui pourrait être utilisé pour établir le diagnostic. Il existe un ensemble de variables morphologiques et cliniques utilisées pour le diagnostic de la NAFLD, marquée par la présence de stéatose. On peut notamment noter que les transaminases, les Gamma-glutamyl-transpeptidases (GGT), la bilirubine ou des variables indicatrices de l'état du métabolisme du patient (comme la glycémie, l'insuline, l'insulinorésistance et le diabète) sont utilisées (Younossi et al., 2016). De plus, la présence du syndrome métabolique, définie par plusieurs facteurs de risques comme l'obésité androïde, l'hypertension artérielle, l'hyperglycéridémie, de faibles niveaux de lipoprotéines de haute densité-cholestérol et l'hyperglycémie, est aussi recherchée (EASL–EASD–EASO, 2016).

Cependant, il est difficile de différencier la NAFLD de la NASH à partir d'indicateurs sériques, et d'avoir des marqueurs spécifiques de la NASH. Les tests sanguins développés pour le diagnostic de la NASH sont peu utilisés car construits à partir de dosages de molécules non réalisés en routine, souvent coûteux et compliqués à mettre en place en laboratoire. De plus, ces tests montrent généralement des faibles performances (EASL–EASD–EASO, 2016). Enfin, il a été suggéré que la détection de la NAFLD en utilisant uniquement les tests sanguins conduit à une sous-estimation du nombre de cas de NAFLD (Younossi et al., 2016).

Biopsie

Actuellement, la méthode de référence pour permettre le diagnostic de la NASH est le prélèvement de tissus hépatiques (appelé biopsie), qui permet de détecter la présence de lésions hépatiques, de stéatose, d'inflammation et de fibrose. Suite au prélèvement, l'histologie permet notamment de différencier les cas de NAFLD des cas de NASH, et d'évaluer la sévérité de l'atteinte au foie. Pour cela, le NAFLD Activity Score (ou NAS, introduit par Kleiner et al. (2005)) est calculé sur la base de trois critères : stéatose, ballonisation et inflammation. Ces critères sont répartis en stades de gravité, et additionnés pour déterminer la présence de NASH selon la règle suivante : si le score est supérieur ou égal à 5, la NASH est certaine, et si le score est inférieur à 3, il n'y a pas de présence de NASH. Plus récemment, Bedossa et al. (2012) ont développé le score SAF (Steatosis-Activity-Fibrosis) qui reprend les caractéristiques du score NAS mais implique qu'il est nécessaire de retrouver de façon concomitante les trois critères histologiques de stéatose, inflammation et ballonisation pour établir le diagnostic de NASH. Pour les données étudiées dans cette thèse, c'est ce score qui a permis d'établir le diagnostic de NASH. La biopsie est actuellement la seule méthode permettant d'évaluer le niveau d'atteinte au foie et de diagnostiquer et évaluer la sévérité de la NASH. Les différents stades d'atteinte au foie visibles sur lames de tissus hépatiques obtenues à l'issue de biopsies sont décrits en figure 1.1.

La biopsie est cependant une méthode invasive et coûteuse, et expose à des risques de complications. De plus, le temps d'établissement du diagnostic peut être long suite au prélèvement, compte tenu de la préparation et de la lecture histologique. En outre, les prélèvements peuvent ne pas être représentatifs de l'ampleur des lésions, car les grades histologiques ne sont pas répartis de façon homogène dans le foie. Cet aspect

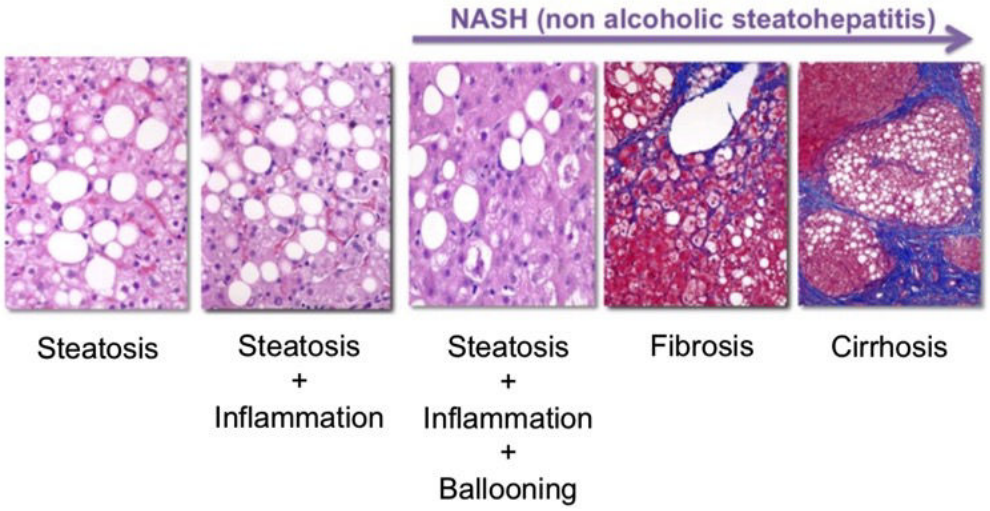

FIGURE 1.1 : **Différents stades d'atteinte au foie montrant l'évolution vers la NASH.** Chaque stade est représenté par une image de l'histologie obtenue après biospie. *(Adams, Riga 2016)*

est notamment illustré par la figure 1.2, montrant la répartition hétérogène des lésions hépatiques. De plus, il existe une variabilité inter-observateur, et des experts différents peuvent évaluer de façon différente un même échantillon. De surcroît, la biopsie ne permet pas de faire un suivi sur le long terme de l'évolution de l'état du foie d'un même patient, puisqu'il est fortement déconseillé de faire plusieurs biopsies du foie sur un même patient. Tous ces inconvénients mettent en exergue l'importance du développement de méthodes de diagnostic non invasif de la NASH.

Actuellement, il existe des méthodes non invasives permettant d'évaluer la stéatose et certains grades de fibrose. Cependant, il n'existe pour l'instant aucune méthode non invasive validée permettant de diagnostiquer la NASH, ou de différencier la NAFLD de la NASH. Avec le nombre croissant de cas de NAFLD dans le monde, la biopsie ne peut être envisagée à large échelle et il y a un réel besoin de développer des outils précis, accessibles, non invasifs et sûrs permettant de diagnostiquer et de surveiller la NASH. De nombreuses recherches sur ce sujet sont en cours (Younossi et al., 2018).

Dans cette thèse, nous nous basons sur des données récoltées sur une cohorte de patients atteints de NASH et de patients témoins, avec pour objectif la construction d'un modèle de prédiction de la NASH, basé sur une méthode non invasive.

1.1.4 Les données NASH de la cohorte de Nice

Les données considérées ici sont issues de mesures cliniques réalisées sur 395 patients provenant de l'hôpital universitaire de Nice. Les patients considérés sont atteints d'obésité morbide, et hospitalisés pour subir une chirurgie bariatrique, c'est-à-dire une chirurgie qui modifie l'anatomie du système digestif dans le but de restreindre l'absorption des aliments.

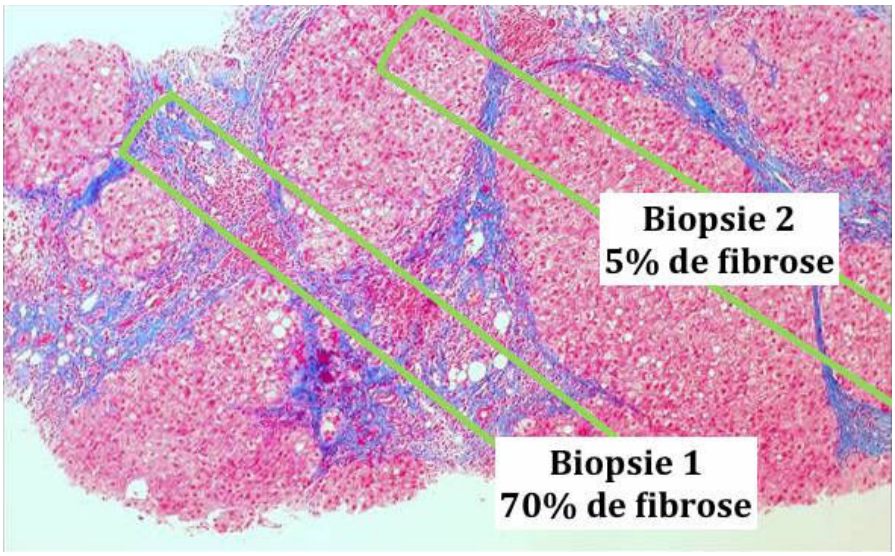

FIGURE 1.2 : **Hétérogénéité des prélèvements possibles de tissu hépatique par biopsie.** La biopsie 1 présente une atteinte au foie avancée, avec une quantité de fibrose importante, alors que la biopsie 2 présente très peu de fibrose. *(Ed Uthman - Cirrhosis of the liver, 2011)*

Des données clinico-biologiques ont été prélevées 2 à 3 semaines avant l'opération et incluent un examen médical physique, des analyses de pression sanguine, des mesures anthropométriques (poids, taille et tour de taille), une évaluation psychiatrique et nutritionnelle et des prélèvements sanguins à jeun. Ces prélèvements permettent l'obtention des variables détaillées dans la table 1.1.

Les tissus hépatiques sont collectés durant l'opération de chirurgie bariatrique et évalués par un pathologiste qui établit le diagnostic de NASH. Parmi les patients, 66 individus sont atteints de NASH, soit environ 17% des observations.

Ces données sont décrites en détail notamment dans Anty et al. (2010). La table 1.2 décrit les caractéristiques cliniques et histologiques des patients de la cohorte par classe de diagnostic. On dispose donc d'un ensemble de variables cliniques mesurées sur les patients, mais il serait intéressant d'avoir un outil permettant de refléter globalement le profil sérique de chaque patient.

Nous cherchons ici à construire un outil de diagnostic non invasif de la NASH. Pour cela, nous souhaitons nous baser sur une méthode permettant d'étudier la signature métabolique d'un échantillon biologique complexe, pour évaluer les différence entre les patients malades et les patients témoins.

Variable	Description
Age	Age
Sex	Sexe
Weight	Poids
BMI	Indice de Masse Corporelle
Height	Taille
AST	Aspartate aminotransferase
ALT	Alanine transaminase
AST.ALT	Ratio Aspartate aminotransferase-Alanine transaminase
GGT	Gamma-glutamyltransferase
Gluc	Glycémie
Insuline	Insuline
HBA1C	Hémoglobine glyquée
chol	Cholestérol total
HDL	High Density Lipoprotein (Cholestérol circulant)
LDL	Low Density Lipoprotein (Cholestérol circulant)
TG	Triglycérides

TABLE 1.1 : **Description des variables cliniques du jeu de données NASH Nice.** Variables morphologiques et issues de prises de sang, disponibles pour les données NASH, et utilisées en combinaison des données de spectrométrie pour l'établissement d'un modèle de diagnostic de la NASH.

1.2 Données de spectrométrie infrarouge

1.2.1 Principe de la spectrométrie infrarouge

La spectrométrie est l'étude de l'interaction entre le rayonnement électromagnétique de la lumière et la matière. Les rayonnements du spectre électromagnétique peuvent être absorbés par la substance étudiée, on parle alors de spectrométrie d'absorption. L'intensité de ces rayonnements en fonction de la longueur d'onde, de la fréquence ou de l'énergie est appelée "spectre".

La spectrométrie infrarouge (IR) est une spectrométrie d'absorption. En effet, lorsqu'une molécule est traversée par un rayonnement IR (c'est-à-dire dont la longueur d'onde est comprise entre 2,5μm et 50μm), certaines liaisons peuvent absorber partiellement et sélectivement ce rayonnement pour changer de fréquence de vibration, ce qui fait apparaître des bandes dans le spectre. Un spectre IR est donc constitué de bandes d'absorption spécifiques des liaisons constituant les molécules étudiées. La position d'une bande dépend des liaisons entre atomes, et de nombreux facteurs externes et internes à la molécule ont une influence sur la fréquence de vibration. L'intensité de la bande dépend de la concentration, de la nature et de la polarité de la liaison, ce qui permet de déterminer la nature des liaisons chimiques présentes dans une molécule, et facilite l'identification spectrale (Lehmann, 1963). Chaque molécule possède des combinaisons de

	NASH	Non NASH	p-value
Age	43 (38-51)	39 (31-47)	4.10^{-3}
Poids	123 (110-135)	114 (106-129)	0.02
IMC	44 (42-48)	43 (41-47)	0.15
AST	35 (26-44)	23 (19-29)	5.10^{-9}
ALT	43 (28-64)	24 (18-36)	2.10^{-5}
GGT	54 (30-81)	28 (19-45)	8.10^{-5}
Glycémie	5.7 (5-8)	5.2 (4.8-5.8)	3.10^{-4}
Insuline	23 (15-33)	18 (11-25)	6.10^{-4}
HBA1C	6 (5.5-7.2)	5.6 (5.3-5.9)	2.10^{-6}
cholestérol	5.2 (4.5-6.1)	5.4 (4.7-6)	0.8
HDL	1.2 (1.1-1.5)	1.4 (1.2-1.6)	5.10^{-3}
LDL	2.9 (2.4-3.7)	3.2 (2.6-3.8)	0.04
Triglycérides	1.7 (1.1-2.9)	1.4 (1.1-1.9)	9.10^{-4}
Tour de taille	33 (26-42)	24 (13-35)	7.10^{-5}
Diabète	38-62	17-83	3.10^{-4}
Syndrome métabolique	72	46	4.10^{-4}
Steatose (%)	0-8-32-60	7-46-26-21	4.10^{-12}
Ballonisation (%)	0-95-5	97-3-0	10^{-16}
Inflammation (%)	0-97-3	99-1-0	10^{-16}
Fibrose (%)	15.4-52.3-16.9-15.4	13.5-63.8-19.6-3.1	5.10^{-3}

TABLE 1.2 : **Caractéristiques des patients obèses de la cohorte Nice, selon le diagnostic de NASH.** Les valeurs quantitatives représentent les médianes et les interquartiles. Pour les variables qualitatives, le pourcentage de chaque catégorie est précisé. La dernière colonne correspond à la p-valeur associée au test de comparaison entre les malades et les non malades pour chaque variable. Un test de Student est effectué pour les variables quantitatives, et un test de Fisher est effectué pour les variables qualitatives.

groupes vibratoires propres, donc le spectre d'absorption dans l'IR est caractéristique et peut être utilisé pour l'identification et l'étude de ces molécules. Les zones du spectre peuvent donc être reliées à des familles de molécules comme les lipides, les glucides ou les protéines par exemple, que l'on retrouve dans les fluides biologiques. Les différentes zones spectrales qui peuvent être reliées à ces types de molécules sont représentées en figure 1.3.

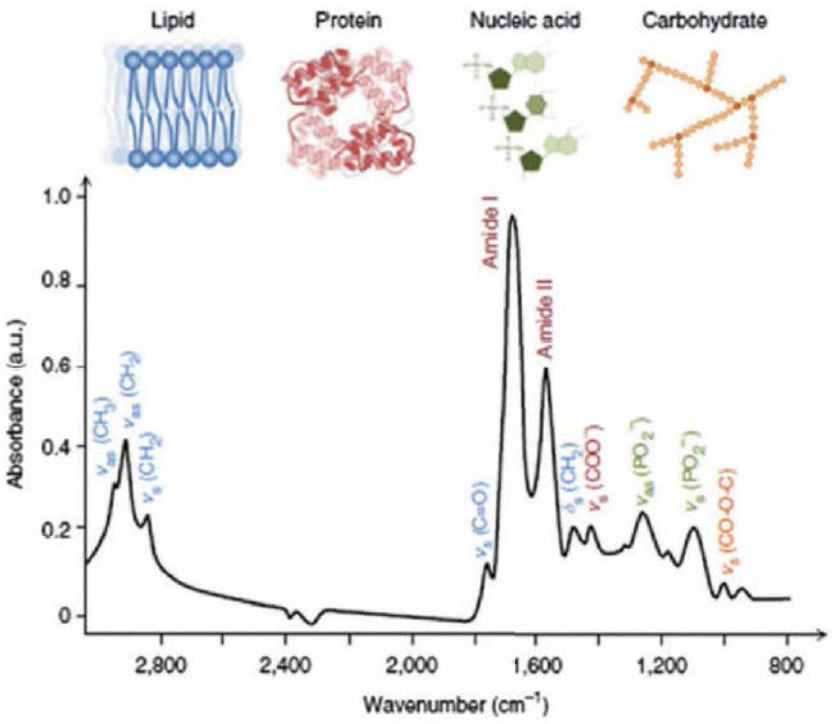

FIGURE 1.3 : **Spectre montrant les correspondances biomoléculaires des bandes de fréquence allant de 3000 à 800 cm^{-1}** (Baker et al., 2016).

Les longueurs d'ondes moyen infrarouge (correspondant à des longueurs d'ondes allant de 2,5μm à 25μm) comprennent les signatures de vibration fondamentales des principales molécules qui composent les tissus et les fluides biologiques. Par conséquent, le spectre d'absorption de la lumière dans le moyen infrarouge d'un fluide biologique est représentatif de sa composition, car il reflète avec précision la structure des molécules constituant l'échantillon.

Technologie d'acquisition spectrale

Les données étudiées dans cette thèse sont obtenues avec la technologie développée par l'entreprise Diafir, utilisant des fibres de verres chalchogénures. Les fibres font partie d'un capteur LS-23 à usage unique développé spécialement, permettant d'enregistrer et d'analyser la signature moléculaire d'un échantillon à partir d'une goutte de liquide.

Les spectres sont mesurés avec un spectromètre DIAFIR SPIDTM FT-IR (pour spectrométrie IR à transformée de Fourier). Le spectromètre mesure un interférogramme, puis une transformée de Fourier permet d'obtenir le spectre simple faisceau. Il correspond au rapport entre l'absorbance mesurée sur l'échantillon et le simple faisceau mesuré sur l'air. Pour chaque échantillon, 64 spectres simples faisceaux sont moyennés pour obtenir le spectre étudié, ce qui permet de réduire le bruit de mesure. Le spectromètre utilisé étant un spectromètre commercial, nous n'avons pas accès aux interférogrammes, mais seulement au spectres simples faisceaux moyennés. Les spectres étudiés ne sont donc pas des spectres bruts. Pour chaque échantillon mesuré, on obtient une courbe d'absorbance, qui représente la valeur de l'absorbance en fonction de la longueur d'onde λ mesurée en μm. On peut choisir d'exprimer le spectre en nombres d'ondes σ exprimé en cm^{-1}, selon $\sigma = 1/\lambda$. Les spectres d'absorption FTIR ont été acquis dans une fenêtre de fréquence allant de 4000 à 600 cm^{-1}, avec une discrétisation entre les points du spectre de 2 cm^{-1}. Lors de l'acquisition, un capteur jetable est placé dans le spectromètre puis le signal simple faisceau est mesuré sur l'air. Une goutte de 7μL d'échantillon à mesurer est ensuite placée sur le capteur, au contact de la fibre de verre chalcogénure. Un temps de séchage permettant d'éliminer l'excès d'eau de l'échantillon est alors nécessaire avant la mesure, afin d'obtenir un signal suffisamment informatif et d'éviter que certaines bandes d'absorption ne soient masquées par le signal de l'eau. La technologie utilisée pour obtenir les mesures est présentée en figure 1.4, montrant le capteur à usage unique ainsi que le spectromètre permettant l'acquisition spectrale.

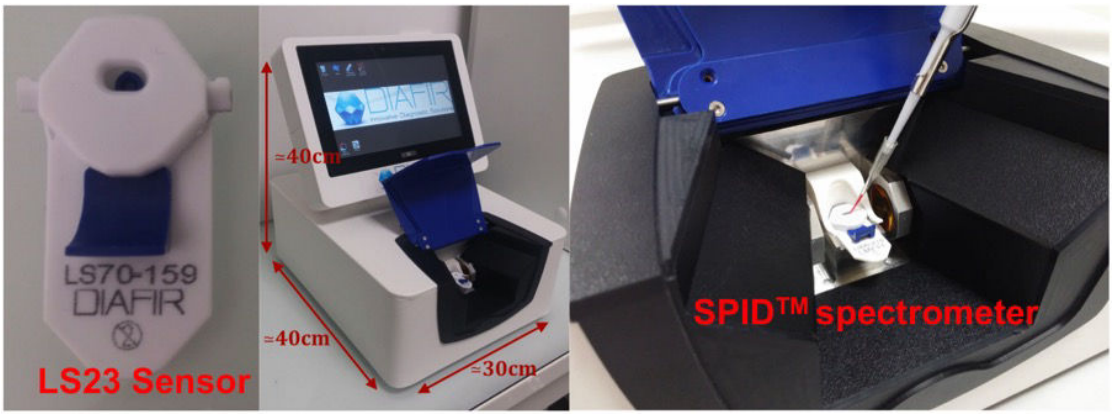

FIGURE 1.4 : **Technologie Diafir.** À gauche : capteur à usage unique contenant la fibre, au milieu : spectromètre permettant la mesure, à droite : détail du dépôt d'un échantillon avant la mesure.

Les avantages de la spectrométrie sont nombreux :

- la rapidité de mesure
- la reproductibilité et la fiabilité
- la haute résolution spectrale

- la simplicité d'utilisation
- pas ou peu de préparations d'échantillon sont nécessaires avant la mesure
- la sensibilité de la mesure.

1.2.2 Les données obtenues

Prétraitements

Des prétraitements sont appliqués aux spectres bruts obtenus, ce qui permet d'améliorer l'efficacité des analyses statistiques réalisées par la suite, de faciliter l'interprétation des données spectrales, de détecter et d'éliminer les données aberrantes et de réduire la dimension des données. Pour plus de détails sur ces prétraitements, le lecteur peut se référer à Lasch (2012).

Ce prétraitements, qui peuvent être combinés, comprennent :

- des tests qualité qui permettent d'apprécier la qualité du signal et d'identifier les spectres aberrants lorsqu'ils ne respectent pas les critères définis, puis de les éliminer. Ces test sont notamment basés sur l'intensité du signal et le rapport signal sur bruit.
- des corrections de ligne de base permettant de corriger les variations de ligne de base, qui peuvent notamment être dues aux conditions d'acquisition.
- une normalisation pour minimiser les différences d'intensité du signal qui ne sont pas liées à l'échantillon mais à l'instrumentation ou à des variations de l'épaisseur de l'échantillon, et ramener les spectres à la même échelle pour pouvoir les comparer. Certaines méthodes prennent en compte seulement le spectre étudié, et d'autres considèrent un ensemble de spectres.
- la filtration qui consiste à choisir les fenêtres spectrales d'intérêt selon l'application et éliminer les zones inutiles ou redondantes, ce qui permet de réduire la dimension des données.
- la dérivation : en spectrométrie, il est très courant de travailler sur les dérivées secondes normalisées des courbes d'absorbance, cela permet notamment :
  - d'améliorer la résolution des spectres
  - de faire ressortir les bandes d'absorbance qui se chevauchent
  - de lisser les petites variations que l'on peut considérer comme du bruit
  - de corriger la ligne de base.

  L'algorithme de Stavisky-Golay (Savitzky and Golay, 1964) est le plus utilisé en spectrométrie infrarouge.

La figure 1.5 présente des spectres mesurés sur des échantillons de sérum, avant dérivation, puis en dérivées secondes.

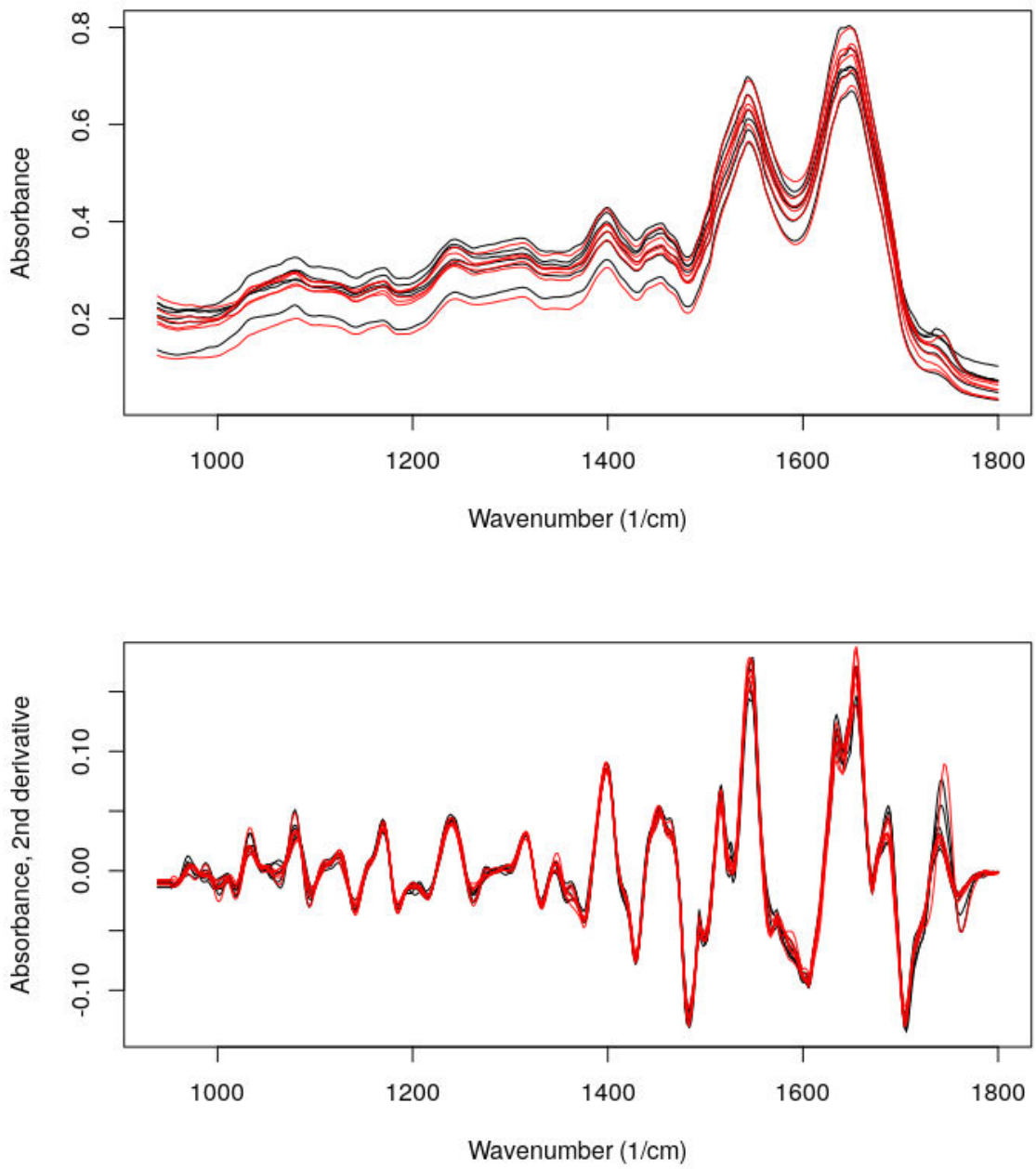

FIGURE 1.5 : **Spectres IR mesurés sur des échantillons de sérum.** Mesures sur des échantillons de sérum prélevés sur des patients atteints de stéatohépatite non alcoolique (rouge) et sur des patients témoins (noirs) : avant dérivation (haut) et après dérivation (bas). Pour plus de clarté, seule la fenêtre spectrale allant de 800 cm^{-1} à 1800 cm^{-1} est représentée.

1.2.3 Justification de l'utilisation des données de spectrométrie IR en diagnostic médical

Certaines maladies sont associées à des modifications biochimiques des cellules, tissus et organes, qui peuvent se refléter dans les fluides biologiques circulant dans l'organisme. Des composés biologiques mesurables dans ces fluides peuvent être utilisés pour détecter ces maladies, et sont appelés biomarqueurs. Une pathologie peut affecter plusieurs biomarqueurs de façon simultanée, et les médecins souhaitent dans ce cas étudier une combinaison de biomarqueurs pour l'établissement du diagnostic. La spectrométrie permet d'obtenir une mesure reflétant la composition moléculaire d'un échantillon. Cette technologie permet de caractériser les modifications biochimiques sur un fluide complexe, et peut donc être utilisée comme méthode de diagnostic. Les modifications biochimiques et donc spectrales provoquées par une pathologie sont spécifiques et uniques, ce qui permet de constituer une empreinte valable même si une variabilité inter-individuelle est présente. Enfin, les biofluides peuvent refléter des changements biochimiques apparaissant au début du développement d'une pathologie, avant des manifestations morphologiques et symptomatiques. L'étude des biofluides pourrait donc permettre de détecter certaines pathologies à des stades peu avancés. Plusieurs études montrent le potentiel de la spectrométrie IR dans le contexte de diagnostics médicaux utilisant des biofluides (comme par exemple Ollesch et al., 2014; Hands et al., 2014).

Lors de cette thèse, plusieurs types de données médicales ont été analysées. On notera par exemple l'analyse d'échantillons d'urine pour établir le diagnostic du cancer de la vessie. la figure 1.6 représente le deuxième et troisième quartiles des dérivées secondes des spectres mesurés sur des individus atteints de cancer de la vessie et des individus témoins. On remarque quelques différences entre les deux groupes, notamment dans la zone spectrale allant de 1300 à 1450 cm^{-1}. Sur ces données, des bonnes performances de prédictions sont obtenues, avec une aire sous la courbe ROC égale à 0.87. Ce critère est un indicateur de la capacité de prédiction d'un classifieur binaire.

L'étude de liquides articulaires pour la prédiction d'arthrite septique a aussi permis de montrer que la signature spectrale des infections est très marquée. Les analyses effectuées sur la cohorte Synofast ont par exemple permis d'obtenir une aire sous la courbe ROC de 0.977, indiquant une très bonne performance de prédiction (Albert et al., 2016). Cela montre la capacité de la spectrométrie à détecter des infections dans les biofluides. Ces problèmes médicaux sont simples ; mais certaines pathologies sont bien plus complexes à diagnostiquer, notamment du fait de l'existence de grades d'évolution dans la gravité de la maladie, ou de typologies de patients différents.

Intérêt de la spectrométrie pour le diagnostic de la NASH

La NASH étant une maladie affectant le métabolisme, des répercussions sont visibles sur la composition du sérum collecté sur des patients atteints. Dans cette situation, nous faisons l'hypothèse que la spectrométrie, qui reflète la composition moléculaire de l'échantillon mesuré, pourrait permettre de mettre en évidence les différences entre les

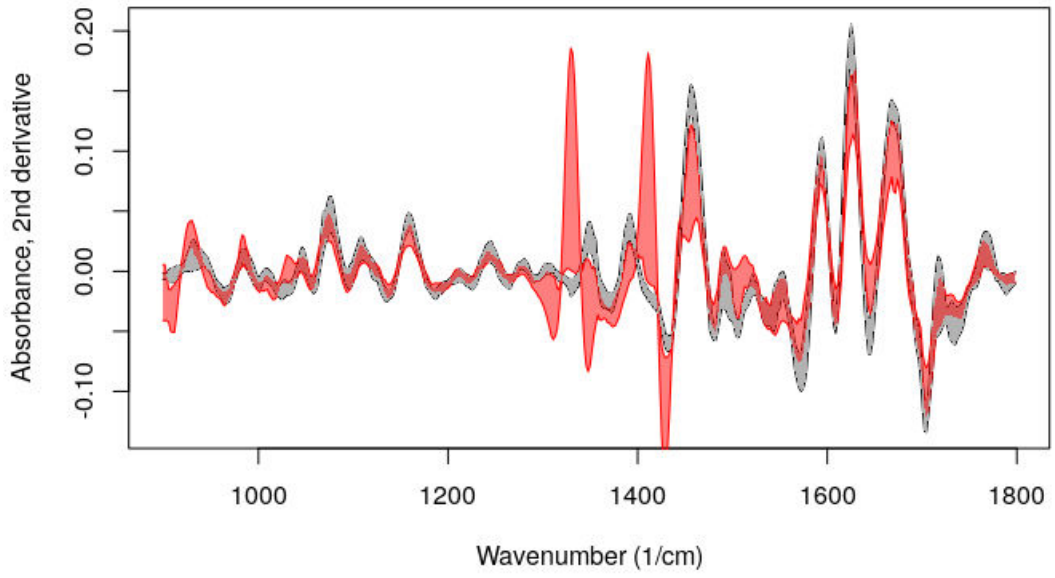

FIGURE 1.6 : **Deuxième et troisième quartiles des dérivées secondes des spectres mesurés sur des échantillons d'urine.** Mesures sur des échantillons prélevés sur des individus atteints de cancer de la vessie (rouge) et des individus témoins (gris). Pour plus de clarté, la zone spectrale représentée est réduite à la fenêtre allant de 800 à 1800 cm^{-1}.

patients atteints de NASH et les patients témoins. Nous souhaitons donc utiliser les données de spectrométrie pour l'étude des patients atteints de NASH, et notamment pour la meilleure compréhension des processus moléculaires d'évolution de la maladie, ainsi que pour l'établissement du diagnostic.

En hépatologie, des études ont montré l'utilité de la technologie de spectrométrie MIR pour le diagnostic de certains cancers du foie ou de la cirrhose décompensée (Thumanu et al., 2014; Zhang et al., 2013; Le Corvec et al., 2012). Nous souhaitons donc utiliser cette technologie pour construire un modèle de diagnostic de la NASH, en nous basant notamment sur les patients de la cohorte NASH-Nice décrite précédemment.

Données spectrométriques NASH-Nice

Des mesures spectrométriques ont été réalisées sur des échantillons prélevés sur les patients de la cohorte NASH-Nice.

Le sérum est collecté pour chaque patient avant l'opération et est immédiatement stocké à une température de -80°C. Les mesures spectrométriques sont effectuées sur le sérum décongelé, 8 minutes après le dépôt de la goutte de liquide dans le capteur,

pour permettre l'évaporation de l'excès d'eau. Sur les données de spectrométrie, les prétraitements classiques ont été effectués. Les spectres sont analysés dans la fenêtre de fréquence allant de 3800 à 950 cm-1, domaine où la plupart des biomolécules sont apparentes. Les critères de qualité classiques ont été mesurés et validés. Le domaine de fréquence allant de 2800 à 1800 cm-1, correspondant à la contribution du CO2 présent dans l'air environnant lors de la mesure, a été éliminé des données. Les dérivées secondes sont calculées en utilisant un algorithme de lissage de Savitzky-Golay (Savitzky and Golay, 1964) sur 13 points, et une normalisation est réalisée sur toute la fenêtre spectrale.

La figure 1.7 représente les deuxième et troisième quartiles des dérivées secondes des spectres par classe de diagnostic des patients. En comparant cette figure à la figure 1.6 obtenue avec des données spectrométriques prélevées sur des urines, nous remarquons que dans le cas des données NASH, il est difficile de discriminer à l'œil nu les patients malades des patients témoins, ainsi que de déterminer les portions du spectre potentiellement prédictives de la maladie. Il est donc nécessaire d'utiliser des méthodes de sélection de variables permettant de déterminer les zones discriminantes du spectre. L'objectif, sur ces données, est d'utiliser les spectres obtenus de façon non invasive, pour établir un modèle de diagnostic de la NASH, les variables cliniques étant utilisées uniquement pour l'interprétation du modèle et des résultats.

Dans la suite, nous nous concentrons sur les données NASH-Nice pour présenter les méthodes d'analyse de données de spectrométrie, avec un objectif de construction d'un modèle de diagnostic.
Même si les méthodes développées dans cette thèse se basent sur une application spécifique, elles sont applicables à de nombreuses situations.

1.3 Prédiction d'une variable binaire à partir de données de spectrométrie

Après le prétraitement des spectres, des analyses statistiques peuvent être réalisées sur les données de spectrométrie afin d'étudier un problème spécifique. Si on se place dans un contexte médical, on cherche par exemple à construire un modèle de diagnostic, c'est-à-dire un modèle permettant de prédire une variable binaire représentant la présence d'une maladie, à partir du spectre.

1.3.1 Caractéristiques des données analysées

Nous faisons face à un signal complexe par plusieurs aspects. Si l'on travaille sur des biofluides, le spectre reflète un échantillon complexe constitué de nombreuses molécules en interaction. Les bandes d'absorption des biomolécules peuvent se chevaucher et les interactions des biomolécules entre elles peuvent décaler les maxima d'absorption, par rapport à des biomolécules mesurées individuellement. De plus, du bruit de mesure lié à l'environnement, à l'opérateur ou au spectromètre utilisé peut faire partie du signal. Enfin, le signal ne dépend pas uniquement de l'objectif considéré, qui peut être de prédire

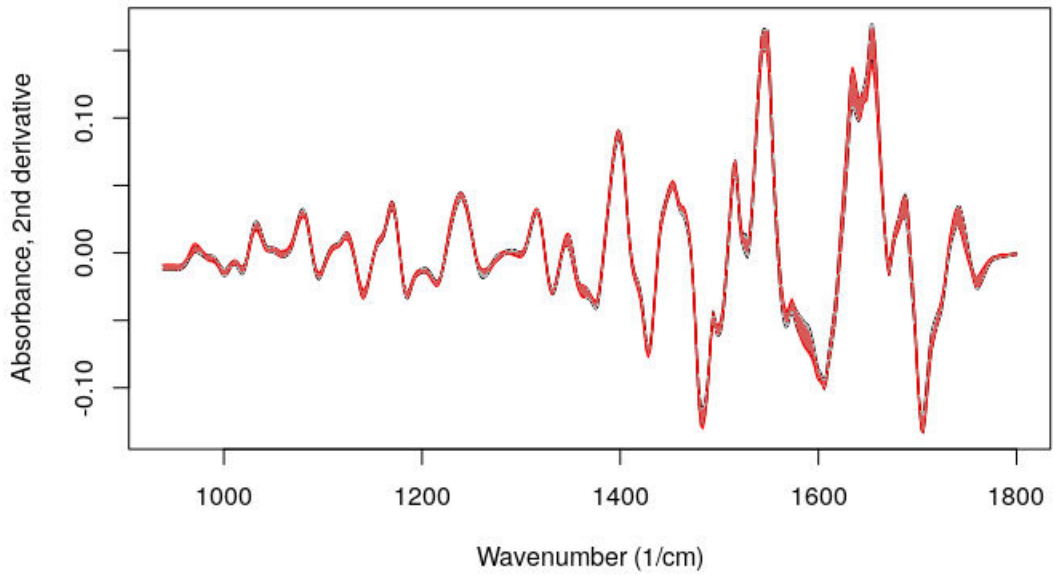

FIGURE 1.7 : **Deuxième et troisième quartiles des dérivées secondes des spectres mesurés sur des échantillons de sérum de la cohorte NASH Nice.** Mesures sur des échantillons prélevés sur des patients atteints de NASH (rouge) et des patients non malades (gris). Pour plus de clarté, la zone spectrale représentée est réduite à la fenêtre allant de 800 à 1800 cm^{-1}.

une maladie ou une concentration en composé à partir du spectre, par exemple. Nous supposons que seule une partie du spectre est informative pour le problème étudié, et nous souhaitons donc connaître les portions du spectre à retenir pour l'étude.

Chaque spectre est constitué d'un ensemble de valeurs d'absorbance associées à des nombres d'ondes. La valeur d'absorbance à chaque nombre d'onde représente une variable, et un spectre est donc constitué de plusieurs centaines de variables indexées et ordonnées par nombres d'ondes. Nous nous trouvons dans le cas où le nombre de variables mesurées est du même ordre de grandeur que le nombre d'observations, voire supérieur. En effet, chaque spectre est constitué de plusieurs centaines de variables, en général plus que d'individus mesurés. Si on considère les mesures de façon discrète, alors on se place dans le cadre de l'analyse de données multivariées, nécessitant l'utilisation de méthodes statistiques permettant de gérer la grande dimension.

Il est aussi possible de considérer la nature fonctionnelle des données, car chaque échantillon mesuré correspond à une courbe. Dans ce cas, on souhaite prendre en compte la forme générale des spectres plutôt que des mesures discrètes. D'un point de vue statistique, on travaille alors sur des données en dimension infinie.

Nous utilisons donc des méthodes d'analyse incluant une étape de réduction de la dimension pour construire un modèle de prédiction. Il existe plusieurs manières de réduire la dimension. On peut citer d'une part la sélection des variables les plus pertinentes pour le problème étudié, et donc le travail sur un sous-ensemble des variables disponibles, et d'autre part, la projection des données sur un espace de dimension plus petit et le travail sur des représentations factorielles. Lorsque l'objectif de l'analyse est de construire un modèle de diagnostic, les méthodes d'analyse de données de spectrométrie sont combinées à l'utilisation d'un modèle de prédiction d'une variable binaire représentant la présence ou l'absence d'une maladie. La régression logistique est une des méthodes les plus utilisées pour la prédiction d'une variable binaire, et est décrite dans cette section.

1.3.2 Régression logistique

Modèle

Soit Y une variable réponse binaire à valeurs dans $\{0, 1\}$, représentant par exemple la présence ($Y = 1$) ou l'absence ($Y = 0$) d'une maladie, et $\mathbf{X} \in \mathbb{R}^p$ un ensemble de p covariables. Y suit une loi de Bernoulli de paramètre $p(\mathbf{x})$, avec $\mathbf{x}$ une réalisation de $\mathbf{X}$ et $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ la probabilité *a posteriori* que la variable $\mathbf{Y}$ prenne la valeur 1, sachant l'observation $\mathbf{x}$:

$$Y|\mathbf{X} = \mathbf{x} \sim \mathcal{B}\big(p(\mathbf{x})\big).$$

Les prédicteurs sont liés à la variable réponse Y grâce à la fonction de lien logistique

$$\text{logit}\big(p(\mathbf{x})\big) = \mathbf{x}^T\boldsymbol{\beta},$$

où logit : $x \mapsto \log(\frac{x}{1-x})$ et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients de régression. On peut réécrire ce modèle sous la forme suivante :

$$p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T\boldsymbol{\beta}}}.$$

Estimation

L'estimation des paramètres du modèle logistique est réalisée avec l'approche du maximum de vraisemblance. Pour un échantillon de n réalisations indépendantes $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ de $(\mathbf{X}, Y)$, si on note $\mathbb{P}(Y_i = y_i)$ la probabilité de Y_i, on peut alors écrire la vraisemblance du modèle logistique :

$$\mathcal{L} = \prod_i \mathbb{P}(Y_i = y_i).$$

La fonction de log-vraisemblance du modèle logistique s'écrit alors :

$$\ln \mathcal{L}_{\boldsymbol{\beta}} = \sum_i \left(y_i \mathbf{x}_i^T\boldsymbol{\beta} - log(1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta}))\right).$$

On cherche le paramètre $\boldsymbol{\beta}$ permettant de maximiser cette log-vraisemblance, ce qui peut notamment se faire en annulant son gradient :

$$\nabla \ln \mathcal{L}_{\boldsymbol{\beta}} = \left(\frac{\partial \ln \mathcal{L}_{\boldsymbol{\beta}}}{\partial \beta_1}, \ldots, \frac{\partial \ln \mathcal{L}_{\boldsymbol{\beta}}}{\partial \beta_p} \right).$$

On peut montrer que

$$\nabla \ln \mathcal{L}_{\boldsymbol{\beta}} = \sum_i [\mathbf{x}_i (y_i - p(\mathbf{x}_i))].$$

S'il existe, l'estimateur du maximum de vraisemblance est solution de l'équation

$$\nabla \ln \mathcal{L}_{\boldsymbol{\beta}} = 0.$$

Il n'y a pas de solution explicite, et l'estimation est basée sur des méthodes itératives qui convergent vers la solution. La méthode la plus utilisée en pratique est l'algorithme de Newton-Raphson.

La méthode d'estimation des paramètres du modèle logistique peut être combinée à de la régularisation (Lasso par exemple), ce qui permet d'effectuer une sélection de variables.

Critères d'évaluation des performances de prédiction

Il existe des outils classiques d'évaluation des performances de prédiction d'une variable binaire. Nous nous appuierons notamment sur les différents critères qui sont couramment utilisés dans les publications médicales, ce qui permet une comparaison facile de nos résultats avec les autres méthodes de la littérature :

- La sensibilité est la probabilité que le test soit positif si la maladie est présente, et est définie par $Se = VP/(VP + FN)$, avec VP le nombre d'individus correctement prédits comme malades et FN le nombre d'individus prédits comme non malades à tort. Ce critère mesure la capacité du classifieur binaire à détecter les individus malades, et on cherche à le faire tendre vers 1.

- La spécificité est la probabilité d'obtenir un test négatif chez les non-malades, et est définie par $Sp = VN/(VN + FP)$, avec VN le nombre d'individus correctement prédits comme non malades et FP le nombre d'individus prédits comme malades à tort. Ce critère mesure la capacité d'un classifieur binaire à éliminer les individus non malades, et on cherche à le faire tendre vers 1.

- L'aire sous la courbe ROC (pour Receiver Operating Characteristic), notée par la suite AUROC, est la courbe qui donne le taux de vrais positifs (Sensibilité) en fonction du taux de faux positifs (1-Spécificité), et correspond à une mesure globale de la capacité de diagnostic d'un classifieur binaire à tout seuil de discrimination. On cherche à faire tendre ce critère vers 1.

- La valeur prédictive positive est la probabilité que la maladie soit présente lorsque le test est positif, et est définie par $VPP = VP/(VP + FP)$. Ce critère mesure la qualité de diagnostic d'un test, et on cherche à le faire tendre vers 1.

- La valeur prédictive négative est la probabilité que la maladie ne soit pas présente lorsque le test est négatif, et est définie par $VPN = VN/(VN + FN)$. Ce critère mesure la qualité de dépistage d'un test, et on cherche à le faire tendre vers 1.

- Le taux de bon classement représente le pourcentage d'individus correctement classés comme malades et comme non malades. Il est défini par $TBC = (VP + VN)/n$, avec n le nombre d'individus classés par le classifieur.

Dans le cas de la NASH, nous souhaitons détecter les individus malades, et donc nous cherchons à maximiser la sensibilité des modèles de diagnostic, tout en conservant une bonne spécificité pour éviter un nombre trop important de faux positifs.

1.3.3 Méthodes usuelles en chimiométrie

Les méthodes d'analyse statistique habituellement utilisées en chimiométrie sont basées sur des techniques de projection sur des composantes principales comme l'analyse en composante principale (ACP) et la méthode des moindres carrés partiels (PLS) (Frank and Friedman, 1993).

La régression en composantes principales (PCR) consiste à effectuer une ACP sur les covariables, puis à utiliser les premières composantes principales obtenues comme régresseurs dans un modèle de régression, qui peut être un modèle de régression logistique dans le cadre de la prédiction d'un diagnostic (Jolliffe, 1982; Cowe and McNicol, 1985). Les composantes principales de l'ACP sont construites uniquement sur les covariables, et ne prennent pas en compte la variable réponse.

L'analyse discriminante par les moindres carrés partiels consiste en une régression PLS classique où on considère une variable catégorielle. Les composantes de la PLS sont construites à la fois pour décrire l'ensemble des covariables en maximisant leur variance, et pour maximiser la corrélation entre les covariables et la réponse (Frank and Friedman, 1993).

En pratique, ces méthodes ont de bonnes performances de prédiction sur les données de spectrométrie (Biancolillo and Marini, 2018). Cependant, un des inconvénients de ces méthodes est le fait de travailler sur des données transformées, puisque projetées sur des nouvelles composantes. On ne travaille alors plus sur les variables brutes et les modèles obtenus sont plus difficilement interprétables. En effet, il est plus difficile de localiser les zones du spectre importantes pour la prédiction, qui sont reliées à des types de molécules pouvant permettre d'interpréter biologiquement le modèle. Dans les méthodes développées dans cette thèse, nous choisissons donc de travailler sur les variables des données non transformées.

1.3.4 Sélection de variables

L'identification des molécules présentes dans un échantillon nécessite le travail sur des variables spécifiques du spectre liées à des fonctions biologiques. De plus, le type d'échantillon mesuré pouvant être très complexe, une multitude d'informations est représentée sous forme d'empreinte moléculaire par le spectre. Lorsqu'à partir de cette courbe d'absorbance,

on souhaite construire un modèle de prédiction, il y a potentiellement une grande partie de la courbe qui ne permet pas de discriminer les patients sains des patients malades. La figure 1.6 met en évidence des différences importantes entre les patients malades et les patients témoins, principalement dans la zone de fréquence allant de 1300 à 1450 cm^{-1}. Ces zones peuvent être utilisées pour la construction d'un modèle de diagnostic du cancer de la vessie. L'objectif est alors de déterminer les parties de la courbe qui sont pertinentes pour prédire la variable considérée. Pour cela, et dans le cadre de la grande dimension en général, il existe de nombreuses méthodes de sélection de variables.

Certaines approches consistent à pré-sélectionner un sous-ensemble de variables, en utilisant des méthodes de sélection de variables comme l'algorithme génétique ou les forêts aléatoires, qui peut ensuite être utilisé dans la construction d'un modèle avec les méthodes classiques comme la régression en composantes principales (PCR) ou l'analyse discriminante par les moindres carrés partiels (PLS-DA) (voir par exemple Menze et al., 2007).

Intérêt de la sélection de variables

La sélection de variables a de nombreux intérêts en modélisation. Tout d'abord, elle permet le travail sur un sous-ensemble de variables, ce qui permet l'estimation des paramètres du modèle lorsque le nombre d'individus est faible. Cela permet aussi une estimation plus rapide du modèle puisque le nombre de paramètres est réduit. De plus, le travail sur les variables les plus pertinentes pour la prédiction de la réponse permet d'éviter le surapprentissage, et d'avoir un modèle plus performant si les bonnes variables sont sélectionnées. Enfin, le modèle construit sur un sous-ensemble de variables est moins complexe et plus interprétable, ce qui permet de mieux comprendre le problème étudié (James et al., 2013).

Méthodes de sélection de variables

Il existe trois types de méthodes de sélection de variables (Guyon and Elisseeff, 2006) :

- les méthodes "filter" : un score, basé par exemple sur la corrélation ou des p-valeurs obtenues à la suite de tests d'hypothèse, est calculé pour chaque variable, ce qui permet un classement de l'ensemble des variables et la sélection des variables à meilleur score. Même si ces méthodes sont peu coûteuses en temps de calculs, les relations entre variables ne sont pas considérées, et ces méthodes peuvent donc entraîner la sélection de variables redondantes.

- les méthodes "wrapper" : des sous-ensembles de variables sont évalués et comparés par rapport aux performances de prédiction obtenues avec le modèle statistique estimé pour chaque sous-ensemble. Ces méthodes peuvent être coûteuses en temps de calcul.

- les méthodes "embedded" : la sélection de variables est intégrée dans le processus d'apprentissage du modèle. Les méthodes les plus communes sont basées sur la régularisation.

Les méthodes de sélection de variables utilisées dans cette thèse sont les suivantes :

- la méthode Lasso (pour Least Absolute Shrinkage and Selection Operator) introduite par Tibshirani (1994) : lors de l'étape d'estimation, la vraisemblance est pénalisée par une pénalité en norme ℓ_1, ce qui entraîne une annulation des coefficients associés aux variables qui apportent peu d'information. Un terme de régularisation est ajouté à la vraisemblance à maximiser et le problème d'optimisation à résoudre est alors :

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \left(\ln \mathcal{L}_{\boldsymbol{\beta}} - \lambda \sum_{j}^{p} |\beta_j| \right),$$

  avec λ le paramètre de régularisation à fixer et $\sum_j^p |\beta_j|$ la norme ℓ_1 des coefficients.

- le Fused-Lasso (Tibshirani et al., 2005) : cette méthode dérivée du Lasso introduit un second terme de pénalisation prenant en compte l'ordre des variables, et forçant les coefficients associés à des variables voisines à être égaux. Le problème à résoudre est alors :

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \left(\ln \mathcal{L}_{\boldsymbol{\beta}} - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right),$$

  avec λ_1 et λ_2 les paramètres de régularisation à fixer par l'utilisateur. Cette méthode permet une sélection de variables par blocs.

- les forêts aléatoires : l'algorithme des forêts aléatoires (Breiman, 2001) consiste à combiner des arbres de décisions. Lors de l'apprentissage, il est facile de calculer un score d'importance des variables (Genuer et al., 2010). Ce score permet d'ordonner les variables selon leur impact sur la prédiction et donc de sélectionner un sous-ensemble de variables.

Stabilité de la sélection

Quand on traite des données observées sur une population complexe et hétérogène, on peut faire face à un manque de stabilité dans la sélection des variables, c'est-à-dire qu'une légère modification de l'échantillon des observations étudié entraîne une sélection de variables différente. Dans le cas de dépendance forte entre les prédicteurs, ce qui est le cas en spectrométrie où les variables proches sont très corrélées, les conditions permettant la consistance de la sélection de variables par Lasso ne sont pas respectées et ses propriétés sont affectées (Fan and Li, 2001). La capacité à déterminer le support du vecteur des régresseurs, correspondant à l'ensemble des indices des coordonnées non nulles, est notamment diminuée (Van De Geer, 2010).

Pour faire face au problème de stabilité de la sélection de variables, les méthodes d'ensemble peuvent être utilisées. Ces méthodes consistent à perturber plusieurs fois les données, à appliquer une procédure de sélection sur les données perturbées, puis à agréger

l'ensemble des résultats obtenus. Par exemple, ce type de procédure a été suggéré pour la sélection de variables par des forêts aléatoires ou par régression régularisée de type Lasso. Classiquement, les méthodes d'ensemble pour la sélection de variables ré-échantillonnent plutôt les individus que les variables, mais il est aussi possible de ré-échantillonner les variables, ce qui permet notamment de gérer la grande dimension. Meinshausen and Bühlmann (2010) ont décrit une procédure de sélection stable très utilisée depuis. On peut la résumer de la façon suivante, lorsque l'on travaille sur un échantillon de calibration d'un jeu de données fixé, avec n individus et p variables :

- Répéter B fois les deux étapes suivantes :
  - Tirage aléatoire d'un sous-échantillon parmi les individus.
  - Sur ce sous-échantillon, une procédure de sélection (de type Lasso par exemple) est effectuée et permet l'obtention d'un ensemble de variables, sélectionnées pour ce sous-échantillon. Avec certaines méthodes de sélection, comme les forêts aléatoires, on fixe le nombre de variables à sélectionner.
- Un ensemble de variables sélectionnées est obtenu pour chaque répétition. On a donc, pour chaque variable du jeu de données, le nombre de sélection sur les B itérations.
- Pour chaque variable, la probabilité de sélection est calculée selon $ps = \frac{\text{Nombre de sélections}}{\text{Nombre d'itérations } B}$. Les variables dont la valeur de ps dépasse un certain seuil fixé sont conservées.

Cette procédure de sélection nécessite de fixer à l'avance la taille des sous-échantillons, le nombre d'itérations de la procédure B ainsi que le seuil de sélection de variable s. Ces paramètres peuvent être fixés de façon théorique, selon les outils détaillés dans Meinshausen and Bühlmann (2010).

On obtient à l'issue de cette procédure un ensemble de variables d'intérêt, sur lesquelles il est possible de construire un modèle de prédiction. Ensuite, on peut valider les résultats avec un échantillon de validation, qui n'aura jamais servi à la procédure de sélection de variables.

Les méthodes décrites précédemment sont appliquées à l'analyse de données concernant des patients atteints de NASH, ce qui permet de mieux comprendre les données ainsi que la maladie étudiée, et de poser les problématiques étudiées dans cette thèse.

1.3.5 Analyses des données spectrométriques NASH-Nice

Les méthodes d'analyse détaillées en sections 1.3.3 et 1.3.4 sont comparées sur les données NASH Nice.

Nous comparons les performances de modèles établis sur des sélections de variables faites par Fused-Lasso et par procédure de sélection stable utilisant le Lasso et les forêts

	Lasso	Forêts aléatoires	Fused-Lasso	PCR	PLS-DA
AUROC	0.61	0.57	0.61	0.7	0.57
Sensibilité	0.46	0.85	0.62	0.85	0.15
Spécificité	0.92	0.39	0.64	0.47	0.98
VPP	0.54	0.2	0.25	0.24	0.67
VPN	0.9	0.93	0.89	0.94	0.86
TBC	0.85	0.47	0.63	0.43	0.85

TABLE 1.3 : **Comparaison des performances de prédiction obtenues avec différentes méthodes sur les données NASH Nice.** Les premières colonnes correspondent respectivement aux performances des modèles logistiques construits sur les variables sélectionnées par la procédure de sélection stable utilisant le Lasso et les forêts aléatoires. La troisième colonne correspond aux performances du modèle logistique construit sur les variables sélectionnées par Fused-Lasso. Les quatrième et cinquième colonnes correspondent aux performances obtenues par PCR et PLSDA.

aléatoires. Nous comparons aussi les performances issues de ces méthodes de sélection de variables aux performances obtenues avec des méthodes de projection de données classiquement utilisées en spectrométrie, comme l'analyse discriminantes par les moindres carrés partiels (PLS-DA), ainsi que la régression en composantes principales (PCR). Pour cela, les données sont séparées de façon aléatoire en échantillons d'apprentissage et de validation. La sélection de variables et la construction du modèle sont effectuées sur l'échantillon d'apprentissage. Le nombre de composantes pour les méthodes de projection ainsi que les paramètres de régularisation du Fused-Lasso sont choisis par validation croisée sur l'échantillon d'apprentissage, avec pour objectif de maximiser l'AUROC. L'évaluation des performances de chaque méthode est effectuée sur l'échantillon de validation. Les outils d'évaluation des performances d'un classifieur binaire présentés précédemment sont utilisés, et présentés en table 1.3.

La valeur d'AUROC obtenue pour la méthode PCR est correcte, mais cette méthode ne permet pas d'obtenir un taux de bon classement élevé, et la spécificité est faible, ce qui souligne une mauvaise capacité à détecter les individus non malades. Concernant la PLS-DA, la performance globale selon l'AUROC est mauvaise. Cette méthode a une spécificité élevée mais une sensibilité très faible, et classe trop d'individus comme non malades à tort. Le taux de bon classement élevé provient du fait que la proportion de non malades dans la cohorte est élevé, et pour cette méthode qui détecte bien les non malades, le nombre d'individus bien classés est élevé, malgré des mauvaises performances de diagnostic. Les mêmes conclusions sont obtenues avec le modèle logistique construit sur les variables sélectionnées par la procédure de sélection stable utilisant la régression Lasso, avec un modèle menant à trop de faux négatifs. Le modèle logistique construit avec les variables sélectionnées par procédure de sélection stable utilisant les forêts aléatoires entraîne des mauvaises performances globales, avec une valeur d'AUROC, une spécificité et un taux de bon classement faibles.

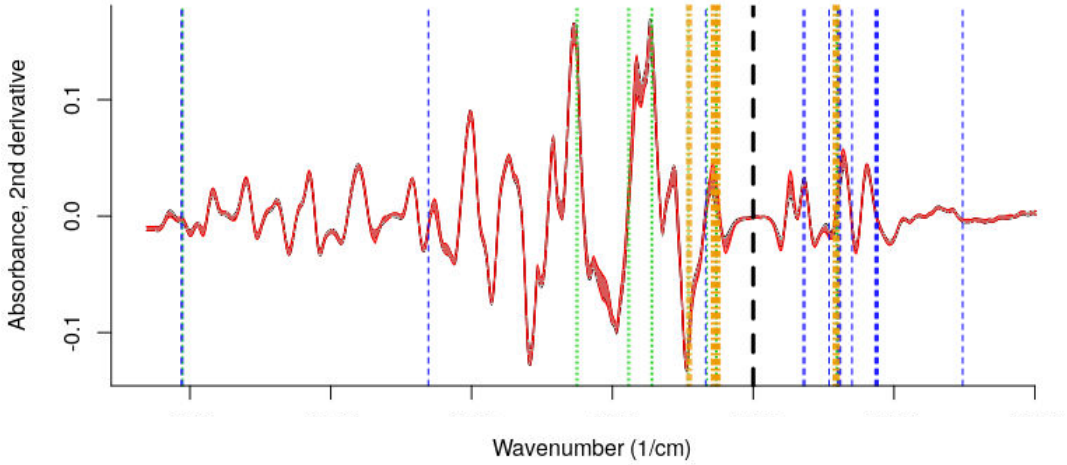

FIGURE 1.8 : **Variables sélectionnées par différentes méthodes de sélection de variables sur les données spectrométriques NASH.** Les deuxième et troisième quartiles des dérivées secondes de spectres mesurés sur des sérums de patients atteints de NASH (rouge) et de patients non malades (gris) sont représentés sur la fenêtre spectrale allant de 800 cm^{-1} à 3200cm^{-1}. La coupure de la fenêtre spectrale allant de 1800 cm^{-1} à 2800cm^{-1} est représentée par la ligne pointillée noire. Les variables sélectionnées sont représentées par les lignes pointillées colorées : la procédure de sélection stable utilisant le Lasso est représentée en vert, la procédure de sélection stable utilisant les forêts aléatoires en bleu et le Fused-Lasso en orange.

Les variables sélectionnées par les procédures de sélection stable ainsi que par le Fused-Lasso sont représentées sur la figure 1.8. Il est important de noter que ces méthodes ne sélectionnent pas les mêmes variables pour la prédiction de la variable réponse.

Finalement, une des conclusions tirées de ces analyses, ainsi que d'analyses menées sur d'autres cohortes de patients atteints de NASH, est l'instabilité des résultats. Les performances de prédiction sont très variables, et chutent lorsque des nouvelles observations provenant de cohortes externes sont soumises aux modèles. On se trouve alors dans le cas du surapprentissage, ce qui peut être lié à la variabilité de la sélection de variables. Cela souligne la complexité de l'information portée par le spectre, et la difficulté à retrouver les variables pertinentes pour la prédiction. Cela nous pousse à envisager des méthodes permettant de mieux sélectionner les variables ou les zones du spectres pertinentes pour la prédiction.

1.4 Problématiques et objectifs de la thèse

Les analyses menées sur la cohorte de patients NASH-Nice montrent la difficulté à établir un modèle de diagnostic de la NASH performant avec les méthodes statistiques classiques. Plusieurs axes sont mis en avant pour expliquer ces difficultés et y faire face.

Tout d'abord, les premières analyses soulignent la présence de bruit dans le signal. Le bruit correspond notamment au fait que la spectrométrie mesure un échantillon complexe, constitué de nombreuses molécules en interaction qui n'entrent pas forcément en jeu dans le diagnostic de la maladie, mais qui peuvent masquer sa signature. Des zones non informatives sont donc à éliminer du spectre, pour pouvoir mieux prédire la maladie. Cela pourrait permettre de limiter le surapprentissage observé dans les analyses précédentes. **Il serait donc pertinent d'effectuer un prétraitement des données permettant d'éliminer les zones du spectre non liées à la maladie.**

De plus, on remarque une hétérogénéité de la population atteinte de NASH qui repose notamment sur la diversité des patients étudiés. Différentes analyses préliminaires menées sur des données mesurées sur des cohortes de patients NASH, ainsi que les discussions avec les experts (biologistes et médecins) suggèrent la présence de groupes de patients méconnus, que l'on n'arrive *a priori* pas à caractériser. Ces groupes de patients présentent des caractéristiques morphologiques, métaboliques et cliniques différentes, qu'il faut prendre en compte dans l'analyse des données. Les patients issus de ces groupes, indépendamment du diagnostic, vont évoluer de façon plus ou moins rapide vers un état grave et/ou ont un diagnostic plus ou moins facile à établir. De plus, la maladie étudiée est complexe et évolutive, des patients peuvent donc être à des stades intermédiaires d'évolution de la maladie, et plus difficiles à diagnostiquer. **Nous souhaitons donc considérer une structure latente, c'est-à-dire non observée, de groupes au sein des données lors de la construction du modèle de diagnostic.**

De plus, de nombreux processus moléculaires liés à l'évolution de la maladie sont mal connus, et il y a un réel besoin de mieux comprendre la maladie et son évolution, ainsi que les caractéristiques des patients atteints. Pour cela, les médecins souhaitent comprendre les modèles statistiques mis en place pour établir le diagnostic de la maladie. Les méthodes de type "boîte noire" sont donc à proscrire au maximum, et les analyses statistiques doivent fournir des outils performants en termes de prédiction mais aussi en termes d'interprétabilité et de compréhension de la maladie. Nous souhaitons aussi inclure les analyses dans un contexte global où la décision ne dépend pas uniquement d'un modèle, mais fait partie d'un ensemble de critères évalués sur des informations extérieures (comme des informations obtenues par questionnaire par exemple). Il serait intéressant de pouvoir mettre en avant les zones du spectre permettant d'établir le diagnostic de la maladie, et de les lier à des types de molécules pour pouvoir interpréter biologiquement le modèle de diagnostic. **Nous nous concentrons sur l'étude des interactions entre zones du spectre pour mettre en évidence d'éventuelles dépendances conditionnelles.** Ce travail peut se faire grâce à l'étude des matrices de covariance estimées sur les données.

L'objectif est de construire un modèle de prédiction de la NASH prenant en compte la complexité des données et les différents groupes de patients, tout en sélectionnant l'information pertinente dans les données pour permettre une interprétation biologique des résultats.

Chapitre 2

Modélisation de classes latentes

Contents

Le métabolisme d'une personne peut dépendre fortement de son mode de vie, de ses habitudes alimentaires et de ses antécédents médicaux globaux. Contraindre un modèle unique à s'ajuster à une cohorte de patients, lorsque l'on s'intéresse à une maladie complexe

liée à un dérèglement du métabolisme, ne paraît donc pas adapté dans cette situation. Les analyses réalisées sur les cohortes de patients NASH, ainsi que les différentes études sur cette maladie suggèrent que l'hétérogénéité des patients rend difficile l'établissement de modèles de diagnostic. Une approche de plus en plus utilisée est de décomposer la cohorte en plusieurs profils de référence résumant au mieux les comportements métaboliques. On retrouve cette approche sous le terme de "disease trajectories" (Ross and Dy, 2013), ce qui fournit aux experts des informations interprétables sur l'état d'un patient, et peut faciliter l'établissement du diagnostic. Ces différents groupes peuvent être liés à des caractéristiques génétiques, métaboliques, morphologiques, ou à une combinaison de ces caractéristiques, et peuvent être très difficiles à décrire *a priori*. De plus, il est difficile de connaître le nombre de groupes ainsi que la partition structurant les données. Il est alors important d'utiliser des méthodes statistiques permettant de regrouper les patients similaires, que l'on suppose provenir d'une même sous-population. Cela pourrait permettre d'obtenir de meilleures performances de diagnostic de certaines maladies complexes. La présence de différentes typologies de patients oriente donc cette étude vers des modèles permettant de trouver des groupes latents, c'est-à-dire non observés, parmi les observations, tout en permettant la prédiction de la maladie.

En classification, on cherche à partitionner les données en plusieurs groupes, tels que les individus au sein d'un même groupe se ressemblent plus que les individus venant de groupes différents. Dans le cas de la classification non supervisée, aucune information sur la classe n'est disponible *a priori*. On la distingue de la classification supervisée, où on dispose d'une information sur les classes et où on cherche à comprendre leur structure et à construire une règle de prédiction permettant de classer de nouveaux individus. Dans le cadre non supervisé, la structure de classe est supposée et le résultat de la classification dépend de la méthode employée et des modèles utilisés (loi des fonctions de distribution composantes du mélange, métrique utilisée ou méthode d'estimation utilisée par exemple). On ne connaît pas la partition théorique, il est donc difficile d'évaluer la qualité de la classification, et de comparer les méthodes de partitionnement. Les méthodes de classification non supervisée permettent d'approcher la structure affectant des données, et de mieux comprendre les comportements des patients.

Les modèles de mélange ont été introduits il y a plus d'une centaine d'années par Newcomb (1886) et Pearson (1894), et permettent d'estimer la structure en groupe qui affecte des données. Ces modèles permettent d'effectuer une classification non supervisée grâce à une approche probabiliste en considérant les labels inconnus des observations comme des variables latentes. On parle dans ce cas de partitionnement par modèles génératifs, retrouvé dans la littérature sous le nom de "model-based clustering". Ils sont devenus très populaires dans divers domaines statistiques, puisqu'ils sont très flexibles et facilement extensibles. L'objectif des modèles de mélange est de décrire la distribution avec une méthode semi-paramétrique en prenant en compte l'hétérogénéité non observée présente au sein des données étudiées. On se trouve dans le cadre de la classification **non supervisée**.

On distingue plusieurs approches dans les modèles de mélange. Tout d'abord, le modèle de mélange peut être construit uniquement sur un ensemble de variables et a dans

ce cas pour but d'évaluer leur structure. Le mélange de régressions est construit sur la loi conditionnelle d'une variable réponse sachant les covariables et a pour but d'évaluer la structure de groupe affectant les données et d'estimer les modèles de régression par groupe de façon simultanée. Les modèles de mélange permettent d'estimer la structure latente existant au sein des données mais sont peu abordés dans le cadre de la prédiction. On retrouve cet objectif dans le cadre des mélanges d'experts qui consistent à construire un modèle de prédiction à l'aide de variables structurées en groupes latents. Ces types de modèles sont détaillés par la suite, ainsi que leurs différences, en termes de spécification de modèles ainsi qu'en termes d'utilisation et de contexte. Dans ce chapitre, nous présentons en détail la construction des modèles de mélange dans le cadre d'un design aléatoire. Ce cadre, peu développé dans la littérature, permet de construire des outils de prédiction basés sur des mélanges de régressions.

2.1 Modèles de mélange

Dans le cadre des modèles de mélange, le partitionnement est formulé comme un problème d'estimation des paramètres d'un mélange de distribution multivariée. Nous commençons par expliciter les différentes notions évoquées dans cette thèse.

Une composante correspond à une distribution de probabilité faisant partie du modèle de mélange. Un groupe correspond à un ensemble d'observations plus similaires les unes des autres par rapport aux observations d'un autre groupe, selon la mesure de distance choisie. Il n'existe pas de vraie définition pour désigner un groupe. On parlera de groupe dans le cas de la classification non supervisée, où l'on cherche à rassembler les individus proches, sans information *a priori*. Finalement, on utilisera la notion de classe pour faire référence à une variable de label associée aux observations, lorsque l'on se place dans le cadre supervisé. Dans ce cas, l'objectif est de prédire cette variable pour des nouvelles observations. La classe peut par exemple correspondre à un diagnostic, et dans ce cas est une variable binaire associée à la présence ou à l'absence d'une maladie.

2.1.1 Définition du modèle

On considère $\mathbf{X} \in \mathbb{R}^p$ un ensemble de p variables aléatoires. Dans le cadre des modèles de mélange, on suppose que les individus proviennent de différentes sous-populations, et sont répartis en K classes latentes de proportions $(\pi_k)_{k=1,...,K}$ avec $0 < \pi_k$ et $\sum_k \pi_k = 1$. Les paramètres qui définissent les variables $\mathbf{X}$ sont notés $\mathbf{\Phi}_k$ pour $k = 1, \ldots, K$. On note $\mathbf{\Phi}$ le paramètre global du mélange : $\mathbf{\Phi} = (\pi_1, ..., \pi_K, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_K)$. On cherche à déterminer la distribution de $\mathbf{X}$ comme un mélange de distributions. La loi du modèle de mélange est caractérisée par sa densité :

$$f(\mathbf{x}; \mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}; \mathbf{\Phi}_k),$$

avec $\mathbf{x}$ une réalisation de $\mathbf{X}$ et $f(\mathbf{x}; \mathbf{\Phi}_k) = f_k(\mathbf{x})$ la densité de $\mathbf{x}$ paramétrée par $\mathbf{\Phi}_k$ pour chaque groupe $k = 1, \ldots, K$. Cette loi peut être une loi de probabilité usuelle comme une

loi normale multivariée ou multinomiale par exemple (Everitt, 1984).

On peut caractériser le modèle de mélange en considérant une variable aléatoire latente $\mathbf{Z}$. La variable latente $\mathbf{Z}$ représente la variable de groupe, distribuée selon une loi multinomiale de probabilités $(\pi_k)_{k=1,...,K}$:

$$\mathbf{Z} \sim \mathcal{M}(\pi_1, ..., \pi_K).$$

Si on suppose que $\mathbf{X}|\mathbf{Z} = k \sim f_k$, alors la densité jointe de $\mathbf{X}$ et $\mathbf{Z}$ peut s'écrire :

$$f(\mathbf{x}, z; \mathbf{\Phi}) = \prod_{k=1}^{K} \pi_k f_k(\mathbf{x}),$$

avec z une réalisation de $\mathbf{Z}$.

Dans le cas des modèles de mélange, on souhaite connaître la structure de groupes affectant $\mathbf{X}$ et on s'intéresse donc au lien entre $\mathbf{X}$ et $\mathbf{Z}$. Quand $\mathbf{Z}$ est connue, on se place dans le cadre de l'analyse discriminante, où le problème est de prédire un vecteur indicateur z_{n+1} pour une nouvelle observation $\mathbf{x}_{n+1}$. Lorsque z est inconnue, on est dans le contexte de l'estimation de densité, ou de partitionnement si l'estimation des z_i est l'objectif premier. Dans ce cas, le vecteur des paramètres à estimer est $\mathbf{\Phi} = (\pi_1, ..., \pi_K, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_K)$, avec $\mathbf{\Phi}_k$ le paramètre de la loi de $\mathbf{X}$ dans la classe k.

2.1.2 Estimation par maximum de vraisemblance

On cherche à estimer le paramètre $\mathbf{\Phi}$ à partir de l'échantillon $\mathbf{x}$. Pour cela, on cherche habituellement le paramètre qui maximise la log-vraisemblance calculée sur les données observées $\mathbf{x}$, appelée la log-vraisemblance observée, définie par :

$$\ln \mathcal{L}(\mathbf{x}; \mathbf{\Phi}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \mathbf{\Phi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \mathbf{\Phi}_k).$$

On souhaite donc résoudre :

$$\hat{\mathbf{\Phi}} = \text{argmax}_{\mathbf{\Phi}} \ln \mathcal{L}(\mathbf{x}; \mathbf{\Phi}). \tag{2.1}$$

Cependant, il n'y a pas de solution analytique à ce problème, et un algorithme Espérance-Maximisation, noté par la suite algorithme EM, est utilisé pour optimiser (2.1).

L'algorithme EM proposé par Dempster et al. (1977) est spécifiquement utilisé pour approcher les estimateurs de maximum de vraisemblance dans les problèmes de données incomplètes. Dans le cas des modèles de mélange, les données observées $(\mathbf{x}_i)_{i=1,...,n}$ peuvent être vues comme des données incomplètes, en considérant que l'information latente de groupe est manquante, pour tout $i = 1, \ldots, n$. On note $(\mathbf{x}_i, \mathbf{z}_i)$ les données augmentées ou complétées, avec $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iK})$ référant aux données non observées, avec $z_{ik} = 1$ si l'observation i appartient au groupe k, et 0 sinon. En présence de groupes (non observés), il est commun de considérer la vraisemblance des données complétées, qui grâce à la règle de Bayes peut se décomposer en

$$\ln \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \mathbf{\Phi}) = \ln \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathbf{z}_1, \ldots, \mathbf{z}_n) + \ln \mathcal{L}(\mathbf{z}_1, \ldots, \mathbf{z}_n).$$

L'algorithme EM consiste en la maximisation de l'espérance conditionnelle de la log-vraisemblance complétée (conditionnellement aux données observées), sachant un ensemble de paramètres fixé $\mathbf{\Phi}_K^*$. Le problème à résoudre est alors

$$\arg\max_{\mathbf{\Phi}} \quad \{\mathbb{E}_{\mathbf{\Phi}^*}\left[\ln \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \mathbf{\Phi}) | \mathbf{x}_1, \ldots, \mathbf{x}_n\right]\}.$$

Ainsi, l'algorithme alterne entre deux étapes, appelées l'étape E et l'étape M, jusqu'à convergence. Commençons par décrire les deux étapes dans le contexte du modèle de mélange dont la densité de $\mathbf{X}$ est notée $f_{\mathbf{X}}$.

L'étape E de l'algorithme permet d'estimer les probabilités *a posteriori* d'appartenance aux groupes latents $\tau_{ik} = \mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \mathbf{\Phi}_k^{[h]})$ pour tout individu $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, K$, sachant les paramètres courants $\mathbf{\Phi}_k^{[h]}$ à l'itération $[h]$. Pour tout $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, K$, elles sont données par

$$\tau_{ik}^{[h+1]} = \frac{\pi_k^{[h]} f_{\mathbf{X}}\left(\mathbf{x}_i; \mathbf{\Phi}_k^{[h]}\right)}{\sum_{\ell=1}^{K} \pi_\ell^{[h]} f_{\mathbf{X}}\left(\mathbf{x}_i; \mathbf{\Phi}_\ell^{[h]}\right)}.$$

Le plus souvent, la densité $f_{\mathbf{X}}$ est connue et le calcul de τ_{ik} est explicite.

L'obtention des probabilités *a posteriori* permet de calculer l'espérance de la log-vraisemblance complétée $Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]})$ sachant les données observées $(\mathbf{x}_i)_{i=1,\ldots,n}$ et $\mathbf{\Phi}^{[h]}$,

$$\mathbb{E}\left[\ln \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \mathbf{\Phi}) | \mathbf{\Phi}^{[h]}\right].$$

L'étape M de l'algorithme consiste en la maximisation de l'espérance conditionnelle calculée dans l'étape E, par rapport au paramètre $\mathbf{\Phi}$, pour obtenir $\mathbf{\Phi}^{[h+1]}$, selon

$$\mathbf{\Phi}^{[h+1]} = \arg\max_{\mathbf{\Phi}} \quad \left\{Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]})\right\}.$$

L'algorithme s'arrête lorsque la log-vraisemblance observée se stabilise ou bien après un nombre prédéfini d'itérations. Plus de détails sur les étapes de l'algorithme EM, notamment les formules explicites dans le cadre gaussien multivarié, sont donnés par la suite.

Même si le modèle de mélange ne fournit pas directement de partition, une fois le modèle estimé, une classification floue des données en K groupes peut être obtenue en termes de probabilités *a posteriori* τ_{ik} que le point $\mathbf{x}_i$ appartienne aux composantes $1, \ldots, K$ du mélange.

Règle du Maximum *a posteriori* : À l'issue de l'estimation, chaque point des données peut être affecté au groupe de plus grande probabilité *a posteriori* estimée, suivant le principe du maximum *a posteriori*. On obtient donc le vecteur des labels de groupes pour chaque individu :

$$\hat{z}_{ik} = \begin{cases} 1 \text{ si } k = \arg\max_l \hat{\tau}_{il} \\ 0 \text{ sinon.} \end{cases}$$

L'estimation des probabilités conditionnelles τ_{ik} permet d'évaluer le risque de classement de chaque individu.

Remarque : l'estimation des paramètres d'un modèle de mélange peut se faire grâce à d'autres méthodes que l'approche du maximum de vraisemblance avec l'algorithme EM. Des méthodes bayésiennes peuvent être utilisées avec des algorithmes de Monte-Carlo par Chaînes de Markov, décrites notamment dans Diebolt and Robert (1994) et Richardson and Green (1997).

2.1.3 Exemple gaussien multivarié

On définit un modèle de mélange gaussien dans le cas où la densité de chaque composante est gaussienne. On a alors :

$$\mathbf{X}|\{Z_k = 1\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k),$$

avec $\boldsymbol{\mu}_k$ et Σ_k la moyenne et la matrice de covariance associées à la composante k, et :

$$f(\mathbf{x}, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k),$$

avec

$$f(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) = f(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{|2\pi\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)\Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

L'algorithm EM se décompose alors selon les deux étapes suivantes :

- À l'étape E, sachant un paramètre $\boldsymbol{\Phi}^*$, on a :

$$\begin{aligned} Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^*) &= E(\ln \mathcal{L}(\boldsymbol{\Phi})|\mathbf{x}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \log f(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k), \end{aligned}$$

  où $\tau_{ik} = P(Z_{ik} = 1|\mathbf{x})$ est la probabilité *a posteriori* que $\mathbf{x}_i$ appartienne au groupe k :

$$\tau_{ik} = \frac{\pi_k f_{\mathbf{X}}(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l f_{\mathbf{X}}(\mathbf{x}_i; \boldsymbol{\mu}_l, \Sigma_l)}.$$

- À l'étape M, les paramètres sont mis à jour selon :

$$\pi_k = \frac{1}{n}\sum_{k=1}^{K} \tau_{ik},$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} \tau_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} \tau_{ik}},$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} \tau_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)'}{\sum_{i=1}^{n} \tau_{ik}}.$$

2.2 Mélange de régressions à design fixe

2.2.1 Définition du modèle

Dans le cadre des mélanges de régressions, une variable réponse est prise en compte dans la modélisation des données (voir par exemple McLachlan and Peel, 2000; Grün and Leisch, 2007). On considère les variables $(\mathbf{x}, Y)$ avec Y la variable à expliquer et $\mathbf{x} \in \mathbb{R}^p$ l'ensemble des p covariables.

Si on fait l'hypothèse d'une population homogène, la régression linéaire multiple permet de modéliser la relation entre Y et $\mathbf{x}$:

$$Y = \boldsymbol{\beta}\mathbf{x} + \epsilon,$$

avec $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ le vecteur des coefficients de régression inconnus et ϵ un terme d'erreur tel que $\epsilon \sim \mathcal{N}(0, \Sigma)$.

Dans certains cas, l'hypothèse que les coefficients de régression sont fixés pour toute réalisation de $Y_1, ..., Y_n$ n'est pas appropriée et on se trouve en présence de groupes non observés. Comme pour les modèles de mélange, on peut supposer qu'il existe une variable latente $\mathbf{Z}$ structurant les données. Cependant, l'objectif ici est aussi de modéliser la dépendance entre Y et $\mathbf{x}$ en se basant sur des données provenant d'une population hétérogène. Le but est donc de prédire Y comme un mélange de régressions. Ce type de modèle a été introduit dans le cadre d'un design fixe, où les covariables $\mathbf{x}$ correspondent à des observations. Dans ce cas, les covariables ne portent pas d'information sur l'appartenance aux groupes.

On considère alors que les n individus dont on dispose des observations $\mathbf{x}$ sont répartis en K classes latentes de proportions $(\pi_k)_{k=1,...,K}$ avec $0 < \pi_k$ et $\sum_k \pi_k = 1$, et que l'appartenance à une classe k influe sur la distribution de $Y|\mathbf{x}$. Dans le cadre d'un design fixe, les observations $\mathbf{x}$ ne sont pas aléatoires, et considérer $Y|\mathbf{x}$ revient à considérer Y. Dans le cas des mélanges de régressions finis d'ordre K, la densité conditionnelle de Y sachant $\mathbf{X}$, correspondant ici à la densité de Y, s'écrit :

$$f(y|\mathbf{x}; \boldsymbol{\Phi}) = f(y; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f(y; \boldsymbol{\beta}_k, \boldsymbol{\Phi}_k),$$

avec y une réalisation de Y, $f(y; \boldsymbol{\Phi}_k)$ sa densité de paramètre $\boldsymbol{\Phi}_k$, pour le groupe k et $\boldsymbol{\Phi} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \pi_1, \ldots, \pi_K, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K)$ le vecteur des paramètres à estimer. Pour chaque groupe k, $\boldsymbol{\beta}_k = (\beta_{1,k}, \ldots, \beta_{p,k})$ est le vecteur des p coefficients de régression

caractérisant le lien entre $\mathbf{x}$ et Y. Selon ces notations, on aurait alors dans le cas linéaire, la relation entre Y et $\mathbf{x}$:

$$Y = \boldsymbol{\beta}_k \mathbf{x} + \epsilon,$$

avec $\epsilon \sim \mathcal{N}(0, \sigma_k^2)$.

La variable réponse peut être de différents types, ce qui nécessite de considérer le cadre des modèles linéaires généralisés et des formes particulières pour la distribution $f(y; \boldsymbol{\Phi}_k)$. On parle alors de mélange de modèles linéaires généralisés.

2.2.2 Les modèles linéaires généralisés

Les modèles linéaires généralisés permettent d'étudier la liaison entre une variable aléatoire dépendante ou réponse Y et un ensemble de prédicteurs $\mathbf{x}$. Ils comprennent notamment la régression linéaire et la régression logistique. La variable réponse aléatoire Y est associée à une loi de probabilité dont les paramètres dépendent des prédicteurs $\mathbf{x}$.

Les mélanges de régressions finis peuvent être basés sur des familles bien connues de distribution pour la densité conditionnelle $f(y|\mathbf{x}; \boldsymbol{\Phi}_k) = f(y; \boldsymbol{\Phi}_k)$. Le lien entre la variable réponse Y et les covariables $\mathbf{x}$ n'est pas obligatoirement linéaire. On peut par exemple avoir une distribution gaussienne pour la variable réponse et travailler avec une fonction de lien linéaire, ou une distribution binomiale et travailler avec une fonction de lien logistique.

On peut considérer les modèles linéaires généralisés dans le cadre du mélange. On a alors, pour une variable réponse Y à valeurs dans un espace $\mathcal{Y} \subseteq \mathbb{R}$, la densité conditionnelle f pour le mélange de régressions :

$$f(y; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f(y; \boldsymbol{\Phi}_k),$$

avec π_k la probabilité *a priori* du groupe k, avec $\pi_k \geq 0$ et $\sum_{k=1}^{K} \pi_k = 1$ et $\boldsymbol{\Phi}_k$ le vecteur de paramètres liés au groupe k, avec $\boldsymbol{\Phi} = (\pi_1, ..., \pi_K, \boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_K)$.

Pour pouvoir prendre en compte différents types de variables réponses, on suppose que pour chaque composante du mélange, la distribution conditionnelle $f(y; \boldsymbol{\Phi}_k)$ appartient à la famille exponentielle. On peut citer par exemple le mélange de régressions logistiques et le mélange de régressions de Poisson.

Réponse binomiale On suppose que la variable réponse Y est à valeurs dans $\mathcal{Y} = \{0, 1\}$ et que la loi de Y est binomiale, paramétrée par $p^{(k)}(\mathbf{x})$:

$$Y \sim \mathcal{B}(p^{(k)}(\mathbf{x})),$$

avec $p^{(k)}(\mathbf{x}) = \mathbb{P}(Y = 1 | Z_k = 1)$. Le lien entre la variable réponse et les prédicteurs est modélisé par la fonction de lien logistique $\text{logit}(p^{(k)}(\mathbf{x})) = \mathbf{x}^T \boldsymbol{\beta}_k$. On a alors :

$$f(y; \boldsymbol{\beta}_k) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_k)}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta}_k)}.$$

Finalement, le modèle de mélange de régressions logistiques s'écrit :

$$f(y;\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_k)}{1+\exp(\mathbf{x}^T\boldsymbol{\beta}_k)}.$$

Réponse de Poisson On suppose que Y est à valeurs dans $\mathcal{Y} = \mathbb{N}$ et que la loi de Y est une loi de Poisson de paramètre $p^{(k)}(\mathbf{x})$:

$$Y \sim \mathcal{P}(p^{(k)}(\mathbf{x})),$$

avec $p^{(k)}(\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}_k)$. On a alors :

$$f(y;\beta_k) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_k)(\mathbf{x}^T\boldsymbol{\beta}_k)^y}{y!}.$$

Finalement, le modèle de mélange de régressions de Poisson s'écrit :

$$f(y;\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_k)(\mathbf{x}^T\boldsymbol{\beta}_k)^y}{y!}.$$

Dans le mélange de régressions décrit précédemment, les covariables $\mathbf{x}$ sont considérées comme déterministes, et leur distribution est invariante selon le groupe. Les covariables ne portent donc pas d'information sur l'appartenance aux groupes latents. Dans ce cas, si l'on dispose d'une nouvelle observation et que l'on souhaite prédire la variable réponse Y pour cette observation, on ne peut pas estimer l'appartenance aux groupes latents pour cet individu. Pour effectuer la prédiction, il est alors possible d'utiliser les modèles de régression estimés pour chaque groupe latent, et de les pondérer par les probabilités *a priori* $\hat{\pi}_k$ estimées. Cependant, Hoshikawa (2013) montre que cette approche ne permet généralement pas d'obtenir de meilleures prédictions qu'un modèle linéaire généralisé classique. De plus, le coût d'estimation des paramètres, au nombre plus important dans le cas du mélange, entraîne une variance plus importante des estimateurs. Le mélange de régressions n'est donc pas un modèle adapté à la prédiction.

2.2.3 Mélange de régressions avec variables concomitantes

Jusqu'à présent, les modèles de mélange présentés comprenaient des pondérations fixées. En effet, les probabilités *a priori* π_k sont indépendantes des données observées $\mathbf{x}$. Dans le cas du mélange de régressions avec variables concomitantes, qui est une extension du mélange de régressions, proposé par Dayton and Macready (1988), les pondérations de chaque composante du mélange sont des fonctions des variables concomitantes, qui peuvent correspondre aux covariables $\mathbf{x}$. On a alors le modèle suivant, selon les mêmes notations que précédemment :

$$p(y;\mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k(\mathbf{x};\boldsymbol{\eta}) f(y;\mathbf{\Phi}_k),$$

avec, pour tout $k = 1, \ldots, K$, $\pi_k(\mathbf{x}; \boldsymbol{\eta})$ les poids du mélange fonctions de $\mathbf{x}$ paramétrées par $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$. Un modèle logistique multinomial modélise les poids du mélange :

$$\pi_k(\mathbf{x}; \boldsymbol{\eta}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\eta}_k)}{\sum_{l=1}^{K} \exp(\mathbf{x}^T \boldsymbol{\eta}_l)}.$$

Le mélange de régressions modélise seulement la distribution de $Y|\mathbf{x}$, qui dans le cadre d'un design fixe revient à considérer la distribution de Y, alors que le mélange de régressions avec variables concomitantes modélise la distribution de Y et un modèle logistique des variables concomitantes. Dans ce cas, les probabilités *a priori* des groupes latents dépendent donc des covariables observées, mais qui sont toujours considérées comme déterministes, et ne sont donc pas modélisées par une distribution de probabilité. Ce modèle n'est donc toujours pas adapté à la prédiction, car l'observation disponible pour un nouvel individu ne permet pas de prédire le groupe latent de cet individu.

2.2.4 Estimation du modèle

L'estimation des paramètres du mélange de régressions peut se faire par maximum de vraisemblance. La fonction de log-vraisemblance du mélange de régressions est donnée par :

$$\ln \mathcal{L}(y, \mathbf{x}; \boldsymbol{\Phi}) = \sum_{i=1}^{n} \log f(y, \boldsymbol{\Phi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(y; \boldsymbol{\Phi}_k).$$

Un algorithme EM peut être utilisé pour l'estimation des paramètres du modèle. Lors de l'étape E, on obtient les probabilités *a posteriori* d'appartenance aux groupes latents pour chaque observation $i = 1, \ldots, n$. Cela permet le calcul de l'espérance conditionnelle de la log-vraisemblance des données complétées, sachant les données observées. L'étape de maximisation permet de mettre à jour les paramètres du modèle maximisant la vraisemblance calculée à l'étape E.

Les formes non linéaires pour la variable réponse ainsi que pour les poids du mélange dans le cas du mélange de régressions avec variables concomitantes (définis par une fonction softmax) nécessitent d'utiliser un algorithme d'optimisation numérique dans l'étape M ; on peut par exemple utiliser les IRLS (pour Iteratively reweighted least squares, voir notamment Burrus (2012)), puisqu'il n'existe pas de résolution analytique pour la maximisation de $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{[h]})$.

Même si le mélange de régressions permet de modéliser le lien entre une variable réponse Y et un ensemble de covariables $\mathbf{x}$ dans le cas de données structurées en groupes inconnus, le but ici est l'estimation de la structure de dépendance entre variables, et non la prédiction. Lorsque l'on dispose d'un nouvel individu, l'impossibilité de déterminer son appartenance aux groupes latents provenant du fait que seule la relation $Y|\mathbf{x}$ est modélisée, entraîne de mauvaises performances de prédiction. Les mélanges de régressions ne sont donc pas adaptés à la prédiction. Dans le cadre de la modélisation de données hétérogènes, les mélanges d'experts, détaillés par la suite, sont adaptés à la prédiction d'une variable réponse.

2.3 Mélanges d'experts

Les mélanges d'experts peuvent être considérés comme une généralisation des mélanges de régressions. La différence apportée ici porte sur la métrique appliquée à chaque classe du mélange, qui est de la même forme que celle utilisée dans les mélanges de régressions avec variables concomitantes. L'autre différence par rapport aux modèles de mélange de régressions tient dans le fait que l'on travaille maintenant dans le cadre supervisé, et l'objectif est de prédire la variable réponse Y.

Le modèle de mélange d'experts original a été introduit Jacobs et al. (1991), et repose sur trois principaux éléments :

- plusieurs experts, qui peuvent être des fonctions de régression ou des classifieurs
- un pont qui permet une partition floue des données observées, et qui définit les régions où chaque expert agit
- un modèle probabiliste qui combine les experts et le pont

Le modèle est une somme pondérée des experts, où les poids dépendent des données observées $\mathbf{x}$. Sous cette forme, les experts peuvent se spécialiser sur des petites parties d'un problème plus large. De plus, ces modèles utilisent une partition souple des données. Ces propriétés permettent la prise en compte de données continues par morceaux dans des processus complexes de régression, et d'identifier la non-linéarité dans les problèmes de classification. Les mélanges d'experts ont été largement étudiés ces dernières années, notamment dans le domaine du machine learning. Ici, nous présentons la version du modèle la plus couramment utilisée, mais ce type de modèle a été revisité en termes de métriques utilisées pour pondérer les experts, de forme pour les modèles d'experts ou d'algorithmes d'estimation. Le lecteur peut notamment se référer à Yuksel et al. (2012) pour plus de détails sur les développements des mélanges d'experts.

2.3.1 Définition du modèle

On considère une réalisation du couple $(\mathbf{X}, Y)$, notée $(\mathbf{x}, y)$, avec Y la variable à expliquer et $\mathbf{X}$ l'ensemble des p covariables. On note aussi $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_g, \boldsymbol{\Phi}_e)$ l'ensemble des paramètres, avec $\boldsymbol{\Phi}_g$ correspondant à l'ensemble des paramètres associés au pont et $\boldsymbol{\Phi}_e$ à l'ensemble

des paramètres associés aux experts. Sachant un vecteur d'entrée $\mathbf{x}$ et une variable à expliquer y, la probabilité d'observer y peut s'écrire sous forme d'experts :

$$\begin{aligned} f(y|\mathbf{x}, \mathbf{\Phi}) &= \sum_{k=1}^{K} P(y, k|\mathbf{x}, \mathbf{\Phi}) \\ &= \sum_{k=1}^{K} P(k|\mathbf{x}, \mathbf{\Phi}_g) P(y|k, \mathbf{x}, \mathbf{\Phi}_e) \\ &= \sum_{k=1}^{K} g_k(\mathbf{x}, \mathbf{\Phi}_g) P(y|k, \mathbf{x}, \mathbf{\Phi}_e), \end{aligned}$$

avec K le nombre d'experts, $g_k(\mathbf{x}, \mathbf{\Phi}_g) = P(k|\mathbf{x}, \mathbf{\Phi}_g)$ les proportions du pont, correspondant à la probabilité de l'expert k sachant $\mathbf{x}$, et $P(y|k, \mathbf{x}, \mathbf{\Phi}_e)$ la probabilité de générer y à partir de l'expert k, sachant $\mathbf{x}$. La proportion du pont pour le k-ième expert est définie par la fonction softmax suivante :

$$g_k(\mathbf{x}, \mathbf{\Phi}_g) = \frac{\exp(\mathbf{x}^T \boldsymbol{\eta}_k)}{\sum_{l=1}^{K} \exp(\mathbf{x}^T \boldsymbol{\eta}_l)},$$

avec $\mathbf{\Phi}_g = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$ les paramètres du pont.

Les fonctions définissant les experts sont différentes selon l'objectif de l'utilisateur, qui peut être d'effectuer une régression ou une classification.

Modèle d'experts de régression : on note $\mathbf{\Phi}_e = (\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_K)$, avec $\mathbf{\Phi}_k = (\boldsymbol{\beta}_k, \sigma_k)$, les paramètres de l'expert. Dans le modèle original, les experts suivent une loi gaussienne de paramètres $(\boldsymbol{\beta}_k, \sigma_k^2)$. On peut écrire, pour le k-ième expert

$$P(y|\mathbf{x}; \mathbf{\Phi}_k) = \mathcal{N}(\mathbf{x}^T \boldsymbol{\beta}_k, \sigma_k^2),$$

avec $\boldsymbol{\beta}_k$ le régresseur associé au k-ième expert.

Pour faire une prédiction pour un nouvel individu, on a :

$$\hat{y} = \sum_k g_k(\mathbf{x}, \mathbf{\Phi}_g) P(y|\mathbf{x}; \mathbf{\Phi}_k).$$

Modèle d'experts de classification : on considère un problème de classification à L classes. La variable réponse Y est de longueur L, avec $y_l = 1$ si $\mathbf{x}$ appartient à la classe l, et 0 sinon. Il y a alors L paramètres $\boldsymbol{\beta}_{kl}$ à estimer pour chaque expert k, correspondant aux paramètres de chaque classe l. Pour le k-ième expert et la classe l, l'expert est donné par la fonction softmax :

$$\hat{y}_{kl} = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_{kl})}{\sum_{r=1}^{L} \exp(\mathbf{x}^T \boldsymbol{\beta}_{kr})}.$$

Pour un nouvel individu, la variable réponse est alors calculée par classe selon :

$$\hat{y}_l = \sum_k g_k(\mathbf{x}, \Phi_g) \hat{y}_{kl}.$$

La classe d'appartenance de l'individu est celle à plus forte valeur de $\hat{y}_l$. Ce type de modèle peut notamment être utilisé pour la prédiction d'une maladie, en considérant la variable réponse binaire correspondant à la présence ($Y = 1$) ou à l'absence ($Y = 0$) d'une maladie.

Même si ce type de modèle a été proposé avec un objectif de prédiction, les covariables $\mathbf{x}$ sont toujours considérées comme déterministes. Là encore, seule la distribution de $\mathbf{Y}|\mathbf{x}$ est modélisée.

2.3.2 Estimation du modèle

L'estimation des paramètres des modèles de mélange d'experts peut notamment se faire grâce à un algorithme EM (Yuksel et al., 2012). On note Z la variable indicatrice des experts, avec z_{ik} égale à 1 si l'individu i appartient à l'expert k, et 0 sinon. La log-vraisemblance des données complétées s'écrit :

$$\ln \mathcal{L}(\mathbf{x}, y, z; \mathbf{\Phi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \big(\log g_k(\mathbf{x}, \mathbf{\Phi}_g) + \log P(y|k, \mathbf{x}; \mathbf{\Phi}_e) \big).$$

L'étape E permet de calculer les probabilités *a posteriori* d'appartenance aux experts $\tau_{ik} = \mathbb{E}(z_{ik}|y, \mathbf{x})$ pour tout individu i et tout expert k. On peut alors calculer l'espérance de la log-vraisemblance complétée $Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]})$, avec $[h]$ l'indice d'itération de l'algorithme :

$$\begin{aligned} Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \big(\log g_k(\mathbf{x}, \mathbf{\Phi}_g) + \log P(y|k, \mathbf{x}; \mathbf{\Phi}_e) \big) \\ &= \sum_{k=1}^{K} Q_{kg} + Q_{ke}, \end{aligned}$$

avec $Q_{kg} = \sum_{i=1}^{n} \tau_{ik} \log g_k(\mathbf{x}, \mathbf{\Phi}_g)$ et $Q_{ke} = \sum_{i=1}^{n} \tau_{ik} \log P(y|k, \mathbf{x}; \mathbf{\Phi}_e)$.

Lors de l'étape M, la maximisation de la vraisemblance calculée à l'étape E permet de mettre à jour les paramètres selon :

$$\boldsymbol{\eta}_k^{[h+1]} = \arg\max_{\boldsymbol{\eta}_k} \quad Q_{kg},$$

$$\mathbf{\Phi}_k^{[h+1]} = \arg\max_{\mathbf{\Phi}_k} \quad Q_{ke}.$$

Dans le cas des mélanges d'experts, lors de l'étape M, $g_{ik} = g_k(\mathbf{x}_i, \mathbf{\Phi}_g)$ est estimé à τ_{ik} constant, calculé lors de l'étape E. Comme on a $0 \leq \tau_{ik} \leq 1$ et $0 \leq g_{ik} \leq 1$, alors le maximum de Q_{kg} est atteint pour $\tau_{ik} = 1$ et $g_{ik} = 1$, avec $\tau_{il} = 0$ et $g_{il} = 0$ quel que soit $l \neq k$. Les experts qui partagent des observations sont donc pénalisés.

2.4 Sélection de modèles

2.4.1 Nombre de groupes - nombre de composantes

La spécification du modèle a jusqu'ici été faite en supposant K fixé. En réalité, on travaille dans le cadre non supervisé, donc on ne connaît pas *a priori* le nombre de groupes structurant les données. Le nombre de groupes K est un paramètre crucial car il est lié à l'hétérogénéité de la population. Cependant, c'est un paramètre latent qui doit être sélectionné. On cherche un nombre de groupes permettant une partition interprétable et pratique pour l'utilisateur, qui a aussi été sélectionné théoriquement comme étant le plus adapté aux données étudiées.

La méthode habituelle consiste à estimer le modèle pour différents nombres de groupes possibles et essayer de déterminer quel modèle est le plus adapté au problème étudié. On peut donc considérer que choisir le nombre de groupes revient à faire de la sélection de modèles.

Généralement, le nombre de groupes est sélectionné de façon automatique grâce à certains critères que l'on cherche à optimiser. Le principe le plus répandu pour ces critères de sélection de modèles dans le cadre du partitionnement est basé sur la vraisemblance pénalisée.

2.4.2 Critères de sélection de modèles

On note $\mathcal{M}_K$ un modèle de mélange estimé avec K composantes. Une collection de modèles $\mathcal{M}_1, ..., \mathcal{M}_{K_{max}}$ est estimée, et le modèle sélectionné est celui qui minimise le critère choisi. D'après Bozdogan (1993), lorsque aucune information n'est disponible sur K_{max}, il est recommandé de faire varier le nombre de groupes de 1 au plus petit entier supérieur à $n^{0.3}$.

Le Critère d'Information d'Akaike (AIC) introduit par Akaike (1973) mesure la qualité d'un modèle estimé, en prenant en compte l'ajustement aux données et le nombre de paramètres. Ce critère permet de sélectionner le modèle qui minimise la perte d'information, parmi un ensemble de modèles testés. L'AIC est un critère généralement utilisé dans les problèmes dont l'objectif est la prédiction, mais est aussi adapté au partitionnement par modèles génératifs (Bozdogan and Sclove, 1984; Shmueli, 2010). Pour un nombre de groupes K, l'AIC est défini par

$$AIC_K = -2 \ln \mathcal{L}(\hat{\mathbf{\Phi}}) + 2\nu_K,$$

avec $\hat{\mathbf{\Phi}}$ le maximiseur de la fonction de log-vraisemblance et ν_K le nombre de paramètres du modèle estimé avec K groupes.

Tout comme le critère AIC, le Critère d'Information Bayésien ou BIC (Schwarz, 1978) permet de faire un compromis entre le nombre de paramètres du modèle et l'ajustement aux données (grâce à la fonction de vraisemblance), et sélectionne un modèle parcimonieux bien ajusté aux données. Des études (par exemple Koehler and Murphree, 1988) montrent que l'AIC a tendance à surestimer le nombre de composantes K du mélange. Dans le cadre des modèles de mélange, le critère BIC donne en pratique de meilleurs résultats

que l'AIC (Celeux and Soromenho, 1996; McLachlan and Peel, 2000). Ce critère est donc communément utilisé pour le choix du nombre de groupes (Keribin, 2000). Selon les notations précédentes, pour un nombre de groupes K, pour un ensemble de paramètres estimés $\hat{\mathbf{\Phi}}$, le BIC est défini par

$$BIC_K = -2 \ln \mathcal{L}(\hat{\mathbf{\Phi}}) + \nu_K \ln(n).$$

Cependant, les critères AIC et BIC ont été proposés pour l'estimation de densité dans le cas de données non structurées, et Biernacki et al. (2000) montrent que dans certain cas, ce critère peut mener au mauvais choix de K. Pour cette raison, le critère de Vraisemblance Classifiante Intégrée (ou ICL pour Integrated Completed Likelihood) est développé (Biernacki et al., 2000; McLachlan and Peel, 2000) et ajoute au BIC un terme d'entropie prenant en compte la concentration des groupes. Pour un modèle estimé avec K groupes, il est défini par

$$ICL_K = BIC_K - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \ln \tau_{ik},$$

avec τ_{ik} les probabilités *a posteriori* estimées pour $i = 1, \ldots, n$ et $k = 1, \ldots, K$. Ce critère est plus adapté au cadre des modèles génératifs. Ce critère est bien adapté au choix du nombre de groupes mais aussi au choix de la forme du modèle et à la paramétrisation. Il est moins sensible que le critère BIC à un mauvais ajustement du mélange aux données, et parvient à retenir les modèles donnant lieu à une classification pertinente des données. Dans le cas de groupes bien séparés, le terme d'entropie est proche de zéro, et le critère ICL a une valeur proche de celle du BIC. Dans le cas de groupes mal séparés, le terme d'entropie est fortement négatif et la valeur de l'ICL augmente. Ainsi, le critère ICL favorise les modèles aux groupes bien séparés.

Comme évoqué précédemment, le cadre des mélanges de régressions considère les covariables observées $\mathbf{x}$ comme fixes. Cependant, dans certains cas d'application, on ne peut pas considérer les observations comme déterministes, et les données $\mathbf{x}$ peuvent se comporter de manière différentes selon les groupes latents. Dans ce cas, le modèle doit prendre en compte l'hétérogénéité des données portée par les covariables en considérant un design aléatoire. Cette approche permet d'obtenir un modèle mieux adapté à la prédiction.

2.5 Mélange de régressions à design aléatoire

Dans cette section, nous proposons une généralisation des modèles de mélange de régressions adaptés à la prédiction, dans le cas des familles de distribution de probabilité usuelles. Nous considérons une modélisation de la dépendance entre des variables issues d'une distribution gaussienne et une variable issue d'une autre distribution de probabilité, comme par exemple une distribution binomiale ou de gamma. Nous prenons en compte un

design aléatoire pour les covariables, ce qui permet de considérer que leur comportement diffère selon le groupe latent. Dans notre situation, l'information de groupe portée par la variable latente est modélisée via la distribution jointe de $(Y, \mathbf{X})$. Cette approche a l'avantage de permettre l'estimation des propabilités *a posteriori* d'appartenance aux groupes latents pour une nouvelle observation, et l'adaptation de ce modèle à la prédiction d'une variable réponse. Étant donné que les groupes latents dépendent des covariables $\mathbf{x}$, les probabilités d'appartenances aux groupes latents peuvent être calculées pour un nouvel individu dont on dispose d'une mesure de $\mathbf{x}$. Cela permet de prédire la variable réponse de cet individu comme une somme des modèles de régression estimés, pondérée par les probabilités *a posteriori* estimées. L'information d'appartenance aux groupes latents de cet individu est alors réellement prise en compte, contrairement au cas des mélanges de régressions, où cette information ne peut pas être évaluée. Cette approche permet alors d'améliorer les performances de prédiction dans le cas de données structurées en groupes latents.

2.5.1 Spécification du modèle

Soit $Y \in \mathcal{Y} \subset \mathbb{R}$ une variable réponse à expliquer par un vecteur de variables aléatoires $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$. On suppose que les individus sont répartis dans K groupes latents de proportions $(\pi_k)_{k=1,\dots,K}$ avec $0 < \pi_k$ et $\sum_{k=1}^{K} \pi_k = 1$. La variable latente discrète Z à valeurs dans $\{1, \cdots, K\}$ représente l'information de groupe et est modélisée par une distribution multinomiale de paramètres $(\pi_k)_{k=1,\dots,K}$:

$$Z \sim \mathcal{M}(\pi_1, \dots, \pi_K).$$

On considère la distribution conditionnelle $(Y|\mathbf{X})$ dans un modèle de mélange de régressions selon les notations précédentes :

$$f(y|\mathbf{x}; \mathbf{\Phi}) = \sum_{k=1}^{K} \pi_k f(y|\mathbf{x}; \mathbf{\Phi}_k).$$

Dans chaque groupe, défini par $Z = k$, on suppose que la réponse Y de la variable dépendante est générée à partir d'une distribution particulière de la famille exponentielle, incluant la plupart des lois usuelles : gaussienne, binomiale, Poisson et gamma. La forme générale de la densité f_k de Y est donnée par

$$f_k(y; \theta_k, \mathbf{\Phi}_k) = \exp\left(\frac{y\theta_k - b(\theta_k)}{a(\mathbf{\Phi}_k)} + c(y, \mathbf{\Phi}_k)\right),$$

avec $\mathbf{\Phi}_k$ le paramètre de dispersion et θ_k le paramètre canonique.

Dans le cadre de la régression, les paramètres de la distribution dépendent des variables $\mathbf{X}$. La distribution de Y sachant $Z = k$ et $\mathbf{X} = \mathbf{x}$ appartient à la famille exponentielle et

$$\begin{aligned} E(Y|\mathbf{X} = \mathbf{x}, Z = k) &= b'(\theta_k) = g^{-1}(\mathbf{x}\boldsymbol{\beta}_k) \\ var(Y|\mathbf{X} = \mathbf{x}, Z = k) &= \mathbf{\Phi}_k b''(\theta_k), \end{aligned}$$

avec $g = (b')^{-1}$ une fonction de lien, appelée la fonction de lien canonique.

La distribution de $\mathbf{X}$ sachant $Z = k$ est une distribution gaussienne multivariée de moyenne $\boldsymbol{\mu}_k$ et de variance Σ_k :

$$\mathbf{X}|\{Z = k\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k).$$

Un des cas particuliers bien connu est le mélange de régressions linéaires où la distribution de Y dans chaque classe est gaussienne. La fonction de lien canonique est la fonction identité, donc la moyenne conditionnelle est égale à $\mathbf{x}^T\boldsymbol{\beta}_k$ et la variance conditionnelle est ϕ_k (plus classiquement notée σ_k^2 pour ce modèle particulier). On a alors, selon les notations précédentes,

$$Y|\{\mathbf{X} = \mathbf{x}, Z_k = 1\} \sim \mathcal{N}\left(\mathbf{x}^T\boldsymbol{\beta}_k, \sigma_k^2\right),$$

avec $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})$ le vecteur des coefficients de régression dans le k-ième groupe. Les prédicteurs sont donc liés à la variable réponse Y grâce au modèle linéaire

$$\mathbb{P}(Y = 1|\{\mathbf{X} = \mathbf{x}, Z_k = 1\}) = \mathbf{x}^T\boldsymbol{\beta}_k.$$

Finalement, la loi de probabilité de Y sachant $\mathbf{X} = \mathbf{x}$ peut s'écrire comme un modèle de mélange de régressions linéaires. La loi de probabilité est donnée par

$$\begin{aligned}\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}; \boldsymbol{\Phi}) &= \sum_{k=1}^{K} \pi_k \mathbb{P}(Y = y|\{\mathbf{X} = \mathbf{x}, Z_k = 1; \boldsymbol{\Phi}_k\}) \\ &= \sum_{k=1}^{K} \pi_k \mathbf{x}^T\boldsymbol{\beta}_k,\end{aligned}$$

avec $\boldsymbol{\Phi}_k = (\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\beta}_k)$ le vecteur des paramètres du groupe k et l'ensemble complet des paramètres à estimer pour le modèle de mélange à K groupes est noté $\boldsymbol{\Phi} = (\pi_1, ..., \pi_K, \boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_K)$.

2.5.2 Prédiction

Dans le cadre des modèles de mélange de régressions tels que décrits précédemment, où le mélange repose sur la distribution conditionnelle, on cherche à modéliser une structure de groupe au sein de modèles de régression, en se concentrant sur l'estimation des paramètres. Cependant, un de nos objectifs est de prédire la variable réponse Y et d'adapter le modèle pour faire de la prédiction. D'autre part, les modèles de mélange d'experts ont pour objectif la prédiction, et notre modèle peut être comparé à ce type de modèles pour l'étape de prédiction. Cependant, en mélange d'experts, la loi de $\mathbf{x}$ n'est généralement pas estimée. Dans notre cas, le modèle de mélange est déterminé par la distribution multinomiale pour la variable de groupe $\mathbf{Z}$, définissant les probabilités *a posteriori* comme poids. On a alors un cas particulier des modèles de mélange d'experts utilisant des formes paramétriques gaussiennes pour pondérer les experts (Yuksel et al., 2012).

L'information des groupes latents étant en partie portée par les covariables, il est possible de calculer l'appartenance aux groupes latents pour une nouvelle observation. Cela permet de prédire la réponse pour un nouvel individu comme une somme des modèles de régression estimés, pondérée par les probabilités *a posteriori* calculées grâce aux covariables.

La règle de prédiction est la suivante

$$\begin{aligned}\mathbb{E}(y|\mathbf{X}=\mathbf{x}) &= \sum_{k=1}^{K} \mathbb{P}(y, Z=k|\mathbf{X}=\mathbf{x}) \\ &= \sum_{k=1}^{K} \mathbb{P}(y|Z=k, \mathbf{X}=\mathbf{x})\mathbb{P}(Z=k|\mathbf{X}=\mathbf{x}). \end{aligned} \tag{2.2}$$

En réalité, la règle de prédiction ne dépend pas de la variable réponse y non observée, et pour un objectif de prédiction, seule l'information concernant l'appartenance au groupe contenue dans $\mathbf{X}$ est nécessaire. Cependant, lors de l'étape d'estimation, l'information observée $(\mathbf{X}_i, Y_i)$ est utilisée pour prendre en compte le lien entre les prédicteurs et la variable réponse structurant les données en groupes. Par conséquent, les paramètres du modèle estimé dépendent des probabilités *a posteriori* $\tau_{ik} = \mathbb{P}(Z=k|\mathbf{X}=\mathbf{x}, Y=y)$ qui prennent en compte implicitement la structure de groupe portée par la distribution conditionnelle. L'estimation des paramètres détaillée en section 2.5.3 permet d'obtenir un estimateur pour (2.2), pour prédire y. Il est possible d'estimer les probabilités d'appartenance de l'observation $\mathbf{x}$ aux groupes latents selon

$$\mathbb{P}(Z=k|\mathbf{X}=\mathbf{x}) = \frac{\pi_k f_{\mathbf{X}}\left(\mathbf{x}; \mu_k, \Sigma_k\right)}{\sum_{\ell=1}^{K} \pi_\ell f_{\mathbf{X}}\left(\mathbf{x}; \mu_\ell, \Sigma_\ell\right)}.$$

On obtient une prédiction pour y en utilisant les paramètres estimés pour les modèles linéaires généralisés associés à chaque groupe.

2.5.3 Estimation par maximum de vraisemblance

L'estimation des paramètres du modèle est basée sur l'approche du maximum de vraisemblance, où l'on considère la variable latente de classe comme une donnée manquante. Nous utilisons donc un algorithme EM pour l'inférence du modèle. La décomposition des étapes est la même que celle détaillée dans le cas des modèles de mélange. L'exemple du mélange de régressions linéaires avec une variable réponse suivant une loi de probabilité gaussienne est détaillé.

De la même façon que précédemment, on cherche à résoudre le problème de vraisemblance suivant :

$$\hat{\mathbf{\Phi}} = \arg\max_{\mathbf{\Phi}} \quad \left(\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{\Phi})\right), \tag{2.3}$$

selon les même notations que précédemment. Là encore, on ne peut pas maximiser directement cette vraisemblance et on utilise un algorithme EM pour optimiser (2.3). On peut décomposer la log-vraisemblance grâce à la règle de Bayes :

$$\begin{aligned}\ln \mathcal{L}(y_1,\dots,y_n,\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\mathbf{\Phi}) = &\ln \mathcal{L}(y_1,\dots,y_n|\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\beta})\\ &+\ln \mathcal{L}(\mathbf{x}_1,\dots,\mathbf{x}_n|\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\mu},\mathbf{\Sigma})\\ &+\ln \mathcal{L}(\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\pi}),\end{aligned}$$

avec $\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_K)$ et $\mathbf{\Sigma}=(\Sigma_1,\dots,\Sigma_K)$. La distribution de $\mathbf{x}|zbf$ est bien considérée dans notre situation. On considère l'espérance conditionnelle de la log-vraisemblance pénalisée des données complétées :

$$\arg\max_{\mathbf{\Phi}} \quad \left(\mathbb{E}_{\mathbf{\Phi}^*}\left[\ln \mathcal{L}(y_1,\dots,y_n,\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\mathbf{\Phi})|\mathbf{x}_1,\dots,\mathbf{x}_n,y_1,\dots,y_n\right]\right).$$

L'étape E de l'algorithme est inchangée dans ce cas, et consiste à calculer les probabilités *a posteriori* $\tau_{ik}=\mathbb{P}(Z_{ik}=1|\mathbf{X}_i=\mathbf{x}_i,Y_i=y_i;\mathbf{\Phi}_k^{[h]})$, sachant les paramètres courants $\pi_k^{[h]},\boldsymbol{\mu}_k^{[h]},\Sigma_k^{[h]},\boldsymbol{\beta}_k^{[h]}$, à l'itération $[h]$. Pour tout $i=1,\dots,n$ et pour tout $k=1,\dots,K$, elles sont données par

$$\tau_{ik}^{[h+1]}=\frac{\pi_k^{[h]} f_{\mathbf{X},Y}\left(\mathbf{x}_i,y_i;\boldsymbol{\mu}_k^{[h]},\Sigma_k^{[h]},\boldsymbol{\beta}_k^{[h]}\right)}{\sum_{\ell=1}^K \pi_\ell^{[h]} f_{\mathbf{X},Y}\left(\mathbf{x}_i,y_i;\boldsymbol{\mu}_\ell^{[h]},\Sigma_\ell^{[h]},\boldsymbol{\beta}_\ell^{[h]}\right)},$$

avec $f_{\mathbf{X},Y}(.)$ la fonction de densité jointe de $(\mathbf{X},Y)$. La densité jointe est calculée grâce à la relation $f_{\mathbf{X},Y}(.)=f_{Y|\mathbf{X}}(.)f_{\mathbf{X}}(.)$ pour laquelle les distributions sont connues comme étant une distribution normale de paramètres $\mathbf{X}\boldsymbol{\beta}_k$ et σ^2 pour $Y|\mathbf{X}=\mathbf{x}$ et une distribution normale de paramètres $\boldsymbol{\mu}_k$ et Σ_k pour $\mathbf{X}$.

On calcule alors l'espérance de la log-vraisemblance complétées $Q(\mathbf{\Phi},\mathbf{\Phi}^{[h]})$ sachant les données observées $(y_i,\mathbf{x}_i)_{i=1,\dots,n}$ et $\mathbf{\Phi}^{[h]}$,

$$\begin{aligned}Q(\mathbf{\Phi},\mathbf{\Phi}^{[h]})=\sum_{i=1}^n\sum_{k=1}^K \tau_{ik}^{[h+1]}\Bigg(&\ln\pi_k^{[h]}-\frac{d+1}{2}\ln 2\pi-\frac{1}{2}\ln(\sigma_k^{[h]})^2-\frac{(y_i-\mathbf{x}_i^t\boldsymbol{\beta}_k^{[h]})^2}{2(\sigma_k^{[h]})^2}+\frac{1}{2}\ln|\Theta_k^{[h]}|\\ &-\frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{[h]}\right)^T\Theta_k^{[h]}\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{[h]}\right)\Bigg),\end{aligned}$$

avec $\Theta_k=\Sigma_k^{-1}$ l'inverse de la matrice de covariance pour chaque groupe, appelée matrice de précision.

Lors de l'étape M, on cherche à maximiser l'espérance conditionnelle calculée à l'étape E, par rapport au paramètre $\mathbf{\Phi}$, pour obtenir $\mathbf{\Phi}^{[h+1]}$, selon

$$\mathbf{\Phi}^{[h+1]}=\arg\max_{\mathbf{\Phi}} \quad \left\{Q(\mathbf{\Phi},\mathbf{\Phi}^{[h]})\right\}.$$

La mise à jour de chaque paramètre à l'itération $[h+1]$, pour tout $k = 1, \ldots, K$, est alors donnée par

$$\widehat{\pi}_k^{[h+1]} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{[h+1]},$$

$$\widehat{\boldsymbol{\mu}}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{x}_i,$$

$$\hat{\Sigma}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \right]^{-1} \sum_{i=1}^{n} \tau_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T,$$

$$\hat{\boldsymbol{\beta}}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{x}_i \mathbf{x}_i^T \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{x}_i y_i,$$

$$(\hat{\sigma}_k^{[h+1]})^2 = \frac{\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k^{[h+1]})^2 \tau_{ik}^{[h+1]}}{\sum_{i=1}^{n} \tau_{ik}^{[h+1]}}.$$

2.5.4 Sélection de modèles

Les méthodes de sélection décrites précédemment sont utilisées pour la sélection du nombre de groupes dans le mélange. Une collection de modèle est donc estimée pour un nombre de groupes allant de 1 à K_{max}, et le modèle minimisant le critère choisi, parmi l'AIC, le BIC et l'ICL, est sélectionné.

2.5.5 Application sur données réelles

Les données Vin

Le jeu de données Vin est typiquement utilisé dans les problèmes de classification, et disponible dans le répertoire de données de machine learning UCI (Lichman, 2013). Ces données contiennent les résultats d'analyses chimiques effectuées sur des vins provenant de la même région d'Italie, de trois variétés différentes. Ces analyses ont permis de déterminer les quantités de 13 constituants trouvés dans chacun des trois types de vin. On considère un sous-ensemble des variables de ce jeu de données avec pour objectif de prédire la variable "Flavonoids". Les données étudiées contiennent 178 observations, décrites par les quatre variables suivantes : alcohol, malic acid, color intensity et OD280. Ces observations sont structurées en trois groupes correspondant aux différentes variétés de vigne. Les variables montrant une structure de groupe en fonction de la variable à expliquer sont conservées pour l'analyse. La structure de groupe est illustrée en figure 2.1. Cette figure

permet de remarquer que la structure de groupe est portée à la fois par les variables explicatives, mais aussi par le lien $Y|\mathbf{x}$. Le but de notre méthode est de prédire une variable réponse, à l'aide de données structurées, qui est influencée par les groupes.

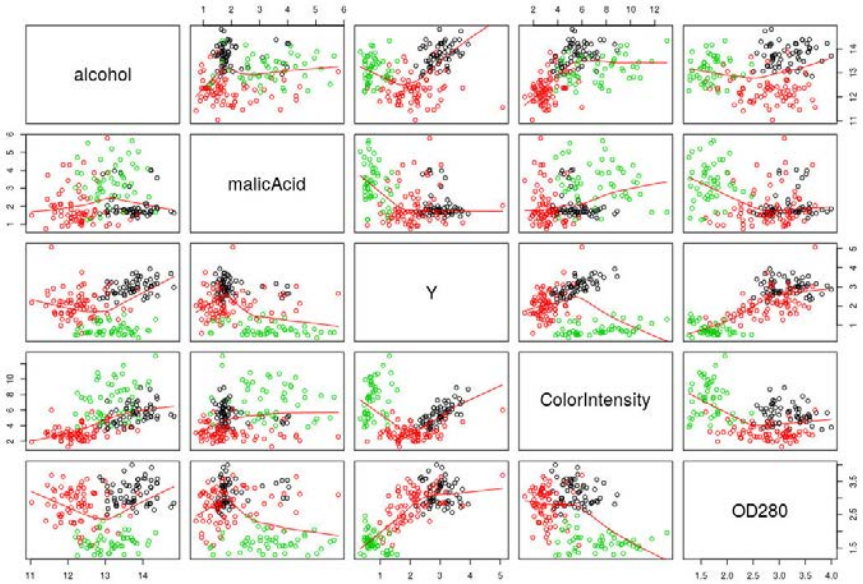

FIGURE 2.1 : **Représentation du lien entre les variables du jeu de données Vin.** Les liens entre la variable réponse et les variables explicatives sont représentés par la courbe rouge. Les groupes correspondant aux variétés de vigne sont symbolisés par la couleur des points : les observations noires correspondent à la première variété, les observations rouges à la deuxième et les observations vertes à la troisième variété. La variable réponse correspond à la variable Flavonoids.

Procédure d'analyse

Les performances de prédiction obtenues avec notre méthode (appelée RM) sont comparées avec les performances de prédiction obtenues avec une modèle linéaire classique (appelé LM), et avec les performances obtenues avec un modèle par groupe appelé GMM. Pour cette dernière méthode, un modèle de mélange gaussien estimé sur les covariables permet d'estimer des groupes grâce à la règle de classification du maximum *a posteriori*. Ensuite, un modèle linéaire est ajusté sur chacun des groupes. Si l'on dispose d'une nouvelle observation, son groupe est prédit grâce au mélange, puis la prédiction est effectuée avec le modèle correspondant. La différence entre cette méthode et notre méthode porte tout d'abord sur l'estimation de la partition qui est basée sur $Y|\mathbf{X}$ dans notre cas, et sur $\mathbf{X}$ dans le cas de la méthode GMM. Ensuite, la prédiction est réalisée suite à une classification dure de la nouvelle observation dans le cas de la méthode GMM, alors que notre méthode considère une classification floue de la nouvelle observation lors de l'étape

de prédiction.

La stabilité de la méthode est étudiée grâce à une procédure de validation croisée. Les 178 observations sont d'abord sous-échantillonnées en un échantillon d'apprentissage comprenant $\frac{2}{3}$ des observations. Le modèle de mélange de régressions, le modèle basé sur les composantes GMM et le modèle linéaire sont estimés sur cet ensemble d'apprentissage. Les prédictions sont ensuite effectuées sur les 59 individus restant. Cette procédure est répétée 100 fois et permet de connaître la stabilité des prédictions en étudiant la répartition des critères d'évaluation.

La qualité de prédiction de la variable réponse est étudiée grâce à l'erreur quadratique moyenne, définie par

$$EQM = \sqrt{\frac{\sum_{i \in E_v} (y_i - \hat{y}_i)^2}{v}},$$

avec y_i la réponse observée et $\hat{y}_i$ la réponse prédite pour l'individu i dans l'échantillon de validation E_v de taille v.

Résultats

Pour les trois méthodes alternatives, le nombre de groupes correspondant au nombre de variétés de vigne est trouvé dans environ 30% des itérations. Cela peut être dû à une structure de groupes caractérisant $\mathbf{X}$ et $Y|\mathbf{X}$, qui n'est pas seulement liée à la variété. En effet, même dans les cas où 3 groupes sont trouvés, ils ne correspondent pas forcément aux variétés. Cela peut s'expliquer par le fait que les variétés n'affectent pas forcément la règle de prédiction pour tous les groupes. Par exemple, l'acide malique semble être lié de façon spécifique à Y pour la première variété, mais le lien semble être le même pour les variétés 2 et 3. Cela est illustré par la pente de la régression entre Y et l'acide malique qui n'est pas modifiée pour ces deux groupes sur la figure 2.1. Pour 69 des 100 itérations, deux groupes sont trouvés par notre méthode, ce qui confirme que la pente entre la variable réponse et les prédicteurs semble avoir un seul point de rupture, ce qui montre deux groupes potentiels. Ce lien doit être plus fort que la structure de groupe présente sur les covariables. Pour le modèle GMM, deux groupes sont trouvés pour seulement 19 itérations, ce qui montre que la partition sur $\mathbf{X}$ ne met pas en évidence la même structure que la partition réalisée grâce à $\mathbf{X}$ et $Y|\mathbf{X}$. Une structure spécifique existe entre la variable Flavonoids et les quatre covariables, et n'est pas trouvée par le modèle GMM. Cela confirme la spécificité de notre méthode pour prendre en compte les groupes affectant à la fois $\mathbf{X}$ et $Y|\mathbf{X}$.

Pour les performances de prédiction, l'erreur quadratique moyenne est plus faible avec notre méthode qu'avec les deux méthodes alternatives, et les résultats sont plus stables, d'après la figure 2.1 et la table 2.1.

Méthode	RM	LM	GMM
EQM	0.47	0.55	0.56
Corrélation	0.89	0.84	0.83

TABLE 2.1 : **Performances de prédiction des méthodes RM, LM et GMM sur les données Vin** La méthode RM correpond au mélange de régressions à design aléatoire, la méthode LM correspond au modèle linéaire et la méthode GMM correspond à la prédiction par groupes trouvés par GMM. Les critères de performance sont l'erreur quadratique moyenne et la corrélation entre la réponse observée et la réponse prédite.

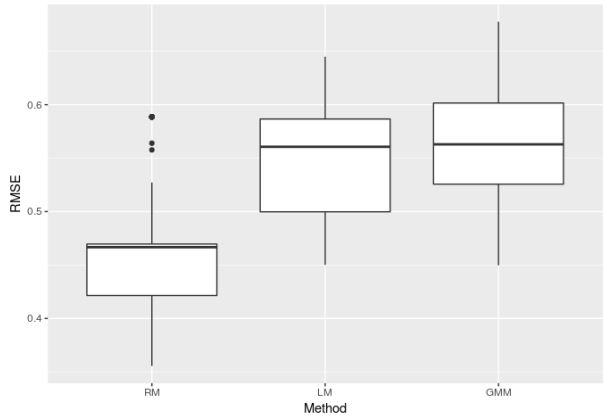

FIGURE 2.2 : **Comparaison des performances de prédiction des trois méthodes alternatives sur les données Vin.** Les méthodes comparées sont le mélange de régressions à design aléatoire (RM), le modèle linéaire (LM) et la régression par classes issues de GMM (GMM). Les boîtes à moustache représentent l'erreur quadratique moyenne calculée sur les 100 répétitions de validation croisée, pour chaque modèle.

Chapitre 3

Mélange de régressions logistiques pénalisées avec design aléatoire

Contents

Ce chapitre fait l'objet de deux articles.

Comme décrit dans le chapitre 1, la NAFLD et la NASH sont des pathologies de plus en plus répandues dans les pays occidentaux. Aujourd'hui, le diagnostic de la NASH reste une problématique importante et il n'existe aucune méthode non invasive et communément

admise permettant de différencier la NAFLD de la NASH, et d'évaluer l'atteinte au foie. De plus, il existe différentes typologies de patients, mal connues, affectant l'évolution de la maladie. Nous souhaitons donc utiliser un modèle de prédiction permettant de diagnostiquer la maladie tout en estimant la structure latente des données. Pour cela, nous nous concentrons sur les modèles de mélange et les modèles d'experts, décrits dans le chapitre précédent. Dans ce chapitre, nous considérons une approche à l'intersection de ces deux modèles, où la distribution de la réponse conditionnellement aux prédicteurs est définie comme un mélange. Nous nous plaçons là encore dans le cadre d'un design aléatoire, ce qui fait l'originalité de notre modèle, et nous permet d'effectuer de la prédiction. Par conséquent, notre modèle a l'avantage d'exploiter l'information de groupes structurant les données et portée par la distribution conditionnelle, ainsi que par les prédicteurs. De plus, la prédiction de la variable réponse peut être calculée explicitement puisque les probabilités *a posteriori* d'appartenance aux groupes ne dépendent pas de la réponse non observée pour la nouvelle observation. Lorsque l'on souhaite établir le diagnostic d'une maladie, on se trouve face à une réponse binaire. Par conséquent, le modèle considéré est un mélange de régressions logistiques, dans lequel le mélange est défini pour la distribution conditionnelle, et les covariables sont supposées gaussiennes. Comme décrit au chapitre 2, l'inférence est effectuée grâce à l'algorithme EM qui est adapté au cadre des variables latentes.

La spectrométrie permettant d'obtenir une empreinte moléculaire d'un échantillon mesuré, elle permettrait l'étude des différents biomarqueurs présents dans le sérum de patients atteints de NASH. Cette technologie pourrait mettre en évidence les différences métaboliques entre patients et potentiellement l'avancée de la maladie ou l'atteinte au foie. Cependant, comme mentionné précédemment, les données de spectrométrie portent toute l'information moléculaire contenue dans l'échantillon biologique mesuré, alors que seuls quelques nombres d'ondes sont supposés informatifs pour prédire la maladie. Le vecteur des coefficients de régression est supposé parcimonieux et une sélection de variables est considérée pour estimer de façon précise le modèle. Nous utilisons une approche de pénalisation de type ℓ_1 de la vraisemblance pour effectuer simultanément la sélection de variables et l'estimation des paramètres. Une telle approche peut s'inclure directement dans l'algorithme EM et bénéficier des garanties théoriques du cadre des modèles de mélange de régressions (Khalili and Chen, 2007; Städler et al., 2010). De façon similaire, un estimateur Graphical-Lasso est considéré pour la matrice de précision des prédicteurs au sein des groupes. Cette seconde pénalisation met en évidence les dépendances conditionnelles entre les covariables et réduit la dimension.

Dans ce chapitre, nous présentons le modèle de mélange de régressions logistiques, ainsi que l'étape de prédiction. Ensuite, l'estimation des paramètres par maximum de vraisemblance est détaillée et la régularisation utilisée pour réduire la dimension et permettre une meilleure interprétabilité est exposée. Les critères de sélection de modèles sont ensuite décrits. Une étude de simulation permettant d'étudier les performances d'estimation et de prédiction de la méthode est ensuite exposée. Enfin, les performances du modèle présenté dans ce chapitre sont évaluées sur les données de spectrométrie mesurées sur la cohorte NASH-Nice, présentées dans le chapitre 1.

L'annexe A présente les codes utilisés pour l'étude de simulation présentée dans ce chapitre.

3.1 Modèle

3.1.1 Modèle de mélange de régressions logistiques

Soit $(\mathbf{X}, Y)$ des variables aléatoires, avec Y une variable réponse binaire à valeurs dans $\{0, 1\}$ et $\mathbf{X} \in \mathbb{R}^p$ un ensemble de p covariables. On suppose que les individus sont répartis dans K groupes inconnus, de proportions $(\pi_k)_{k=1,...,K}$ avec $0 < \pi_k$ et $\sum_{k=1}^{K} \pi_k = 1$. On note $\mathbf{Z} = (Z_1, ..., Z_K)$ la variable aléatoire de groupes, avec Z_k égal à 1 si l'individu appartient au groupe k, et 0 sinon. On considère que les covariables ainsi que le modèle de régression dépendent de la structure en groupes. Le modèle logistique pour la réponse Y dans le mélange est alors défini par

$$Y|\{\mathbf{X} = \mathbf{x}, Z_k = 1\} \sim \mathcal{B}\left(p^{(k)}(\mathbf{x})\right),$$

avec $\mathbf{x}$ une réalisation de $\mathbf{X}$ et $p^{(k)}(\mathbf{x}) = \mathbb{P}(Y = 1|\{\mathbf{X} = \mathbf{x}, Z_k = 1\})$. Les prédicteurs sont liés à la variable réponse Y grâce à la fonction de lien logistique

$$\text{logit}\big(p^{(k)}(\mathbf{x})\big) = \mathbf{x}\boldsymbol{\beta}_k,$$

où logit : $x \mapsto \log(\frac{x}{1-x})$ et $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})$ est le vecteur des coefficients de régression dans le k-ième groupe. De plus, sachant $\{Z_k = 1\}$, la covariable $\mathbf{X}$ est modélisée par une distribution gaussienne multivariée selon

$$\mathbf{X}|\{Z_k = 1\} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k),$$

avec $\boldsymbol{\mu}_k \in \mathbb{R}^p$ et Σ_k la moyenne et la matrice de covariance des prédicteurs du groupe k.

Finalement, la loi de probabilité de Y sachant $\mathbf{X} = \mathbf{x}$ peut s'écrire comme un modèle de mélange de régressions logistiques. La réponse étant binaire, cette loi de probabilité est donnée par

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k \mathbb{P}(Y = 1|\{\mathbf{X} = \mathbf{x}, Z_k = 1; \boldsymbol{\Phi}_k\}) = \sum_{k=1}^{K} \pi_k p^{(k)}(\mathbf{x}),$$

avec $\boldsymbol{\Phi}_k = (\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\beta}_k)$ le vecteur des paramètres du groupe k et $\boldsymbol{\Phi} = (\pi_1, ..., \pi_K, \boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_K)$ l'ensemble des paramètres à estimer pour le modèle de mélange à K groupes.

Par la suite, on notera bien la différence entre la notion de classe représentée par la variable réponse Y binaire, composée de la classe des malades $(Y = 1)$ et de la classe des non malades $(Y = 0)$, et la notion de groupe représentée par la variable $\mathbf{Z}$, qui correspond aux groupes latents parmi les individus, portés par $\mathbf{X}$ et $\mathbf{Y}|\mathbf{X}$.

3.1.2 Prédiction

Comme pour le modèle de mélange de régressions avec design aléatoire présenté en section 2.5 du chapitre précédent, notre objectif est d'effectuer la prédiction y_0 à partir d'une nouvelle observation $\mathbf{x}_0$. Là encore, on se place dans le cas d'un design aléatoire, et on utilise uniquement l'information de structure des covariables pour prédire l'appartenance de la nouvelle observation aux groupes latents. La règle de prédiction est inchangée et définie par

$$\begin{aligned}\mathbb{E}(Y_0|\mathbf{X}_0 = \mathbf{x}_0) &= \sum_{k=1}^{K} \mathbb{P}(Y_0 = 1, Z_0 = k|\mathbf{X}_0 = \mathbf{x}_0) \\ &= \sum_{k=1}^{K} \mathbb{P}(Y_0 = 1|Z_{0k} = 1, \mathbf{X}_0 = \mathbf{x}_0)\mathbb{P}(Z_{0k} = 1|\mathbf{X}_0 = \mathbf{x}_0). \end{aligned} \tag{3.1}$$

L'estimation des paramètres détaillée en sections 3.2 et 3.3 permet d'obtenir un estimateur pour (3.1), pour prédire y_0. On note $\tau'_{0k} = \mathbb{P}(Z_{0k} = 1|\mathbf{X}_0 = \mathbf{x}_0)$, alors

$$\hat{\tau}'_{0,k} = \frac{\hat{\pi}_k f_{\mathbf{X}_0}(\mathbf{x}_0; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{K} \hat{\pi}_l f_{\mathbf{X}_0}(\mathbf{x}_0; \hat{\boldsymbol{\mu}}_l, \hat{\Sigma}_l)},$$

et on obtient une prédiction pour y_0 selon

$$\hat{y}_0 = \sum_{k=1}^{K} \hat{\tau}'_{0,k} \frac{\exp(\mathbf{x}_0^t \hat{\boldsymbol{\beta}}_k)}{1 + \exp(\mathbf{x}_0^t \hat{\boldsymbol{\beta}}_k)}.$$

3.2 Estimation par maximum de vraisemblance pénalisée

3.2.1 Maximum de vraisemblance pénalisée

Soit $\{(y_i, \mathbf{x}_i)_{i=1,\dots,n}\}$ un échantillon de taille n de réalisations des variables aléatoires $(\mathbf{X}, Y)$ introduites précédemment. L'estimation des paramètres est faite par maximum de vraisemblance. De nombreuses variables sont considérées dans les données d'application, et on suppose que certaines d'entre elles ne sont pas pertinentes. Par conséquent, une sélection de variables est utilisée pour faire ressortir les variables pertinentes et obtenir un modèle interprétable. Deux types de paramètres doivent être estimés pour chaque groupe, les coefficients de régression ainsi que les matrices de covariance. Une estimation consistante d'une matrice de covariance non structurée est difficile même en dimension modérée. De plus, pour notre application, certaines covariables sont conditionnellement indépendantes des autres, mais il existe une structure de corrélation forte entre les covariables, ce qui est représenté dans le modèle gaussien de matrice de covariance Σ par la matrice de précision Θ, correspondant à l'inverse de la matrice de covariance : $\Theta = \Sigma^{-1}$. Ainsi, une pénalisation de type Graphical Lasso (comme proposé dans Friedman et al., 2008) est utilisée pour contraindre certaines valeurs à zéro dans la matrice de précision associée aux covariables. De manière similaire, dans notre situation, on suppose que seul

un sous-ensemble de covariables est pertinent pour la prédiction de la variable réponse, et on s'attend à de la parcimonie dans les coefficients de régression. Une pénalité Lasso (Tibshirani, 1994) est donc utilisée pour contraindre certains éléments du vecteur $\boldsymbol{\beta}_k$ à valoir exactement zéro. Par conséquent, une estimation avec régularisation de type ℓ_1 est utilisée pour obtenir des estimateurs parcimonieux des matrices de précision des prédicteurs $\Theta_1, \ldots, \Theta_K$ et des coefficients de régression $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$. Selon l'échantillon $(y_i, \mathbf{x}_i)_{i=1,...,n}$, le problème de vraisemblance pénalisée que l'on cherche à résoudre est donné par, pour $\lambda_k \geq 0, \rho_k \geq 0$, pour tout $k = 1, \ldots, K$,

$$\hat{\boldsymbol{\Phi}}^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = \arg\max_{\boldsymbol{\Phi}} \left\{ \ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\Phi}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k\|_1 \right\}, \tag{3.2}$$

avec $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{p} |\beta_{k,j}|$, $\Theta_k = \Sigma_k^{-1}$ la matrice de précision du groupe k et $\|\Theta_k\|_1$ la somme des valeurs absolues de Θ_k. Les quantités λ_k et ρ_k correspondent aux paramètres de régularisation réglant la contrainte sur les paramètres $\boldsymbol{\beta}_k$ et Θ_k pour chaque groupe k, $k = 1, \ldots, K$. Une méthode de sélection de ces paramètres est décrite en section 3.2.2.

3.2.2 Sélection des paramètres de régularisation

Les paramètres de régularisation $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ et $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)$ déterminent la quantité de régularisation, et leur choix est important dans l'approche de vraisemblance pénalisée. Des fortes valeurs de paramètres de régularisation tendent à sélectionner un modèle simple dont les paramètres estimés ont une variance faible, alors que des valeurs faibles de paramètres de régularisation mènent à des modèles complexes, avec un biais faible mais avec une variance élevée. Le choix des paramètres de régularisation optimaux est basé sur un compromis entre biais et variance. Dans notre situation, il faut fixer un paramètre λ_k et un paramètre ρ_k pour chaque groupe k, ce qui correspond à $2K$ paramètres de régularisation à régler. La validation croisée est une méthode populaire pour la sélection des paramètres de régularisation dans le cadre de la vraisemblance pénalisée (voir par exemple Fan and Li, 2001; Khalili and Chen, 2007). Les paramètres de régularisation sont choisis un par un parmi un ensemble de valeurs possibles, en minimisant une fonction de perte par validation croisée. Cependant, les méthodes de validation croisée sont coûteuses numériquement. De plus, une étude de Wang et al. (2007) met en évidence que la validation croisée peut mener à la sélection de variables non pertinentes. Différentes études (Wang et al., 2007; Jiang et al., 2018; Lloyd-Jones et al., 2018; Khalili and Lin, 2013) suggèrent l'utilisation du BIC pour régler les paramètres de régularisation. Le BIC, présenté au chapitre précédent, permet un compromis entre le nombre de paramètres du modèle et l'ajustement aux données (grâce à la fonction de vraisemblance), et sélectionne un modèle parcimonieux bien ajusté aux données. Pour un nombre de groupe K, pour un ensemble de paramètres estimés $\hat{\boldsymbol{\Phi}}^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ obtenu avec des paramètres de régularisation $\boldsymbol{\lambda}$ et $\boldsymbol{\rho}$, le BIC est défini par

$$BIC^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = -2 \ln \mathcal{L}(\hat{\boldsymbol{\Phi}}^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + \nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})} \ln(n),$$

avec $\hat{\boldsymbol{\Phi}}^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ le maximiseur de la fonction de log-vraisemblance pénalisée et $\nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ le nombre de paramètres du modèle, correspondant au nombre de coefficients du modèle différents de zéro. Dans notre cadre, deux vecteurs de paramètres de régularisation différents $\boldsymbol{\lambda}$ et $\boldsymbol{\rho}$ doivent être choisis.

Une procédure automatique est proposée pour construire une grille de valeurs possibles de paramètres de régularisation pour chaque groupe. Cette procédure comprend plusieurs étapes :

- L'initialisation correspondant à une classification par K-means suivie de quelques itérations EM, permettant l'estimation des paramètres pour le calcul d'une classification initiale.
- Cette classification initiale permet la construction d'une grille de ($\boldsymbol{\lambda}$, $\boldsymbol{\rho}$) :
  - calcul des valeurs maximales λ_{max} et ρ_{max} de $\boldsymbol{\lambda}$ et $\boldsymbol{\rho}$ pour chaque groupe.
  - extension à une grille régulière de paramètres de régularisation possibles.
  - réduction de la grille de valeurs de $\boldsymbol{\rho}$ aux trois valeurs permettant le meilleur BIC sur des estimations obtenues par une étape M.
- Pour chaque combinaison de ($\boldsymbol{\lambda}$, $\boldsymbol{\rho}$), estimation d'un modèle grâce à quelques itérations de l'algorithme EM.
- Pour les 10 valeurs de ($\boldsymbol{\lambda}$, $\boldsymbol{\rho}$) menant aux meilleurs modèles selon le critère BIC, ré-estimation des modèles sans limitation du nombre d'itérations de l'algorithme EM.
- Sélection, parmi ces 10 modèles, du meilleur modèle selon le critère BIC.

Cette procédure a plusieurs avantages. Premièrement, les paramètres de régularisation évalués sont spécifiques aux données et en particulier à chaque groupe, puisque les valeurs à tester sont fixées pour chaque groupe de la classification initiale. De plus cette procédure permet la sélection simultanée des hyperparamètres réglant la sélection au sein des régresseurs, et des hyperparamètres introduisant la parcimonie au sein des matrices de précision. Cependant, cette procédure est numériquement coûteuse, puisque pour un modèle à K groupes et l valeurs de λ testées pour chaque groupe, il faut estimer $3 * l^K$ modèles.

3.2.3 Sélection du nombre de groupes

La sélection du nombre de groupes est effectuée selon les critères AIC, BIC et ICL décrits au chapitre précédent. Le modèle est estimé pour différentes valeurs de K, et le modèle sélectionné est celui qui minimise le critère choisi. Par la suite, les trois critères sont comparés. L'AIC étant particulièrement adapté aux problèmes de prédiction, on portera une attention particulière sur ce critère.

3.3 Algorithme Espérance-Maximisation

3.3.1 Formulation

Un algorithme Espérance-Maximisation (EM) est utilisé pour optimiser (3.2). On note $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ les données complétées, avec $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$ référant aux données non observées, où $z_{ik} = 1$ si l'observation i appartient au groupe k, 0 sinon. La vraisemblance des données complétées peut se décomposer en

$$\begin{aligned}\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \boldsymbol{\Phi}) = &\ln \mathcal{L}(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \boldsymbol{\beta}) \\ &+ \ln \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_n | \mathbf{z}_1, \ldots, \mathbf{z}_n; \boldsymbol{\mu}, \boldsymbol{\Theta}) \\ &+ \ln \mathcal{L}(\mathbf{z}_1, \ldots, \mathbf{z}_n; \boldsymbol{\pi}).\end{aligned}$$

L'algorithme EM classique consiste en la maximisation de l'espérance conditionnelle de la log-vraisemblance complétée (conditionnellement aux données observées), sachant un ensemble de paramètres fixé $\boldsymbol{\Phi}^*$, au lieu de maximiser la vraisemblance seule. Dans notre situation, nous considérons l'espérance conditionnelle de la log-vraisemblance complétée pénalisée telle que

$$\begin{aligned}\arg\max_{\boldsymbol{\Phi}} \quad \Bigg\{ &\mathbb{E}_{\boldsymbol{\Phi}^*} \left[\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \boldsymbol{\Phi}) | \mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n \right] \\ &- \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k\|_1 \Bigg\}.\end{aligned}$$

L'étape E de l'algorithme reste inchangée dans notre contexte, et permet de prédire les groupes latents non observés par leur espérance conditionnelle pour tous les individus $i = 1, \ldots, n$ en utilisant les probabilités *a posteriori* $\tau_{ik} = \mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i, Y_i = y_i; \boldsymbol{\Phi}_k^{[h]})$ pour tout $k = 1, \ldots, K$, sachant les paramètres courants $\pi_k^{[h]}, \boldsymbol{\mu}_k^{[h]}, \Sigma_k^{[h]}, \boldsymbol{\beta}_k^{[h]}$, à l'itération $[h]$. Pour tout $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, K$, elles sont données par

$$\tau_{ik}^{[h+1]} = \frac{\pi_k^{[h]} f_{\mathbf{X},Y} \left(\mathbf{x}_i, y_i; \boldsymbol{\mu}_k^{[h]}, \Sigma_k^{[h]}, \boldsymbol{\beta}_k^{[h]} \right)}{\sum_{\ell=1}^{K} \pi_\ell^{[h]} f_{\mathbf{X},Y} \left(\mathbf{x}_i, y_i; \boldsymbol{\mu}_\ell^{[h]}, \Sigma_\ell^{[h]}, \boldsymbol{\beta}_\ell^{[h]} \right)},$$

avec $f_{\mathbf{X},Y}(.)$ la fonction de densité jointe de $(\mathbf{X}, Y)$. La densité jointe est calculée grâce à la relation $f_{\mathbf{X},Y}(.) = f_{Y|\mathbf{X}}(.) f_{\mathbf{X}}(.)$ pour laquelle les distributions sont connues comme étant une distribution binomiale de paramètre $p^{(k)}(\mathbf{x})$ pour $Y|\mathbf{X} = \mathbf{x}$ et une distribution gaussienne de paramètres $\boldsymbol{\mu}_k$ et Σ_k pour $\mathbf{X}$.

On calcule alors l'espérance de la log-vraisemblance complétée **pénalisée** $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{[h]})$

sachant les données observées $(y_i, \mathbf{x}_i)_{i=1,\dots,n}$ et $\mathbf{\Phi}^{[h]}$,

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{[h+1]} \left(\ln\pi_k^{[h]} - \frac{1}{2}\left(y_i \mathbf{x}_i^t \boldsymbol{\beta}_k^{[h]} - \mathbf{x}_i^t \boldsymbol{\beta}_k^{[h]} - 1 \right) + \frac{1}{2}\ln|\Theta_k^{[h]}| \right.$$
$$\left. -\frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{[h]}\right)^T \Theta_k^{[h]} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{[h]}\right) \right) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k^{[h]}\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k^{[h]}\|_1.$$

Ensuite, l'étape M consiste à mettre à jour chaque paramètre à l'itération $[h+1]$ et on a, pour tout $k = 1, \dots, K$,

$$\widehat{\pi}_k^{[h+1]} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{[h+1]},$$

$$\widehat{\boldsymbol{\mu}}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]}\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{x}_i,$$

$$\hat{\Theta}_k^{[h+1]} = \arg\max_{\Theta_k} \left\{ \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(\log\det\Theta_k - \frac{1}{2}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h+1]}\right)^T \Theta_k \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h+1]}\right) \right) - \rho_k \|\Theta_k\|_1 \right\},$$

$$\hat{\boldsymbol{\beta}}_k^{[h+1]} = \arg\max_{\boldsymbol{\beta}_k} \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(y_i \mathbf{x}_i^T \boldsymbol{\beta}_k - \ln\left(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)\right)\right) \right] - \lambda_k \|\boldsymbol{\beta}_k\|_1.$$

Pour chaque groupe, les régresseurs $\boldsymbol{\beta}_k$ sont estimés grâce à une régression logistique pénalisée de type Lasso avec un paramètre de régularisation λ_k, où les observations sont pondérées par leur probabilité *a posteriori* d'appartenance au groupe k. De la même façon, les matrices de précision sont estimées par groupe k grâce à un algorithme Graphical Lasso avec un paramètre de régularisation ρ_k.

3.3.2 Réglage de l'algorithme EM

La convergence de l'algorithme EM vers la solution optimale peut dépendre fortement des paramètres initiaux. Par conséquent, l'algorithme doit démarrer à partir de paramètres initiaux raisonnables pour éviter la convergence vers un maximum local de la fonction de vraisemblance. Ici, nous adoptons une stratégie Research/Run/Selection, comme celle développée dans Biernacki et al. (2003) qui consiste à :

- Trouver t positions initiales des paramètres : obtenir une partition de l'ensemble des observations des variables explicatives $(\mathbf{x}_i)_{i=1,\dots,n}$ en K groupes avec un algorithme K-means (Macqueen, 1967). Suivant cette partition, calculer des estimateurs de la régression logistique dans chaque groupe. Répéter ces deux étapes pour chacun des t essais.

- Lancer un faible nombre, fixé, d'itérations de l'algorithme EM pour chacune des t positions initiales trouvées précédemment.

- Parmi ces t valeurs de départ possibles, sélectionner celle qui maximise la log-vraisemblance pour démarrer l'algorithme EM.

Les étapes E et M sont répétées tant que la log-vraisemblance n'augmente pas plus qu'un seuil fixé, ou qu'un nombre maximum d'itérations n'est atteint.

3.4 Expérimentations sur données simulées

Une étude de simulation est effectuée pour évaluer les performances de notre modèle de mélange de régressions pénalisées. Nos objectifs sont (i) d'évaluer la qualité de l'estimation des paramètres, (ii) d'évaluer la performance de prédiction de notre modèle et (iii) d'évaluer l'intérêt d'utiliser notre modèle pour traiter des données non homogènes, où des groupes latents modulent la règle de prédiction.

3.4.1 Méthodes alternatives et critères d'évaluation

Les méthodes sont comparées sur 30 répétitions de chaque scénario de simulation. La méthode est évaluée selon deux points de vue : estimation et prédiction.

Pour évaluer les estimateurs, on compare les performances en estimation de la méthode proposée, appelée PMLR (pour Penalized Mixture of Logistic Regression) avec la régression logistique pénalisée, appelée PLR, et avec un mélange de régressions logistiques noté MLR. Le modèle PLR est défini en section 1.3.2 du chapitre 1, ainsi que la vraisemblance pénalisée à maximiser. Le modèle MLR correspond au mélange de régressions dans le cas des modèles linéaires généralisés, décrit en section 2.2 du chapitre 2. On note que les principales différences entre cette méthode et la nôtre sont la pénalisation, mais aussi le design aléatoire permettant la prédiction.

Le nombre de groupes varie pour l'étape d'estimation, et les performances de la procédure de sélection sont étudiées pour l'AIC, le BIC et l'ICL.

Le biais et la variance des estimateurs $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\beta}}$ sont estimés à partir des 30 répétitions de chaque scénario.

Les supports de $\boldsymbol{\beta}$ et $\boldsymbol{\Theta}$ sont aussi comparés avec les vrais supports, et résumés grâce aux critères de Relevant Variable Detection (RVD) et Irrelevant Variable Elimination (IVE) [1] des coefficients de régression. Le RVD d'un estimateur $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,p})$ est défini par

$$RVD_k = \frac{TP_k}{TP_k + FN_k},$$

[1] Ces critères correspondent à la sensibilité et spécificité en évaluation d'un classifieur binaire, renommés ici pour éviter toute confusion.

avec TP_k le nombre de coefficients $(\beta_{k,j})_{j=1,...,p}$ correctement prédits comme non-zéro et FN_k le nombre de coefficients $(\beta_{k,j})_{j=1,...,p}$ prédits comme zéro à tort. L'IVE d'un estimateur $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,p})$ est défini par

$$IVE_k = \frac{TN_k}{TN_k + FP_k},$$

avec TN_k le nombre de coefficients $(\beta_{k,j})_{j=1,...,p}$ correctement prédits à zéro et FP_k le nombre de coefficients $(\beta_{k,j})_{j=1,...,p}$ prédits non-zéro à tort. Le RDV est égal à 1 si toutes les variables pertinentes sont retenues dans le modèle. L'IVE est égal à 1 si toutes les variables non-pertinentes sont éliminées du modèle.

La performance de la classification est évaluée grâce à l'index de Rand ajusté (ARI), qui mesure la similarité entre deux partitions (Hubert and Arabie, 1985). Si on considère un ensemble de n observations, groupées selon deux partitions P et P' comprenant respectivement r et s groupes, avec $P = \{P_1, P_2, \ldots, P_r\}$ et $P = \{P'_1, P'_2, \ldots, P'_s\}$, on peut résumer le chevauchement entre les deux partitions dans la table de contingence suivante

	P'_1	P'_2	...	P'_s	Somme
P_1	n_{11}	n_{12}	...	n_{1s}	a_1
P_2	n_{21}	n_{22}	...	n_{2s}	a_2
...	...	...	...	...	...
P_r	n_{r1}	n_{r2}	n_{rs}	n_{11}	a_r
Somme	b_1	b_2	...	b_s	

avec n_{rs} le nombre de fois où une observation appartient simultanément au groupe r de P et au groupe s de P'.

L'index de Rand ajusté est défini par :

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{n}{2}}.$$

Une valeur d'ARI égale à 1 signifie que la partition prédite est exactement la même que la partition théorique. Ce critère a pour avantage d'être adapté à la comparaison de partitions avec des nombres de groupes différents, et n'est pas sensible aux permutations.

Pour évaluer la qualité de prédiction de la réponse binaire, on compare les résultats obtenus par PMLR avec ceux obtenus par PLR et par régression logistique (notée LR), qui sont des méthodes adaptées à la prédiction. On considère comme critère de qualité de prédiction l'aire sous la courbe ROC, notée AUROC. Les performances de prédiction sont évaluées pour un nombre fixé de groupes K.

3.4.2 Cadre de simulation

Génération des données Tout d'abord, n observations issues d'une distribution multinomiale sont générées, avec $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ le vecteur des proportions de chaque groupe k : $\mathbf{Z} \sim \mathcal{M}(n, \boldsymbol{\pi})$. Ensuite, sachant la variable simulée $\mathbf{Z}$, pour chaque individu $i = 1, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$ est généré selon une distribution gaussienne multivariée telle que $X_i \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, avec $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$ la moyenne et la matrice de covariance associées au groupe k. Enfin, la variable réponse Y est générée selon une distribution binomiale de paramètres n et p, avec $p^{(k)}(\mathbf{x}_i) = \mathbb{P}(Y_i = 1 | \{\mathbf{X}_i = \mathbf{x}_i, Z_{ik} = 1\})$ pour $i = 1, \ldots, n$.

Paramètres de simulation Les données sont simulées à partir d'un mélange à $K = 3$ groupes de proportions $\boldsymbol{\pi} = (0.3, 0.3, 0.4)$, avec $n = 250$ ou 500 observations, et $p = 40$ variables, avec seulement 8 variables pertinentes (dont les coefficients sont différents de zéro). Quatre scenarii de simulation sont étudiés, et présentés en table 3.1.

Dans les deux premiers cas, les coefficients différents de zéro concernent les mêmes variables dans tous les groupes (même support), mais la concentration des groupes et l'équilibre entre les deux classes dans les groupes sont différents (ce qui correspond à un cas (1) facile et un cas (2) difficile). Pour les cas trois et quatre, les coefficients différents de zéro concernent des variables différentes selon le groupe (supports différents). Dans le cas 4 (cas difficile), la partition repose sur $Y|\mathbf{X}$ alors que les moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_3$ sont égales, ce qui mène à deux groupes très similaires. Dans le cas 3 (cas facile), la partition repose sur $Y|\mathbf{X}$ et $\mathbf{X}$, mais les moyennes des groupes sont toutes différentes.

3.4.3 Résultats et interprétations

Sélection de modèles

Le vote majoritaire de la sélection de modèles selon la valeur la plus faible des critères est calculé sur les 30 répétitions pour chaque scénario pour l'AIC, le BIC et l'ICL, pour des modèles estimés avec 1 à 4 groupes. Pour chaque critère et chaque scénario, le nombre de groupes le plus fréquemment sélectionné est résumé dans la table 3.2. Pour notre méthode, l'AIC mène à la sélection du bon nombre de groupes dans presque tous les cas, sauf pour le cas 4 où deux groupes sont très similaires. L'information la plus importante pour le partitionnement est portée par $\mathbf{X}$ et dans ce cas, le modèle à deux groupes est sélectionné. Le critère ICL et le critère BIC mènent à la sélection du mauvais nombre de groupes pour 5 cas sur 8. Cependant, ces critères mènent à une meilleure sélection de modèle pour notre méthode que pour un mélange de régressions classique, où les critères mènent à un modèle à 1 groupe dans presque tous les cas. Pour conclure, la sélection de modèles est meilleure avec notre méthode selon ces critères, et on se concentrera par la suite sur le critère AIC qui a les meilleures performances dans le cas de la sélection du nombre de groupes latents.

	k	Mêmes supports - Cas facile (1)	Mêmes supports - Cas difficile (2)
	1	$(-1, 1, 0.5, 1, \mathbf{0}_{32}, -1, -1, 0.5, 0.2)$	$(-2, 1, 0.5, 1, \mathbf{0}_{32}, -1, -0.2, 0.5, 0.2)$
$\boldsymbol{\beta}$	2	$(1, 0.5, 0.5, -1, \mathbf{0}_{32}, 1, -0.2, 1, -0.5)$	$(1, -0.5, -0.5, -1, \mathbf{0}_{32}, 1, -0.2, 1, -0.5)$
	3	$(-1, 0.5, 1, -1, \mathbf{0}_{32}, 1, 1, 2, -1)$	$(-2, 0.5, -2, -1, \mathbf{0}_{32}, 1.5, 1, 2, -1)$
	1	$(-\mathbf{1}_{20}, \mathbf{1}_{20})$	$(-\mathbf{2}_{20}, \mathbf{1}_{20})$
$\boldsymbol{\mu}$	2	$\mathbf{1}_{40}$	$\mathbf{1}_{40}$
	3	$(\mathbf{1}_{20}, \mathbf{2}_{20})$	$\mathbf{3}_{40}$
	1	diag(0.5)	diag(2/3)
$\boldsymbol{\Sigma}$	2	band$(0.5, 0.2, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.5, 0.1, 0.1, \mathbf{0}_{36})$
	3	band$(0.8, 0.4, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.4, 0.1, 0.1, \mathbf{0}_{36})$
	k	Supports différents - Cas facile (3)	Supports différents - Cas difficile (4)
	1	$(-2, 1, 0.5, 1, \mathbf{0}_{32}, -1, -0.2, 0.5, 0.2)$	Même paramètre que le cas 3
$\boldsymbol{\beta}$	2	$(\mathbf{0}_4, 1, -0.5, -0.5, -1, \mathbf{0}_{16}, 1, -0.2, 1, -0.5, \mathbf{0}_4)$	Même paramètre que le cas 3
	3	$(\mathbf{0}_8, -2, 0.5, -2, -1, 1.5, 1, 2, -1, \mathbf{0}_{16})$	Même paramètre que le cas 3
	1	$(-\mathbf{1}_{20}, \mathbf{1}_{20})$	$(\mathbf{1}_{20}, \mathbf{2}_{20})$
$\boldsymbol{\mu}$	2	$\mathbf{1}_{40}$	$\mathbf{1}_{40}$
	3	$(\mathbf{1}_{20}, \mathbf{2}_{20})$	$(\mathbf{1}_{20}, \mathbf{2}_{20})$
	1	diag(0.5)	diag(2/3)
$\boldsymbol{\Sigma}$	2	band$(0.8, 0.3, 0.1, 0.1, \mathbf{0}_{36})$	diag(2/3)
	3	band$(0.6, 0.3, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.4, 0.1, 0.1, \mathbf{0}_{36})$

TABLE 3.1 : **Paramètres utilisés pour les 4 scenarii de simulation.** Pour définir un modèle, $\boldsymbol{\beta}, \boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ doivent être définis. Comme on se place dans un cas où le modèle de mélange a 3 groupes, il y a 3 paramètres différents pour chaque cas (le groupe est représenté par k). Dans les cas 1 et 2, les variables pertinentes sont les mêmes pour les groupes. Les différences portent sur la concentration des groupes et l'équilibre entre les deux classes (représentées par Y) dans chaque groupe. Dans les cas 3 et 4, les variables pertinentes sont différentes selon les groupes. Dans le cas 3, la partition repose sur $Y|\mathbf{X}$ et $\mathbf{X}$. Dans le cas 4, la partition repose sur $Y|\mathbf{X}$ alors que $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_3$ ont les mêmes valeurs. Les matrices de covariance $\boldsymbol{\Sigma}$ sont des matrices diagonales et bande-diagonales.

Performances d'estimation

La qualité de la partition est évaluée grâce à l'index de Rand ajusté (ARI), détaillé dans la table 3.3. Les valeurs d'ARI sont plus élevées pour les partitions obtenues avec notre méthode que pour les partitions obtenues avec un mélange de régressions logistiques réalisé sur la distribution conditionnelle de la variable réponse sachant les covariables (méthode MLR), même pour des modèles avec le mauvais nombre de groupes (K=2). Les performances sont similaires pour $n = 250$ et $n = 500$, l'asymptotique est donc déjà atteinte. Les données ont été générées selon le modèle PMLR dans lequel la distribution jointe $(\mathbf{X}, Y)$ est considérée dans le mélange. L'information portée par $\mathbf{X}$ est utilisée explicitement pour le partitionnement. Dans le cas du MLR, seule la distribution conditionnelle est utilisée dans la tâche d'estimation, donc le partitionnement repose sur la relation entre $\mathbf{X}$ et Y. Ces résultats montrent que si les covariables sont

Critère	AIC				BIC				ICL			
Méthode	PMLR		MLR		PMLR		MLR		PMLR		MLR	
Taille d'échantillon	250	500	250	500	250	500	250	500	250	500	250	500
Mêmes supports												
Cas facile (1)	**3**	**3**	1	1	2	2	1	1	2	2	1	1
Cas difficile (2)	**3**	**3**	1	1	**3**	**3**	1	1	**3**	**3**	1	1
Supports différents												
Cas facile (3)	**3**	**3**	1	2	**3**	2	1	1	**3**	2	1	1
Cas difficile (4)	2	2	1	2	1	2	1	1	1	1	1	1

TABLE 3.2 : **Nombre de groupes sélectionnés pour chaque cas de simulation.** Les méthodes PMLR (Penalized Mixture of Logistic Regression) et MLR (Mixture of Logistic Regression) sont comparées. Pour chaque méthode, le nombre de groupes permettant d'obtenir la plus faible valeur des critères AIC, BIC et ICL le plus grand nombre de fois parmi les 30 répétitions pour chacun des 4 cas de simulation est donné. Les simulations sont réalisées pour deux tailles d'échantillon (n=250 et n=500). Le vrai nombre de groupes ($K = 3$) est représenté en gras.

Méthode	PMLR		MLR	
Taille d'échantillon n	250	500	250	500
Mêmes supports - cas facile (1)	0.89	0.91	0	0.01
Mêmes supports - cas difficile (2)	1	1	0.01	0.01
Supports différents - cas facile (3)	0.96	0.97	0.01	0.01
Supports différents - cas difficile (4)	0.42	0.42	0	0.01

TABLE 3.3 : **Performances de partitionnement évaluées par l'index de Rand ajusté (ARI) pour les 4 cas de simulation.** Les méthodes PMLR (Penalized Mixture of Logistic Regression) et MLR (Mixture of Logistic Regression) sont comparées. Pour chaque méthode, pour un nombre de groupe fixé à $K = 3$, l'ARI est calculé, plus la valeur est proche de 1, meilleure est la partition. Le calcul est effectué pour les 30 répétitions de chacun des 4 cas de simulation. Les simulations sont réalisées pour deux tailles d'échantillon ($n = 250$ et $n = 500$).

aussi structurées en groupes, la méthode MLR ne montre pas de bonnes performances, puisque le partitionnement ne se base pas sur les données portant la structure. Comme indiqué dans le chapitre précédent, le mélange de régression ne permet pas de considérer l'information portée sur les covariables pour évaluer la structure des données, et n'est donc pas adapté à la prédiction d'un nouvel individu dont on ne connaît pas la classe d'appartenance.

La figure 3.1 montre la capacité à retrouver les variables pertinentes pour tous les scenarii de simulation. On remarque que la méthode proposée permet de meilleurs résultats que la régression logistique pénalisée. La capacité à éliminer les variables non pertinentes est légèrement plus faible pour notre méthode dans les cas 1 et 2, et similaire à la régression logistique pénalisée dans les cas 3 et 4. Ces résultats montrent que notre

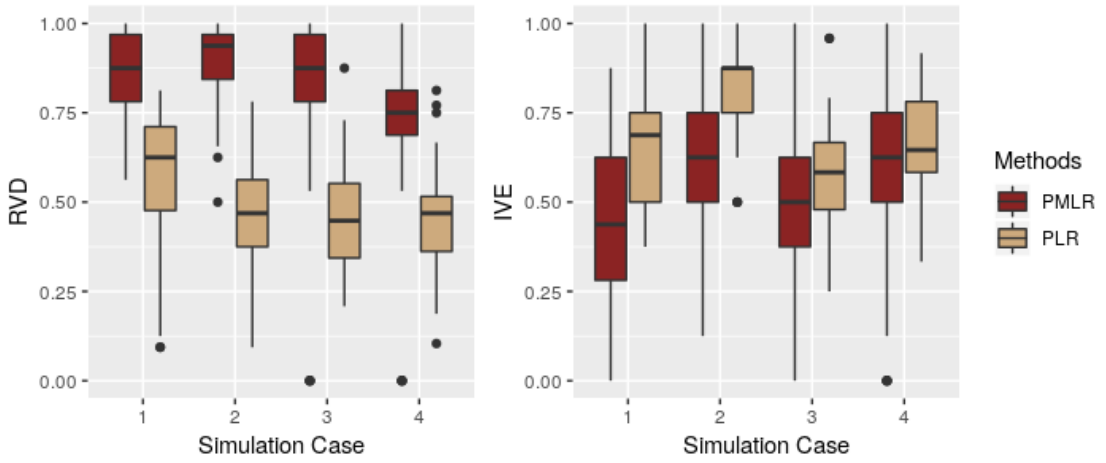

FIGURE 3.1 : **Performances de la sélection de variables pour $\hat{\beta}$ pour les 4 cas de simulation.** Les boîtes à moustaches pour la Relevant Variable Detection (RVD) sont représentées à gauche et les boîtes à moustaches pour la Irrelevant Variable Elimination (IVE) sont représentées à droite, pour chacun des 4 cas de simulation. Notre méthode PMLR est comparée à la régression logistique pénalisée (PLR), qui sélectionne les variables dans la matrice de régression avec une pénalisation ℓ_1. Les paramètres de régularisation sont sélectionnés avec le BIC.

méthode sélectionne trop de variables, mais réussit à trouver les variables pertinentes, alors que la régression logistique pénalisée élimine trop de variables, dont des variables importantes. En médecine, il est particulièrement intéressant de ne pas éliminer des variables pertinentes, car l'interprétation du modèle est importante.

On prête aussi attention au biais et à la variance des estimateurs. Les performances sont bonnes pour notre méthode, et meilleures avec un échantillon plus grand (ce qui est lié au fait que le maximum de vraisemblance a de bonnes performances asymptotiquement).

Performances de prédiction

Les performances de prédiction sont montrées en figure 3.2. Notre méthode permet clairement de meilleures performances pour les scenarii 2 et 3. Le cas 1 étant facile, toutes les méthodes ont des bonnes performances. Concernant le cas 4, les meilleures performances sont obtenues avec notre méthode mais le mauvais nombre de groupes est choisi à cause des fortes similarités entre deux groupes. On peut aussi conclure que notre méthode a une faible variabilité dans les AUROC sur les 30 répétitions.

Pour illustrer ces résultats, les courbes ROC obtenues pour les 30 répétitions des scenarii de simulation 2 et 3 pour les méthodes PMLR (avec 3 groupes), PLR et LR sont représentées en figure 3.3. Pour ces cas, les courbes ROC obtenues avec notre méthode de prédiction sont au dessus des courbes ROC obtenues avec les deux autres méthodes alternatives, montrant de meilleures performances de prédiction dans ces cas avec notre

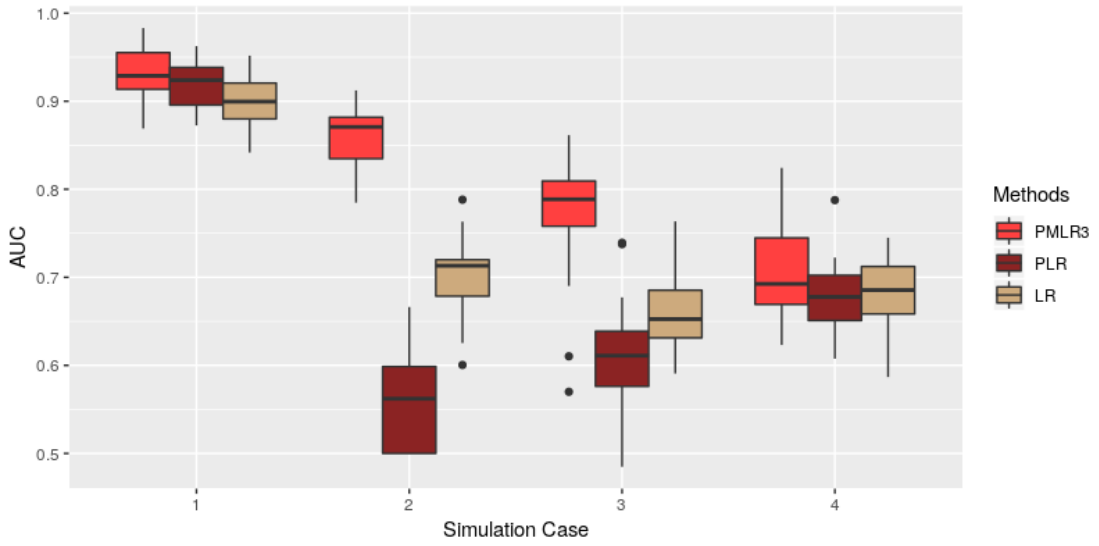

FIGURE 3.2 : **Performances de prédiction pour les 4 cas de simulation.** Les boîtes à moustaches représentant la répartition des valeurs d'aire sous la courbe ROC (AUROC) obtenues pour les 30 répétitions de chacun des 4 cas de simulation sont représentées pour chaque méthode : régression logistique (LR), régression logistique pénalisée (PLR) et mélange de régressions logistiques pénalisées à trois groupes (PMLR).

méthode. On remarque que PLR et LR ont des courbes très proches, alors que leurs AUROC sont différentes.

3.5 Application aux données NASH

Sur des applications biologiques, on fait habituellement face à des effets individuels changeant la règle de prédiction. Dans le cadre de la NASH, nous supposons qu'il existe des groupes homogènes au sein des observations, reposant sur des similarités biologiques ou génétiques. Ces similarités peuvent être indépendantes de la sévérité de la maladie que l'on cherche à diagnostiquer. Une meilleure prédiction de la maladie peut être obtenue en considérant cette structure en groupes, et de manière plus importante, une meilleure compréhension des processus d'évolution de la maladie est possible avec notre modèle. Nous appliquons donc dans cette partie la méthode PMLR décrite précédemment pour construire un modèle de diagnostic de la NASH.

3.5.1 Description des données

Pour étudier les performances de la méthode PMLR sur données réelles, on considère les données NASH décrites en section 1.1.4 du chapitre 1. Des portions des courbes de spectrométrie ont été pré-sélectionnées *a priori* par des experts, pour leur capacité à décrire les variations métaboliques qui pourraient être liées à l'atteinte hépatique des

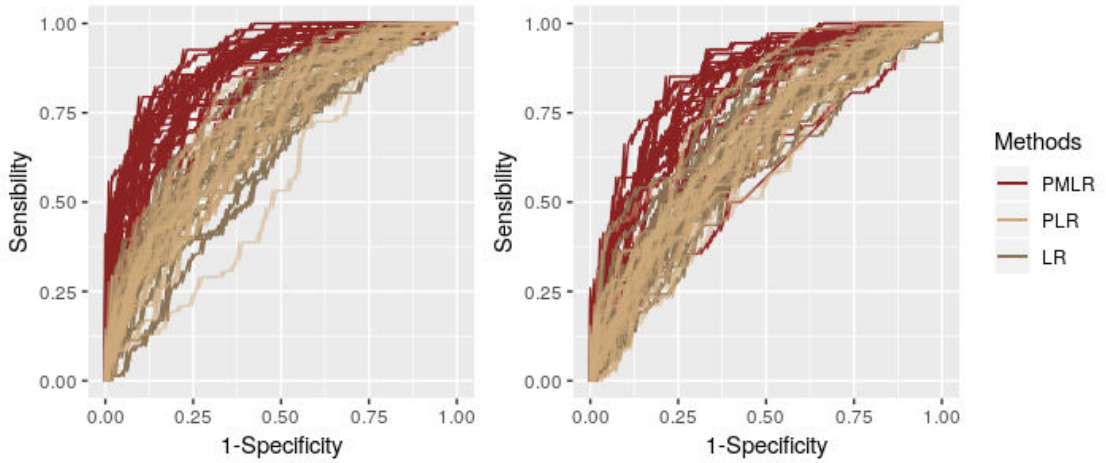

FIGURE 3.3 : **Performances de prédiction pour l'ensemble des répétitions des cas de simulation 2 et 3.** Les courbes ROC sont représentées pour chacune des méthodes : régression logistique (LR), régression logistique pénalisée (PLR) et mélange de régressions logistiques pénalisées à trois groupes (PMLR). Le graphe de gauche correspond au cas de simulation 2, et le graphe de droite correspond au cas de simulation 3. Chaque cas de simulation est répété 30 fois, et une courbe est représentée pour chaque répétition pour mettre en évidence la variance.

patients. Les variables spectrales sont utilisées pour construire le modèle de prédiction, alors que les variables cliniques et biologiques sont utilisées pour interpréter les résultats. Les variables cliniques disponibles sont décrites dans la table 1.1. Dans cette partie, les résultats obtenus avec notre procédure PMLR sont décrits, en termes d'estimation et d'interprétation statistique et biologique.

3.5.2 Analyses et résultats

Sélection de modèles

Les données sont séparées aléatoirement en un échantillon d'apprentissage contenant 4/5 des individus (316 individus incluant 53 patients atteints de NASH) et un échantillon de validation (contenant 79 individus dont 13 patients NASH). Ces ensembles sont choisis de manière aléatoire, mais contiennent la même proportion de patients NASH, et aucune différence significative n'est relevée sur les variables cliniques entre ces deux ensembles. Le modèle est estimé sur l'échantillon d'apprentissage, pour 1 à 3 groupes. Les critères de sélection de modèles sont calculés pour chaque modèle et présentés en table 3.4. Les plus faibles valeurs d'AIC et de BIC sont obtenues pour le modèle estimé avec deux groupes. La plus faible valeur d'ICL est obtenue pour le modèle à un seul groupe, mais avec une légère différence avec la valeur d'ICL correspondant au modèle à deux groupes. Suivant les conclusions détaillées en section 3.4.3, nous gardons le modèle à deux groupes selon la valeur d'AIC.

	K = 1	K = 2	K = 3
AIC	-50180	**-50459**	-30957
BIC	-49936	**-49970**	-30627
ICL	**-49936**	-49901	-30365

TABLE 3.4 : **Sélection de modèles sur les données NASH.** Comparaison des critères de sélection de modèles AIC, BIC et ICL pour les modèles estimés par PMLR sur l'échantillon d'apprentissage du jeu de données NASH, pour 1 à 3 groupes. Les valeurs en gras indiquent la meilleure valeur de critère obtenue.

Estimateurs et modèles

Les modèles graphiques obtenus à partir des matrices de précision parcimonieuses estimées pour chaque groupe sont présentés en figure 3.4. On observe des relations différentes entre les variables selon le groupe. En considérant seulement les relations entre les variables, on note que pour le premier groupe, il y a un groupe de variables allant de X2 à X11 très liées. Pour le second groupe, on observe deux groupes de variables fortement liées : d'une part les variables X2, X3, X4, X6, X7, X8 et d'autre part les variables X1, X5, X12, X14, X17 et X19. Les liens entre les variables sont complètement différents selon le groupe. La couleur du nœud représente la valeur du coefficient estimé pour la variable pour le modèle qui s'applique au groupe considéré. Pour le premier groupe, on observe que beaucoup de coefficients de régression sont proches ou égaux à zéro. Pour le second groupe, les coefficients ont des valeurs plus extrêmes. Les liens entre variables et les effets complètement différents sur la prédiction pour chaque groupe suggèrent qu'il existe différents mécanismes métaboliques chez les patients, selon le groupe. On remarque que les variables qui servent à trouver les groupes sont différentes de celles qui entrent dans le modèle logistique. On peut supposer que les différences entre les matrices de précision servent à former les groupes.

Interprétations statistiques du modèle construit

Les proportions des groupes sont égales à 0.66 et 0.34. La proportion de malades est différente selon le groupe : 19% dans le groupe 1 et 12% dans le groupe 2. Les performances obtenues avec PMLR1 et PLR sont similaires. En effet, quand on considère seulement un groupe, PMLR1 consiste en une régression logistique pénalisée.

Dans la table 3.5, on observe que les plus fortes valeurs d'AUROC et les meilleurs taux de bon classement sont obtenus pour le modèle à deux groupes, qui correspond aussi au modèle montrant les plus faibles valeurs d'AIC et BIC. La sélection de modèles avec les critères AIC et BIC est consistante avec les performances de validation croisée obtenues. Comparé aux méthodes compétitives, le modèle choisi montre les meilleures performances avec la plus forte valeur d'AUROC, le meilleur taux de bon classement, ainsi qu'une valeur prédictive négative élevée indiquant un bon test de dépistage de la NASH. La répartition des scores prédits selon la vraie classification binaire des individus de l'ensemble d'apprentissage est représentée en figure 3.5.

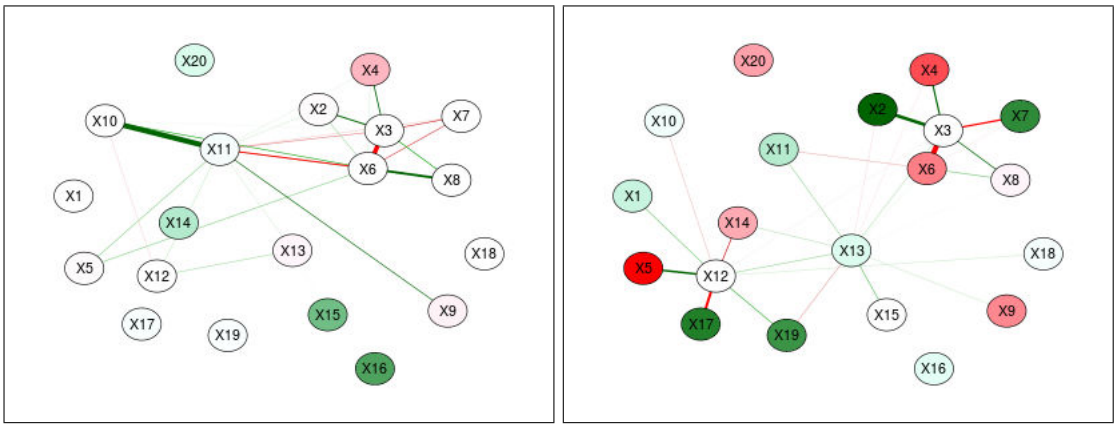

FIGURE 3.4 : **Modèles graphiques construits sur les matrices de précision estimées pour chaque groupe par PMLR.** Le modèle sélectionné a 2 groupes. Le réseau correspondant au premier groupe est représenté à gauche, et le réseau correspondant au second groupe est représenté à droite. Les matrices de précision sont estimées avec parcimonie, donc les réseaux sont parcimonieux. La couleur des flèches correspond au signe de la corrélation partielle (vert pour une corrélation positive et rouge pour une corrélation négative) et l'intensité des flèches correspond à la valeur de la corrélation (plus la couleur est intense, plus la corrélation est forte). La couleur des nœuds correspond à la valeur du coefficient de régression pour chaque variable pour le modèle appliqué au groupe considéré. Un nœud sans couleur indique une valeur de coefficient de régression égale à zéro.

	PMLR-1	PMLR-2	PMLR-3	PLR	LR
AUROC	0.64	**0.75**	0.68	0.64	0.67
Se	0.62	**0.77**	0.85	0.62	0.69
Sp	0.62	**0.76**	0.5	0.62	0.7
NPV	0.89	**0.94**	0.94	0.89	0.92
PPV	0.24	**0.38**	0.25	0.24	0.31
CR	0.62	**0.76**	0.56	0.62	0.7

TABLE 3.5 : **Comparaison des performances de prédiction obtenues sur les données NASH avec différentes méthodes.** Notre méthode estimée pour des modèles allant de 1 à 3 groupes (PMLR-1, PMLR-2, PMLR-3) est comparée à la régression logistique pénalisée (PLR) et à la régression logistique (LR). Le modèle choisi est indiqué en gras. Les quantités utilisées pour la comparaison sont les suivantes : aire sous la courbe ROC (AUROC), sensibilité (Se), spécificité (Sp), valeur prédictive négative (NPV), valeur prédictive positive (PPV), taux de bon classement (CR). Pour tous ces critères, plus la valeur est forte, meilleur est le modèle.

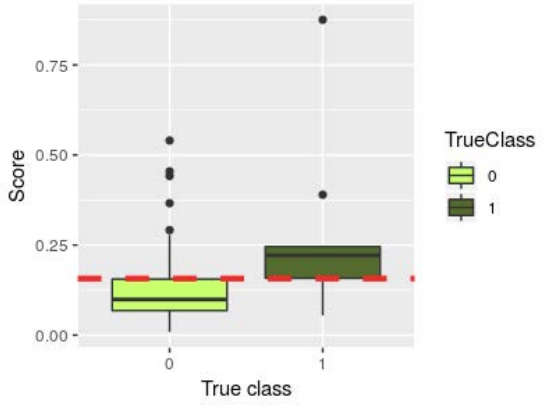

FIGURE 3.5 : **Performances de prédiction du modèle PMLR estimé sur les données NASH.** Boîtes à moustaches représentant la répartition des scores prédits par la méthode de mélange de régressions logistiques pénalisées selon la vraie classe (Y = 0 ou Y = 1). La ligne pointillée rouge correspond au seuil choisi de manière automatique sur l'échantillon d'apprentissage.

Le seuil à partir duquel un patient est prédit comme atteint de la NASH est choisi de façon automatique pour maximiser la somme de la sensibilité et de la spécificité. Ce seuil est représenté en figure 3.5 et permet une bonne séparation entre les patients NASH et les patients témoins.

Interprétations biologiques du modèle construit

On caractérise les groupes estimés par le modèle choisi avec les variables cliniques disponibles et résumées en table 3.6. Les valeurs correspondent à la moyenne calculée sur les individus pour chaque groupe. On compare les moyennes de chaque groupe avec des tests de Student, et on reporte les p-valeurs et leur significativité dans la table 3.6. La description des variables cliniques est présente en table 1.1. On note qu'il y a une différence significative entre les deux groupes pour les variables ALT, AST.ALT, GGT, HBA1C, chol et TG. Dans le premier groupe, les variables liées au diabète (Gluc et HBA1C) ont des valeurs plus élevées tout comme les variables indiquant la sévérité de l'atteinte au foie (ALT, GGT). Plus généralement, les patients du premier groupe semblent avoir des complications au foie plus avancées que les patients du second groupe, selon les indicateurs sériques. De plus, il n'y a pas de différence significative entre les deux groupes concernant les variables morphologiques (age, poids, taille, IMC), ce qui montre que le modèle permet de reconnaître la sévérité de l'atteinte au foie quand les patients ne sont pas différents morphologiquement.

Finalement, on représente la distribution des scores prédits en fonction des différents grades histologiques de stéatose, ballonisation, inflammation et fibrose en figure 3.6. On note que le score prédit est un bon indicateur des caractéristiques histologiques des patients. En effet, le score augmente avec le grade de stéatose, de ballonisation et

	Groupe 1	Groupe 2	p-value	Significativité
Age	40	39	0.6	
Sex	0.84	0.88	0.4	
Weight	119	120	0.4	
BMI	44	45	0.6	
Height	164	164	0.7	
AST	28	26	0.2	
ALT	38	29	0.001	**
AST.ALT	0.88	1	6.10^{-4}	**
GGT	47	34	10^{-3}	**
Gluc	6.2	5.7	0.06	
Insuline	24	21	0.2	
HBA1C	6.1	5.7	0.01	*
chol	5.5	5.2	4.10^{-3}	**
HDL	1.4	1.4	0.5	
LDL	3.2	3.1	0.4	
TG	2	1.4	4.10^{-7}	**

TABLE 3.6 : **Caractérisation des groupes obtenus par PMLR sur les données NASH grâce aux variables cliniques.** Le modèle sélectionné a deux groupes. Les valeurs représentées correspondent à la moyenne calculée sur tous les individus de chaque groupe. Pour chaque variable clinique, les moyennes par groupes sont comparées à l'aide d'un test de Student, dont la significativité de la p-valeur est reportée en dernière colonne. Pour la variable Sexe, le pourcentage de femmes est indiqué, et un test de Fisher est utilisé pour la comparaison.

d'inflammation, indicateurs utilisés pour établir le diagnostic à partir de l'échantillon de foie prélevé par biopsie. La fibrose n'entre pas en compte dans la définition de la NASH, ce qui explique l'absence de lien entre le score prédit par le modèle et le grade de fibrose.

3.6 Conclusion

Dans ce chapitre, nous avons présenté une méthode prédictive qui permet de construire un modèle sur des données structurées en groupe, incluant des variables non pertinentes. Cette méthode permet l'obtention d'outils interprétables pour aider à mieux comprendre les données, avec des performances de prédiction plus élevées que les méthodes compétitives. Ce travail a été mené sur des données réelles issues de spectrométrie infrarouge, pour développer un outil de diagnostic non invasif de prédiction de la NASH. Les résultats obtenus sont encourageant à la fois en termes de prédiction ainsi qu'en termes d'interprétation, avec des groupes caractérisés par des variables cliniques et un score de prédiction lié aux variables histologiques. De plus, il est commun dans les problèmes médicaux de faire face à des données structurées en groupes de patients aux profils inconnus. La méthode décrite dans ce chapitre ne prend pas en compte les spectres entiers

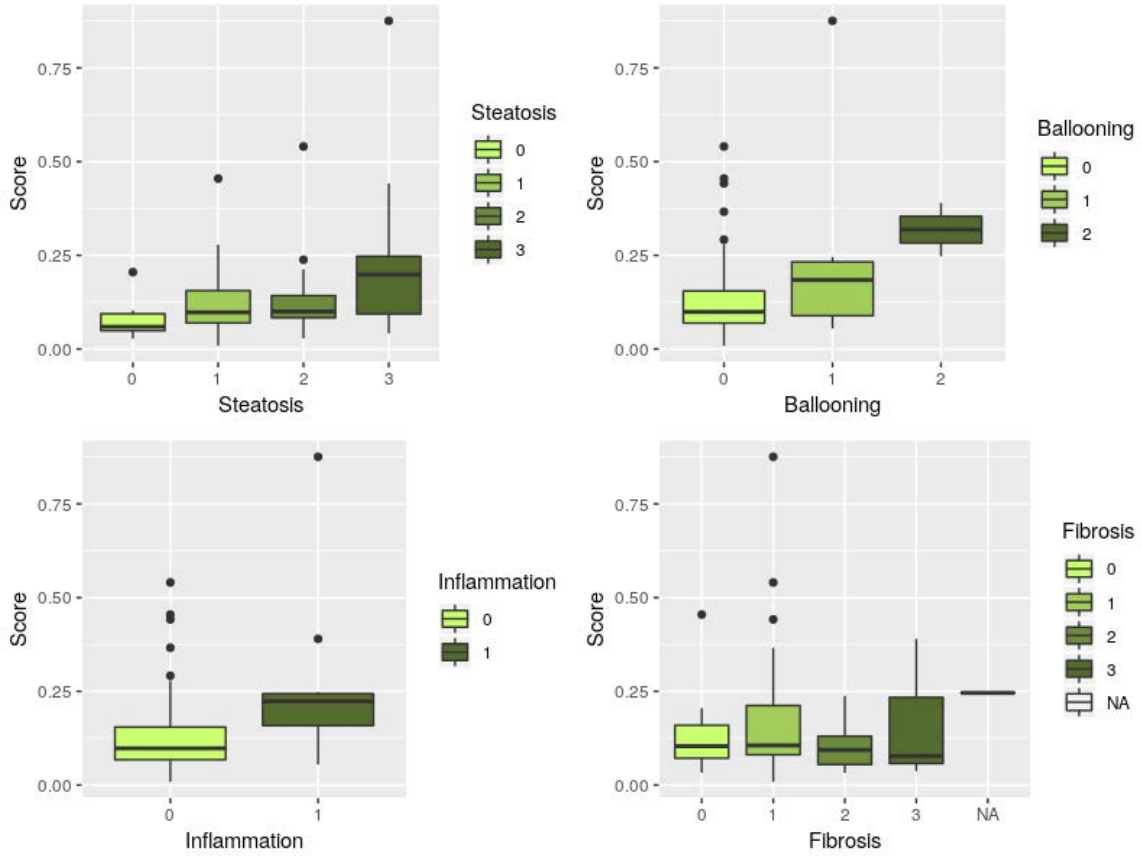

FIGURE 3.6 : **Boîtes à moustaches représentant la répartition des scores prédits par le modèle estimé sur les données NASH selon le grade de chaque variable histologique.** En haut à gauche : stéatose, en haut à droite : ballonisation, en bas à gauche : inflammation, en bas à droite : fibrose.

pour la construction du modèle, mais seulement des variables pré-sélectionnées dans le spectre. Ainsi, notre méthode peut être plus généralement utilisée pour traiter des données multivariées structurées en groupes.

Chapitre 4

Modèle appliqué aux données fonctionnelles

Contents

Dans le chapitre précédent, l'analyse se concentre sur une sélection de nombres d'ondes effectuée par des experts. Cependant, les données de spectrométrie sont des signaux mesurés sur un intervalle de nombres d'ondes, et il est plus cohérent de considérer

tout le spectre dans son entièreté, comme une courbe. Par conséquent, il apparaît pertinent d'adapter cette méthode pour la prise en compte de données fonctionnelles et la sélection de zones spécifiques du spectre. Cela pourrait mettre en évidence le type de molécules impliquées dans la maladie et offrir la possibilité de lier les différentes zones du spectre. De plus, l'information discriminante permettant la meilleure prédiction pourrait être portée à la fois par l'intensité des valeurs du spectre à des nombres d'ondes particuliers et par la forme du spectre à ces zones spécifiques. La généralisation du modèle proposé dans le chapitre 3 à l'analyse de données fonctionnelles est donc développée dans ce chapitre, ainsi qu'une méthode de pré-sélection des zones intéressantes du spectre sous forme de portions de courbes.

Après une rapide présentation de l'analyse de données fonctionnelles, nous détaillerons l'adaptation du modèle de mélange de régressions logistiques pénalisées à ce type de données. Une nouvelle procédure de pré-sélection des zones du spectre prenant en compte la particularité fonctionnelle des données et basée sur des tests multiples sera ensuite exposée. Dans un dernier temps, ces méthodes seront appliquées aux données NASH.

4.1 Les données fonctionnelles

Les outils de mesures actuels permettent de récolter de plus en plus de données sous forme de signaux, par exemple mesurés sur un intervalle de temps ou de longueurs d'ondes comme dans le cas de la spectrométrie. Les données obtenues représentent donc dans ces situations des courbes. Lorsque les données sont récoltées sur une grille de discrétisation fine et régulière, et lorsque chaque observation correspond à une courbe, alors on parle de données fonctionnelles. L'idée générale de cette approche, introduite par Ramsay and Silverman (1997), est de considérer les fonctions observées comme des entités singulières, plutôt que comme des successions d'observations, même si en pratique les données sont observées à un nombre fini de points de mesures.

Le traitement statistique de ce type de données est appelé analyse de données fonctionnelles, et se développe de plus en plus, suivant l'évolution des outils de récolte de données. On peut notamment se référer aux ouvrages de Ramsay and Silverman (1997) ou de Ferraty and Vieu (2006) qui ont largement développé les méthodes de traitement de données fonctionnelles.

Dans ce travail, les données de spectrométrie peuvent être considérées comme fonctionnelles puisque correspondant à des courbes mesurant l'absorbance en fonction du nombre d'onde. Suivant le contexte détaillé précédemment, nous nous plaçons dans le cas où les données observées correspondent à des variables explicatives fonctionnelles (les courbes de spectrométrie), et à une variable réponse scalaire (la variable binaire de diagnostic). Nous cherchons toujours à estimer la structure en groupe des données observées, en considérant maintenant leur caractère fonctionnel. De nombreuses méthodes sont utilisées pour gérer les données fonctionnelles, mais une des plus communes est de travailler sur des coefficients issus de la projection des données, détaillée par la suite.

4.1.1 Projection des données

On considère l'échantillon de courbes $(f_i)_{i=1,\dots,n}$ mesurées sur une grille de discrétisation $\{t_1, \dots, t_p\}$. Il est possible d'analyser l'échantillon $(f_i(t_j))$ pour tout $i = 1, \dots, n$ et tout $j = 1, \dots, p$, ce qui correspond à l'application de méthodes adaptées aux données multivariées en grande dimension. C'est l'approche qui a été étudiée dans les chapitres précédents.

En réalité, les observations f_i sont mesurées sous forme de courbes, dont on dispose des valeurs $(f_i(t_j))$ aux points $(t_j)_{j=1,\dots,p}$. Dans le cas de données fonctionnelles, c'est-à-dire correspondant à des fonctions, il serait possible d'obtenir une valeur de $f_i(t)$ pour tout t. Les données sont alors dans un espace de dimension infinie. Une des approches classiques est d'utiliser des combinaisons linéaires de fonctions de base pour représenter ces fonctions.

Le fait de projeter les données sur une base de fonctions permet de reconstruire la forme fonctionnelle des données supposées de dimension infinie, mais mesurées à des points de discrétisation. On considère alors réellement les données sous forme de courbes, en les modélisant dans un espace engendré par un nombre fini de fonctions de base. Cela permet de les lisser, mais aussi de réduire la dimension correspondant aux points de discrétisation.

Si on considère les projections sur une base de fonctions $B(t) = h_r(t)_{r\in\mathbb{N}^*}$, alors on peut approcher chaque observation (courbe) comme une combinaison linéaire de ces fonctions de base :

$$f_i(t) = \sum_{r=1}^{\infty} s_r(f_i) h_r(t) \approx \sum_{r=1}^{R} s_r(f_i) h_r(t),$$

avec $s_r(f_i)$ le coefficient de la base correspondant à l'observation i pour la fonction h_r et R la dimension de la base de fonctions retenue pour approcher f_i. Les méthodes d'estimation des coefficients de projection dans la base de fonctions, ainsi que le choix du nombre de fonctions de base à considérer, qui peut être difficile, sont notamment discutés dans Ramsay and Silverman (1997) et Ferraty and Vieu (2006). Si on note $\mathbf{S}_i = (\mathbf{s}_j(f_i))_{1\leq j\leq R}$ les coefficients de la projection pour l'individu i, on peut rassembler l'ensemble des coefficients dans une matrice $\mathbf{S}$ de taille $n \times R$ et on obtient alors l'écriture matricielle

$$f(t) = B(t)^T \mathbf{S}.$$

Dans le cas d'une base orthonormale, si on suppose que les $(f_i(t_j))_{i=1,\dots,n}$ suivent une loi normale pour un j donné, alors les coefficients $\mathbf{S}_i$ suivent une loi normale.

Choix de la base de fonctions Il existe de nombreuses bases de fonctions possibles, chacune ayant des propriétés la rendant plus ou moins adaptée à l'application considérée. Le choix de la base de fonctions doit donc se faire en accord avec le type de données étudiées ainsi que la problématique posée, et est une question cruciale en analyse de données fonctionnelles. Il existe des bases guidées par les données comme les bases construites à partir des composantes principales de l'ACP fonctionnelle (Besse and Ramsay, 1986) et d'autres dont les fonctions de base sont données *a priori* comme les ondelettes ou les splines. Les bases d'ondelettes sont utilisées dans de nombreux problèmes et sont

particulièrement adaptées dans le cas de discontinuités ou de changements brusques du signal. Elles sont par exemple utilisées dans le cas de courbes contenant des pics. Dans cette thèse, notre choix de base de projection se porte sur les bases de fonctions splines, détaillées par la suite.

Pour définir une fonction spline, on commence par diviser l'intervalle sur lequel la fonction doit être approchée en L sous-intervalles séparés par les valeurs $\iota_1, \ldots, \iota_{L-1}$ appelées nœuds. Sur chaque intervalle, une spline correspond à un polynôme d'ordre fixé m, égal à son degré additionné d'une unité. Les polynômes adjacents sont joints aux nœuds de telle sorte que leurs valeurs sont contraintes à être égales à leurs jonctions. Il est nécessaire que les polynômes soient de dérivées continues aux points de jonction jusqu'à l'ordre $m-2$.

Une base de splines est un système de fonctions de base qui sont chacune des splines, définies par un ordre m et une séquence de nœuds $\iota_1, \ldots, \iota_{L-1}$. Le système de base de fonctions splines défini par de Boor (2001), appelé B-splines, est le plus populaire, et est celui qui est considéré dans cette thèse. La figure 4.1 représente les 15 fonctions splines d'ordre 3, d'une base de B-splines définie par 12 nœuds équidistants, représentés par les lignes pointillées verticales. Les fonctions qui sont éloignées d'au moins deux sous-intervalles du bord de l'intervalle d'étude, sont positives sur 3 intervalles adjacents. Cela illustre la localisation conservée par la base de splines sur l'intervalle étudié. De manière générale, la base de B-splines sur laquelle les données sont projetées comprend $R = m + L - 1$ fonctions. Le nombre de sous-intervalles et l'ordre des splines définissent le nombre de fonctions de la base de projection.

Les bases de B-splines sont particulièrement adaptées dans le cas où les courbes à modéliser sont des fonctions lisses et régulières. Ce sont des bases flexibles et numériquement stables (Ramsay and Silverman, 1997). Ces propriétés en font des fonctions de base bien adaptées à l'étude des données de spectrométrie, qui sont des courbes lisses contenant des bandes d'absorption et non des pics. De plus, un avantage de la base de splines est la conservation de l'information de localisation sur la courbe que l'on peut considérer lorsque l'on interprète les résultats de modélisation. Ce type de base peut donc permettre de relier *a posteriori* les zones du spectre à des familles de molécules. Cette caractéristique peut aider à une meilleure compréhension du problème biologique étudié, ce qui oriente notre choix vers ce type de base dans notre cadre.

Cette étape de projection constitue la première étape de nombreuses méthodes d'analyse de données fonctionnelles. Après projection, l'information fonctionnelle de forme et de dimension est résumée par les coefficients de projection dans la base de fonctions. Ces coefficients sont donc considérés pour représenter les courbes mesurées pour les méthodes de régression ou de classification.

4.1.2 Régression fonctionnelle

Dans notre situation, nous souhaitons construire un modèle de régression sur les prédicteurs fonctionnels $\mathbf{X}(t)$, pour prédire la variable scalaire Y. Nous nous basons notamment sur le modèle logistique fonctionnel décrit par James (2002) pour notre modèle de diagnostic.

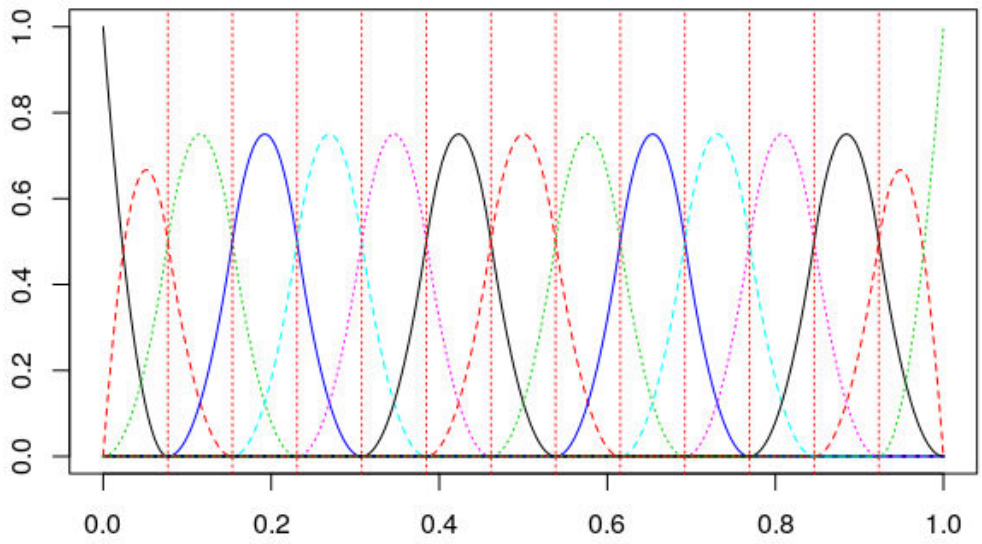

FIGURE 4.1 : **Exemple de base de B-splines utilisée pour approcher des données fonctionnelles.** Base de 15 fonctions splines d'ordre 3, définie par 12 nœuds équidistants, représentés par les lignes pointillées verticales.

Si on considère une variable réponse binaire Y, le modèle logistique s'écrit :

$$\text{logit}\big(p(\mathbf{x})\big) = \int \mathbf{w}(t)\mathbf{X}(t)dt,$$

avec $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$.

Comme évoqué précédemment, $\mathbf{X}(t)$ est seulement observé pour un ensemble fini de points. Nous supposons donc que chaque prédicteur peut être modélisé dans une base de B-splines à R nœuds :

$$\mathbf{X}_i(t) = B(t)^T s_i,$$

avec $B(t)$ la base de fonctions splines et s_i les coefficients de projection dans cette base associés au prédicteur $\mathbf{X}_i$.

En considérant la projection, on peut finalement écrire :

$$\begin{aligned}\text{logit}\big(p(\mathbf{s}_i)\big) &= \int \mathbf{w}(t)B(t)^T\mathbf{s}_i dt \\ &= \boldsymbol{\beta}^T\mathbf{s}_i,\end{aligned}$$

avec $\mathbf{s}_i$ le vecteur des coefficients de projection dans la base de fonctions pour l'individu i et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_R) = \int \mathbf{w}(t)B(t)dt$ où $\mathbf{w}(t)$ est l'analogue fonctionnel des coefficients de régression dans le cas des modèles linéaires généralisés (non fonctionnels).

Dans notre étude, nous souhaitons dans le même temps sélectionner les portions de courbes intéressantes pour la prédiction, tout en estimant la structure en groupes latents des données.

Le modèle de mélange de régressions fonctionnelles a notamment été étudié par Yao et al. (2010).

4.1.3 Partitionnement de données fonctionnelles

Globalement, les techniques de partitionnement de données fonctionnelles sont basées sur trois approches résumées dans Jacques and Preda (2014). On exclut ici le travail sur les données brutes, considérant les discrétisations mesurées des signaux, ce qui correspond plutôt à de l'analyse multivariée utilisant des méthodes permettant de gérer la grande dimension. Les trois approches fonctionnelles peuvent être résumées de la façon suivante :

- les méthodes procédant en deux étapes : réduction de la dimension (approximation des courbes grâce à la projection sur des bases de fonctions) puis utilisation de méthodes de classification usuelles sur les coefficients de projection obtenus.
- les méthodes non paramétriques qui utilisent des techniques de classification automatique classiques en considérant des distances ou dissimilarités spécifiques aux données fonctionnelles.
- les approches basées sur les modèles génératifs correspondant à notre cadre d'étude. La notion de densité de probabilité n'est pas définie dans le cas de variables fonctionnelles, donc la modélisation est basée sur les coefficients de projection, considérés comme des variables aléatoires, ce qui diffère des approches en deux temps.

4.2 Application du modèle PMLR sur données fonctionnelles

4.2.1 Spécification du modèle

Soit deux variables $(\mathbf{X}, Y)$ avec Y une variable réponse binaire à valeurs dans $\{0, 1\}$, et $\mathbf{X}$ un ensemble de courbes correspondant aux spectres. Les spectres sont mesurés à p points de fréquence, que l'on note $t = (t_1, \ldots, t_p)$. On a alors $X_i(t) = (X_i(t_1), ..., X_i(t_p))$ pour

toute observation i, avec $i = 1, \ldots, n$. Dans notre cadre de modèles à classes latentes, on suppose que les n observations sont réparties dans K groupes de proportion $(\pi_k)_{k=1,\ldots,K}$, avec $0 < \pi_k$ et $\sum_k \pi_k = 1$. On considère la variable aléatoire de classe $\mathbf{Z} = (Z_1, \ldots, Z_K)$, avec Z_k égale à 1 si l'individu appartient à la classe k, et 0 sinon. Là encore, nous considérons que l'appartenance à une classe a une influence à la fois sur les courbes $\mathbf{X}$ et sur le modèle de régression.

Tout d'abord, les données spectrales sont projetées sur une base de B-splines. Soit $(\mathbf{X}_1, ..., \mathbf{X}_n)$ un échantillon de $\mathbf{X} = \mathbf{X}(t)_{t\in[0,1]}$. Chaque courbe peut être définie comme une combinaison linéaire de fonctions de base :

$$X_i(t) \approx \sum_{r=1}^{R} s_r(X_i)\, h_r(t), \tag{4.1}$$

avec R le nombre de fonctions de base sélectionnées en utilisant la méthode décrite plus loin, $h(t) = (h_r(t))_{1\leq r\leq R}$ l'ensemble de fonctions de base de B-splines et $s_r(X_i)$ le coefficient de la base correspondant à l'observation i pour la fonction r. On peut rassembler l'ensemble des coefficients de base dans une matrice de taille $n \times R$ appelée $\mathbf{S}$ et écrire l'équation (4.1) sous forme matricielle :

$$\mathbf{X}(t) = B(t)^T \mathbf{S}.$$

Comme détaillé en sections 4.1.1 et 4.1.2, nous travaillons par la suite sur les coefficients $\mathbf{S}$ de la base pour spécifier le modèle de mélange de régressions logistiques fonctionnelles. Le modèle logistique fonctionnel pour la réponse Y dans le cas du mélange de régressions, en considérant l'information de groupe, s'écrit :

$$Y_i|\{\mathbf{S}_i = \mathbf{s}_i, Z_{ik} = 1\} \sim \mathcal{B}\left(p_k(\mathbf{s}_i)\right),$$

avec $\mathbf{s}_i$ une réalisation de $\mathbf{S}_i$ et $p_k(\mathbf{s}_i) = \mathbb{P}(Y_i = 1|\mathbf{S}_i = \mathbf{s}_i, Z_{ik} = 1)$. Les prédicteurs sont liés à la variable réponse grâce à la fonction de lien logistique :

$$\text{logit}\big(p_k(\mathbf{s}_i)\big) = \mathbf{s}_i \boldsymbol{\beta}_k,$$

avec $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,R})$ les coefficients de régression dans le groupe k. La base de B-splines étant orthonormale (de Boor, 2001), si on considère que $\mathbf{X}$ suit une loi normale, alors les coefficients de la base suivent une loi normale. Sachant $\{Z_k = 1\}$, $\mathbf{S}$ est donc modélisée par une distribution gaussienne multivariée :

$$\mathbf{S}_i|\{Z_{ik} = 1\} \sim \mathcal{N}_p(\boldsymbol{\eta}_k, \Sigma_k).$$

Finalement, on peut écrire la distribution de probabilité de Y sachant $\mathbf{S} = \mathbf{s}$ comme un mélange de régressions logistiques :

$$\mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k \mathbb{P}(Y = 1|\{\mathbf{S} = \mathbf{s}, Z_k = 1; \boldsymbol{\Phi}_k\}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{s}),$$

avec $\mathbf{\Phi}_k = (\boldsymbol{\eta}_k, \Sigma_k, \boldsymbol{\beta}_k)$ le vecteur des paramètres du groupe k et $\mathbf{\Phi} = (\pi_1, ..., \pi_K, \mathbf{\Phi}_1, ..., \mathbf{\Phi}_K)$ l'ensemble complet des paramètres à estimer pour le modèle de mélange à K groupes.

4.2.2 Prédiction

Nous considérons la même règle de prédiction que celle décrite au chapitre précédent. Il faut cependant projeter les nouvelles observations sur la base de fonctions utilisée pour construire le modèle et travailler sur les coefficients de la base pour ces nouvelles observations. On considère les coefficients splines S_0 de la projection du nouvel individu X_0 pour effectuer la prédiction de la variable réponse pour cet individu :

$$\mathbb{E}(Y_0|\mathbf{S}_0 = \mathbf{s}_0) = \sum_{k=1}^{K} \mathbb{P}(Y_0 = 1|Z_{0k} = 1, \mathbf{S}_0 = \mathbf{s}_0)\mathbb{P}(Z_{0k} = 1|\mathbf{S}_0 = \mathbf{s}_0).$$

Soit $\tau_{0,k} = \mathbb{P}(Z_{0k} = 1|S = S_0, Y_0 = y_0)$ les probabilités *a posteriori* prenant en compte la structure de groupe. Soit $\tau'_{0k} = \mathbb{P}(Z_{0k} = 1|\mathbf{S}_0 = \mathbf{s}_0)$, alors

$$\hat{\tau}'_{0,k} = \frac{\pi_k f_{\mathbf{S}_0}(\mathbf{s}_0; \boldsymbol{\eta}_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l f_{\mathbf{S}_0}(\mathbf{s}_0; \boldsymbol{\eta}_l, \Sigma_l)},$$

et,

$$\hat{y}_0 = \sum_{k=1}^{K} \hat{\tau}'_{0,k} \frac{\exp(\mathbf{s}_0^t \boldsymbol{\beta}_k)}{1 + \exp(\mathbf{s}_0^t \boldsymbol{\beta}_k)}.$$

Si les paramètres $\pi_k, \boldsymbol{\eta}_k, \Sigma_k$ et $\boldsymbol{\beta}_k$, pour $k = 1, \ldots, K$, sont inconnus, ils sont remplacés par leurs estimateurs respectifs $\hat{\pi}_k, \hat{\boldsymbol{\eta}}_k, \hat{\Sigma}_k$ et $\hat{\boldsymbol{\beta}}_k$.

4.2.3 Estimation

Le problème de vraisemblance pénalisée à résoudre est alors donné par :

$$\arg\max_{\boldsymbol{\beta}, \mathbf{\Theta}} \quad \{\ln \mathcal{L}(\mathbf{Y}, \mathbf{S}; \boldsymbol{\beta}, \mathbf{\Theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\mathbf{\Theta}_k\|_1\},$$

avec $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{R} |\beta_{k,j}|$, $\mathbf{\Theta}_k = \mathbf{\Sigma}_k^{-1}$ la matrice de précision du groupe k et $\|\mathbf{\Theta}_k\|_1$ la somme des valeurs absolues de $\mathbf{\Theta}_k$. Cette vraisemblance peut être décomposée de la même manière que celle détaillée en section 3.3 du chapitre précédent, en utilisant les données complétées. On a alors :

$$\arg\max_{\mathbf{\Phi}} \quad \left\{ \mathbb{E}_{\mathbf{\Phi}^*} \left[\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{s}_1, \ldots, \mathbf{s}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n; \mathbf{\Phi})|\mathbf{s}_1, \ldots, \mathbf{s}_n, y_1, \ldots, y_n\right] \right.$$
$$\left. - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\mathbf{\Theta}_k\|_1 \right\}.$$

Là encore, un algorithme EM est considéré pour estimer les paramètres du modèle. Lors de l'étape E, les probabilités conditionnelles $\tau_{ik} = \mathbb{P}(Z_{ik} = 1 | \mathbf{S}_i = \mathbf{s}_i, Y_i = y_i; \mathbf{\Phi}_k^{[h]})$ permettant de prédire le groupe latent non observé sont calculées pour tout individu $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, K$, à l'itération $[h]$. Cela permet le calcul de l'espérance de la log-vraisemblance complétée $Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]})$ sachant les données observées $(y_i, \mathbf{s}_i)_{i=1,\ldots,n}$ et $\mathbf{\Phi}^{[h]}$:

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{[h]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{[h+1]} \left(\ln \pi_k^{[h]} - \frac{1}{2} \left(y_i \mathbf{s}_i^t \boldsymbol{\beta}_k^{[h]} - \mathbf{s}_i^t \boldsymbol{\beta}_k^{[h]} - 1 \right) + \frac{1}{2} \ln |\mathbf{\Theta}_k^{[h]}| \right.$$
$$\left. - \frac{1}{2} \left(\mathbf{s}_i - \boldsymbol{\eta}_k^{[h]} \right)^T \Theta_k^{[h]} \left(\mathbf{s}_i - \boldsymbol{\eta}_k^{[h]} \right) \right) - \sum_{k=1}^{K} \lambda_k \| \boldsymbol{\beta}_k^{[h]} \|_1 - \sum_{k=1}^{K} \rho_k \| \mathbf{\Theta}_k^{[h]} \|_1.$$

La mise à jour des paramètres maximisant cette vraisemblance lors de l'étape M permet l'obtention de $\mathbf{\Phi}^{[h+1]}$, et on a alors pour tout $k = 1, \ldots, K$:

$$\widehat{\pi}_k^{[h+1]} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{[h+1]},$$

$$\widehat{\boldsymbol{\eta}}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{s}_i,$$

$$\hat{\mathbf{\Theta}}_k^{[h+1]} = \arg\max_{\mathbf{\Theta}_k} \left\{ \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(\log \det \mathbf{\Theta}_k - \frac{1}{2} \left(\mathbf{s}_i - \widehat{\boldsymbol{\eta}}_k^{[h+1]} \right)^T \Theta_k \left(\mathbf{s}_i - \widehat{\boldsymbol{\eta}}_k^{[h+1]} \right) \right) - \rho_k \| \mathbf{\Theta}_k \|_1 \right\},$$

$$\hat{\boldsymbol{\beta}}_k^{[h+1]} = \arg\max_{\boldsymbol{\beta}_k} \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(y_i \mathbf{s}_i^T \boldsymbol{\beta}_k - \ln \left(1 + \exp(\mathbf{s}_i^T \boldsymbol{\beta}_k) \right) \right) \right] - \lambda_k \| \boldsymbol{\beta}_k \|_1.$$

L'algorithme EM, et notamment son initialisation, est réglé de la même façon que dans le chapitre précédent.

4.2.4 Sélection de modèles

Comme dans le chapitre précédent, il est nécessaire de fixer les paramètres de régularisation permettant la sélection de coefficients dans les matrices de précision et les vecteurs de régresseurs dans chaque groupe. Nous considérons les mêmes outils que ceux détaillés dans le chapitre précédent, en nous basant sur le critère BIC pour la sélection des paramètres de régularisation. De la même façon, le nombre de groupes du mélange K doit être fixé. Nous utilisons le critère AIC ainsi que la validation croisée pour la sélection K, selon les conclusions tirées de l'étude de simulation menées dans le chapitre 3.

Procédure de sélection du nombre de fonctions de base Il est aussi nécessaire de sélectionner le nombre de nœuds de la base de splines. Pour cela, une méthode qui permet d'obtenir un bon compromis entre la minimisation de l'erreur de reconstruction des données et la complexité du modèle de projection est utilisée. La sélection du nombre

de nœuds dans la base de B-splines est réalisée par validation croisée pour minimiser l'erreur de modélisation. Cette procédure se déroule selon les étapes suivantes :

- Projection des données étudiées sur une base de splines à t_{R_1} fonctions
- Reconstruction des courbes avec les coefficients de projection, et la base de fonctions
- Calcul de l'erreur quadratique entre la reconstitution après projection et le spectre de départ, pour chaque point de discrétisation du spectre mesuré. Ces erreurs sont sommées pour tous les points, et tous les individus observés. Cette valeur d'erreur $E_{t_{R_1}}$ est enregistrée.
- Répétition des étapes précédentes pour un ensemble de nombres de nœuds possibles dans la base $t_{R_1}, t_{R_2}, \ldots, t_{R_T}$
- Représentation des erreurs quadratiques $E_{t_{R_1}}, \ldots, E_{t_{R_T}}$ en fonction du nombre de fonctions de base
- Sélection du nombre de fonctions de base présentant un coude sur la courbe. Un exemple de représentation du coude sur la courbe des erreurs quadratiques est présenté en figure 4.3.

Nous avons détaillé un modèle permettant de combiner le partitionnement et la prédiction d'une variable réponse à partir de prédicteurs fonctionnels, tout réalisant de la sélection de variables. Cependant, pour faciliter l'estimation du modèle de mélange de régressions logistiques fonctionnelles, nous souhaiterions nous baser sur des fenêtres spectrales plus restreintes portant l'information sur la maladie étudiée. Nous détaillons donc par la suite une procédure de sélection de portions de courbes pouvant être appliquée aux données de spectrométrie.

4.3 Sélection de zones du spectre

Dans cette partie nous cherchons à pré-sélectionner des zones du spectre pertinentes pour la discrimination des individus malades et des individus non malades. La figure 4.2, déjà présentée en section 1.3.5 du chapitre 1, illustre les variables sélectionnées par les procédures de sélection stable ainsi que par le Fused-Lasso. Même si ces méthodes sélectionnent des variables de façon isolée dans le spectre, nous pouvons noter que ces variables se trouvent regroupées dans des zones restreintes. Par exemple, un nombre important de variables est sélectionné dans la zone spectrale allant de 1550 cm^{-1} à 1750 cm^{-1}, ainsi que dans la zone allant de 2850 cm^{-1} à 3000 cm^{-1}. À l'inverse, la zone spectrale allant de 1000 cm^{-1} à 1350 cm^{-1} ne semble pas porter d'information importante pour la prédiction de la variable réponse. Nous ne souhaitons donc plus sélectionner des variables isolées mais bien des bandes larges dans le spectre, pouvant être reliées à des molécules.

Un autre aspect de ce type de méthode de sélection dans notre cadre est le manque de stabilité de la sélection de variables. En effet, la dépendance forte entre les variables

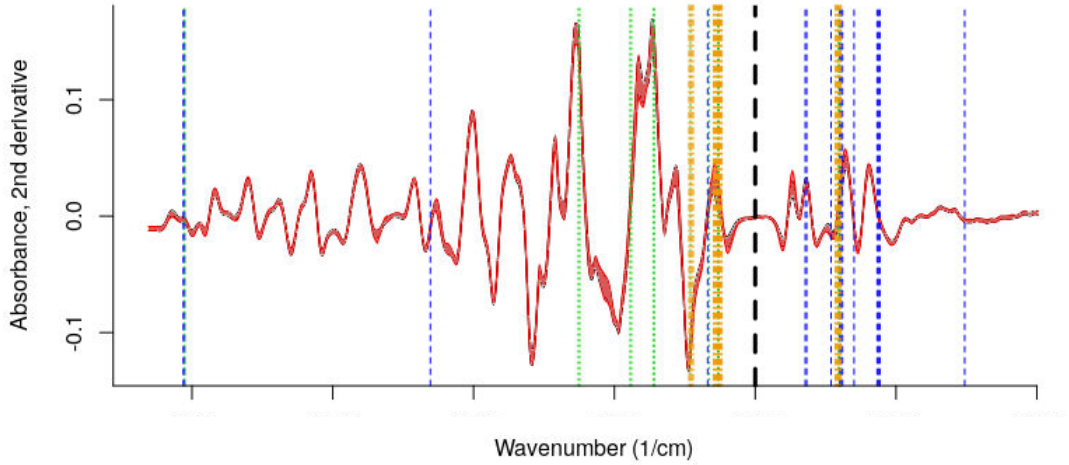

FIGURE 4.2 : **Variables sélectionnées par différentes méthodes de sélection de variables sur les données spectrométriques NASH.** Les deuxième et troisième quartiles des dérivées secondes de spectres mesurés sur des sérums de patients atteints de NASH (rouge) et de patients non malades (gris) sont représentés sur la fenêtre spectrale allant de 800 cm^{-1} à 3200cm^{-1}. La coupure de la fenêtre spectrale allant de 1800 cm^{-1} à 2800cm^{-1} est représentée par la ligne pointillée noire. Les variables sélectionnées sont représentées par les lignes pointillées colorées : la procédure de sélection stable utilisant le Lasso est représentée en vert, la procédure de sélection stable utilisant les forêts aléatoires en bleu et le Fused-Lasso en orange.

proches dans le spectre affecte les propriétés de la méthode Lasso, notamment sa capacité à déterminer le support du vecteur des régresseurs. On propose dans la suite une autre approche qui s'appuie sur la représentation fonctionnelle des données et permet de sélectionner des bandes ou portions du spectre. Cette étape de pré-sélection a l'avantage de faciliter l'estimation du modèle PMLR fonctionnel, mais aussi d'améliorer la compréhension du problème étudié.

Dans le cadre fonctionnel, il existe de nombreuses méthodes permettant la sélection de variables discriminantes, principalement basées sur des approches d'estimation par vraisemblance pénalisée (voir par exemple Gertheiss et al., 2013; Matsui, 2014). Concernant la sélection de domaines, c'est-à-dire d'intervalles de variables fonctionnelles, on peut par exemple citer Picheny et al. (2019); Fauvel et al. (2015); Fraiman et al. (2016), qui sont basées sur des méthodes intégrées dans l'algorithme d'estimation (ou "embedded") pour la sélection d'intervalles. Dans notre cadre, nous considérons une méthode basée sur des tests d'hypothèse, ce qui permet un contrôle du risque d'erreur de première espèce, c'est-à-dire du risque de rejeter l'hypothèse nulle à tort et de sélectionner une

variable non informative. L'adaptation des méthodes basées sur les tests d'hypothèse dans le cadre fonctionnel a notamment été étudié dans Collazos et al. (2016); Kong et al. (2016); Yang and Nie (2008); Pomann et al. (2013), mais concerne principalement la sélection de variables et n'est pas développée pour la sélection de portions de courbes. La méthode proposée dans cette section a pour objectif la sélection de portions de courbes, correspondant à des intervalles de prédicteurs fonctionnels voisins. Dans ce travail, nous considérons les tests de permutation, permettant d'approcher la loi de la statistique de test par une estimation empirique et d'estimer les p-valeurs. Dans notre cadre, nous nous basons sur les travaux de Pini and Vantini (2017), dans lesquels les tests d'hypothèse sont utilisés pour tester l'égalité de la distribution de deux populations caractérisées par des données fonctionnelles et mettre en avant des zones de courbes discriminantes entre les deux groupes de courbes. Cette méthode permet de chercher les zones du spectre qui présentent des différences selon le diagnostic, tout en prenant en compte la caractéristique fonctionnelle des données. Dans notre cas, nous modifions une partie de la procédure pour l'adapter aux données étudiées et à notre problématique, et nous considérons la classification hiérarchique pour construire les zones du spectre à tester.

4.3.1 Tests par intervalles sur données fonctionnelles

La procédure sur laquelle se base notre approche est décrite en détail dans Pini and Vantini (2017), ainsi que les propriétés théoriques associées. La prise en compte du caractère fonctionnel des données repose sur la considération des coefficients de projection sur la base de fonctions pour effectuer les tests. On peut regrouper les coefficients des observations par classe, selon la variable de diagnostic Y, et on a $\{\mathbf{s}_{i,j}\}$ pour tout $i = 1, \ldots, n$ et $j \in \{m, t\}$, avec m l'indice représentant les individus malades et t l'indice représentant les individus témoins. On suppose que les observations de la classe m sont indépendantes des observations de la classe t et on peut écrire $\mathbf{s}_{i \in C_m^{(r)}} \sim \mathbf{S}_m$ et $\mathbf{s}_{i \in C_t^{(r)}} \sim \mathbf{S}_t$ avec $C_m^{(r)}$ et $C_t^{(r)}$ les distributions (inconnues) du coefficient r de la base, pour les classes 1 et 2. Chaque coefficient est alors testé selon

$$H_0^{(r)} : C_m^{(r)} = C_t^{(r)}.$$

Un test bivarié est ensuite effectué sur chaque couple de coefficients successifs selon

$$H_0^{(r,r+1)} : H_0^{(r)} \cap H_0^{(r+1)},$$

puis un test trivarié sur chaque triplet de coefficients successifs et ainsi de suite jusqu'à un test global multivarié

$$H_0^{(1,\ldots,R)} : \cap_{r=1}^{R} H_0^{(r)}.$$

On obtient une famille de tests ainsi que leur p-valeurs associées. Une fois la p-valeur calculée pour l'ensemble des coefficients constituant nos données, une correction multiple permet le calcul de la p-valeur ajustée.

Dans cette thèse, l'originalité consiste à adapter cette procédure à la sélection de portions de courbes en modifiant les indices des intervalles testés. Nous nous basons sur les groupes de coefficients de la base construits par classification ascendante hierarchique (CAH). Nous supposons que les groupes de coefficients construits par CAH pourraient correspondre à des groupements moléculaires reflétés par les données spectrales. Les caractéristiques de ce type de données suggèrent que ces groupements moléculaires correspondent à des ensembles de nombres d'ondes voisins sur le spectre, mais il est quand même possible que les groupes soient non concomitants. La matrice de distance sur laquelle la CAH est construite permet de prendre en compte l'ordre des coefficients et donc leur localisation, ce qui pousse les groupes formés à être constitués de coefficients voisins. La matrice de distance euclidienne entre les coefficients est d'abord calculée sur les données normalisées. Un terme prenant en compte l'éloignement entre les coefficients en terme d'ordre dans le signal y est ensuite ajouté, grâce au calcul de la distance entre la position des coefficients :

$$D = D_c + \gamma D_p,$$

avec D_c la matrice de distance euclidienne entre coefficients calculée sur les données normalisées, γ un hyperparamètre permettant de régler l'importance de la prise en compte de l'ordre des coefficients et D_p la matrice de distance calculée sur les indices de position des coefficients de la base.

La CAH est ensuite réalisée sur la matrice D, en utilisant la méthode "average" comme stratégie d'agrégation. Les intervalles testés ensuite correspondent à l'ensemble des groupes obtenus pour chaque coupure possible $k = 1, \ldots, K$ de l'arbre. Ces groupes peuvent correspondre à des portions non connexes.

Les tests d'hypothèse effectués ici reposent sur des tests de permutation, décrits par la suite.

4.3.2 Principe des tests de permutation

Nous souhaitons tester l'indépendance entre deux populations représentées par leur diagnostic, pour des portions de courbes représentées par des intervalles de coefficients. Pour ce faire, nous nous basons sur l'approche des tests de permutation dans le contexte de la comparaison de deux distributions.

Les tests de permutation sont des tests statistiques basés sur le ré-échantillonnage et permettent de fournir un contrôle de l'erreur de type 1. Cette erreur correspond au fait de rejeter à tort l'hypothèse nulle, ce que l'on nomme faux positif. Ces tests ne sont plus basés sur une distribution théorique de la statistique pour évaluer la significativité, mais sur une distribution empirique, construite directement à partir des données. L'intérêt des tests de permutation réside dans leur flexibilité et dans le fait qu'ils laissent le choix de la statistique de test, que l'on peut adapter au problème étudié. Ces tests sont basés sur l'interchangeabilité : sous l'hypothèse nulle choisie, toutes les combinaisons possibles des données sont équiprobables.

On suppose que l'on observe deux échantillons issus de deux variables aléatoires X_m et X_t. On cherche à développer un test pour tester l'égalité entre les distributions des

deux populations. On veut donc tester : $H_0 : f_m = f_t$ contre $H_1 : f_m \neq f_t$, avec f_i la distribution de la population i, pour $i \in \{m, t\}$. Dans le cadre non paramétrique des tests de permutation, aucune supposition n'est faite sur la distribution des données, et le test développé est basé sur la distribution des données observées associées à X_m et X_t. Sous l'hypothèse nulle, les deux populations ont la même distribution $f_m = f_t = f$. Dans ce cas, les observations sont interchangeables, puisqu'échantillonnées à partir de la même population de distribution f, et chaque réarrangement des données est alors équivalent. L'information portée par n'importe quelle permutation est donc la même.

On note T la statistique de test définie pour le problème étudié. Dans les tests de permutation, on suppose toujours que la statistique de test a de plus fortes valeurs sous l'hypothèse alternative. Au lieu d'essayer d'obtenir la distribution de T, on estime la distribution de T conditionnellement aux données, soit la distribution de T pour chaque permutation, que l'on note T^*. On note T_0 la statistique de test calculée sur les données observées non permutées. La p-valeur du test est ensuite calculée comme la proportion de statistiques de test T^* qui sont supérieures à T_0.

Finalement, les étapes du test de permutation peuvent être résumées de cette façon :

- Évaluer la statistique de test T_0 sous les données observées
- Répeter B fois :
  - choisir une permutation aléatoire X_b
  - évaluer la statistique de test sur les données permutées, notée T_b
  - enregistrer la statistique de test T_b
- évaluer la p-valeur du test comme la proportion des scenarii permutés pour lesquels la statistique de test T_b est supérieure à T_0.

Lorsque la distribution empirique sous l'hypothèse nulle est basée sur toutes les permutations possibles, on parle de test exact, ce qui fournit un contrôle de l'erreur de type I au seuil désiré. Cependant, le nombre de permutations possibles est généralement très élevé, et évaluer l'ensemble des permutations peut être numériquement coûteux. Pour cette raison, dans les applications, un sous-échantillon des permutations possibles est évalué grâce à des approches de type Monte Carlo. On parle alors de test de permutation approximatif. Le contrôle de l'erreur de type I est alors conservé de façon approximative, et peut être approché au plus proche du seuil désiré si le nombre de permutation est suffisamment grand.

4.3.3 Correction multiple

La correction de la p-valeur permet le contrôle de la probabilité de rejeter à tort l'hypothèse nulle par intervalle, ce qui correspond au contrôle de l'erreur de première espèce entraînant la mise en évidence de faux positif. Avec cette procédure, on effectue un contrôle de l'erreur de première espèce par intervalle du domaine. Cela signifie que la probabilité

de détecter à tort une différence significative entre les deux populations dans n'importe quel intervalle où il n'y a pas de différence est contrôlée au seuil choisi (Pini and Vantini, 2017).

En pratique, on calcule une p-valeur ajustée prenant en compte la multiplicité et la structure en intervalle des tests.

Pour une coupure de l'arbre menant à k groupes, la p-valeur associée aux coefficients est calculée pour chacun des groupes allant de 1 à k. Au sein d'un groupe, tous les coefficients ont la même valeur de p-valeur. Pour un groupe g_k de la partition obtenue pour une coupure de l'arbre menant à k groupe, la statistique de test $T_0^{(g_k)}$ est définie par

$$T_0^{(g_k)} = \sum_{r \in g_k} T_0^{(r)},$$

avec $T_0^{(r)}$ la statistique de test calculée pour le coefficient r sur les données observées non permutées.

La statistique de test associée au groupe g_k pour chacune des B permutations est ensuite calculée. Pour une permutation b, elle est donnée par :

$$T^{(b,g_k)} = \sum_{r \in g_k} T^{(r)}.$$

On obtient alors B valeurs de $T^{(b,g_k)}$ pour $b = 1, \ldots, B$ pour le groupe g_k. La p-valeur associée à chaque coefficient du groupe g_k est la proportion de statistique de test $T^{(b,g_k)}$ supérieure à $T_0^{(g_k)}$. La p-valeur associée au coefficient r pour la coupure de l'arbre de CAH menant à k groupes est notée $\lambda^{(r,k)}$.

Cette étape est répétée pour la coupure menant à $k+1$ groupes, et permet d'obtenir les p-valeurs $\lambda^{(r,k+1)}$ pour tout $r = 1, \ldots, R$, associées à la coupure menant à $k+1$ groupes.

La p-valeur ajustée associée au coefficient r, notée $\lambda^{(r)}$, correspond alors à la p-valeur maximale obtenue pour ce coefficient entre la p-valeur calculée pour la coupure menant à k groupes et pour la coupure menant à $k+1$ groupes :

$$\lambda^{(r)} = \max(\lambda^{(r,k)}, \lambda^{(r,k+1)}).$$

Cette opération est répétée jusqu'à la coupure menant à K groupes et permet l'obtention de la p-valeur ajustée associée à chaque coefficient. La p-valeur ajustée associée à tous les coefficients d'un même intervalle est ensuite calculée comme étant la moyenne des p-valeurs des coefficients de cet intervalle. La p-valeur ajustée a donc la même valeur sur tout l'intervalle considéré, défini par la CAH.

4.3.4 Procédure de tests par intervalles

Finalement, la procédure de sélection de portions de courbes de spectrométrie est résumée de la façon suivante :

- Projection des données sur une base de B-splines, en choisissant le nombre de nœuds permettant un compromis entre la minimisation de l'erreur et le faible nombre de composantes

- Test sur intervalles sur les coefficients de la base :
  - test de permutation univarié pour chaque composante-coefficient de la base : obtention d'une p-valeur pour chaque coefficient
  - Tests par intervalles :
    * calcul de la matrice de distance particulière entre les coefficients de la base
    * classification ascendante hiérarchique sur cette matrice de distance (méthode "average")
    * calcul de la p-valeur ajustée en se basant sur les intervalles obtenus par CAH.

    $\rightarrow$ obtention de la p-valeur ajustée pour chaque coefficient. Cette p-valeur à la même valeur pour tous les coefficients d'un même intervalle.
- Sélection des intervalles dont les p-valeurs ajustées sont inférieures au seuil α choisi.

Cette procédure implique le choix de plusieurs hyperparamètres, comme le paramètre γ permettant de prendre en compte la distance entre les coefficient le long de la fenêtre spectrale ainsi que le nombre de groupes considérés par la CAH. Ces paramètres sont choisis par validation croisée, avec pour objectif d'optimiser les performances de prédiction de la méthode de prédiction appliquée par la suite.

4.4 Application aux données NASH

4.4.1 Analyses

On considère les données décrites au chapitre 1 en se basant sur la même séparation en échantillon d'apprentissage et de validation que dans le chapitre 3.

La sélection du nombre de nœuds considéré dans la base de B-splines et la pré-sélection des zones du spectre d'intérêt grâce à la procédure de tests par intervalles sont effectuées sur l'échantillon d'apprentissage. Ensuite, la procédure PMLR fonctionnelle est appliquée sur les zones spectrales sélectionnées. Le nombre de groupes K est sélectionné grâce au critère AIC, et les paramètres de régularisation réglant la sélection de variables sont sélectionnés grâce au critère BIC. La qualité du modèle sélectionné est ensuite évaluée sur l'échantillon de validation.

4.4.2 Sélection de modèles

Choix de la base de fonctions La figure 4.3 représente les erreurs calculées lors de la procédure de sélection de la base de projection détaillée en section 4.2.4 pour les données NASH. Le point rouge représente le coude observé sur la courbe des erreurs absolues de projection pour une base de B-splines à 150 nœuds, et correspond au nombre de nœuds choisi pour la base de projection.

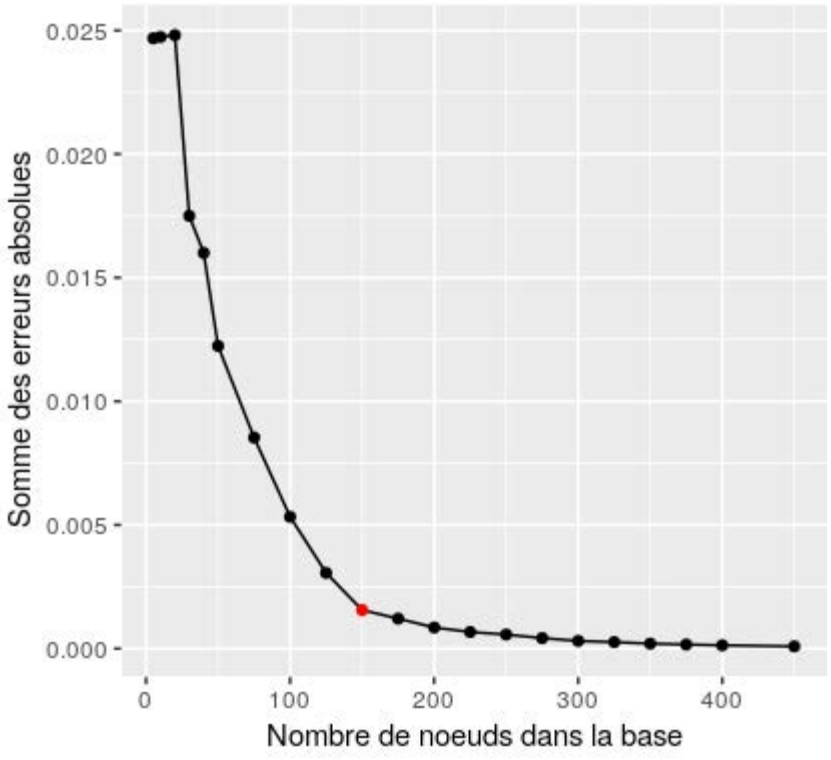

FIGURE 4.3 : **Erreur de reconstitution spectrale après projection, en fonction du nombre de nœuds considéré dans la base de B-splines, pour les données de spectrométrie NASH.** L'erreur correspond à l'erreur quadratique pour chaque point de discrétisation, entre la reconstitution du spectre après projection sur la base de splines et le spectre de départ. Les spectres de départ sont constitués de 621 points. Le point rouge représente le coude observé et correspond au nombre de nœuds choisi dans la base de fonctions B-splines. Le nombre de nœuds choisi est égal à 150.

Choix du nombre de groupes latents La table 4.1 montre les critères utilisés pour la sélection du nombre de groupes du mélange. D'après les analyses précédentes, nous choisissons comme critère de sélection de modèles l'AIC, le nombre de groupes retenu est donc 2. On nommera par la suite le modèle retenu PMLRF-2.

4.4.3 Estimateurs et modèles

Le seuil à partir duquel un patient est prédit comme atteint de la NASH à partir du score prédit par le modèle est choisi de façon automatique pour maximiser la somme de la sensibilité et de la spécificité. Les performances de prédiction obtenues sur l'échantillon de validation, pour les modèles PMLR fonctionnels estimés pour 1 à 3 groupes latents sont présentées en table 4.2. On remarque que le modèle à 2 groupes sélectionné grâce au critère AIC permet d'obtenir les meilleures performances de prédiction, ce qui est cohérent

	K = 1	K = 2	K = 3
AIC	-95565.83	**-96090.82**	-25158.2
BIC	**-94435.35**	-93889.95	-24891.54
ICL	**-94435.35**	-93874.73	-24822.49

TABLE 4.1 : **Sélection du modèle PMLR fonctionnel estimé sur les données NASH.** Comparaison des critères de sélection de modèles AIC, BIC et ICL pour les modèles estimés par PMLR fonctionnel sur l'échantillon d'apprentissage du jeu de données NASH réduit aux zones du spectre pré-sélectionnées, pour 1 à 3 groupes. Les valeurs en gras indiquent la meilleure valeur de critère obtenue.

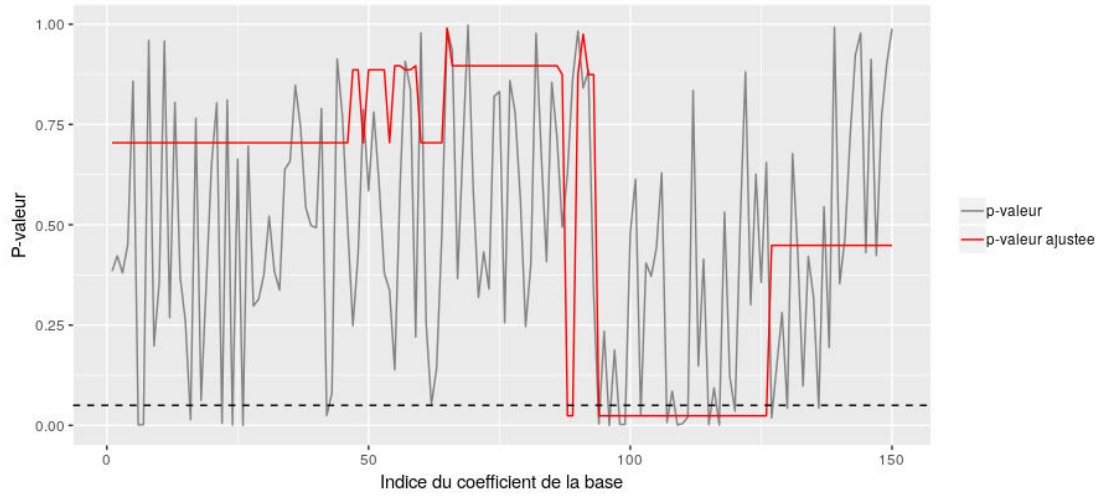

FIGURE 4.4 : **p-valeurs et p-valeurs ajustées pour chaque composante de la base de B-splines modélisant les données NASH, obtenues avec la procédure de tests de permutation par intervalles construits par CAH.** La p-valeur est représentée en gris, la p-valeur ajustée est représentée en rouge, et le seuil α de significativité égal à 0.05 est représenté par la ligne pointillée noire.

avec les conclusions tirées de l'étude de simulation menée au chapitre 3. Les performances obtenues lors des analyses précédentes dans le cadre multivarié sont présentées dans la table 4.3. On remarque que la performance globale du modèle retenu (PMLRF-2) est plus élevée que la performance du modèle estimé par notre méthode dans le cadre multivarié (voir le modèle PMLR-2 du chapitre 3). Les valeurs d'AUROC, de spécificité et de taux de bon classement obtenues avec le modèle PMLRF-2 sont nettement supérieures aux valeurs obtenues avec le modèle construit avec l'approche multivariée. La prise en compte de l'aspect fonctionnel a dans ce cas permis une amélioration de la qualité de prédiction du modèle. Globalement, le modèle PMLRF-2 permet d'obtenir un taux de bon classement élevé tout en menant à des bonnes valeurs de spécificité et sensibilité, contrairement aux méthodes issues de l'approche multivariée et aux méthodes usuelles de spectrométrie

présentées en chapitre 1. Les performances sont globalement meilleures lorsque la structure en groupes latents est prise en compte, comme l'indiquent les performances des modèles PMLRF-2 et PMLR-2 qui sont meilleures que celles obtenues avec les autres méthodes présentées.

	PMLRF-1	PMLRF-2	PMLRF-3
AUROC	0.7	**0.77**	0.65
Se	0.62	**0.7**	0.69
Sp	0.67	**0.82**	0.55
PPV	0.23	**0.43**	0.23
NPV	0.9	**0.93**	0.9
TBC	0.66	**0.8**	0.57

TABLE 4.2 : **Comparaison des performances de prédiction obtenues pour des modèles allant de 1 à 3 groupes sur les données NASH.** Les modèles estimés sont des mélanges de régressions logistiques pénalisées pour données fonctionnelles (notés PMLRF-K). Les performances sont évaluées sur l'échantillon de validation des données NASH. Les critères utilisés pour la comparaison sont les suivants : aire sous la courbe ROC (AUROC), sensibilité (Se), spécificité (Sp), valeur prédictive négative (NPV), valeur prédictive positive (PPV), taux de bon classement (TBC). Le modèle sélectionné grâce au critère AIC est indiqué en gras.

	PMLRF-2	PMLR-2	Lasso	FA	Fused-Lasso	PCR	PLS-DA
AUROC	**0.77**	0.75	0.61	0.57	0.61	0.7	0.57
Se	**0.7**	0.77	0.46	0.85	0.62	0.85	0.15
Sp	**0.82**	0.76	0.92	0.39	0.64	0.47	0.98
PPV	**0.43**	0.38	0.54	0.25	0.25	0.24	0.67
NPV	**0.93**	0.94	0.9	0.93	0.89	0.94	0.86
TBC	**0.8**	0.76	0.85	0.47	0.63	0.43	0.85

TABLE 4.3 : **Comparaison des performances de prédiction obtenues avec différentes méthodes sur les données NASH.** Les méthodes comparées sont le mélange de régressions logistiques pénalisées pour données fonctionnelles, estimé pour un modèle à 2 groupes (PMLRF-2), le meilleur modèle obtenu par le mélange de régressions logistiques pénalisées détaillé dans le chapitre précédent (PMLR-2), ainsi que les méthodes usuelles utilisées en spectrométrie et présentées au chapitre 1. Ces méthodes correspondent à des modèles construits sur des sélections de variables basées sur le Lasso, les forêts aléatoires (FA), et le Fused-Lasso, ainsi que la PCR et la PLS-DA. Le modèle choisi est indiqué en gras. Les critères utilisés pour la comparaison sont les suivants : aire sous la courbe ROC (AUROC), sensibilité (Se), spécificité (Sp), valeur prédictive négative (NPV), valeur prédictive positive (PPV), taux de bon classement (TBC).

4.4.4 Interprétations du modèle construit

Zones pré-sélectionnées Les zones du spectre sélectionnées par la procédure de pré-sélection sont représentées par la figure 4.5. On remarque d'une part que, même si la méthode permet de prendre en compte l'ordre des coefficients le long de la fenêtre spectrale, certaines portions considérées sont non connexes, c'est-à-dire que des coefficients non consécutifs peuvent faire partie d'un même intervalle construit par CAH. Cela souligne des liens entre des coefficients non voisins dans le spectre. La correction de la p-valeur, qui est représentée en figure 4.4, entraîne ensuite une égalité des p-valeurs ajustées pour tous les coefficients d'un même intervalle, et l'on remarque des pics sur la représentation de la p-valeur ajustée le long de la fenêtre d'étude. Certains intervalles considérés concernent un ensemble important de coefficients, que l'on pourrait relier à des signatures de groupements moléculaires particuliers. On peut notamment relier la fenêtre spectrale allant de 2800 à 3000 cm^{-1} à la famille des lipides en spectrométrie IR, ce qui est cohérent avec les dysfonctionnements observés dans le cas de maladie métabolique.

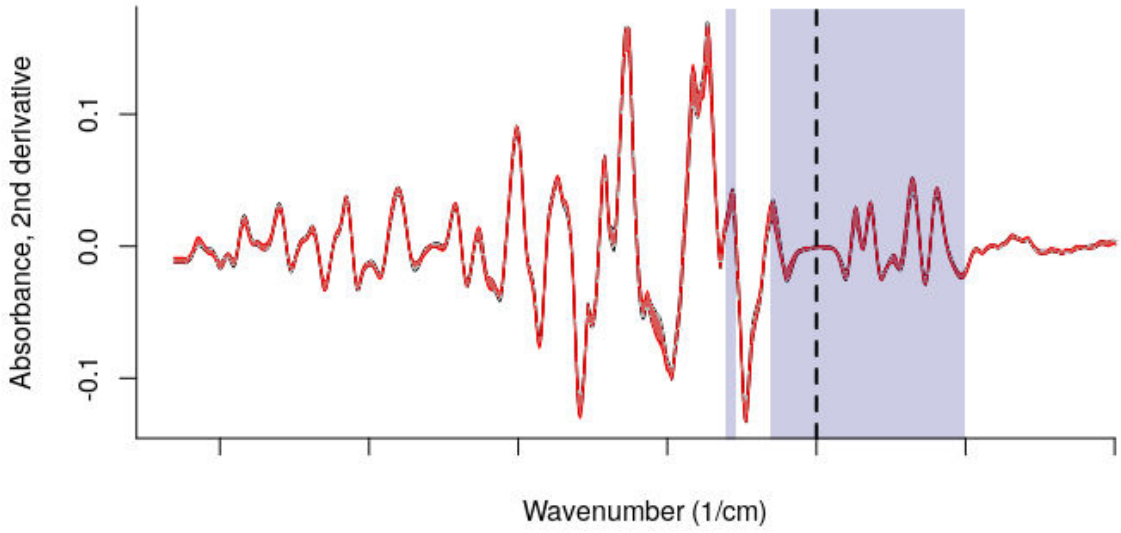

FIGURE 4.5 : **Zones spectrales sélectionnées par la procédure de pré-sélection par tests sur intervalles, sur les données NASH.** Deuxième et troisième quartiles des spectres des individus NASH (en rouge) et non NASH (en noir), représentés sur le domaine de nombre d'onde allant de 1800 à 3800 cm-1. Une coupure de 1800 à 2800 cm-1 est représentée par la ligne verticale pointillée noire. Les zones du spectre sélectionnées sont représentées par les lignes verticales bleues.

Interprétations biologiques La table 4.4 présente les caractéristiques cliniques des groupes estimés par le modèle, ainsi que la significativité des différences entre ces variables selon le groupe. On remarque que le nombre de malades est similaire dans chaque groupe, ce qui indique que les profils de patients ne dépendent pas du diagnostic. Ces profils semblent reposer sur des différences métaboliques, comme l'indiquent les différences significatives retrouvées pour les variables AST.ALT, GGT, Gluc, Chol et TG. Ces indicateurs sériques semblent montrer des dérèglements métaboliques plus importants chez les patients du groupe 1.

	Groupe 1	Groupe 2	P-value	Signif
NASH	16.9	16.3	0.9	
Age	40	39	0.2	
BMI	44	44	0.45	
AST	28	28	0.47	
ALT	37	30	0.18	
AST.ALT	0.91	1	7.10^{-4}	**
GGT	49	35	0.01	*
Gluc	6.2	5.6	0.03	*
Insuline	21	23	0.59	
HBA1C	6	5.8	0.29	
chol	5.5	5.1	1.10^{-3}	**
HDL	1.4	1.4	0.21	
LDL	3.3	3.1	0.08	
TG	1.9	1.5	0.01	*

TABLE 4.4 : **Caractérisation des groupes obtenus par PMLRF sur les données NASH grâce aux variables cliniques.** Le modèle sélectionné a deux groupes. Les valeurs représentées correspondent à la moyenne calculée sur tous les individus de chaque groupe. Pour chaque variable clinique, les moyennes par groupes sont comparées à l'aide d'un test de Student, dont la significativité de la p-valeur est reportée en dernière colonne. Pour la variable NASH, le pourcentage de malades est indiqué, et un test de Fisher est utilisé pour la comparaison.

De plus, la figure 4.6 montre la répartition des scores prédits par le modèle, par groupe latent et selon le diagnostic observé. On remarque que les différences entre les scores selon le diagnostic sont plus marquées dans le groupe 2, ce qui indique un diagnostic plus facile à établir par le modèle pour ce profil de patients. Cela pourrait être lié aux dérèglements métaboliques plus importants chez les patients du groupe 1, ce qui perturbe la capacité à établir le diagnostic. On peut supposer que les patients du groupe 2, aux dérèglements métaboliques moins importants, ont un signal plus marqué en présence de NASH, ce qui facilite le diagnostic.

La figure 4.7 représente la répartition des scores prédits selon les grades histologiques et les groupes latents. On note que la valeur du score prédit augmente avec la gravité du stade histologique du patient. Le score prédit est donc un bon indicateur de l'ampleur de

l'atteinte au foie. On note aussi que les grades histologiques les plus élevés sont observés chez les patients du groupe 1, ce qui suggère des lésions plus importantes au foie chez ces patients. Cet état plus grave peut expliquer l'importance des dérèglements métaboliques observés chez ces patients.

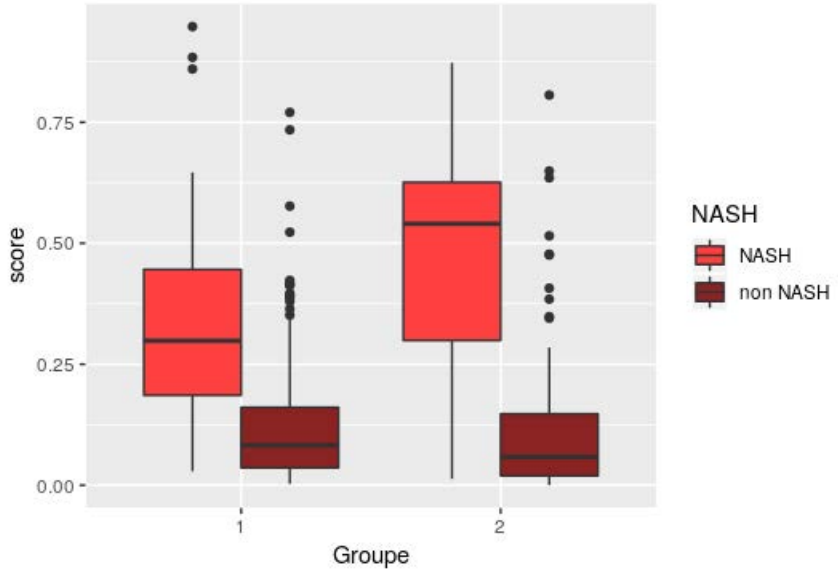

FIGURE 4.6 : **Performances de prédiction du modèle PMLRF estimé sur les données NASH.** Boîtes à moustaches représentant la répartition des scores prédits par la méthode de mélange de régressions logistiques pénalisées selon le vrai diagnostic (NASH ou non NASH), en fonction du groupe latent estimé.

Conclusion Le modèle obtenu dans cette section est performant pour le diagnostic de la NASH. De plus, le fait que la prise en compte du caractère fonctionnel des données améliore la qualité du modèle de prédiction confirme notre hypothèse selon laquelle la forme du spectre peut être plus informative que l'absorbance "brute".

Cette section a aussi permis de déterminer des groupes de patients aux profils métaboliques différents, construits grâce à des zones discriminantes du spectre. Il serait intéressant d'essayer d'évaluer les interactions entre ces zones pré-sélectionnées dans le spectre. Cela a plusieurs avantages, d'abord d'un point de vue de l'interprétation des résultats et d'une meilleure compréhension du problème étudié, puisque les différentes zones du spectre peuvent refléter des groupements moléculaires en interaction. D'autre part, une prise en compte de la structure des données pourrait améliorer leur modélisation, notamment l'estimation complexe des paramètres dans le cadre de données en dimension modérée.

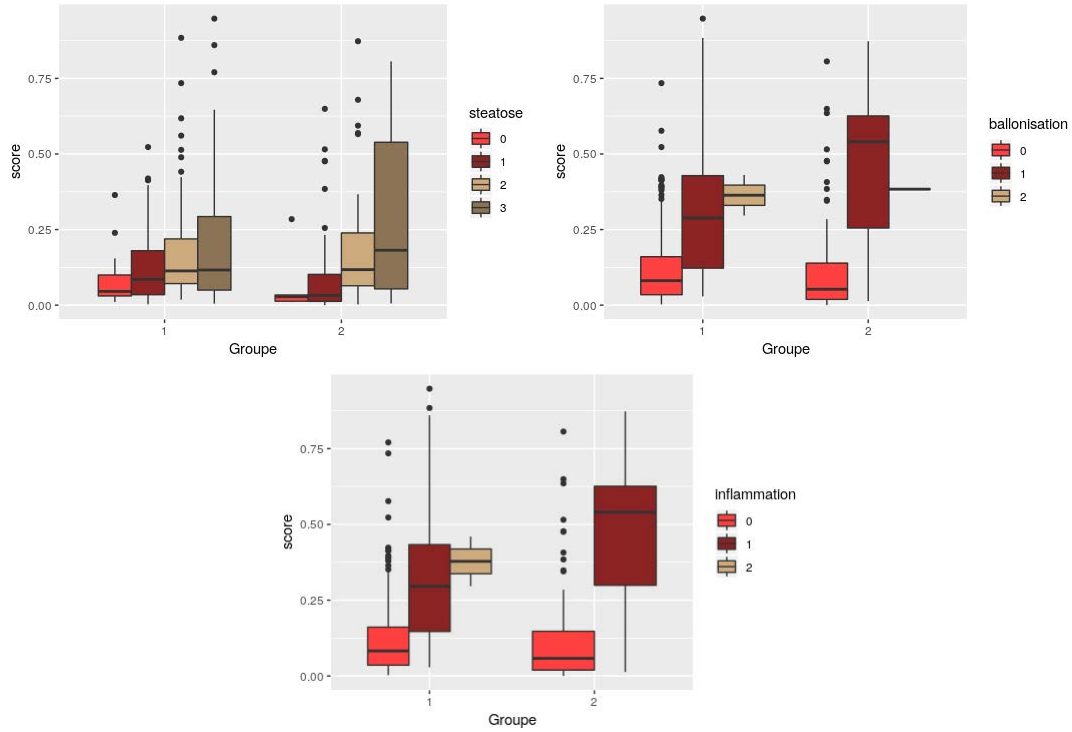

FIGURE 4.7 : **Boîtes à moustaches représentant la répartition des scores prédits par le modèle PMLRF estimé sur les données NASH selon le grade de chaque variable histologique.** En haut à gauche : stéatose, en haut à droite : ballonisation, en bas : inflammation.

Chapitre 5

Tests par blocs au sein des matrices de corrélation

Ce travail est réalisé en collaboration avec Alessia Pini (Université Catholique Du Sacré-Coeur, Milan).

Contents

5.1 Introduction

Le dernier travail de cette thèse porte sur l'étude de la structure des dépendances entre variables, caractérisées par les matrices de covariance. Dans le cas des données de spectrométrie où la dimension est élevée, le nombre de paramètres à estimer est important,

ce qui rend difficile l'estimation des matrices de covariance, même lorsque l'on met en œuvre une méthode pénalisée de type Graphical Lasso. On se pose alors la question d'aborder le problème de l'identification des covariances significatives par une autre approche, en l'occurrence par des tests statistiques. Si on réussit à mettre en évidence que seul un petit nombre de coefficients des matrices de covariance sont significatifs, on pourra limiter l'estimation à ces coefficients et donc implicitement poser un problème d'estimation dans un espace de faible dimension.

Jusqu'à présent, les matrices de covariance estimées à partir des données sont supposées non structurées. Cependant, dans notre cadre, il paraît pertinent de considérer une structure au sein de la matrice de covariance. En effet, en spectrométrie, il existe des bandes d'absorbance liées à des types moléculaires sur le spectre. On peut supposer que ces bandes composées d'un ensemble de nombres d'ondes reflètent des groupements moléculaires. Il apparaît alors cohérent de considérer que des zones du spectre, correspondant à ces bandes, structurent les données et entraînent une organisation de la matrice de covariance en blocs. Cet aspect se retrouve notamment sur les matrices de covariance empiriques estimées sur les données spectrales NASH pour les patients atteints de NASH et pour les patients témoins, présentées en figure 5.1, montrant des structures qui ont l'air de se répartir en blocs. On suppose donc que ces blocs correspondent aux groupements moléculaires mesurés par le spectre. L'ordre des variables est donc capital dans cette approche, puisque chaque bloc est constitué d'un ensemble d'absorbances mesurées à des nombres d'ondes voisins. Un des objectifs de ce chapitre est donc d'étudier les dépendances entre blocs de variables associés à des groupements moléculaires, grâce aux matrices de covariance.

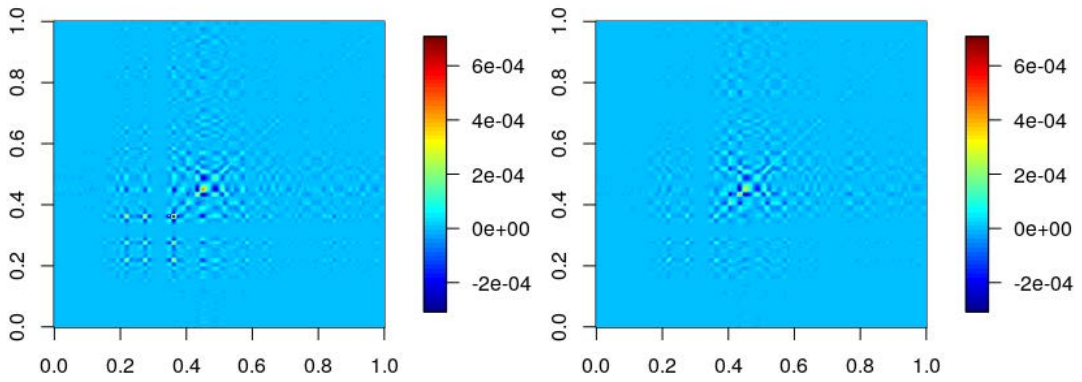

FIGURE 5.1 : **Matrices de covariance empiriques calculées sur les données de spectrométrie NASH.** La matrice de covariance à gauche est calculée sur les patients atteints de NASH et la matrice de covariance de droite est calculée sur patients non atteints de NASH.

Dans notre cadre, nous nous intéressons à la structure des dépendances entre les

variables, et plus précisément à la significativité des liens entre les zones spectrales. Dans ce cas, on cherche seulement à déterminer les blocs non nuls dans la matrice de covariance. Il est alors équivalent de travailler sur la matrice de corrélation, correspondant à une forme standardisée de la matrice de covariance. Les coefficients de la matrice étant sur une même échelle, le test est plus puissant. Par la suite, nous considérons donc l'étude des matrices de corrélation associées aux données étudiées.

Dans ce chapitre, nous exposons une méthode permettant d'étudier les dépendances entre des zones du spectre pré-sélectionnées selon le diagnostic, grâce à une procédure basée sur des tests d'hypothèse effectués sur les matrices de corrélation. Nous commencerons par exposer le cadre de travail de ce chapitre avant de détailler une contribution de cette thèse, la procédure de tests, d'abord par l'étude de la dépendance entre deux blocs de variables puis avec la généralisation à l'étude de la matrice complète. Une étude de simulation permettant l'évaluation des performances de la procédure sera ensuite présentée. Un autre apport de cette thèse est présenté dans cette section, avec une méthode de génération de matrices de covariance parcimonieuses par blocs. L'application de la procédure à l'analyse des données NASH sera enfin exposée.

5.2 Procédure de tests par blocs

On considère des données gaussiennes fonctionnelles $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ définies sur le domaine $D \subset \mathbb{R}^p$. Selon les notations définies en section 4.1.1, on note $(\mathbf{S}_1, \ldots, \mathbf{S}_n)$ les vecteurs des R coefficients de projection dans la base de fonctions selon $X_i(t) \approx \sum_{r=1}^{R} s_r(X_i)\, h_r(t)$, pour $i = 1, \ldots, n$. Nous travaillons par la suite sur les coefficients de projection $\mathbf{S}$ pour l'étude des matrices de corrélation associées aux données. L'objectif est alors d'estimer le type de structure des coefficients de la matrice de corrélation estimée sur les données projetées. On suppose qu'on dispose de portions de courbes formant des blocs ordonnés, et que les coefficients de projection $\{s_r(X_i)\}_{1 \leq i \leq n}$ dans la base de fonctions sont répartis en M blocs, associés à ces différentes portions de courbes. Ces portions de courbes peuvent, par exemple, être déterminées grâce à la méthode de sélection de zones spectrales présentée en section 4.3. À l'issue de cette méthode, on obtient $\mathcal{J}_1, \ldots, \mathcal{J}_M \subset \{1, \ldots, R\}$ une partition de $\{1, \ldots, R\}$ correspondant à l'index d'appartenance au bloc discriminant. On suppose aussi que les blocs sont ordonnés selon l'indice $m = 1, \ldots, M$, avec le bloc m associé à la m-ième portion de courbe.

On note $\mathbf{C} = (c_{jj'})_{j,j'=1,\ldots,R}$ la matrice de corrélation des coefficients de la base, de taille $R \times R$. Pour explorer les dépendances entre zones spectrales, on souhaite tester la significativité de chaque bloc de la matrice. On travaille donc sur des sous-matrices de matrice, chacune correspondant à la corrélation d'un couple d'intervalles de coefficients, eux même associés à des bandes spectrales. Pour $1 \leq m < m' \leq M$, on note $\mathbf{C}_{\mathcal{J}_m \times \mathcal{J}_{m'}}$ la sous-matrice de la matrice de corrélation $\mathbf{C}$ restreinte au couple de blocs $\mathcal{J}_m$ et $\mathcal{J}_{m'}$.

L'objectif de cette section est de détailler la procédure permettant d'étudier les interactions entre les blocs structurant une matrice de corrélation, grâce à des tests d'hypothèse. On commencera par détailler les tests d'hypothèse effectués sur une unique sous-matrice de la matrice de corrélation, associée à un couple de blocs, avec la spécification

des hypothèses testées et du calcul de la statistique de test basé sur les tests de permutation. Dans un second temps, le test sera répété pour étudier la significativité de plusieurs couples de blocs de façon à identifier la structure de la matrice de corrélation complète. Dans notre cas, il y a une structure de proximité à prendre en compte, puisque les blocs sont ordonnés. On peut exploiter cette structure de proximité pour définir des séquences - ou intervalles - de blocs consécutifs, et tester l'indépendance entre eux. Les tests de significativité sont donc appliqués à l'ensemble des blocs structurant la matrice de corrélation, ce qui nécessite une prise en compte de la multiplicité des tests.

5.2.1 Tests de permutation sur une sous-matrice

On souhaite tester l'hypothèse d'indépendance entre deux intervalles de variables, appelés blocs, correspondant à une sous-matrice au sein de la matrice de corrélation. Dans le cas gaussien, cela revient à effectuer un test de corrélation nulle. En particulier, pour un couple de blocs $\mathcal{J}_m, \mathcal{J}_{m'}$ avec $1 \leq m < m' \leq M$, on teste :

$$H_{0,m,m'} : \mathbf{C}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0 \quad \text{versus} \quad H_{1,m,m'} : \mathbf{C}_{\mathcal{J}_m \times \mathcal{J}'_m} \neq 0, \tag{5.1}$$

où $\mathbf{C}_{\mathcal{J}_m \times \mathcal{J}'_m} \neq 0$ signifie que la sous-matrice contient au moins un coefficient non nul.

On utilise des tests de permutation pour tester les hypothèses (5.1). Un test de permutation est défini par un ensemble de permutations préservant la vraisemblance sous l'hypothèse nulle, et une statistique de test qui est stochastiquement plus élevée sous l'hypothèse alternative.

Comme les données fonctionnelles $\mathbf{X}_i$, pour un individu $i = 1, \ldots, n$, sont indépendantes et identiquement distribuées (iid), les vecteurs de coefficients de la projection sur une base $\mathbf{S}_i$ sont iid (et donc échangeables par rapport aux observations). Faisant partie des vecteurs $\mathbf{S}_i$, les sous-vecteurs de coefficients $\mathbf{S}_{i_{\mathcal{J}_m}}$ et $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$ sont échangeables par rapport aux observations. De plus, sous l'hypothèse nulle, de tels sous-vecteurs sont indépendants les uns des autres pour tout i. Par conséquent, sous $H_{0,m,m'}$, $\mathbf{S}_{i_{\mathcal{J}_m}}$ est échangeable par rapport aux observations en gardant $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$ fixé, et vice-versa. En particulier, $n!$ permutations invariantes en vraisemblance peuvent être trouvées en permutant les observations du premier vecteur en gardant le deuxième vecteur fixe, et vice-versa.

On définit la statistique de test comme la somme des corrélations standardisées entre les observations de $\mathcal{J}_m$ et $\mathcal{J}_{m'}$:

$$T_{m,m'}(\mathbf{S}_{\mathcal{J}_m}, \mathbf{S}_{\mathcal{J}_{m'}}) = \sum_{(j,j') \in \mathcal{J}_m \times \mathcal{J}_{m'}} \frac{\hat{c}^2_{jj'}}{1 - \hat{c}^2_{jj'}},$$

avec $\hat{c}_{jj'}$ l'estimateur empirique de $c_{jj'}$.

On note que la statistique de test est invariante par permutations jointes (permutations avec les mêmes indices) de $\mathbf{S}_{i_{\mathcal{J}_m}}$ et $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$. Par conséquent, pour trouver la distribution de permutation, on peut considérer seulement les $n!$ permutations de $\mathbf{S}_{i_{\mathcal{J}_m}}$ au sein des observations en gardant $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$ fixé. On note $\mathbf{S}^*{}_{\mathcal{J}_{m_b}}$, avec $b = 1, \ldots, n!$ l'ensemble des vecteurs permutés de $\mathbf{S}_{\mathcal{J}_m}$. La p-valeur non ajustée $p_{m,m'}$ du test est alors définie comme

le nombre de permutations (sur le total de permutations $n!$) menant à une valeur de statistique de test supérieure ou égale à la statistique de test observée sur les données non permutées :

$$p_{m,m'} = \frac{\sum_{b=1}^{n!} \mathbb{I}(T_{m,m'}(\mathbf{S}^*{}_{\mathcal{J}_{m_b}}, \mathbf{S}_{\mathcal{J}_{m'}}) \geq T_{m,m'}(\mathbf{S}_{\mathcal{J}_m}, \mathbf{S}_{\mathcal{J}_{m'}}))}{n!},$$

avec $\mathbb{I}$ la fonction indicatrice. Comme les permutations sont de vraisemblance invariante sous $H_{0,m,m'}$, le test est exact. En pratique, le nombre de permutations possibles $n!$ est très élevé. Pour éviter d'explorer toutes les permutations possibles, on peut estimer la p-valeur en considérant un sous-échantillon des permutations possibles grâce à la méthode de simulation de Monte Carlo, en échantillonnant aléatoirement un ensemble de B permutations et en calculant

$$p_{m,m'} = \frac{\sum_{b=1}^{B} \mathbb{I}(T_{m,m'}(\mathbf{S}^*{}_{\mathcal{J}_{m_b}}, \mathbf{S}_{\mathcal{J}_{m'}}) \geq T_{m,m'}(\mathbf{S}_{\mathcal{J}_m}, \mathbf{S}_{\mathcal{J}_{m'}}))}{B}. \tag{5.2}$$

Comme $\hat{c}_{jj'}$ est un estimateur consistant de $c_{jj'}$, la statistique de test est stochastiquement supérieure sous $H_{1,m,m'}$ que sous $H_{0,m,m'}$, et le test est donc aussi consistant (Pesarin and Salmaso, 2010).

5.2.2 Prise en compte de la multiplicité des tests

L'étude des dépendances est étendue à la matrice de corrélation complète, et chaque couple de blocs consécutifs est testé, ainsi que les groupes de blocs consécutifs, ce qui permet de considérer le caractère fonctionnel des données avec l'ordre des variables. On effectue donc de façon conjointe $M(M-1)/2$ tests, et une fois que le test pour chaque sous-matrice est effectué, il est important d'ajuster les résultats pour prendre en compte la multiplicité des tests. En pratique, on calcule une p-valeur ajustée prenant en compte la multiplicité et la structure en intervalle des tests. Cependant, un nombre différent de tests est effectué selon la position des blocs dans la matrice complète, ce qui est à prendre en compte lors du calcul de la p-valeur ajustée. En effet, les couples de blocs situés aux "bords" de la matrice entrent moins souvent en compte dans les groupes de blocs successifs, et sont moins souvent testés, contrairement aux couples de blocs centraux hors diagonale de la matrice.

La correction de la p-valeur permet le contrôle de la probabilité de rejeter à tort l'hypothèse nulle par intervalle, ce qui correspond au contrôle de l'erreur de première espèce entraînant la mise en évidence de faux positif. Avec cette procédure, on effectue un contrôle de l'erreur de première espèce par intervalle du domaine. Cela signifie que la probabilité de détecter à tort une dépendance significative entre les deux populations dans n'importe quel intervalle où il n'y a pas de dépendance est contrôlée au seuil choisi (Pini and Vantini, 2017).

On propose de définir la p-valeur ajustée en se basant sur la procédure suivante.

- Effectuer les tests (5.1) d'indépendance entre chaque couple de blocs $\mathcal{J}_m$ et $\mathcal{J}_{m'}$ de la partition avec le test de permutation décrit dans le dernier paragraphe. On note $p_{m,m'}$ la p-valeur obtenue pour chaque couple de blocs.

- Effectuer un test d'indépendance entre chaque couple d'intervalles non chevauchant de blocs $\mathcal{J}_{\mathcal{I}} = \cup_{i=m}^{m+h}\{\mathcal{J}_i\}$ et $\mathcal{J}_{\mathcal{I}'} = \cup_{i=m'}^{m'+h'}\{\mathcal{J}_i\}$ avec $1 \leq m \leq m+h < m' \leq m'+h' \leq M$. De tels tests peuvent aussi être effectués avec le test de permutation décrit dans la section précédente, en remplaçant $\mathcal{J}_m$ et $\mathcal{J}_{m'}$ par $\mathcal{J}_{\mathcal{I}}$ et $\mathcal{J}_{\mathcal{I}'}$, respectivement. Remarquons qu'avec $h = h' = 0$, on trouve le test sur les blocs $\mathcal{J}_m$ et $\mathcal{J}_{m'}$ de nouveau. On note $p_{\mathcal{I},\mathcal{I}'}$ la p-valeur obtenue.

- Pour chaque couple de blocs $\mathcal{J}_m$ et $\mathcal{J}_{m'}$, on calcule la p-valeur ajustée selon :

$$\tilde{p}_{m,m'} = \max_{\mathcal{I}:m\in\mathcal{I},\mathcal{I}':m\in\mathcal{I}'} p_{\mathcal{I},\mathcal{I}'}.$$

  On note que les p-valeurs des tests sur les blocs $\mathcal{J}_m$ et $\mathcal{J}_{m'}$ sont aussi incluses dans la maximisation.

5.2.3 Procédure de tests pour les matrices de corrélation

La procédure complète d'étude de la structure des dépendances entre zones spectrales peut être résumée par plusieurs étapes. À partir d'un ensemble de spectres mesurés, les étapes suivantes sont effectuées :

- Projection des spectres sur une base de splines. La dimension de la base est sélectionnée pour permettre de conserver suffisamment d'information tout en travaillant sur un faible nombre de fonctions de base, selon la méthode détaillée en section 4.2.4. Par la suite, les analyses sont menées sur les coefficients de projection dans la base de fonctions.

- Sélection de blocs discriminants de variables (la discrimination se fait selon la variable réponse y), grâce à la procédure de tests détaillée en section 4.3. Par la suite, les analyses sont menées sur les données restreintes aux intervalles de variables sélectionnés.

- Estimation de la corrélation sur les données étudiées.

- Sur la matrice de corrélation estimée, la significativité des coefficients est étudiée par blocs :

  - Tests d'hypothèse par sous-matrices de la matrice de corrélation, basés sur les tests de permutation, permettant l'obtention d'une p-valeur non ajustée associée à chaque bloc de la matrice selon (5.2).
  - Prise en compte de la multiplicité des tests et ajustement de la p-valeur selon le nombre de tests effectués sur chaque sous-matrice considérée.

➔ Obtention de la p-valeur ajustée pour chaque bloc structurant la matrice de corrélation, illustrant la significativité des dépendances entre intervalles de variables (et donc zones spectrales).

5.2.4 Extension aux matrices de précision

Il est possible d'étendre la procédure de tests à l'étude de la matrice de précision $\mathbf{\Omega}$, pour tester l'indépendance conditionnelle. Les blocs que l'on a introduits sont, là encore, associés aux sous-matrices de la matrice $\mathbf{\Omega}$. Il faut noter que l'information portée par la matrice de précision est différente de celle portée par la matrice de corrélation, et les conclusions tirées de ces tests portent sur l'indépendance conditionnelle entre les blocs. De façon plus générale, si on décide de réaliser les tests à la fois sur la matrice de corrélation et sur la matrice de précision, les résultats des deux tests ne seront pas forcément cohérents entre eux.

Plusieurs difficultés entrent en jeu lors de l'étude des matrices de précision. Tout d'abord, les estimateurs de la matrice de précision sont souvent non consistants et ne permettent pas de définir facilement une version standardisée de l'estimation. Ici, nous nous basons sur les travaux de Cai and Liu (2016) et de Xia et al. (2018) pour son estimation.

Une fois la matrice de précision estimée, on souhaite tester l'hypothèse d'indépendance conditionnelle entre les blocs. Dans le cas gaussien, cela revient à tester si les coefficients de la matrice de précision sont égaux à zéro. En particulier, pour tout couple d'intervalles de variables $\mathcal{J}_m, \mathcal{J}_{m'}$ avec $1 \leq m < m' \leq M$, on teste :

$$H_{0,m,m'} : \mathbf{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0 \quad \text{versus} \quad H_{1,m,m'} : \mathbf{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} \neq 0. \tag{5.3}$$

Une autre difficulté est alors la nécessité de définir une statistique de test dont la distribution soit stochastiquement supérieure sous l'hypothèse alternative que sous l'hypothèse nulle.

On utilise là encore des tests de permutation pour tester l'hypothèse (5.3). Dans ce cas, l'hypothèse nulle $\mathbf{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0$ signifie que les sous-vecteurs $\mathbf{S}_{i_{\mathcal{J}_m}}$ et $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$ sont conditionnellement indépendants sachant $\mathbf{S}_{i_{\{1,\dots,p\}\setminus(\mathcal{J}_m \cup \mathcal{J}_{m'})}}$. Dans le cas d'indépendance conditionnelle, le sous-vecteur $\mathbf{S}_{i_{\mathcal{J}_m}}$ n'est pas échangeable par rapport aux observations en gardant $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$ fixé. Par conséquent, il faut changer le type de permutations que l'on utilise pour le test. La seule exception est le cas $\mathcal{J}_m \cup \mathcal{J}_{m'} = \{1, \dots, p\}$, où l'hypothèse nulle $\mathbf{\Omega}_{\mathcal{J}_m \times \mathcal{J}_{m'}} = 0$ est équivalente à l'indépendance (inconditionnelle) entre $\mathbf{S}_{i_{\mathcal{J}_m}}$ et $\mathbf{S}_{i_{\mathcal{J}_{m'}}}$.

Si $\mathcal{J}_m \cup \mathcal{J}_{m'} \subset \{1, \dots, p\}$, il est possible de calculer les résidus non corrélés et de les permuter. Cela mène à un test exact asymptotiquement. Dans le cas gaussien, l'indépendance conditionnelle est équivalente à l'indépendance conditionnelle linéaire :

$$\begin{aligned}
\mathbf{S}_{\mathcal{J}_m} &= \mathbf{S}_{\{1,\dots,p\}\setminus(\mathcal{J}_m \cup \mathcal{J}_{m'})}\mathbf{A} + \boldsymbol{\varepsilon}_{\mathcal{J}_m} \\
\mathbf{S}_{\mathcal{J}_{m'}} &= \mathbf{S}_{\{1,\dots,p\}\setminus(\mathcal{J}_m \cup \mathcal{J}_{m'})}\mathbf{A}' + \boldsymbol{\varepsilon}_{\mathcal{J}_{m'}},
\end{aligned}$$

avec $\mathbf{A}$ et $\mathbf{A}'$ deux matrices de dimension $n \times p - \text{card}(\mathcal{J}_m) - \text{card}(\mathcal{J}_{m'})$, et $\boldsymbol{\varepsilon}_{\mathcal{J}_m}$ et $\boldsymbol{\varepsilon}_{\mathcal{J}_{m'}}$ des résidus mutuellement indépendants. On utilise ici la notation matricielle pour les régressions linéaires multivariées de $\mathbf{S}_{\mathcal{J}_m}$ et $\mathbf{S}_{\mathcal{J}_{m'}}$ sur un ensemble de coefficients $\mathbf{S}_{\{1,\dots,p\}\setminus(\mathcal{J}_m \cup \mathcal{J}_{m'})}$. En détail, $\mathbf{S}_{\mathcal{J}_m}$ est une matrice de dimension $(n \times \text{card}(\mathcal{J}_m))$ dont les lignes sont les vecteurs $\mathbf{S}_{i_{\mathcal{J}_m}}$ avec $i = 1, \dots, n$, et de même pour les matrices $\mathbf{S}_{\mathcal{J}_{m'}}$

et $\mathbf{S}_{\{1,\ldots,p\}\setminus(\mathcal{J}_m\cup\mathcal{J}_{m'})}$; de façon similaire, $\boldsymbol{\varepsilon}_{\mathcal{J}_m}$ (respectivement $\boldsymbol{\varepsilon}_{\mathcal{J}_{m'}}$) est une matrice de dimension $n \times \text{card}(\mathcal{J}_m)$ (respectivement $n \times \text{card}(\mathcal{J}_{m'})$) dont les lignes sont les vecteurs résiduels $\boldsymbol{\varepsilon}_{i_{\mathcal{J}_m}}$ ($\boldsymbol{\varepsilon}_{i_{\mathcal{J}_{m'}}}$). Sous l'hypothèse nulle d'indépendance conditionnelle, les résidus $\boldsymbol{\varepsilon}_{i\mathcal{J}_m}$ sont échangeables par rapport aux observations en gardant $\boldsymbol{\varepsilon}_{i\mathcal{J}_{m'}}$ fixé, et vice-versa. On peut donc estimer les résidus et utiliser des permutations de ces résidus. Puisque les résidus estimés sont seulement asymptotiquement échangeables, le test est dans ce cas exact asymptotiquement (Freedman and Lane, 1983).

5.3 Étude de simulation

L'objectif de l'étude de simulation est d'évaluer la capacité de notre méthode à détecter les blocs significatifs structurant la matrice de corrélation estimée sur les données. On cherche à évaluer les paramètres agissant sur les performances de notre méthode. Par exemple, le nombre de tests effectués sur un couple de bloc étant différent selon la zone des blocs au sein de la matrice, on cherche à savoir si certains blocs sont plus ou moins faciles à détecter selon leur localisation. On cherche aussi à déterminer si les valeurs de corrélation au sein de ces blocs ont une influence sur la capacité de détection de ceux-ci.

Pour évaluer la capacité à retrouver les blocs significatifs, on évalue le pourcentage de rejet de l'hypothèse nulle, correspondant au pourcentage de significativité de la dépendance entre deux blocs, ainsi que la capacité à retrouver les dépendances significatives et les coefficients non significatifs grâce aux critères de sensibilité et de spécificité.

5.3.1 Modélisation de matrices de covariance structurées

Dans notre cadre, il est nécessaire de simuler des matrices de covariance structurées en blocs pour générer les données de l'étude de simulation. Une des difficultés lors de la simulation de matrice de covariance est l'obtention de matrice définie positive. De plus, dans notre étude, on souhaite générer des données dont les matrices de covariance ont une structure en groupes, avec des blocs de covariance nulle hors de la diagonale. Pour cela, nous développons une méthode de modélisation de matrices de covariance structurées, basée sur la diagonalisabilité des matrices de covariance. En effet, pour une matrice de covariance M, la diagonalisation permet d'écrire :

$$M = P^{-1}DP, \tag{5.4}$$

avec P la matrice des vecteurs propres de M et D la matrice diagonale des valeurs propres de M. Pour modéliser les matrices de covariance structurées, notre approche se base sur la modélisation de la matrice P avec une structure en blocs.

Nous commençons par définir les blocs structurant la matrice, c'est-à-dire le nombre d'intervalles de variables consécutives regroupées en blocs et leur bornes. Ensuite, la matrice des vecteurs propres P est modélisée. Les couples blocs significatifs sont générés par des matrices de rotation aléatoires, qui sont orthogonales par définition, et intégrées dans la matrice de covariance comme les sous-matrices correspondant aux blocs étudiés. Les coefficients de la matrice de covariance correspondant à des couples de blocs non

significatifs sont fixés à zéro. On obtient donc une matrice P diagonale par blocs, avec des blocs hors diagonale non nuls. La matrice M est ensuite calculée grâce à l'équation (5.4). Cette méthode a l'avantage de générer des matrices de covariance définies positives, symétriques et facilement inversibles avec une structure parcimonieuse par blocs.

5.3.2 Cadre de simulation

Génération des données Tout d'abord, la matrice de covariance M de dimension $p \times p$, structurée en b blocs, est générée. Ensuite, n observations issues d'une distribution gaussienne multivariée de paramètres μ et de covariance M sont générées.

Paramètres de simulation Plusieurs cas de simulation sont étudiés. Deux cas de figure principaux sont évalués, un cas avec une matrice de covariance où seul un bloc hors diagonal est significatif (cas 1), et un cas où plusieurs blocs hors diagonale sont significatifs (cas 2). Les paramètres que l'on fait varier sont le nombre d'observations n, le nombre de variables p, le nombre de blocs b, l'emplacement des blocs ainsi que les valeurs de covariance non nulles. Les différents paramètres de simulation étudiés pour chaque cas sont présentés en table 5.1.

Selon le nombre de blocs et de variables, nous considérons différentes configurations pour l'emplacement des blocs significatifs de façon à évaluer l'effet du nombre de tests effectués sur le bloc. Lorsque l'on évalue des données simulées avec 100 variables, dans le premier cas de simulation, le bloc significatif est évalué en bord de matrice (configuration 1), proche de la diagonale (configuration 2) ou en position centrale hors diagonale (configuration 3). Les configurations 2 et 3 du cas à un bloc significatif (cas 1) sont représentées par les matrices représentant la structure de la covariance de simulation en figure 5.2. De la même façon, dans le deuxième cas de simulation, les blocs sont évalués en bord de matrice, proche de la diagonale et en position centrale à la fois (configuration 1) et sur une même "ligne" où un bloc a des dépendances avec tous les autres blocs (configuration 2). Les configurations du cas à plusieurs blocs significatifs (cas 2) sont représentées par les matrices représentant la structure de la covariance de simulation en figure 5.3. L'effet de l'ordre de grandeur des valeurs de covariance est aussi évalué en considérant des matrices de covariance à faibles valeurs (cas D1), d'autres à fortes valeurs (cas D2), des matrices à covariance plus élevée pour les blocs significatifs (cas D3) et des matrices à covariance plus faible pour les blocs significatifs (cas D4). On effectue 50 répétitions de chaque combinaison de paramètres de simulation.

5.3.3 Résultats

La capacité de la procédure à détecter les dépendances significatives est évaluée grâce au critère de sensibilité, correspondant ici au rapport entre le nombre de coefficients de corrélation non nuls détectés comme significatifs par la procédure et le total des coefficients de corrélation non nuls. On évalue la capacité de la procédure à éliminer les dépendances

	1 bloc significatif	Plusieurs blocs significatifs
p (nombre de variables)	20, 100	100
n (nombre d'observations)	pour $p = 20 : 100, 200$ pour $p = 100 : 500, 1000$	 500, 1000
b (nombre de blocs)	Pour $p = 20 : 3, 5$ Pour $p = 100 : 5, 10$	 10
Configuration (blocs significatifs)	1 : bloc au bord 2 : bloc proche diagonale 3 : bloc central hors diagonale	1 : blocs au bord, au milieu, proche diagonale 2 : une ligne de blocs
Valeurs de covariance	D1 : valeurs faibles D2 : valeurs élevées D3 : valeurs plus élevées aux blocs significatifs D4 : valeurs plus faibles aux blocs significatifs	

TABLE 5.1 : **Paramètres utilisés pour les deux cas de simulation.**

non significatives grâce au critère de spécificité, correspondant au rapport entre le nombre de coefficients de corrélation nuls détectés comme non significatifs par la procédure et le total des coefficients de corrélation nuls. Les valeurs moyennes de sensibilité et de spécificité calculées sur les 50 répétitions pour le cas de simulation 1 sont détaillées dans les tables 5.2 et 5.3. Les valeurs moyennes de sensibilité et de spécificité calculées sur les 50 itérations pour le cas de simulation 2 sont détaillées dans la table 5.4. Des résultats supplémentaires obtenus par cette étude de simulation sont présentés en annexe B.

Il est possible de visualiser de façon globale la capacité de la procédure à retrouver les structures de dépendances au sein de la matrice de corrélation. La probabilité de détection d'un coefficient significatif peut être calculée sur les 50 répétitions de chaque combinaison de paramètres, pour chaque coefficient et représentée dans une matrice. Ces matrices sont représentées en figures 5.2 et 5.3 pour certains cas de simulation, et permettent de mettre en évidence les différences entre la structure estimée et la vraie structure des données pour ces cas.

Globalement, on observe pour tous les cas de simulation de très bonnes performances à la fois en termes de sensibilité et de spécificité. La procédure permet de retrouver de façon efficace les dépendances significatives au sein de la matrice de corrélation des données, et d'éliminer les dépendances non significatives. Cependant, dans certains cas, les performances sont légèrement moins bonnes. On remarque notamment que dans le cas où la matrice de simulation comporte un bloc significatif très proche de la diagonale, la capacité de la procédure à le trouver peut être diminuée (voir par exemple la configuration 2 en table 5.2 ou la configuration 2 en table 5.3). La figure 5.2 permet de visualiser la difficulté à retrouver un bloc significatif proche de la diagonale. Cela peut être relié au nombre de tests effectués sur les blocs proches de la diagonale qui est plus faible que sur les blocs plus éloignés de la diagonale et en position centrale hors diagonale, représentés sur la même figure par les matrices de gauche. On note que dans ce cas, un bloc de très

		Config. 1		Config. 2	
		Se	Sp	Se	Sp
D1	n = 100	1 (1-1)	0.96 (1-1)	1 (1-1)	0.98 (1-1)
	n = 200	1 (1-1)	0.95 (1-1)	1 (1-1)	0.93 (1-1)
D2	n = 100	0.98 (1-1)	0.96 (1-1)	1 (1-1)	0.91 (1-1)
	n = 200	1 (1-1)	0.98 (1-1)	1 (1-1)	0.97 (1-1)
D3	n = 100	0.99 (1-1)	0.96 (1-1)	0.99 (1-1)	0.98 (1-1)
	n = 200	1 (1-1)	0.97 (1-1)	1 (1-1)	0.95 (1-1)
D4	n = 100	1 (1-1)	0.96 (1-1)	1 (1-1)	0.91 (1-1)
	n = 200	1 (1-1)	0.96 (1-1)	1 (1-1)	0.93 (1-1)

TABLE 5.2 : **Sensibilités et spécificités moyennes obtenues à l'issue de la procédure de tests pour le cas de simulation 1, avec des données simulées à 20 variables réparties en 3 blocs.** La moyenne est calculée sur les 50 répétitions de chaque cas de simulation. Les valeurs entre parenthèses correspondent aux premier et troisième quartiles des valeurs.

		Config. 1		Config. 2		Config. 3	
	n	Se	Sp	Se	Sp	Se	Sp
D1	500	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
	1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D2	500	0.96 (0.95-1)	1 (0.99-1)	0.97 (0.93-1)	1 (0.99-1)	0.99 (0.99-1)	0.99 (0.99-1)
	1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D3	500	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (1-1)
	1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D4	500	0.99 (0.96-1)	0.99 (0.99-0.99)	**0.94 (0.93-0.93)**	**1 (1-1)**	**0.98 (0.93-1)**	**0.99 (0.99-0.99)**
	1000	1 (1-1)	0.99 (1-1)	1 (1-1)	0.98 (1-1)	1 (1-1)	1 (1-1)

TABLE 5.3 : **Sensibilités et spécificités moyennes obtenues à l'issue de la procédure de tests pour le cas de simulation 1, avec des données simulées à 100 variables réparties en 10 blocs.** La moyenne est calculée sur les 50 répétitions de chaque cas de simulation. Les valeurs entre parenthèses correspondent aux premier et troisième quartiles des valeurs. Les valeurs en gras correspondent aux cas de simulation représentés en figure 5.2.

		Config. 1		Config. 2	
		Se	Sp	Se	Sp
D1	n = 500	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
	n = 1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D2	n = 500	**1 (1-1)**	**0.99 (0.99-0.99)**	**0.95 (0.9-1)**	**0.99 (0.99-1)**
	n = 1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D3	n = 500	1 (1-1)	0.99 (0.99-0.99)	0.96 (0.92-1)	0.99 (0.99-0.99)
	n = 1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)
D4	n = 500	1 (1-1)	0.99 (0.99-0.99)	0.96 (0.94-1)	0.99 (0.99-0.99)
	n = 1000	1 (1-1)	0.99 (0.99-0.99)	1 (1-1)	0.99 (0.99-0.99)

TABLE 5.4 : **Sensibilités et spécificités moyennes obtenues à l'issue de la procédure de tests pour le cas de simulation 2, avec des données simulées à 100 variables réparties en 10 blocs.** La moyenne est calculée sur les 50 répétitions de chaque cas de simulation. Les valeurs entre parenthèses correspondent aux premier et troisième quartiles des valeurs. Les valeurs en gras correspondent aux cas de simulation représentés en figure 5.3.

petite dimension n'a pas été détecté, et la procédure a détecté une structure à 9 blocs au lieu de détecter une structure à 10 blocs. Le bloc de petite dimension a été intégré à un bloc voisin.

De la même façon, lorsque le bloc à retrouver est de très petite dimension, la procédure de tests donne dans certains cas de moins bonnes performances, ce qui se retrouve par exemple dans le cas de la configuration 3 en table 5.3. Dans ce cas, la procédure retrouve bien la significativité du bloc au bon emplacement, mais ne retrouve pas la bonne structure de bloc et fait ressortir un bloc de plus grande dimension, et détecte à tort un bloc significatif proche de la diagonale dans certains cas. Ce cas de figure est représenté en figure 5.2. Dans ce cas, le bloc de petite dimension est bien détecté et non intégré dans un bloc voisin et la procédure retrouve bien une structure à 10 blocs, mais certaines variables voisines non significatives sont intégrées à ce bloc, pour former un bloc de plus grande dimension. La bonne performance de détection est liée au nombre élevé de tests effectués sur les blocs situés en position centrale hors de la diagonale.

Dans la configuration 1 du cas de simulation 2, représentée en figure 5.3, comme observé précédemment, le bloc de très petite dimension n'est pas détecté et intégré à un bloc voisin, et la structure de bloc trouvée comporte 9 blocs au lieu de 10 blocs. De plus, le bloc proche de la diagonale est détecté avec une probabilité légèrement plus faible, du fait du nombre de tests plus faible effectués à cet emplacement. La configuration 2 du cas de simulation 2 regroupe les différents cas les plus difficiles, ce qui peut expliquer les performances légèrement moins bonnes pour ce cas de simulation. Les blocs en position centrale sont retrouvés avec une meilleure probabilité que les blocs proches de la diagonale ou des bords de la matrice. Comme observé précédemment, le bloc de très petite dimension est bien détecté, mais avec une dimension plus importante, et la procédure détecte à tort un bloc significatif proche de la diagonale à cet emplacement.

En conclusion, cette étude de simulation révèle que la méthode proposée pour identifier des blocs de corrélation non nulle dans une matrice de corrélation est efficace, même pour des tailles d'échantillon assez faibles (n=100, p=20) avec une sensibilité moyenne toujours égale à 1 et spécificité moyenne égale à 0,91 dans les cas les plus difficiles. On note cependant que si les blocs sont petits, ils peuvent ne pas être détectés ou être assimilés à des blocs plus grands.

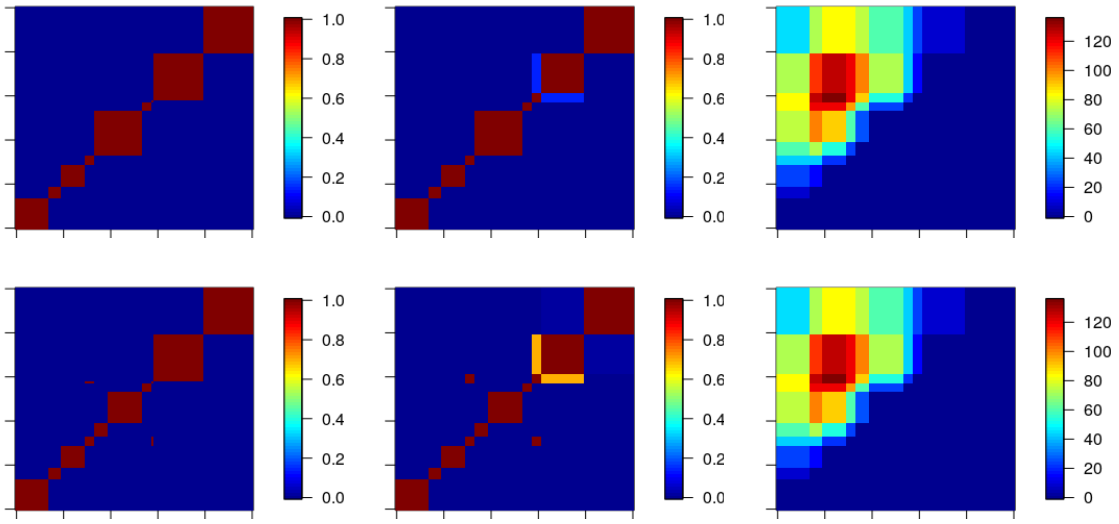

FIGURE 5.2 : **Capacité de la procédure à retrouver la structure de dépendance des matrices de corrélation dans le premier cas de simulation.** Les matrices à gauche représentent la structure des matrices de covariance simulées, que l'on cherche à retrouver. Les matrices au milieu représentent les matrices de probabilité de rejet de l'hypothèse nulle. Chaque coefficient de ces matrices correspond à la probabilité de rejet de l'hypothèse nulle obtenue par la procédure de tests par blocs sur les matrices de covariance, calculée sur les 50 répétitions de chaque cas de simulation. Les matrices de droite représentent le nombre de tests effectués pour chaque bloc, pour une des 50 répétitions du cas de simulation étudié. Les matrices représentées correspondent au premier cas de simulation, pour 500 observations, 100 variables, la matrice de simulation D4 pour la deuxième configuration (en haut) et la troisième configuration (en bas).

5.4 Application sur données réelles

Le travail sur les données d'application NASH présentées dans cette thèse a aussi pour objectif de permettre une meilleure compréhension des processus moléculaires liés à l'évolution de la maladie. Il existe encore des questionnements concernant les interactions entre les molécules impliquées dans le métabolisme et l'évolution de la NASH. La composition des échantillons prélevés sur des patients atteints de NASH et des patients

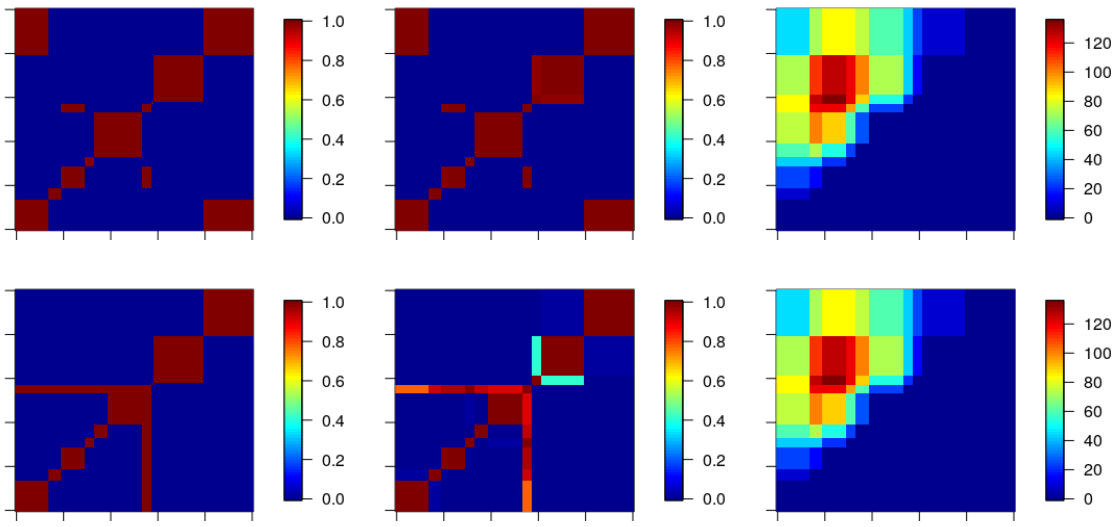

FIGURE 5.3 : **Capacité de la procédure à retrouver la structure de dépendance des matrices de corrélation dans le deuxième cas de simulation.** Les matrices à gauche représentent la structure des matrices de covariance simulées, que l'on cherche à retrouver. Les matrices au milieu représentent les matrices de probabilité de rejet de l'hypothèse nulle. Chaque coefficient de ces matrices correspond à la probabilité de rejet de l'hypothèse nulle obtenue par la procédure de tests par blocs sur les matrices de covariance, calculée sur les 50 répétitions de chaque cas de simulation. Les matrices de droite représentent le nombre de tests effectués pour chaque bloc, pour une des 50 répétitions du cas de simulation étudié. Les matrices représentées correspondent au deuxième cas de simulation, pour 500 observations, 100 variables, la matrice de simulation D2 pour la première configuration (en haut) et la deuxième configuration (en bas).

témoins étant reflétée par les données de spectrométrie, il serait intéressant d'étudier les interactions entre les zones spectrales selon le diagnostic. Ces zones spectrales pouvant être reliées par leurs positions à des types de molécules, leurs interactions peuvent refléter des processus moléculaires mis en jeu dans la maladie. Ces interactions peuvent être considérées grâce aux matrices de covariance qui reflètent les dépendances entre variables. On peut s'attendre à ce que des portions de courbes soient dépendantes, et entraînent une structure en blocs dans la matrice de covariance. Plus particulièrement, la méthode de sélection de portions de courbes présentée au chapitre 4 a permis de mettre en évidence la présence de zones discriminantes dans le spectre, montrant des différences significatives selon le diagnostic de NASH. L'objectif de cette thèse étant la prédiction de la variable de diagnostic, nous nous intéressons à ces bandes discriminantes spécifiques. Dans cette partie, nous faisons donc l'hypothèse que ces zones discriminantes se reflètent sur la structure en blocs des matrices de covariance. On suppose que chaque zone discriminante correspond à un bloc structurant la matrice de covariance. Nous faisons donc le choix de

nous restreindre à l'étude des dépendances entre blocs de variables discriminants, sur des matrices de covariance réduites à un ensemble de variables pré-sélectionnées.

5.4.1 Étude des dépendances moléculaires selon le profil métabolique

L'étude des données de spectrométrie recueillies sur la cohorte NASH présentée dans le chapitre 4 a permis d'estimer des groupes latents au sein des données. La table 4.4 présentant les caractéristiques cliniques des groupes indique que la proportion de patients malades n'est pas significativement différente selon le groupe, contrairement à certaines variables cliniques reflétant l'état métabolique des patients. On se trouve donc face à des patients dont l'évolution de la maladie est marquée par des caractéristiques métaboliques différentes. Les groupes latents estimés semblent donc correspondre à deux types de profils métaboliques distincts.

Il est intéressant d'étudier ces différences et d'essayer de comprendre les dépendances entre groupements moléculaires selon le profil du patient. Pour cela, nous nous concentrons sur les dépendances entre variables par groupe latent, et nous étudions donc les matrices de corrélations des données pour chacun des groupes latents. Les zones spectrales étudiées étant restreintes aux zones discriminantes, l'information de diagnostic et l'information de profil métabolique sont prises en compte de façon conjointe. Les matrices de covariance empiriques estimées sur les données de spectrométrie pour chaque groupe latent sont présentées en figure 5.4. On note la présence de blocs de variables, qui semblent différents selon le groupe considéré. Les matrices de corrélation représentées sur cette même figure sont les matrices sur lesquelles la procédure de tests est effectuée.

5.4.2 Analyses

Nous considérons les données recueillies sur la cohorte NASH, décrites au chapitre 1.

La sélection du nombre de nœuds considéré dans la base de B-splines est effectuée selon la méthode détaillée en section 4.2.4. La sélection des blocs discriminants structurants le spectre est ensuite réalisée sur les coefficients de projection dans la base grâce à la procédure de sélection d'intervalles présentée en section 4.3. Nous considérons ensuite les données par groupe latent selon les groupes estimés en section 4.4. La procédure décrite en section 5.2.3 est ensuite appliquée à chaque groupe latent, sur les données restreintes aux portions de courbes discriminantes sélectionnées grâce à la procédure décrite au chapitre 4.

5.4.3 Résultats et interprétations

Les matrices représentant les p-valeurs ajustées obtenues à l'issue de la procédure de tests sont représentées en figure 5.5. Pour chaque groupe latent, toute la matrice est détectée comme significative par la procédure. Cela peut être dû au fait que de trop nombreuses dépendances sont réparties partout dans les matrices et dans chaque bloc défini par la CAH, ce qui entraîne uniquement des blocs significatifs. Il est alors difficile d'identifier des zones où les différences sont plus importantes. On note à la figure 5.4 que les dépendances

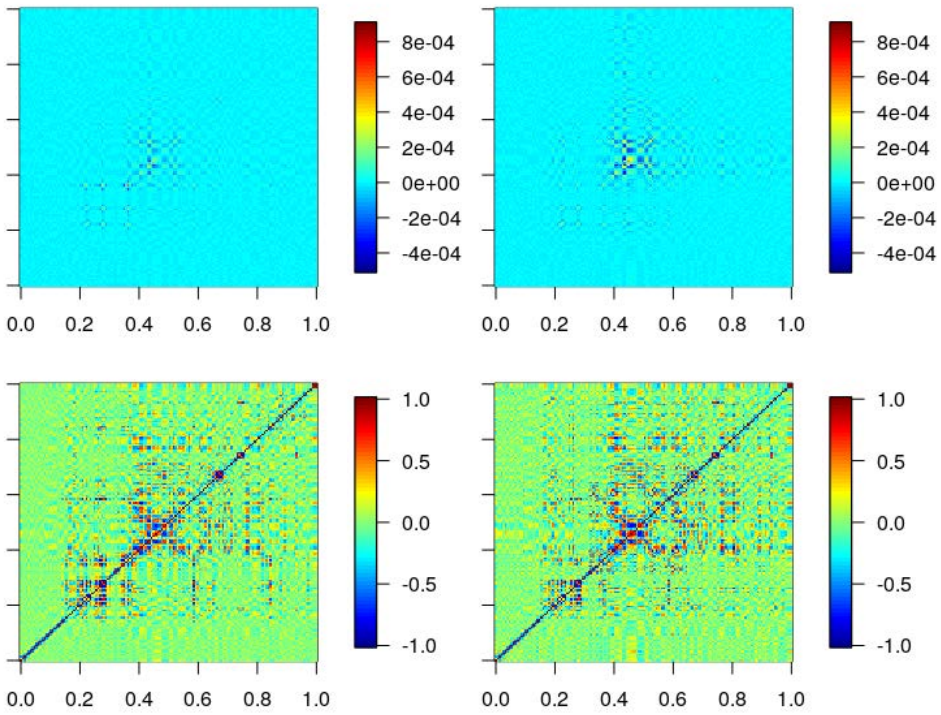

FIGURE 5.4 : **Matrices de covariance empiriques et matrices de corrélation calculées sur les données de spectrométrie NASH par groupe latent.** La matrice de covariance en haut à gauche et la matrice de corrélation en bas à gauche sont calculées sur les patients du premier groupe latent correspondant au profil aux dérèglements métaboliques les plus importants. La matrice de covariance en haut droite et la matrice de corrélation en bas à droite sont calculées sur patients du deuxième groupe latent.

visibles sur les matrices de covariance sont réparties sur des blocs de très petites tailles, ne correspondant pas aux blocs prédéfinis par la CAH avant la procédure. Ces dépendances sont présentes dans tous les blocs spécifiés de la matrice, et mènent à une détection de blocs de dépendances significatifs partout dans les matrices de corrélation. La présence de dépendances nombreuses et réparties dans tous les blocs a donc pour effet de mettre en avant trop de dépendances significatives, comprenant aussi des coefficients normalement non significatifs. Dans le cadre d'application, on ne retrouve pas les caractéristiques du cadre de simulation, où des blocs, généralement de grande dimension, représentent une dépendance très significative par rapport à des blocs de corrélation nulle. Comparées aux matrices de covariance de l'étude de simulation, les matrices calculées sur les données NASH ne semblent donc pas avoir une structure assez marquée pour permettre de mettre en avant certains blocs, ce qui pourrait expliquer les résultats obtenus par la procédure sur ces données. Les résultats obtenus pour l'instant correspondent aux résultats obtenus avec la procédure standard, et une étude plus poussée pourrait être menée pour mieux

comprendre les dépendances entre les variables spectrales.

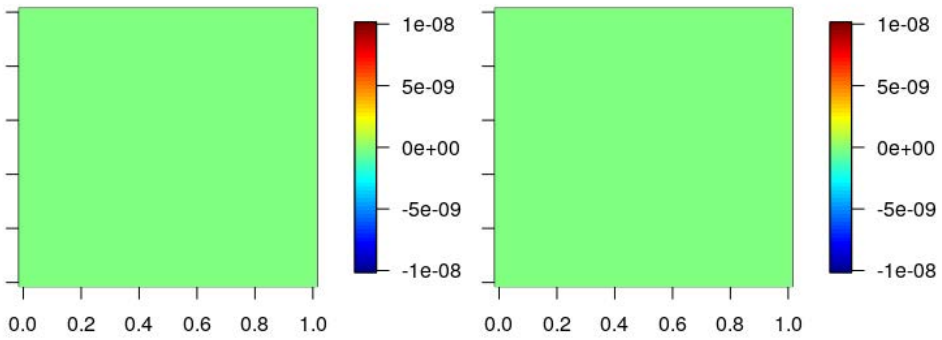

FIGURE 5.5 : **Matrices de p-valeurs ajustées obtenues par la procédure de tests sur les données de spectrométrie NASH par groupe latent.** La matrice à gauche est estimée sur les patients du premier groupe latent correspondant au profil aux dérèglements métaboliques les plus importants. La matrice de droite est estimée sur les patients du deuxième groupe latent.

Chapitre 6

Bilan et perspectives

Contents

6.1 Conclusions

Dans cette thèse, nous avons cherché à répondre à des attentes pratiques concernant la modélisation de données particulières, correspondant à des courbes, dans un contexte médical. L'objectif de ce travail s'articulait autour de plusieurs axes. Nous souhaitions proposer une méthode permettant l'estimation de profils de patients inconnus *a priori*, dans le but de construire un modèle de prédiction d'une variable de diagnostic performant. Les données étudiées étant complexes, l'enjeu était de considérer des méthodes permettant l'obtention de modèles interprétables.

Premièrement, nous avons proposé une méthode permettant de coupler le partitionnement et la prédiction d'une variable réponse, basée sur les modèles génératifs permettant d'estimer des groupes latents au sein d'une population. La flexibilité du type de modèles utilisé permet une généralisation aux lois de probabilités conditionnelles usuelles, et donc à la prédiction d'une variable binaire de diagnostic. Deuxièmement, nous avons adapté cette méthode au cadre de la modélisation de données en dimension modérée, nécessitant une sélection de l'information pertinente pour la prédiction. Une méthode de sélection de paramètres basée sur la régularisation a été intégrée à l'algorithme d'estimation du modèle. Ensuite, nous avons étendu cette méthode à la modélisation de données fonctionnelles, c'est-à-dire dans le cas où les données observées correspondent à des fonctions, mesurées dans un espace de dimension infinie. Les méthodes développées ont été appliquées à l'analyse de données de spectrométrie infrarouge.

Tout au long de cette thèse, l'accent a été mis sur l'interprétabilité des modèles obtenus,

dans le but de mieux comprendre les problèmes médicaux étudiés. Cet aspect a notamment été pris en compte par l'utilisation d'outils de visualisation des dépendances entre variables comme les modèles graphiques, mais aussi par des méthodes de pré-sélection de portions de courbes, permettant de mettre en avant les zones spectrales informatives dans notre cadre d'étude. Dans un dernier temps, l'étude des interactions entre les portions de courbes a été menée sur les matrices de corrélation estimées sur les données. Ce dernier travail avait pour objectif de faire ressortir les dépendances entre les zones spectrales prédictives de la maladie étudiée.

D'autre part, la complexité du contexte applicatif était un des enjeux de ce travail. En effet, le diagnostic de la maladie étudiée, la stéatohépatite non-alcoolique, est difficile et invasif, et la méthode proposée offre une alternative qui pourrait être utilisée en pré-diagnostic et éventuellement permettre de diminuer le nombre de biopsies. La méthode développée dans cette thèse et appliquée aux données NASH permet de déterminer un score illustrant la gravité de l'atteinte au foie des patients ainsi que leur profil métabolique, ce qui peut être utilisé pour évaluer les patients dont la maladie est la plus avancée.

En dernier lieu, un aspect de ce travail porte sur l'utilisation des données de spectrométrie dans le cadre médical. Les méthodes d'analyse des données de spectrométrie recueillies selon la technologie étudiée dans cette thèse sont très variables selon les études (voir par exemple Albert et al., 2016; Le Corvec et al., 2012). Dans notre approche, nous apportons un cadre général flexible permettant la modélisation de ce type de données, ce qui permet une homogénéisation des procédures d'analyses et leur application à divers problèmes médicaux.

6.2 Perspectives

Les développements possibles à apporter aux méthodes présentées portent d'abord sur des aspects d'analyse de données fonctionnelles de manière générale, ainsi que sur la spécificité des données de spectrométrie.

Dans le cadre de l'étude de la sélection de zones discriminantes dans les courbes, il serait intéressant de pouvoir comparer les différentes méthodes de sélection de variables étudiées dans cette thèse. Deux approches ont été employées, d'une part les méthodes de régression pénalisée de type Lasso et d'autre part les méthodes de type "filtre" basées dans notre cas sur les tests d'hypothèse. Le principal problème des approches de sélection basées sur les tests d'hypothèse est le recouvrement du support. En effet, comme il n'est pas possible de tester toutes les combinaisons possibles de variables, il est possible de ne pas considérer complètement la "bonne" combinaison de variables, même si cette approche a l'avantage de permettre un contrôle du risque d'erreur de type I, contrairement à l'approche de type régression pénalisée. Un moyen de procéder serait d'utiliser une méthode de régression pénalisée de type Lasso et ensuite une méthode basée sur les tests pour valider la combinaison sélectionnée et avoir un contrôle précis de l'erreur de première espèce.

Dans notre cadre, un des aspects limitant ici est la considération de l'aspect fonctionnel

utilisé dans la procédure de sélection de portions de courbes basée sur les tests de permutation. L'aspect fonctionnel n'est pas considéré par les autres procédures de sélection de variables, qui sont des méthodes utilisées dans le cadre multivarié classique. Il serait néanmoins possible de considérer la sélection de blocs de variables obtenue par Fused-Lasso. Dans ce cas, le travail consisterait en la comparaison d'une méthode de sélection de variables de type "filtre" avec une méthode de type "embedded", dans le cas de données fonctionnelles.

D'un point de vue de la modélisation des données fonctionnelles, nous avons fait le choix de considérer la projection des données sur une base de fonctions splines réparties de façon équidistante sur la fenêtre spectrale. Il est aussi possible de considérer un espacement des nœuds irrégulier dans la base de projection. En effet, le fait de considérer un nombre plus important de nœuds à certaines zones peut permettre d'approcher de façon plus précise les régions du spectre comprenant des variations complexes. Cela implique de connaître *a priori* les zones du spectres les plus complexes qui peuvent être informatives pour le problème étudié.

Un autre aspect de la projection des données fonctionnelles concerne le choix de la base. Dans cette thèse, nous avons fait le choix de travailler sur une base de fonctions splines, permettant de conserver l'information de localisation des coefficients sur la fenêtre spectrale. Le choix de ce type de base a aussi été guidé par sa facilité d'utilisation et ses bonnes propriétés, notamment pour l'approximation de données de spectrométrie. Cependant, il existe des fonctions classiques utilisées en chimiométrie pour modéliser les données spectrales, et il serait cohérent de considérer ce type de fonctions pour la modélisation. Il est notamment envisagé de travailler avec des fonctions lorentziennes et voigtiennes. Une des difficultés lorsque l'on souhaite travailler avec ce type de fonctions concerne l'étape d'ajustement de la projection, qui n'est pas automatique, car il est nécessaire de spécifier la localisation des fonctions le long de la fenêtre spectrale. L'utilisation de ces fonctions adaptées à la modélisation des données de spectrométrie pourrait permettre d'améliorer les performances des modèles estimés.

D'autre part, les méthodes développées et appliquées à l'analyse de la NASH ont permit de soulever de nouveaux axes d'étude. Il serait par exemple intéressant de considérer le caractère non invasif de ces méthodes pour faire un suivi longitudinal des patients et étudier la possibilité d'établir des diagnostics précoces.

Annexe A

Code PMLR

Cette annexe présente le code du mélange de régressions logistiques pénalisées. Un exemple de simulation simple est d'abord présenté, suivi du code avec les fonctions internes ainsi que les fonctions principales de l'algorithme.

A.1 Exemple de simulation

L'utilisation de la méthode PMLR est présentée dans cette partie sur des données simulées.

```
library(mvtnorm)
library(glmnet)
library(matrixcalc)
library(glasso)
library(MASS)
library(brglm2)
```

L'exemple présenté ici correspond au scénario de simulation 4 du chapitre 3, et est basé sur des données simulées avec 3 groupes. Deux groupes ont la même moyenne mais des coefficients de régression différents. La différence entre les groupes est alors seulement portée par $\mathbf{Y}|\mathbf{X}$.

On commence par fixer les paramètres de simulation :

```
K      = 3
p      = 40
N      = 500

mu1    = c(rep(1, p/2), rep(2, p/2))
mu2    = c(rep(1, p))
mu3    = c(rep(1, p/2), rep(2, p/2))
mu     = as.matrix(data.frame(mu_1 = mu1, mu_2 = mu2, mu_3 = mu3))
```

```
beta1  = c(-2, 1, 0.5, 1, rep(0, 32), -1, -0.2, 0.5, 0.2)
beta2  = c(rep(0, 5), 1, -0.5, -0.5, -1, rep(0, 22), 1, -0.2, 1, -0.5, rep(0, 5))
beta3  = c(rep(0, 10), -2, 0.5, -2, -1, 1.5, 1, 2, -1, rep(0, 22))
beta   = as.matrix(data.frame(beta_1 = beta1, beta_2 = beta2, beta_3 = beta3))

SIGMA1 = 2*diag(p)/3
SIGMA2 = 2*diag(p)/3
SIGMA3 = toeplitz(c(1, 0.4, 0.1, 0.1, rep(0, 36)))
SIGMA  = list(SIGMA1, SIGMA2, SIGMA3)
pik    = c(0.3, 0.3, 0.4)
```

Les données sont ensuite simulées selon un modèle de mélange de régressions.

```
set.seed(1)
classe = c()
pY     = c()
Y      = c()
X      = matrix(NA, ncol = p, nrow = N)

for (i in 1:N) {
  k         = sample(1:K, 1, prob = pik)
  classe[i] = k
  X[i, ]    = rmvnorm(1, mean = mu[, k], sigma = SIGMA[[k]])
  pY[i]     = exp(X[i, ] %*% beta[, k])/ (1 + exp(X[i, ] %*% beta[, k]))
  U = runif(length(pY))
  Y = as.numeric(U < pY)
}
```

La procédure d'estimation avec la sélection des paramètres de régularisation peut être exécutée avec la fonction suivante :

```
t0  = Sys.time()
res = procedure_selec_model(Y = Y, X = X, Mmin = 2, Mmax = 2, length.lambda = 5,
      length.rho = 5)
t1  = Sys.time()
```

Les paramètres estimés sont listés dans l'objet "mod". Cet objet contient les cinq meilleurs modèles estimés avec différents paramètres λ et ρ, ainsi que le modèle retenu avec les paramètres de régularisation choisis, dans l'objet "res", contenant tous les paramètres du modèle estimé :

```
mod    = res$res
mod$pik
```

```r
mod$clustering
image.plot(mod$THETA[[1]])
```

Ensuite, une prédiction peut être effectuée pour un nouvel échantillon de données simulées, et les performances de prédiction peuvent être évaluées.

```r
classet = c()
pYt     = c()
Yt      = c()
Xt      = matrix(NA, ncol = p, nrow = N)

for (i in 1:N) {
  k          = sample(1:K, 1, prob = pik)
  classet[i] = k
  Xt[i, ]    = rmvnorm(1, mean = mu[, k], sigma = SIGMA[[k]])
  pYt[i]     = exp(Xt[i, ] %*% beta[, k])/ (1 + exp(Xt[i, ] %*% beta[, k]))
  U  = runif(length(pYt))
  Yt = as.numeric(U < pYt)
}

pred   = prediction(Xt, mod, logit = TRUE)
boxplot(pred$Yhnew ~ Yt)
```

A.1.1 Algorithme EM

Dans cette partie, les fonctions utilisées pour la méthode PMLR sont présentées.

A.1.2 Fonctions internes

EM_converged : fonction de vérification de la convergence de l'algorithme avec la croissance de la log-vraisemblance.

```r
EM_converged <- function(loglik, previous_loglik, threshold = 1e-4) {

  converged = 0;
  decrease  = 0;
  if (!(previous_loglik == -Inf)) {
    if (loglik - previous_loglik < -1e-3) # allow for a little imprecision
        {
        print(c("******likelihood decreased from ", previous_loglik,
                " to ", loglik), quote = FALSE)
        decrease = 1;
        }
```

```
    delta_loglik = abs(loglik - previous_loglik);
    avg_loglik   = (abs(loglik) + abs(previous_loglik) + threshold)/2;
    bb = ((delta_loglik/avg_loglik) < threshold)
    if (bb) {converged = 1}
    }

  res <- NULL
  res$converged <- converged
  res$decrease  <- decrease
  return(res)

}
```

NR_logit_w : calcul des coefficients de régression logistique avec un algorithme Newton-Raphson, lorsqu'aucune pénalisation Lasso n'est appliquée.

```
NR_logit_w <- function(y, X, beta = NULL, w = rep(1, length(y))/length(y),
                       maxiter = 100, epsilon = 1e-5){
    options(warn = -1)
    if (is.null(beta)){
        beta = coefficients(lm(y ~ X-1, weights = w))
    }
    beta0 = beta
    options(warn = 0)
    err  = 2*epsilon
    iter = 0
    yw   = y*w
    ll   = NULL
    while ((err > epsilon) & (iter < maxiter)){
        iter = iter + 1
        beta.old = beta
        e  = exp(X %*% beta)
        p  = e/(1+e)
        pr = prod(c(p * (1 - p) * w))
        if (pr < 1e-30 | is.na(pr)){
            beta = coefficients(lm(y ~ X-1, weights = w))
            warning("Completely separable pb...")
            break
        }
        V    = diag(c(p*(1-p)*w))
        mu   = p*w
        beta = beta + solve((t(X) %*% V %*% X), t(X) %*% (yw-mu))
        ll   = c(ll, t(yw) %*% X %*% beta - matrix(w, 1, length(y))
```

```
                %*% log(1 + exp(X %*% beta)))
        err  = mean(abs(beta-beta.old)/abs(beta.old))
    }
    return(list(beta = beta, ll = ll, beta0 = beta0))
}
```

NR_logit_w_lasso : fonction pour forcer une pénalisation pour les coefficients de régression si nécessaire, avec un algorithme Newton-Raphson.

```
NR_logit_w_lasso <- function(y, X, beta = NULL, w = rep(1, length(y))/length(y),
                             maxiter = 10, epsilon = 1e-5, lambda = .1){
    if (is.null(beta)){
        beta = coefficients(lm(y ~ X-1, weights = w))
        kp   = which(abs(beta) > lambda)
    } else
    {kp  = which(!(beta == 0))}
    err  = 2*epsilon
    iter = 0
    yw   = y*w
    ll   = NULL

    while ((err > epsilon) & (iter < maxiter)){
        iter     = iter + 1
        beta.old = beta
        e        = exp(X %*% beta)
        p        = e/(1+e)
        pr       = prod(c(p * (1 - p) * w))
        if (pr<1e-30 | is.na(pr)){
            beta   = coefficients(lm(y ~ X-1, weights = w))
            warning("Completely separable pb...")
            break
        }
        V        = diag(c(p*(1-p)*w))
        mu       = p*w
        score1   = t(X) %*% (yw - mu) + lambda*sign(beta)
        score2   = t(X) %*% V %*% X
        kp       = which(abs(beta) > lambda)
        beta[kp] = beta[kp] + solve(score2[kp, kp], score1[kp])
        ll       = c(ll, t(yw) %*% X %*% beta - matrix(w, 1, length(y))
                     %*% log(1 + exp(X %*% beta)))
        err      = mean(abs(beta[kp] - beta.old[kp])/abs(beta.old[kp]))
    }
    beta[-kp]  = 0 * beta[-kp]
```

```
    return(list(beta = beta, ll = ll))
}
```

A.1.3 Fonctions principales

Initialisation de l'algorithme avec des répétitions de partitionnement par K-means, et quelques itérations des étapes E et M. Pour l'étape d'initialisation, trois fonctions sont nécessaires : une fonction d'étape M adaptée à l'initialisation, et les fonctions d'étapes E et M utilisées plus tard dans l'algorithme.

Étape de Maximisation utilisée lors de l'initialisation :

```
MstepMMbinit = function(X, espt, rho = rho, THETA = THETA, maxiter = 10,
                        epsilon = 1e-5){
  n = dim(X)[1]
  p = dim(X)[2]
  M = dim(espt)[2]

  # Z : pi (proportion of each cluster)
  pik = sapply(1:M, function(k) {(1/n) * sum(espt[, k])})

  # X :
  # mu:
  mu = sapply(1:M, function(k) sapply(1:p, function(j)
               sum(espt[, k] * X[, j]) / sum(espt[, k])))
  # SIGMA
  SIGMA   = lapply(1:M, function(k)
             Reduce('+', lapply (1:n, function(i)  sapply(1:p, function(j)
             sapply(1:p, function(l) (X[i, j] - mu[j, k]) * (X[i, l]
             - mu[l, k]))) * espt[i, k])) / sum(espt[, k]))
  if (sum(abs(rho)) != 0){
    SIGMA = lapply(1:M, function(k)
            glasso(SIGMA[[k]], rho = rho[k], wi.init = THETA)$w)
    THETA = lapply(1:M, function(k)
            glasso(SIGMA[[k]], rho = rho[k], wi.init = THETA)$wi)
  } else {THETA = lapply(1:M, function(k) solve(SIGMA[[k]]))}

  # llik
  llik = sum(sapply(1:n, function(i) log(sum(sapply(1:M, function(k)
             pik[k]* dmvnorm(X[i, ], mean = mu[, k], sigma = SIGMA[[k]])))))))

  ddltheta   = sapply(1:M, function(m) sum(abs(THETA[[m]])>0))
  npar.THETA = sum(sapply(1:M, function(m) (ddltheta[m]+p)/2))
```

```
  npar = M*p + M - 1 + npar.THETA
  BIC  = (-2 * llik[length(llik)]) + (npar * log(n))

  return(list(pik = pik, mu = mu, SIGMA = SIGMA, THETA = THETA, llik = llik,
              BIC = BIC))
}
```

Étape de Maximisation : estimation des paramètres

```
MstepMMb = function(Y, X, beta = beta, espt, Intercept = FALSE, lasso = lasso,
       lambda = lambda, rho = rho, THETA = THETA, maxiter = 10, epsilon = 1e-5){
  n = dim(X)[1]
  p = dim(X)[2]
  M = dim(espt)[2]

  classetemp = apply(espt, 1, which.max)

  if (Intercept) {X1 = cbind(1, X)}
  else {X1 = X}

  # Z : pi (proportion of each cluster)
  pik    = sapply(1:M, function(k) {(1/n) * sum(espt[, k])})

  # X :
  # mu:
  mu     = sapply(1:M, function(k) sapply(1:p, function(j)
             sum(espt[, k] * X[ ,j]) / sum(espt[, k])))
  # SIGMA
  SIGMA  = lapply(1:M, function(k) Reduce('+', lapply (1:n, function(i)
            sapply(1:p, function(j) sapply(1:p, function(l) (X[i, j] - mu[j, k])*
            (X[i, l] - mu[l, k]))) * espt[i, k])) / sum(espt[, k]))
  if (sum(abs(rho)) != 0){
    SIGMA = lapply(1:M, function(k)
              glasso(SIGMA[[k]], rho = rho[k], wi.init = THETA)$w)
    THETA = lapply(1:M, function(k)
              glasso(SIGMA[[k]], rho = rho[k], wi.init = THETA)$wi)
  } else {THETA = lapply(1:M, function(k) solve(SIGMA[[k]]))}

  # Y :
  # beta
  if (lasso == FALSE) {
    if (Intercept) {X1 = cbind(1, X)}
    else {X1 = X}
```

```
    beta = sapply(1:M, function(k) NR_logit_w(Y, X1, beta = beta[, k],
                    w = espt[, k])$beta)
    for (k in 1:M){
      if (glm(Y[which(classetemp == k)]~.,
              data = as.data.frame(X[which(classetemp == k),]), family = binomial,
              method = "detect_separation")$separation == TRUE) {
        beta[, k] = NR_logit_w_lasso(Y, X1, beta = beta[, k], w = espt[, k],
                      lambda = .5)$beta
      } else { beta[, k] = NR_logit_w(Y, X1, beta = beta[, k],
                  w = espt[, k])$beta }
    }
  } else {
    if (Intercept) {
      beta = sapply(1:M, function(k) as.matrix(coef(glmnet(X, Y,
                family = "binomial", weights = espt[, k], standardize = TRUE,
                lambda = lambda[k]))))
    } else {
      beta = sapply(1:M, function(k) as.matrix(coef(glmnet(X, Y,
                family = "binomial", weights = espt[, k], standardize = TRUE,
                lambda = lambda[k], intercept = FALSE))))
      beta = as.matrix(beta[-1, ])
    }
  }

  # llik
  pbx  = sapply(1:M, function(k) sapply(1:n, function(i)
           exp(sum(X1[i, ] * beta[, k]))/ (1 + exp(sum(X1[i, ] * beta[, k])))))
  llik = sum(sapply(1:n, function(i) log(sum(sapply(1:M, function(k)
            pik[k]* dbinom(Y[i], size = 1, prob = pbx[i, k]) *
            dmvnorm(X[i, ], mean = mu[, k], sigma = SIGMA[[k]]))))))

  return(list(pik = pik, mu = mu, SIGMA = SIGMA, THETA = THETA, beta = beta,
                llik = llik, lambda = lambda, rho = rho))
}
```

Étape Espérance : calcul des probabilités a posteriori

```
EstepMMb = function(Y, X, pik, mu, SIGMA, beta, M, Intercept = FALSE){
  if (Intercept) {X1 = cbind(1, X)}
  else {X1 = X}

  n = dim(X1)[1]
  p = dim(X1)[2]
```

```
  if (M == 1) {
    espt = as.matrix(rep(1, nrow(X)))
  } else {
    pbx  = sapply(1:M, function(k) sapply(1:n, function(i)
            exp(sum(X1[i, ] * beta[, k]))/ (1 + exp(sum(X1[i, ] * beta[, k])))))
    espt = sapply(1:M, function(k) sapply(1:n, function(i)
             pik[k] * dbinom(Y[i], size = 1, prob = pbx[i, k]) *
             dmvnorm(X[i, ], mean = mu[, k], sigma = SIGMA[[k]]) /
             sum(sapply(1:M, function(z) pik[z] *
             dbinom(Y[i], size = 1, prob = pbx[i, z]) *
             dmvnorm(X[i, ], mean = mu[, z], sigma = SIGMA[[z]]) ))))
  }
  return(list(espt = espt))
}
```

Fonction d'initialisation : choix des paramètres d'initialisation parmi les répétitions de partitionnement par K-means suivies de quelques itérations EM.

```
init.stepMMb = function(Y, X, param.init, M, maxrep = 10, Intercept = FALSE,
                        lasso = TRUE, lambda = rep(0, M), rho = rep(0, M)) {
  if (Intercept) {X1 = cbind(1, X)}
  else {X1 = X}
  n = dim(X1)[1]
  p = dim(X1)[2]

  pii0   = mui0 = SIGMAi0 = THETAi0 = betai0 = espti0 = classes0 = list()
  lliki0 = c()

  if (param.init == 0) {
    for (it in 1:maxrep){
      # kmeans :
      resclas        = kmeans(X, centers = M)
      pii0[[it]]     = resclas$size/n
      mui0[[it]]     = t(resclas$centers)
      # For a cluster with only an individuals
      Ngr  = 0
      Kout = NULL
      Iout = NULL
      for (k in 1:M) {
        if (length(which(resclas$cluster == k)) == 1) {
          Ngr     = Ngr + 1
          Ioutdef = c(Ioutdef, which(resclas$cluster == k))
          Iout    = which(resclas$cluster == k)
```

```
    Xout    = X[Iout, ]
    X       = X[-Iout, ]
    Y       = Y[-Iout]
    n       = n-1
    Kout    = k
    classes = resclas$cluster[-Iout]

  }}
if (Ngr == 0) {
  classes = resclas$cluster
  Kout    = M+1
  Iout    = NULL
}


SIGMAi0[[it]]  = lapply(1:M, function(k) var(X[classes == k, ]))
THETAi0[[it]]  = list()
for (k in 1:M) {
  if (is.singular.matrix(SIGMAi0[[it]][[k]])) {
    THETAi0[[it]][[k]] = glasso(SIGMAi0[[it]][[k]], rho = 0.1)$wi
  } else { THETAi0[[it]][[k]]  = solve(SIGMAi0[[it]][[k]]) }
}


if (lasso == FALSE) {
  lambdai = rep(0, M) ; rhoi = rep(0, M)
  if (M > 1) {
    for (k in 1:M){ if (p*(M+1) > length(which(classes == k))) {
                    lambdai[k] = 0.01; rhoi[k] = 0.01}}
  }
  if (sum(abs(lambdai)) != rep(0, M) | sum(abs(rhoi)) != rep(0, M))
           {lasso = TRUE}
} else {lambdai = lambda; rhoi = rho;
  for (k in 1:M){ if (p*(M+1) > length(which(classes == k))) {
              lambdai[k] = lambda[k]*5; rhoi[k] = rho[k]*5}}
}


if (Intercept) {
  betai0[[it]] = sapply(1:M, function(k)
                  as.matrix(coef(glmnet(X[classes == k,], Y[classes == k],
                  family = "binomial", standardize = TRUE,
                  lambda = lambdai[k]))))
```

```
    } else {
      betai0[[it]] = sapply(1:M, function(k)
                     as.matrix(coef(glmnet(X[classes == k,], Y[classes == k],
                     family = "binomial", standardize = TRUE,
                     lambda = lambdai[k], intercept = FALSE))))
    }


    # Repetition of the EM algorithm

    resE  = EstepMMb(Y, X, pik = pii0[[it]], mu = as.matrix(mui0[[it]]),
                     SIGMA = SIGMAi0[[it]], beta = betai0[[it]], M = M,
                    Intercept = Intercept)
    resM  = MstepMMb(Y, X, beta = betai0[[it]], espt = resE$espt,
                    Intercept = Intercept, lasso = lasso, lambda = lambdai,
                    rho = rhoi, THETA = THETAi0[[it]])

    repet = 1
    while (repet <= maxrep) {
      resE = EstepMMb(Y, X, pik = resM$pik, mu = resM$mu, SIGMA = resM$SIGMA,
                      beta = resM$beta, M = M, Intercept = Intercept)
      resM  = MstepMMb(Y, X, beta = resM$beta, espt = resE$espt,
                      Intercept = Intercept, lasso = lasso, lambda = lambdai,
                      rho = rhoi, THETA = resM$THETA)
      repet = repet + 1
    }

    pii0[[it]]     = resM$pik
    mui0[[it]]     = resM$mu
    SIGMAi0[[it]]  = resM$SIGMA
    THETAi0[[it]]  = resM$THETA
    betai0[[it]]   = resM$beta
    espti0[[it]]   = resE$espt
    lliki0[it]     = resM$llik
    classes0[[it]] = sapply(1:n, function(i) which.max(resE$espt[i, ]))
  }

  choix    = which.max(lliki0)
  pii      = pii0[[choix]]
  mui      = mui0[[choix]]
  SIGMAi   = SIGMAi0[[choix]]
  THETAi   = THETAi0[[choix]]
  betai    = betai0[[choix]]
```

```
    espti    = espti0[[choix]]
    llik     = lliki0[choix]
    classifi = classes0[[choix]]

  } else {
    pii      = param.init[1]
    mui      = param.init[2]
    SIGMAi   = param.init[3]
    THETAi   = param.init[4]
    betai    = param.init[5]

    pbxi     = sapply(1:M, function(k) sapply(1:n, function(i)
          exp(sum(X1[i, ] * betai[, k]))/ (1 + exp(sum(X1[i, ] * betai[, k])))))
    espti    = sapply(1:M, function(k) sapply(1:n, function(i)
          pii[k] * dbinom(Y[i], size = 1, prob = pbxi[i, k]) *
          dmvnorm(X[i, ], mean = mui[, k], sigma = SIGMAi[[k]]) /
          sum(sapply(1:M, function(z) pii[z] * dbinom(Y[i], size = 1,
         prob = pbxi[i, z]) * dmvnorm(X[i, ], mean = mui[, z],
         sigma = SIGMAi[[z]]) ))))

    llik     = sum(sapply(1:n, function(i) log(sum(sapply(1:M, function(k)
                  pii[k]* dbinom(Y[i], size = 1, prob = pbxi[i, k]) *
                  dmvnorm(X[i, ], mean = mui[, k], sigma = SIGMAi[[k]])))))))

    classifi = sapply(1:n, function(i) which.max(espti[i, ]))
  }

  return(list(pik = pii, mu = mui, SIGMA = SIGMAi, THETA = THETAi, beta = betai,
              llik = llik, classifi = classifi, Iout = Iout, M = M,
              lambda = lambdai, rho = rhoi))
}
```

A.2 Fonction principale de l'algorithme

Fonction utilisée pour effectuer l'algorithme EM en combinant toutes les fonctions définies précédemment :

```
EM.MMbinom = function(Y, X, M, param.init = 0, maxit = 500, keep = TRUE,
                      eps = 10^-5, Intercept = FALSE, lasso = TRUE,
                      lambda = rep(0, M), rho = rep(0, M), param.init.res = NULL,
                      first.step = NULL) {
  # Y          : binary vector
```

```
# X          : numeric matrix
# M          : number of clusters
# param.init : initial parameters c(pik, mu, SIGMA, beta)
# maxit      : maximum number of EM iterations
# keep       : boolean, keep results of each iteration or not
# Intercept  : boolean, Consider intecept in the model or not
# lasso      : boolean, perform variable selection or not
# lamda, rho : tunning parameters for the regularization

n  = dim(X)[1]
p  = dim(X)[2]
Md = M

if (is.null(first.step)){
  if (is.null(param.init.res)){
    print("initializing")
    res.init = init.stepMMb(Y, X, param.init, M = M, maxrep = 1,
              Intercept = Intercept, lasso = TRUE, lambda = lambda, rho = rho)
    if (is.null(res.init$Iout)) {
      Xn = X ; Yn = Y ; Iout = res.init$Iout
    } else {
      Xn     = X[-c(res.init$Iout), ]
      Yn     = Y[-res.init$Iout]
      n      = nrow(Xn)
      lambda = res.init$lambda
      rho    = res.init$rho
      Iout   = res.init$Iout
    }

  } else {
    print(paste("Use previous computations to initialize"))
    res.init = param.init.res
  }
  pik      = res.init$pik
  mu       = res.init$mu
  SIGMA    = res.init$SIGMA
  THETA    = res.init$THETA
  beta     = res.init$beta
  llik     = res.init$llik
  M        = res.init$M
  print(paste("Iter", 1, ", llik = ",  res.init$llik))
  resM.old = res.init
  resE     = EstepMMb(Yn, Xn, pik = resM.old$pik, mu = resM.old$mu,
```

```
                        SIGMA = resM.old$SIGMA, beta = resM.old$beta, M = M,
                        Intercept = Intercept)
  resM     = MstepMMb(Yn, Xn, beta = beta, espt = resE$espt,
                        Intercept = Intercept, lasso = lasso, lambda = lambda,
                        rho = rho, THETA = THETA)
  print(paste("Iter", 2, ", llik = ",  resM$llik))
} else {
  resE = first.step$resE
  resM = first.step$resM
  llik = resM$llik
  Iout = NULL
}

resE.old   = resE
resM.old   = resM
if (keep  == TRUE) {
  pik        = rbind(pik, resM$pik)
  mu         = list(mu, resM$mu)
  SIGMA      = list(SIGMA, resM$SIGMA)
  THETA      = list(THETA, resM$THETA)
  beta       = list(beta, resM$beta)
  tau        = list(NULL, resE$espt)
  classif    = list(res.init$classifi, sapply(1:n, function(i)
                    which.max(resE$espt[i, ])))
  npar.beta  = sum(abs(beta[[2]])>0)
  ddltheta   = sapply(1:M, function(m) sum(abs(THETA[[2]][[m]])>0))
  npar.THETA = sum(sapply(1:M, function(m) (ddltheta[m] + p)/2))
} else {
  pik        = resM$pik
  mu         = resM$mu
  SIGMA      = resM$SIGMA
  THETA      = resM$THETA
  beta       = resM$beta
  classif    = sapply(1:n, function(i) which.max(resE$espt[i, ]))
  npar.beta  = sum(abs(beta)>0)
  ddltheta   = sapply(1:M, function(m) sum(abs(THETA[[m]]) > 0))
  npar.THETA = sum(sapply(1:M, function(m) (ddltheta[m]+p)/2))
  tau        = resE$espt
}

llik     = cbind(llik, resM$llik)
epsilon  = 1
it       = 2
```

```
print("EM algorithm")
while (EM_converged(loglik = resM$llik,
         previous_loglik = llik[length(llik)-1])$converged == 0 & it < maxit){
  it       = it + 1

  resE     = EstepMMb(Yn, Xn, pik = resM.old$pik, mu = resM.old$mu,
                      SIGMA = resM.old$SIGMA, beta = resM.old$beta, M = M,
                      Intercept = Intercept)
  resM     = MstepMMb(Yn, Xn, beta = resM.old$beta, espt = resE$espt,
                      Intercept = Intercept, lasso = lasso, lambda = lambda,
                      rho = rho, THETA = resM.old$THETA)
  print(paste("Iter", it, ", llik = ",  resM$llik))

  epsilon  = sum((unlist(resM)-unlist(resM.old))^2)/sum(unlist(resM.old)^2)

  resE.old = resE
  resM.old = resM

  lambda   = resM$lambda
  rho      = resM$rho

  if (keep == TRUE){
    pik           = rbind(pik, resM$pik)
    mu[[it]]      = resM$mu
    SIGMA[[it]]   = resM$SIGMA
    THETA[[it]]   = resM$THETA
    beta[[it]]    = resM$beta
    npar.beta     = sum(abs(beta[[it]])>0)
    ddltheta      = sapply(1:M, function(m) sum(abs(THETA[[it]][[m]])>0))
    npar.THETA    = sum(sapply(1:M, function(m) (ddltheta[m]+p)/2))
    classif[[it]] = sapply(1:n, function(i) which.max(resE$espt[i, ]))
    tau[[it]]     = resE$espt
  } else {
    pik         = resM$pik
    mu          = resM$mu
    SIGMA       = resM$SIGMA
    THETA       = resM$THETA
    ddltheta    = sapply(1:M, function(m) sum(abs(THETA[[m]])>0))
    npar.THETA  = sum(sapply(1:M, function(m) (ddltheta[m]+p)/2))
    beta        = resM$beta
    npar.beta   = sum(abs(beta)>0)
    classif     = sapply(1:n, function(i) which.max(resE$espt[i, ]))
```

```
    tau         = resE$espt
  }
  llik = c(llik, resM$llik)
}
npar = npar.beta + 2*M -1 + npar.THETA
BIC  = (-2 * llik[length(llik)]) + (npar * log(n))
AIC  = (2 * npar) - (2 * llik[length(llik)])

# class order
if (keep == TRUE) {
  if (M == 1) {ordre = lapply(1:it, function(nit) 1)}
  else {
    detSigma     = sapply(1:it, function(nit) sapply(1:M, function(k)
                          det(SIGMA[[nit]][[k]])))
    ordre        = lapply(1:it, function(nit) order(detSigma[nit, ]))
  }
  clustering = lapply(1:it, function(nit) sapply(1:length(classif[[nit]]),
                       function(i) which(ordre[[nit]] == classif[[nit]][i])))
  pik            = lapply(1:it, function(nit) pik[nit, ][ordre[[nit]]])
  mu             = lapply(1:it, function(nit) mu[[nit]][, ordre[[nit]]])
  beta           = lapply(1:it, function(nit) beta[[nit]][, ordre[[nit]]])
  SIGMA          = lapply(1:it, function(nit) lapply(1:M, function(i)
                              SIGMA[[nit]][[ordre[[nit]][i]]]))
  THETA          = lapply(1:it, function(nit) lapply(1:M, function(i)
                              THETA[[nit]][[ordre[[nit]][i]]]))
  tau            = lapply(2:it, function(nit) lapply(1:M, function(i)
                              tau[[nit]][[ordre[[nit]][i]]]))
} else {
  detSigma       = sapply(1:M, function(k) det(SIGMA[[k]]))
  ordre          = order(detSigma)
  clustering = sapply(1:length(classif), function(i) which(ordre==classif[i]))
  pik            = pik[ordre]
  mu             = as.matrix(mu[, ordre])
  beta           = as.matrix(beta[, ordre])
  SIGMA          = lapply(1:M, function(i) SIGMA[[ordre[i]]])
  tau            = sapply(1:M, function(i) tau[,ordre[[i]]])
  if (sum(abs(rho)) != 0) {THETA = lapply(1:M, function(i) THETA[[ordre[i]]])}
}

MAP  = sapply(1:n, function(i) which.max(tau[i, ]))
zikt = matrix(0, nrow = n, ncol = M) ; for (i in 1:n) {zikt[i, MAP[i]] = 1}
ICL  = BIC - 2 * sum(sapply(1:n, function(i) sum(sapply(1:M, function(k)
                     zikt[i, k] %*% log(tau[i, k])), na.rm = TRUE)))
```

```
  reg = list(M = M, lambda = lambda, rho = rho)

  return(list(pik = pik, mu = mu, SIGMA = SIGMA, THETA = THETA, beta = beta,
          llik = llik, BIC = BIC, AIC = AIC, ICL = ICL, clustering = clustering,
          reg = reg, tau = tau, Iout = Iout, Md = Md, Mnew = M))
}
```

A.3 Fonction pour la sélection des paramètres de régularisation

Cette fonction permet la sélection des paramètres de régularisation selon la procédure suivante. Une grille de valeurs possibles pour les paramètres λ et ρ est construite pour chaque groupe. Les modèles sont ajustés pour chaque point de la grille, avec un petit nombre d'itérations de l'algorithme EM. Le critère BIC est ensuite utilisé pour sélectionner les 5 meilleurs modèles, qui sont ré-estimés avec un plus grand nombre d'itérations de l'algorithme EM. Enfin, le meilleur modèle selon le critère BIC est choisi.

```
procedure_selec_model = function(Y, X, Mmin = 2, Mmax = 5, length.lambda = 5,
                                 length.rho = 20, Intercept = TRUE) {
  # Y             : binary vector
  # X             : numeric matrix
  # Mmin          : minimum number of clusters to test
  # Mmax          : maximum number of clusters to test
  # length.lambda : size of the lambda grid to consider
  # length.rho    : size of the rho grid to consider
  # Intercept     : boolean, Consider intecept in the model or not

  resEM = resEM.2 = list()
  it    = 1

  lambda.min = 0
  rho.min    = 0

  for (m in Mmin:Mmax) {
    # Initialization: GMM
    mi       = m
    res.init = init.stepMMb(Y, X, param.init=0, M = m, maxrep = 1,
                            Intercept = Intercept, lasso = FALSE)
    pik      = res.init$pik
    mu       = res.init$mu
    SIGMA    = res.init$SIGMA
    THETA    = res.init$THETA
```

```
beta     = res.init$beta
llik     = res.init$llik
m        = res.init$M
if (is.null(res.init$Iout)) {
  Xn = X ; Yn = Y
} else {
  Xn       = X[-c(res.init$Iout), ]
  Yn       = Y[-res.init$Iout]
}

print(paste("Iter", 1, ", llik = ",  res.init$llik))
resM.old = res.init

resE     = EstepMMb(Yn, Xn, pik = resM.old$pik, mu = resM.old$mu,
                    SIGMA = resM.old$SIGMA, beta = resM.old$beta, M = m,
                    Intercept = Intercept)
classifi = apply(resE$espt, 1, which.max)

# Looking for lambda and rho in a large grid, few iterations of EM
# algorithm
lambda.max = rep(0, m)
rho.max    = rep(0, m)

for (k in 1:m){
  beta.ridge    = lm.ridge(Yn[classifi == k] ~ Xn[classifi == k, ],
                           lambda = 0.01)
  lambda.max[k] = max(abs(beta.ridge$coef))/length(classifi == k) * 100
  S             = cov(Xn[classifi == k, ])
  rho.max[k]    = max(abs(S))/length(classifi == k) * 10
}

# regular grid needed if we don't take too much regularization parameters
grid.lambda = sapply(lambda.max, function(x)
                  seq(0, x, length.out = length.lambda))
# regular grid needed if we don't take too much regularization parameters
grid.rho    = sapply(rho.max, function(x) seq(0,x, length.out = length.rho))

grid.rho.full    = expand.grid(as.data.frame(grid.rho))
grid.lambda.full = expand.grid(as.data.frame(grid.lambda))

# Reduce the size of grid.rho.full
it.init = 1
resMstep.init = list()
```

```
  tot = dim(grid.rho.full)[1]

  for (r in 2:dim(grid.rho.full)[1]){
    print(paste('Rho ', it.init, 'over', tot))
    resMstep.init[[it.init]] = MstepMMbinit(Xn, espt = resE$espt,
                                    rho = grid.rho.full[r, ], THETA = THETA,
                                    maxiter = 5, epsilon = 1e-5)
    it.init = it.init + 1
  }
  BICrho         = sapply(1:length(resMstep.init), function(i)
                                resMstep.init[[i]]$BIC)
  ind.rho        = order(BICrho)[1:3]
  grid.rho.small = as.matrix(grid.rho.full[ind.rho, ])

  # For each regularization parameter, run the EM algorithm
  for (r in 1:dim(grid.rho.small)[1]){
    for (l in 1:dim(grid.lambda.full)[1]){
      print(paste('Model', it, 'over', 3*(length.lambda)^m))
      resM        = MstepMMb(Yn, Xn, beta = beta, espt = resE$espt,
                             Intercept = Intercept, lasso = TRUE,
                             lambda = grid.lambda.full[l, ],
                             rho = grid.rho.small[r, ], THETA = THETA)
      first.step  = list(resE = resE, resM = resM)
      resEM[[it]] = EM.MMbinom(Yn, Xn, M = m, param.init = 0, maxit = 5,
                            eps = 10^-5, keep = FALSE, Intercept = Intercept,
                            lasso = TRUE, lambda = grid.lambda.full[l, ],
                            rho = grid.rho.small[r,], first.step = first.step)
      assign("resEMcurrent", resEM, envir = .GlobalEnv)
      it = it+1
    }
  }
}

print('Selection of best models by BIC')
BIC      = sapply(1:length(resEM), function(i) resEM[[i]]$BIC)
order.bic = order(BIC)[1:5]

# Run again for the 5 best models the EM algorithm with large number of
# iterations
it2 = 1
for (ind in order.bic){
  print(paste('Model',it2))
  reg = resEM[[ind]]$reg
```

```
    resEM.2[[it2]] = EM.MMbinom(Yn, Xn, M = reg$M, param.init = 0, maxit = 50,
                                eps = 10^-5, keep = FALSE, Intercept = Intercept,
                                lasso = TRUE, lambda = reg$lambda, rho = reg$rho)
    it2 = it2 + 1
    assign("resEM2current", resEM.2, envir = .GlobalEnv)
  }

  BIC     = sapply(1:length(resEM.2), function(i) resEM.2[[i]]$BIC)
  bestRes = which.min(BIC)
  res     = resEM.2[[bestRes]]

  return(list(res = res, resEM.2))}
```

A.4 Prédiction

Après l'estimation et la sélection du modèle, il est possible de prédire la réponse binaire pour des nouvelles données, avec la fonction de prédiction :

```
prediction = function(Xnew, mod, logit = TRUE) {
  # Xnew  : numeric matrix, new data to predict
  # mod   : model estimated with procedure_selec_model
  # logit : boolean indicating whether or not the predicted response is binary

  M     = mod$Mnew
  pik   = mod$pik
  mu    = mod$mu
  SIGMA = mod$SIGMA
  beta  = mod$beta
  n     = dim(Xnew)[1]

  if (nrow(beta) == ncol(Xnew)){
    X1new = Xnew
  } else {X1new = cbind(1, Xnew) }

  # computation of the posterior probability
  PZnew = sapply(1:M, function(k) sapply(1:dim(Xnew)[1], function(i)
                  pik[k] * dmvnorm(Xnew[i, ], mean = mu[, k],
                  sigma = SIGMA[[k]]) / sum(sapply(1:M, function(z) pik[z] *
                  dmvnorm(Xnew[i, ], mean = mu[, z], sigma = SIGMA[[z]]) ))))

  knew  = sapply(1:n, function(i) which.max(PZnew[i, ]))
```

```
  # weighted sum of each model
  Yhnew = c()
  for (ind in 1:n) {
    if (logit == FALSE) {
      Yhnew[ind] = sum(sapply(1:M, function(k)
                        PZnew[ind, k] * X1new[ind, ] * beta[, k]))
    } else {
      Yhnew[ind] = sum(sapply(1:M, function(k) PZnew[ind, k] *
                        (exp(X1new[ind, ] %*% beta[, k])/
                          (1 + exp(X1new[ind, ] %*% beta[, k])))))
    }
  }
  $
  return(list(Yhnew = Yhnew, classif = knew))
}
```

Annexe B

Résultats supplémentaires de l'étude de simulation évaluant la procédure de tests sur les matrices de corrélation structurées

Cette annexe comporte des résultats supplémentaires de l'étude de simulation permmetant d'évaluer la procédure de tests par bloc sur les matrices de corrélation structurées, présentée au chapitre 5. Les tables B.1 et B.2 présentent les valeurs moyennes de sensibilités et spécificités obtenues pour certaines configurations de paramètres pour le cas de simulation 1. Pour les deux configurations de paramètres présentées, les performances en sensibilité et en spécificité sont élevées, et montrent une bonne capacité de la procédure à retrouver les blocs de dépendances significatives.

La figure B.1 présente la configuration 1 du premier cas de simulation à 100 variables et 10 blocs. On cherche dans ce cas à évaluer la capacité de la procédure à détecter un bloc de dépendance situé au bord de la matrice. La procédure de tests permet de bien détecter le bloc de dépendance. On note toutefois la difficulté de la procédure à gérer le bloc de très petite dimension. Cela entraîne d'une part la détection de la mauvaise structure en bloc à cet emplacement avec un bloc de trop grande dimension mis en avant, et d'autre part la détection à tort d'un bloc de dépendances significatives proche de la diagonale, à cet emplacement. Les résultats obtenus avec différentes valeurs possibles dans la matrice de covariance de simulation sont présentés en figure B.1. On note que ce paramètre a un effet sur la détection du bloc significatif à tort. En effet, les résultats de simulation avec une matrice de covariance à fortes valeurs (D2) montrent une probabilité de détection faible, tout comme les résultats de simulation avec une matrice de covariance à faibles valeurs aux blocs significatifs (D4), dans une moindre mesure. De la même façon, on note sur la figure B.2 que lorsque les valeurs de covariance sont élevées dans toute la matrice ou seulement aux blocs significaifs, les performances de détection des blocs de la procédure sont moins bonnes.

		Config. 1		Config. 2		Config. 3	
		Se	Sp	Se	Sp	Se	Sp
D1	n = 100	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.99 (1-1)
	n = 200	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	1 (1-1)
D2	n = 100	1 (1-1)	0.99 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)
	n = 200	1 (1-1)	0.99 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)
D3	n = 100	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)
	n = 200	1 (1-1)	0.99 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)
D4	n = 100	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.99 (1-1)
	n = 200	1 (1-1)	0.99 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.99 (1-1)

TABLE B.1 : **Sensibilités et spécificités moyennes obtenues à l'issue de la procédure de tests pour le cas de simulation 1, avec des données simulées à 20 variables réparties en 5 blocs.** La moyenne est calculée sur les 50 répétitions de chaque cas de simulation. Les valeurs entre parenthèses correspondent aux premier et troisième quartiles des valeurs.

		Config. 1		Config. 2		Config. 3	
		Se	Sp	Se	Sp	Se	Sp
D1	n = 500	1 (1-1)	0.99 (1-1)	1 (1-1)	0.98 (1-1)	1 (1-1)	0.98 (1-1)
	n = 1000	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.99 (1-1)
D2	n = 500	1 (1-1)	0.98 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.98 (1-1)
	n = 1000	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)	1 (1-1)
D3	n = 500	1 (1-1)	0.98 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	0.98 (1-1)
	n = 1000	1 (1-1)	1 (1-1)	1 (1-1)	0.99 (1-1)	1 (1-1)	1 (1-1)
D4	n = 500	1 (1-1)	0.98 (1-1)	1 (1-1)	0.98 (1-1)	1 (1-1)	0.98 (1-1)
	n = 1000	1 (1-1)	0.99 (1-1)	1 (1-1)	0.98 (1-1)	1 (1-1)	1 (1-1)

TABLE B.2 : **Sensibilités et spécificités moyennes obtenues à l'issue de la procédure de tests pour le cas de simulation 1, avec des données simulées à 100 variables réparties en 5 blocs.** La moyenne est calculée sur les 50 répétitions de chaque cas de simulation. Les valeurs entre parenthèses correspondent aux premier et troisième quartiles des valeurs.

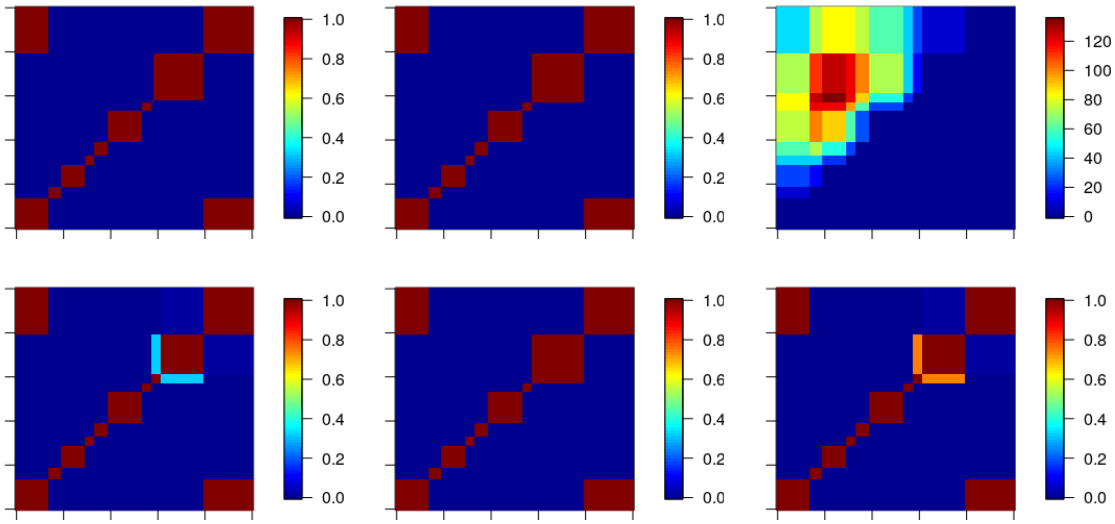

FIGURE B.1 : **Capacité de la procédure à retrouver la structure de dépendance des matrices de corrélation dans le premier cas de simulation pour la configuration 1.** La matrice en haut à gauche représente la structure de la matrice de covariance simulée, que l'on cherche à retrouver. La matrice en haut et au milieu représente la matrice de probabilité de rejet de l'hypothèse nulle, pour la matrice de simulation D1. Chaque coefficient de cette matrice correspond à la probabilité de rejet de l'hypothèse nulle obtenue par la procédure de tests par blocs sur les matrices de covariance calculée sur les 50 répétitions de chaque cas de simulation. La matrice en haut à droite représente le nombre de tests effectués pour chaque bloc, pour une des 50 répétitions du cas de simulation étudié. Sur la deuxième ligne, les matrices représentent les matrices de probabilités de rejet de l'hypothèse nulle, pour les matrices de simulation D2 (à gauche), D3 (au milieu) et D4 (à droite). Les matrices représentées correspondent au premier cas de simulation, pour 500 observations et 100 variables.

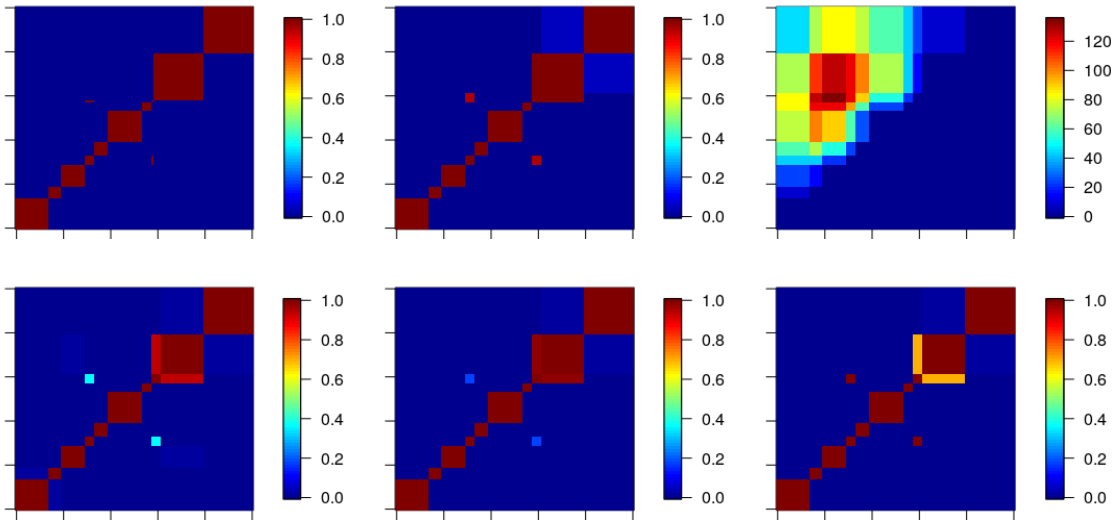

FIGURE B.2 : **Capacité de la procédure à retrouver la structure de dépendance des matrices de corrélation dans le premier cas de simulation pour la configuration 3.** La matrice en haut à gauche représente la structure de la matrice de covariance simulée, que l'on cherche à retrouver. La matrice en haut et au milieu représente la matrice de probabilité de rejet de l'hypothèse nulle, pour la matrice de simulation D1. Chaque coefficient de cette matrice correspond à la probabilité de rejet de l'hypothèse nulle obtenue par la procédure de tests par blocs sur les matrices de covariance calculée sur les 50 répétitions de chaque cas de simulation. La matrice en haut à droite représente le nombre de tests effectués pour chaque bloc, pour une des 50 répétitions du cas de simulation étudié. Sur la deuxième ligne, les matrices représentent les matrices de probabilités de rejet de l'hypothèse nulle, pour les matrices de simulation D2 (à gauche), D3 (au milieu) et D4 (à droite). Les matrices représentées correspondent au premier cas de simulation, pour 500 observations et 100 variables.

Bibliographie

Adams, L., J. Lymp, J. Sauver, S. Sanderson, K. Lindor, A. Feldstein, and P. Angulo (2005). The natural history of nonalcoholic fatty liver disease : A population-based cohort study. *Gastroenterology 129*(1), 113 – 121.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Proceedings, 2nd Internat. Symp. on Information Theory*, 267–281.

Akaike, H. (1978). On newer statistical approaches to parameter estimation and structure determination. *IFAC Proceedings Volumes 11*(1), 1877 – 1884. 7th Triennial World Congress of the IFAC on A Link Between Science and Applications of Automatic Control, Helsinki, Finland, 12-16 June.

Albert, J., V. Monbet, A. Jolivet-Gougeon, N. Fatih, M. Le Corvec, M. Seck, F. Charpentier, G. Coiffier, C. Boussard-Pledel, B. Bureau, P. Guggenbuhl, and O. Loréal (2016). A novel method for a fast diagnosis of septic arthritis using mid infrared and deported spectroscopy. *Joint Bone Spine 83*(3), 318 – 323.

Anty, R., A. Iannelli, S. Patouraux, S. Bonnafous, V. Lavallard, M. Senni-Buratti, I. Ben Amor, A. Staccini-Myx, M. Saint-Paul, F. Berthier, P. Huet, Y. Le Marchand-Brustel, J. Gugenheim, P. Gual, and A. Tran (2010). A new composite model including metabolic syndrome, alanine aminotransferase and cytokeratin-18 for the diagnosis of non-alcoholic steatohepatitis in morbidly obese patients. *Alimentary pharmacology & therapeutics 32 11-12*, 1315–22.

Anty, R., M. Morvan, M. Le Corvec, C. Canivet, S. Patouraux, J. Gugenheim, S. Bonnafous, B. Bailly-Maitre, O. Sire, H. Tariel, et al. (2019). The mid-infrared spectroscopy : A novel non-invasive diagnostic tool for nash diagnosis in severe obesity. *JHEP Reports 1*, 361–368.

Baker, M., S. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. Lukaszewski, G. Thiéfin, and G. Sockalingum (2016). Developing and understanding biofluid vibrational spectroscopy : a critical review. *Chem. Soc. Rev. 45*, 1803–1818.

Bedossa, P., C. Poitou, N. Veyrie, J. Bouillot, A. Basdevant, V. Paradis, J. Tordjman, and K. Clement (2012). Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients. *Hepatology 56*(5), 1751–1759.

Besse, P. and J. Ramsay (1986). Principal components analysis of sampled functions. *Psychometrika 51*(2), 285–311.

Biancolillo, A. and F. Marini (2018). Chemometric methods for spectroscopy-based pharmaceutical analysis. *Frontiers in chemistry 6*(576).

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell. 22*(7), 719–725.

Biernacki, C., G. Celeux, and G. Govaert (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis 41*(3), 561–575.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In O. Opitz, B. Lausen, and R. Klar (Eds.), *Information and Classification*, Berlin, Heidelberg, pp. 40–54. Springer Berlin Heidelberg.

Bozdogan, H. and S. Sclove (1984). Multi-sample cluster analysis using akaike's information criterion. *Annals of the Institute of Statistical Mathematics 36*, 163–180.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Burrus, C. (2012). Iterative reweighted least squares. *OpenStax-CNC document* (module :m45285).

Buzzetti, E., M. Pinzani, and E. Tsochatzis (2016). The multiple-hit pathogenesis of non-alcoholic fatty liver disease (nafld). *Metabolism 65*(8), 1038 – 1048.

Cai, T. and W. Liu (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association 111*(513), 229–240.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*(2), 195–212.

Collazos, J., R. Dias, and A. Zambom (2016). Consistent variable selection for functional regression models. *Journal of Multivariate Analysis 146*, 63–71.

Cowe, I. and J. McNicol (1985). The use of principal components in the analysis of near-infrared spectra. *Applied Spectroscopy 39*(2), 257–266.

Dayton, C. and G. Macready (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association 83*(401), 173–178.

de Boor, C. (2001). *A Practical Guide to Spline*. New York : Springer.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Diebolt, J. and C. Robert (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society : Series B (Methodological) 56*, 363–375.

EASL–EASD–EASO (2016). Clinical practice guidelines for the management of non-alcoholic fatty liver disease. *Journal of Hepatology 64*(6), 1388 – 1402.

Everitt, B. (1984). *An Introduction to LatentVariable Models*. Chapman and Hall, New York.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fauvel, M., C. Dechesne, A. Zullo, and F. Ferraty (2015). Fast forward feature selection of hyperspectral images for classification with gaussian mixture models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8*(6), 2824–2831.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics.

Fraiman, R., Y. Gimenez, and M. Svarc (2016). Feature selection for functional data. *Journal of Multivariate Analysis 146*, 191 – 208. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Frank, l. and J. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Freedman, D. and D. Lane (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics 1*(4), 292–298.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics 9*(3), 432–441.

Genuer, R., J. Poggi, and C. Tuleau-Malot (2010). Variable selection using random forests. *Pattern Recognition Letters 31*(14), 2225 – 2236.

Gertheiss, J., A. Maity, and A. Staicu (2013). Variable selection in generalized functional linear models. *Stat 2*(1), 86–101.

Grün, B. and F. Leisch (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis 51*(11), 5247–5252.

Guyon, I. and A. Elisseeff (2006). *An Introduction to Feature Extraction*, pp. 1–25. Berlin, Heidelberg : Springer Berlin Heidelberg.

Hands, J., K. Dorling, P. Abel, K. Ashton, A. Brodbelt, C. Davis, T. Dawson, M. Jenkinson, R. Lea, C. Walker, and M. Baker (2014). Attenuated total reflection fourier transform infrared (atr-ftir) spectral discrimination of brain tumour severity from serum samples. *Journal of Biophotonics 7*(3-4), 189–199.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics, New York.

Henao-Mejia, J., E. Elinav, C. Jin, L. Hao, W. Mehal, T. Strowig, C. Thaiss, A. Kau, S. Eisenbarth, M. Jurczak, J. Camporez, G. Shulman, J. Gordon, H. Hoffman, and R. Flavell (2012). Inflammasome-mediated dysbiosis regulates progression of nafld and obesity. *Nature 482*, 179–185.

Hoshikawa, T. (2013). Mixture regression for observational data, with application to functional regression models. `https://arxiv.org/pdf/1307.0170.pdf`. [Online].

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Jacobs, R., M. Jordan, S. Nowlan, and G. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*(1), 79–87.

Jacques, J. and C. Preda (2014). Functional data clustering : a survey. *Advances in Data Analysis and Classification 8*(3), 231–255.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 64*(3), 411–432.

James, G., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *The Annals of Statistics 37*(5A), 2083–2108.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Jiang, Y., Y. Conglian, and J. Qinghua (2018). Model selection for the localized mixture of experts models. *Journal of Applied Statistics 45*(11), 1994–2006.

Jolliffe, I. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society : Series C (Applied Statistics) 31*(3), 300–303.

Jordan, M. (2004). Graphical models. *Statistical Science 19*(1), 140–155.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya : The Indian Journal of Statistics, Series A 62*(1), 49–66.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1038.

Khalili, A. and S. Lin (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics 69*(2), 436–446.

Kleiner, D., E. Brunt, M. Van Natta, C. Behling, M. Contos, O. Cummings, L. Ferrell, Y. Liu, M. Torbenson, A. Unalp-Arida, M. Yeh, A. McCullough, A. Sanyal, and N. S. C. R. Network (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology 41*(6), 1313–1321.

Koehler, A. and E. Murphree (1988). A comparison of the akaike and schwarz criteria for selecting model order. *Journal of the Royal Statistical Society : Series C (Applied Statistics) 37*(2), 187–195.

Kong, D., A. Staicu, and A. Maity (2016). Classical testing in functional linear models. *Journal of nonparametric statistics 28*(4), 813–838.

Lasch, P. (2012). Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems 117*, 100 – 114.

Le Corvec, M., C. Jezequel, N. Monbet, V. Fatih, F. Charpentier, H. Tariel, C. Boussard-Plédel, B. Bureau, O. Loréal, O. Sire, and E. Bardou-Jacquet (2012). Mid-infrared spectroscopy of serum, a promising non-invasive method to assess prognosis in patients with ascites and cirrhosis. *PLoS One 12(10)*.

Lehmann, W. (1963). An introduction to infrared spectroscopy (brugel, w.). *Journal of Chemical Education 40*(6), 336.

Lichman, M. (2013). UCI machine learning repository.

Lloyd-Jones, L., H. Nguyen, and G. J. McLachlan (2018). A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics & Data Analysis 119*, 19–38.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Matsui, H. (2014). Variable and boundary selection for functional data via multiclass logistic regression modeling. *Computational Statistics & Data Analysis 78*, 176–185.

McCullough, A. (2006). Pathophysiology of nonalcoholic steatohepatitis. *Journal of Clinical Gastroenterology 40*, S17–S19.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 72*(4), 417–473.

Menze, B., W. Petrich, and F. Hamprecht (2007). Multivariate feature selection and hierarchical classification for infrared spectroscopy : serum-based detection of bovine spongiform encephalopathy. *Analytical and Bioanalytical Chemistry 387*(5), 1801–1807.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics 8*, 343–366.

Ollesch, J., M. Heinze, M. Heise, T. Behrens, T. Brüning, and K. Gerwert (2014). It's in your blood : spectral biomarker candidates for urinary bladder cancer from automated ftir spectroscopy. *Journal of Biophotonics 7*(3-4), 210–221.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philisophical Transactions of the Royal Society of London. A, 185*, 71–110.

Pesarin, F. and L. Salmaso (2010). *Permutation Tests for Complex Data*. John Wiley & Sons.

Pessayre, D., A. Mansouri, and B. Fromenty (2002). V. mitochondrial dysfunction in steatohepatitis. *American Journal of Physiology-Gastrointestinal and Liver Physiology 282*(2), G193–G199.

Picheny, V., R. Servien, and N. Villa-Vialaneix (2019). Interpretable sparse sir for functional data. *Statistics and Computing 29*(2), 255–267.

Pini, A. and S. Vantini (2017). Interval-wise testing for functional data. *Journal of Nonparametric Statistics 29*(2), 407–424.

Pisto, P., M. Santaniemi, R. Bloigu, O. Ukkola, and A. Kesäniemi (2014). Fatty liver predicts the risk for cardiovascular events in middle-aged population : a population-based cohort study. *BMJ Open 4*(3).

Pomann, G., A. Staicu, and S. Ghosh (2013). Two sample hypothesis testing for functional data. Technical report, North Carolina State University. Dept. of Statistics.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis*. Springer Series in Statistics.

Richardson, S. and P. Green (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 59*, 731 – 792.

Ross, J. and J. Dy (2013). Nonparametric mixture of gaussian processes with constraints. *Proc. 30th Int. Conf. Mach. Learn. 28*, 1346–1354.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry 36*(8), 1627–1639.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Seko, Y., K. Yamaguchi, and Y. Itoh (2018). The genetic backgrounds in nonalcoholic fatty liver disease. *Clinical Journal of Gastroenterology 11*(2), 97–102.

Shmueli, G. (2010). To explain or to predict ? *Statistical Science 25*(3), 289–310.

Städler, N., P. Bühlmann, and S. van de Geer (2010). ℓ1-penalization for mixture regression models. *TEST 19*(2), 209–256.

Thumanu, K., S. Sangrajrang, T. Khuhaprema, A. Kalalak, W. Tanthanuch, S. Pongpiachan, and P. Heraud (2014). Diagnosis of liver cancer from blood sera using ftir microspectroscopy : a preliminary study. *Journal of Biophotonics 7*(3-4), 222–231.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 67*(1), 91–108.

Van De Geer, S. (2010). ℓ1-regularization in high-dimensional statistical models. In *International Congress of Mathematicians 2010 (ICM 2010)*, pp. 2351–2369.

Vernon, G., A. Baranova, and Z. M. Younossi (2011). Systematic review : the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. *Alimentary Pharmacology & Therapeutics 34*(3), 274–285.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

White, D., F. Kanwal, and H. ES. (2012). Association between nonalcoholic fatty liver disease and risk for hepatocellular cancer, based on systematic review. *Clinical Gastroenterology and Hepatology 10*(12), 1342 – 1359.e2.

Xia, Y., T. Cai, and T. Cai (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *Journal of the American Statistical Association 113*(521), 328–339.

Yang, X. and K. Nie (2008). Hypothesis testing in functional linear regression models with neyman's truncation and wavelet thresholding for longitudinal data. *Statistics in medicine 27*(6), 845–863.

Yao, F., Y. Fu, and T. Lee (2010). Functional mixture regression. *Biostatistics 12*(2), 341–353.

Younossi, Z., Q. Anstee, M. Marietti, T. Hardy, L. Henry, M. Eslam, J. George, and E. Bugianesi (2018). Global burden of nafld and nash : trends, predictions, risk factors and prevention. *Nature Reviews Gastroenterology & Hepatology 15*, 11–20.

Younossi, Z., A. Koenig, D. Abdelatif, Y. Fazel, L. Henry, and M. Wymer (2016). Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology 64*(1), 73–84.

Younossi, Z., R. Loomba, Q. Anstee, M. Rinella, E. Bugianesi, G. Marchesini, B. Neuschwander-Tetri, L. Serfaty, F. Negro, S. Caldwell, V. Ratziu, K. Corey, S. Friedman, M. Abdelmalek, S. Harrison, A. Sanyal, J. Lavine, P. Mathurin, M. Charlton, Z. Goodman, N. Chalasani, K. Kowdley, J. George, and K. Lindor (2018). Diagnostic modalities for nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, and associated fibrosis. *Hepatology 68*(1), 349–360.

Yuksel, S., J. Wilson, and P. Gader (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems 23*(8), 1177–1193.

Zhang, X., G. Thiéfin, C. Gobinet, V. Untereiner, I. Taleb, B. Bernard-Chabert, A. Heurgué, C. Truntzer, P. Ducoroy, P. Hillon, and G. Sockalingum (2013). Profiling serologic biomarkers in cirrhotic patients via high-throughput fourier transform infrared spectroscopy : toward a new diagnostic tool of hepatocellular carcinoma. *Translational Research 162*(5), 279 – 286.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Modèles de régression pour données fonctionnelles hétérogènes. Application à la modélisation de données de spectrométrie dans le moyen infrarouge.

Mot clés : mélange de régression, prédiction, pénalisation, matrice de covariance parcimonieuse, données fonctionnelles

Résumé : Dans de nombreux domaines d'application, les données récoltées correspondent à des courbes. Ce travail se concentre sur l'analyse de courbes de spectrométrie, constituées de plusieurs centaines de variables ordonnées, correspondant chacune à une valeur d'absorbance associée aux nombres d'ondes mesurés. Dans ce contexte, une méthode de traitement statistique automatique est développée, avec pour objectif la construction d'un modèle de prédiction prenant en compte l'hétérogénéité des données observées. Plus particulièrement, un modèle de diagnostic d'une maladie métabolique est établi à partir de courbes mesurées sur des individus provenant d'une population constituée de profils de patients différents. La procédure développée permet de sélectionner l'information pertinente sous forme de portions de courbes discriminantes, puis de construire de façon simultanée une partition des données et un modèle de prédiction parcimonieux grâce à un mélange de régressions pénalisées adapté aux données fonctionnelles. Ces données étant complexes, tout comme le cas d'application étudié, une méthode permettant une meilleure compréhension et une meilleure visualisation des interactions entre les portions de courbes a par ailleurs été développée. Cette méthode se base sur l'étude de la structure des matrices de covariance, avec pour but de faire ressortir des blocs de dépendances entre intervalles de variables. Un cas d'application médicale est utilisé pour présenter la méthode et les résultats, et permet l'utilisation d'outils de visualisation spécifiques.

Title: Regression models for heterogeneous functional data. Application to the modelization of mid-infrared spectrometric data.

Keywords: mixture of regressions, prediction, penalization, sparse covariance matrix, functional data

Abstract: In many application fields, data corresponds to curves. This work focuses on the analysis of spectrometric curves, composed of hundreds of ordered variables that corresponds to the absorbance values measured for each wavenumber. In this context, an automatic statistical procedure is developped, that aims at building a prediction model taking into account the heterogeneity of the observed data. More precisely, a diagnosis tool is built in order to predict a metabolic disease from spectrometric curves measured on a population composed of patients with differents profile. The procedure allows to select portions of curves relevant for the prediction and to build a partition of the data and a sparse predictive model simultaneously, using a mixture of penalized regressions suitable for functional data. In order to study the complexity of the data and of the application case, a method to better understand and display the interactions between variables is built. This method is based on the study of the covariance matrix structure, and aims to highlight the dependencies between blocks of variables. A medical exemple is used to present the method and results, and allows the use of specific visualization tools.