UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 7 août 2006

Présentée par

Joyce-Madison Giacofci

Thèse dirigée par **Sophie Lambert-Lacroix**
et codirigée par **Franck Picard**

préparée au sein du **Laboratoire Jean Kuntzmann**
et de **Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Classification non supervisée et sélection de variables dans les modèles mixtes fonctionnels. Applications à la biologie moléculaire

Thèse soutenue publiquement le **22 Octobre 2013**,
devant le jury composé de :

M. Anatoli Juditsky
Professeur, Université Grenoble 1, Président
Mme Béatrice Laurent
Professeur, INSA Toulouse, Rapporteur
M. Hervé Cardot
Professeur, Université de Bourgogne, Rapporteur
M. Vincent Rivoirard
Professeur, Université Paris-Dauphine, Examinateur
Mme, Sophie Lambert-Lacroix
Professeur, Université Grenoble 2, Directrice de thèse
M. Franck Picard
Chargé de Recherche, CNRS, Co-Directeur de thèse

Remerciements

Au même titre que la soutenance, le temps des remerciements est un moment souvent fantasmé au cours du long et parfois laborieux parcours d'un thésard comme le symbole d'un point final à tous nos efforts. On se rend bien vite compte que ce n'est en réalité qu'un passage parmi d'autres mais il a l'avantage de me permettre de rendre hommage à toutes les personnes qui ont pu contribuer à la réussite de cette épopée.

Je souhaite remercier en premier lieu mes deux rapporteurs Béatrice Laurent-Bonneau et Hervé Cardot de m'avoir fait l'honneur d'accepter de rapporter ce travail de thèse. Vos relectures attentives et vos critiques constructives m'ont permis d'apporter un autre regard à ce travail et d'en améliorer la finalisation. Je remercie aussi chaleureusement Vincent Rivoirard et Anatoli Juditsky d'avoir accepté de faire partie de mon jury. Enfin, des remerciements tous particuliers vont à mes deux directeurs de thèse, Sophie Lambert-Lacroix et Franck Picard. Vos apports tant sur le plan professionnel qu'humain m'ont permis de devenir la jeune chercheuse que je suis aujourd'hui. Je ne suis pas sûre de pouvoir résumer ma reconnaissance en quelques lignes mais je souhaite sincèrement à tout doctorant de pouvoir avoir la même chance que celle que j'ai eu en entamant cette aventure à vos côtés.

J'adresse aussi un merci particulier à Anestis Antoniadis, pour le professeur qu'il a été et surtout pour le modèle professionnel et humain qu'il reste à mes yeux.

La vie d'un doctorant ne se résume heureusement pas à la lecture d'articles ou au développement de codes. J'ai eu la chance de pouvoir enseigner durant ces quatre années au sein de l'IUT STID et pour cela, je tiens à remercier chaleureusement l'ensemble de l'équipe enseignante de l'IUT STID, qui m'a formé au métier d'enseignant chercheur et auprès de laquelle j'ai toujours pu trouver une oreille attentive lorsque j'en ai eu besoin.

Merci à Manu, Fred et Franck pour les doubles disputés, je vais continuer à travailler mon service et mon jeu au filet, peut-être qu'on remettra ça un de ces jours ! Je pense aussi à tous les doctorants et post-doctorants du labo, toutes ces personnes qui passent et repartent et auprès de qui j'ai toujours beaucoup appris, beaucoup ri, et grandement amélioré mon niveau de coinche, de potins ou de mots fléchés ! En vrac, je citerai Roland, Chloé, Bertie Love, Thomas, Matthias, P-O, Pierre-Jean, Lukas, Vincent, Euriell, Alexandre, Samuel, Christophe, Brice, Meryam, Emilie, Mélanie, Rémi, Claire, Ibrahim, Azmi, Souleymane, David... J'en oublie sûrement beaucoup mais je vous adresse à tous un grand merci pour toutes ces joyeuses années.

Plutôt que de simples remerciements, les suivants peuvent aussi être lus comme de sincères excuses. La rédaction d'une thèse peut être un exercice totalitaire et j'ai bien peur d'en avoir fait subir bien des conséquences à mes proches. Merci à mes poulettes préférées, Pout et Moul, parce que des copines comme vous, on en rencontre

pas souvent! Merci aux potes toujours partants pour une chouille Damien, Mike, Gogo, Manu, Dada, Kaka, Rouf, Chap, Tibs, Betty et tous les autres.

Un merci spécial à mes parents pour leur soutien indéfectible, j'espère être digne de la confiance que vous m'avez toujours accordée.

Enfin, merci à toi, Philippe, qui a été le garant de mon bonheur pendant toutes ces années, tu as été ma famille, mon pilier, mon repère. Je te souhaite à présent de pouvoir en trouver un qui soit moins instable que celui que j'ai pu être.

Table des matières

Chapitre 1

Introduction générale

1.1 Contexte applicatif

Les dernières décennies ont vu le développement rapide du domaine de la biologie moléculaire et, grâce à des progrès techniques constants, l'émergence d'une biologie dite à "haut-débit" se traduisant par une forte augmentation de la quantité de données disponibles. L'analyse de ce type de données offre au statisticien de nombreux challenges : en effet, une de leurs principales caractéristiques est que le nombre d'individus observés (de l'ordre de la centaine) est relativement faible devant le nombre de variables considérées (souvent en dizaines de milliers). Les problématiques d'intérêt restent généralement les mêmes, à savoir, la découverte de groupes, la discrimination, l'estimation et la prédiction pour n'en citer que quelques unes. Cependant, les méthodes classiquement utilisées à ces fins nécessitent d'être adaptées à la grande dimension de ces données car cette caractéristique les rend peu performantes.

Deux types de données issues du domaine de la biologie moléculaire à haut-débit ont motivé ce travail de thèse : les données de microarray CGH (Comparative Genomic Hybridization) et les données de spectrométrie de masse. Les données de microarray CGH sont des données visant à l'étude du génome et plus particulièrement, à la mesure du ratio du nombre de copies des gènes entre un échantillon d'intérêt et un échantillon de référence. La technologie de la spectrométrie de masse vise quant à elle à l'étude du protéome et sert à déterminer, par un procédé d'ionisation, la composition en protéines ou polypeptides d'un échantillon biologique. Une vaste littérature a été développée concernant l'étude de telles données et pour un panorama des approches adoptées, nous renvoyons le lecteur à la revue de van de Wiel et al. (2011) pour les données de microarray CGH et à la revue de Roy et al. (2011) pour les données de spectrométrie de masse.

Les caractéristiques communes de ces technologies sont de produire des données de grande dimension mesurées à haut-débit et présentant des comportements fortement discontinus. À ce titre, notre choix de modélisation s'est porté sur une

approche fonctionnelle de l'étude de telles données. En effet, de par leurs caractéristiques, ces données s'inscrivent naturellement dans le paradigme développé par Ramsay et Silverman (1997). Dans leur cadre, des données sont dites *fonctionnelles* si elles sont mesurées sur une grille de discrétisation fine et régulière et pour lesquelles la notion de courbes représente l'unité idéale d'observation, c'est-à-dire pour lesquelles on souhaite s'intéresser à des quantités n'ayant un sens que dans une approche fonctionnelle, comme la régularité par exemple. De plus, alors que jusqu'à présent, les efforts de recherche ont été concentrés sur la caractérisation de la variabilité directement imputable aux appareils de mesure, une nouvelle voie émerge, visant à étudier la variabilité biologique propre aux individus inhérente à de telles données. En effet, les réactions physiologiques dans une circonstance donnée (face à une maladie par exemple), peuvent se révéler très différentes selon les individus concernés et la modélisation de cette variabilité dans l'optique d'une meilleure compréhension du phénomène étudié est actuellement un enjeu majeur. Dans le cadre de l'étude de données non complexes, les modèles mixtes représentent l'outil d'étude *ad-hoc* pour la modélisation de la variabilité individuelle. Dans le contexte des données complexes, l'extension des modèles mixtes à un cadre fonctionnel en devient alors l'outil d'analyse privilégié.

Le document présent est constitué de trois parties. Dans la première partie, nous nous attachons à la description des principaux outils de modélisation statistique représentant les fondements de ce travail : les modèles mixtes et l'approche fonctionnelle basée sur les ondelettes. Dans une deuxième partie et aussi première contribution de ce travail, nous étudions la problématique de la classification non supervisée au sein des modèles mixtes fonctionnels. Enfin, dans une troisième partie, représentant la deuxième contribution de ce travail, nous nous concentrons sur des problématiques d'estimation dans les modèles mixtes fonctionnels au sein d'un groupe homogène d'individus.

1.2 Outils de modélisation

Résumé de la Partie I

Dans cette partie, notre volonté est d'introduire les principaux concepts associés au développement des contributions proposées dans cette thèse. Nous commençons par décrire dans un premier chapitre, la notion de modèles linéaires mixtes (Laird et Ware 1982). Cette introduction est réalisée dans le cas particulier de la modélisation de données *longitudinales*. Des données sont dites longitudinales lorsque les mesures réalisées sur les différents individus le sont selon une grille de temps ou d'espace, induisant un ordre naturel sur les données, et de ce fait, constituant une base naturelle à une future extension aux données fonctionnelles. Pour une introduction détaillée des modèles linéaires mixtes appliqués aux données longitudinales, le lecteur pourra se référer à l'ouvrage de Verbeke et Molenberghs (2000). L'idée générale des modèles

mixtes est basée sur la notion d'effets fixes et aléatoires permettant de distinguer le comportement moyen commun à une population d'une variabilité propre aux individus. Cette distinction conduit à deux approches distinctes, présentées dans le Chapitre 2. D'une part, l'approche marginale consiste, vis-à-vis des effets fixes, à une formulation du modèle en un modèle linéaire hétéroscédastique, et d'autre part, l'approche jointe, permet la prise en compte explicite des effets aléatoires.

Dans une optique d'extension de la modélisation mixte à un cadre fonctionnel, nous présentons, dans le Chapitre 3, la notion de modélisation fonctionnelle non paramétrique basée sur une représentation du modèle dans une base de fonctions. Notre choix de base se porte tout au long de ce travail sur les ondelettes : en effet, dans le contexte considéré, les ondelettes présentent de nombreux avantages rendant leur utilisation pleinement justifiée. Elles permettent d'une part la modélisation de signaux présentant de fortes discontinuités, qui est une des caractéristiques des données étudiées. De plus, leur propriété clé de bonne localisation aussi bien en temps qu'en fréquence, rendent la représentation des signaux réguliers dans le domaine des ondelettes naturellement parcimonieuse. Les méthodes de régression par ondelettes tirant avantage de leurs propriétés de parcimonie ont été introduites, en statistiques, de manière pionnière par Donoho et Johnstone (1994) et ont, par la suite, été largement développées. Une revue comparative intéressante de ces nombreuses méthodes peut être trouvée dans Antoniadis et al. (2001), nous en donnons, pour notre part, un aperçu au cours du Chapitre 3.

Dans un dernier chapitre, nous présentons la classe des modèles à variables *latentes*. Des variables sont dites latentes lorsqu'elles ne sont pas observées ; on parle aussi de variables non observées ou de données cachées. Cette classe est particulièrement intéressante dans notre contexte car elle constitue un cadre commun à la classification non supervisée basée sur une approche probabiliste où les variables latentes sont les labels inconnus des individus et aux modèles mixtes où les variables latentes sont les effets aléatoires individuels. Dans ce cadre, nous présentons deux modèles à variables latentes particuliers : le modèle de classification de courbes et le modèle mixte fonctionnel. L'accent est mis sur ces modèles particuliers car notre première contribution est basée sur une modélisation des données alliant ces deux objectifs. L'objectif de ce chapitre est alors d'en présenter les principales particularités qui seront utiles au modèle développé en Partie II. Le premier modèle de classification (non supervisée et sans effet aléatoire) de courbes décrit a été proposé par Antoniadis et al. (2008) dans le cadre de la classification d'image fonctionnelle. Leur modélisation est basée sur une représentation du modèle fonctionnel dans une base d'ondelettes, se ramenant ainsi à un modèle de mélange gaussien standard mais, dans un cadre fonctionnel, considéré en grande dimension. La problème de la dimension est géré par une étape préalable de réduction de dimension tirant parti des propriétés de parcimonie des ondelettes, dont nous nous inspirons au cours de notre procédure de classification de courbes dans les modèles mixtes fonctionnels. Remarquons qu'une vaste littérature existe pour la problématique de la classifica-

tion non supervisée de courbes et le lecteur intéressé pourra se référer aux revues de Bouveyron et Brunet (2013) et Jacques et Preda (2013).

En deuxième lieu, nous présentons un modèle mixte fonctionnel proposé par Antoniadis et Sapatinas (2007). Dans un cadre fonctionnel, les effets fixes et aléatoires deviennent des effets fonctionnels caractérisant respectivement un comportement fonctionnel moyen et des comportements fonctionnels propres aux individus modélisant la présence d'une variabilité inter-individuelle. Dans le modèle proposé, l'accent est mis par les auteurs sur la modélisation de la variabilité individuelle : en effet, Antoniadis et Sapatinas (2007) proposent une modélisation faisant l'hypothèse que les effets fixes et aléatoires fonctionnels sont situés dans le même espace fonctionnel. Cela peut se justifier par le fait qu'une telle modélisation permet alors une interprétation plus aisée des déviations individuelles vis-à-vis du comportement. Ainsi, dans le cas de l'étude de données de spectrométrie de masse par exemple, les déviations individuelles seront représentées par des signaux de même nature que le signal moyen, traduisant la présence/absence de certaines protéines suivant l'individu considéré. Dans un cadre de modélisation basée sur les ondelettes, ceci est assuré par une propriété, démontrée par Abramovich et al. (1998), de décroissance exponentielle des variances des coefficients associés aux effets aléatoires en fonction du niveau de résolution, permettant alors de contrôler la régularité du processus sous-jacent. L'originalité de leur approche se situe dans le fait que leur modélisation est développée directement sur les coefficients d'ondelettes. Nous réalisons dans ce contexte une étude de simulation exploratoire afin d'évaluer la diversité des processus engendrés par ce type de modélisation. Le calcul de la fonction de covariance associée au processus dans le cas de la base de Haar nous montre que la classe des processus atteinte est relativement large et permet la modélisation de processus même non stationnaires.

1.3 Classification non supervisée dans les modèles mixtes fonctionnels

Résumé de la Partie II

Cette partie représente la première contribution de ce travail et se concentre sur la problématique de la classification non supervisée dans les modèles mixtes fonctionnels.

La classification au sens large revêt une importance particulière dans le cadre de l'étude de données issues de la biologie moléculaire. Dans un cadre supervisé, c'est-à-dire, lorsque les groupes sont connus et que le problème est de trouver une règle de décision conduisant à ces groupes, cela représente un espoir pour la détection de marqueurs biologiques conduisant au développement de certaines maladies. Dans un contexte non supervisé, c'est-à-dire quand l'objectif est de former des groupes

homogènes, cela ouvre la voie à de potentielles meilleures performances en terme de diagnostic de certaines pathologies. En effet, la question du diagnostic est actuellement basée majoritairement sur l'observation de données cliniques et peut conduire à des erreurs induisant des choix de traitements non adaptés. L'étude des phénomènes au niveau moléculaire pourrait, en ce sens, ouvrir la voie à de meilleures performances de diagnostic et conduire, à terme, à une meilleure prise en charge des patients. Cela ne représente qu'un exemple mais souligne l'importance de la problématique de classification non supervisée dans ce contexte. Néanmoins, contrairement au cadre supervisé, la classification non supervisée dans le cas de données complexes reste un problème relativement peu abordé. Qui plus est, la prise en compte de la variabilité inter-individuelle inhérente à ce type de données dans un tel contexte est aussi largement sous représentée dans la littérature actuelle.

Dans le Chapitre 5, notre objectif est de proposer une procédure efficace d'un point de vue numérique, permettant la découverte de groupes au sein de données complexes en présence de variabilité inter-individuelle. Pour ce faire, nous choisissons d'adopter une approche probabiliste de la classification basée sur un modèle de mélange fonctionnel mixte. Notre modèle est la réunion du modèle de classification de courbes développé par Antoniadis et al. (2008) et du modèle mixte fonctionnel proposé par Antoniadis et Sapatinas (2007), présentés en Partie I. Une représentation du modèle dans le domaine des ondelettes est utilisée afin de se ramener, au sein de chaque groupe homogène, à un modèle linéaire mixte. Dans un cadre fonctionnel, nous proposons à l'image d'Antoniadis et al. (2008), une première étape de réduction de dimension basée sur les techniques de seuillage par ondelettes. L'estimation des paramètres du modèle sur les données réduites est ensuite réalisée par maximum de vraisemblance au moyen de l'algorithme EM, adapté à la prise en compte de variables latentes. Le point difficile ici est que nous sommes confrontés à la présence simultanée de deux types de variables latentes : les labels des individus et les effets aléatoires. Le choix du nombre de groupes est réalisé *a posteriori* grâce à l'utilisation d'un critère de type BIC. Bien que ne prenant pas en compte la présence d'effets aléatoires dans le calcul de la vraisemblance, ce critère montre de bonnes performances sur les simulations effectuées.

Dans le Chapitre 6, nous présentons une étude de simulation approfondie visant à étudier le comportement de notre procédure dans une large variété de configurations. Nous nous sommes particulièrement concentrés ici sur la définition d'un protocole de simulation de jeux de données synthétiques. La définition de cadres de simulations unifiés est un sujet important à plusieurs titres. Cela représente en premier lieu un moyen de comparer objectivement différentes procédures, nouvelles et existantes, grâce à l'utilisation de valeurs de paramètres de simulations canoniques. Par ailleurs, cela conduit à une meilleure exploration de l'univers des simulations permettant ainsi d'étudier les limites des procédures testées. Enfin, nous pouvons souligner l'importance de cette démarche dans un cadre d'analyse fonctionnelle. En effet, les procédures étant appliquées sur les coefficients d'une décomposition en on-

delettes provenant d'un modèle fonctionnel sous-jacent, il est crucial de s'assurer que les données simulées aient un sens dans le cadre fonctionnel. Malgré cela, cette démarche de spécification est relativement peu abordée dans la littérature et en ce sens, ce travail y apporte une contribution. Dans le cadre de cette étude de simulation, nous mettons en avant l'importance de la prise en compte des effets aléatoires pour la classification lorsque les données présentent une variabilité inter-individuelle significative. De plus, en comparant notre procédure à une procédure développée par James et Sugar (2003), basée sur une modélisation mixte similaire mais utilisant les bases de splines, nous montrons le gain de performance, vis-à-vis de la classification, apporté par les ondelettes pour l'étude de données irrégulières de grande dimension.

Nous développons en outre deux applications à des jeux de données réelles. Le premier est issu de la technologie des données de spectrométrie de masse et concerne des sujets féminins atteints ou non d'un cancer de l'ovaire (Petricoin et al. 2002), représentant deux groupes d'individus homogènes. Ces données sont analysées sans la connaissance des labels individuels mais en fixant le nombre de groupes à deux, l'objectif étant de voir si les classes "malades" et "sains" sont retrouvées par notre procédure. L'idée sous-jacente est de se demander si ces données sont pertinentes pour la caractérisation de ces groupes. Dans ce contexte, nous constatons que la prise en compte de la variabilité inter-individuelle est un facteur important vis-à-vis des performances de classification. La présence d'une forte variabilité individuelle est un fait connu pour ce type de données (Antoniadis et al. 2007) et notre étude met en lumière la nécessité de les prendre en compte lors de l'analyse de telles données. Un résultat secondaire concerne la modélisation de la variabilité individuelle : en effet, notre étude fait apparaître de meilleures performances lorsque la variabilité dépend de la position et de la résolution considérée. Dans ce contexte, l'observation des variances estimées des effets aléatoires montre de nombreuses valeurs proches de zéro. Cela nous incite à penser que la représentation de la variabilité individuelle dans le domaine des ondelettes est parcimonieuse, c'est-à-dire qu'une grande partie des variances associées aux effets aléatoires sont nulles. Ce point particulier constitue une motivation pour le développement de la deuxième contribution de ce manuscrit, développée en Partie III.

Enfin, notre deuxième application concerne l'étude d'un jeu de données génomiques issu de la technologie des microarray CGH. Ces données sont mesurées sur une cinquantaine de patientes atteintes d'un cancer du sein pour la mesure d'environ 2000 points le long du génome pour chaque individu (Fridlyand et al. 2006). Le développement de ce cancer peut prendre plusieurs formes différentes associées à plusieurs variantes de la pathologie. La découverte et le diagnostic précis de ces variantes constituent alors un enjeu et nécessitent d'aborder ces données d'un point de vue entièrement non supervisé, c'est-à-dire, sans connaissance des labels individuels et du nombre de groupes. Ces données ont été analysées de nombreuses fois (Fridlyand et al. 2006; Van Wieringen et al. 2008), donnant lieu à des résultats de classification différents suivant les approches considérées. Notre originalité est de proposer une

approche modélisant la présence d'effets aléatoires, ce qui, à notre connaissance, n'a pas encore été réalisé sur ce type de données. Cela fait apparaître, par une estimation *a posteriori*, la présence d'une très forte variabilité inter-individuelle, nous amenant à conclure que la découverte de groupes homogènes d'un point de vue biologique nécessiterait de disposer d'un nombre beaucoup plus important d'individus.

Ces travaux ont fait l'objet d'un article scientifique publié dans la revue *Biometrics* (Giacofci et al. 2013) ainsi que de présentations orales aux congrès IBS Channel (Bordeaux - Avril 2011) ainsi qu'à la première conférence de l'ISNPS (Grèce - Juin 2012). L'ensemble des procédures développées sont de plus implémentées dans le package `R` nommé `curvclust`, disponible sur le site du CRAN [1].

1.4 Estimation dans les modèles mixtes fonctionnels

Résumé de la Partie III

Une fois les groupes formés, se pose naturellement la question de l'estimation des paramètres du modèle mixte fonctionnel au sein d'un groupe d'individus homogène. À l'image des problématiques rencontrées dans les modèles mixtes standards, l'objectif premier est de proposer un estimateur de l'effet fixe, fonctionnel dans notre cadre, car celui-ci est associé au comportement moyen au sein d'un groupe et est donc lié au phénomène sous-jacent étudié. De manière latente, se pose la problématique d'estimation des effets aléatoires : en effet, la qualité d'estimation des effets fixes au sein des modèles mixtes est dépendante de celle des effets aléatoires. De plus, la compréhension de la variabilité inter-individuelle représente un enjeu en soi dans la mesure où elle peut conduire à terme à mieux appréhender les effets d'une pathologie par exemple en terme de variabilité de la réponse individuelle. Ce double objectif est abordé selon deux approches distinctes dans cette troisième partie, qui représente la deuxième contribution de ce travail. Étant donné le caractère fonctionnel des effets fixes et aléatoires, l'estimation basée sur une représentation du modèle dans le domaine des ondelettes, se traduit par une représentation parcimonieuse des effets fixes et aléatoires. Concernant les effets aléatoires, réalisations de processus gaussiens centrés, la parcimonie se traduit alors par une parcimonie du vecteur des variances associées aux effets aléatoires. Notre objectif, dans un cadre fonctionnel, est de retrouver cette parcimonie grâce à l'adaptation des techniques d'estimation non paramétrique par ondelettes au cadre mixte.

Dans le Chapitre 7, nous présentons une première approche basée sur une vision marginale des modèles mixtes. Dans cette approche, notre objectif principal se concentre sur la reconstruction de l'effet fixe fonctionnel. Vis-à-vis de cette problématique, la représentation du modèle mixte fonctionnel dans le domaine des ondelettes se ramène à un modèle de régression non paramétrique par ondelettes.

1. `http://cran.r-project.org/`

L'idée de seuillage dans un cadre non homoscédastique n'est pas nouvelle au sein de la littérature non paramétrique. En effet, à la suite des travaux pionniers de Donoho et Johnstone (1994) sur le seuillage par ondelettes, Johnstone et Silverman (1997), Gao (1997) et von Sachs et MacGibbon (2000) se sont intéressés à la problématique de l'estimation non paramétrique lorsque le bruit perturbant le signal moyen n'est plus un bruit blanc mais respectivement, un processus stationnaire ou un processus non stationnaire. Dans le domaine des ondelettes et sous la propriété de décorrélation de ces dernières (Frazier et al. 1991), ces modélisations se traduisent par des variances dépendant du niveau de résolution dans le cas stationnaire ou du niveau de résolution et de la position dans le cas non stationnaire. Cependant, ces travaux se placent dans le cadre non paramétrique classique, c'est-à-dire quand le nombre d'individus noté N est égal à 1, entraînant la nécessité de mettre des hypothèses supplémentaires sur les variances pour leur estimation (de régularité par exemple). Dans notre cadre, nous disposons de répétitions individuelles, permettant une estimation plus aisée des paramètres de variance. Peu de travaux existent dans le cadre hétéroscédastique avec répétitions et un des articles pionniers est celui de Amato et Sapatinas (2005). Ces derniers proposent une étude empirique concluant qu'en présence de répétitions, la stratégie consistant à effectuer un seuillage de la moyenne des signaux est préférable à celle consistant à faire une moyenne des signaux seuillés individuellement. Néanmoins, dans ce cadre, les propriétés de convergence des estimateurs et la problématique d'estimation des variances ne sont pas abordées. Forts de ces approches existantes, notre stratégie, développée dans un cadre hétéroscédastique en présence de répétitions individuelles est basée sur les idées suivantes :

- En se plaçant à variance connue, nous adoptons la stratégie mise en avant par Amato et Sapatinas (2005) consistant à appliquer une procédure de seuillage sur la moyenne des signaux individuels. Cette stratégie possède l'avantage de conserver les propriétés de convergence des estimateurs de seuillage et le fait de moyenner les signaux permet en outre de diminuer la variabilité d'un facteur N^{-1}. La vitesse de convergence de l'estimateur de l'effet fixe fonctionnel résultant dépend alors principalement de la taille des signaux notée M.
- L'estimation des variances est réalisée de manière séparée en adoptant par la suite une stratégie de type *plug-in* pour la mise en œuvre du seuillage. La présence à chaque position de N répétitions nous permet d'estimer les paramètres de variance dans un contexte hétéroscédastique en atteignant un taux de convergence paramétrique en le nombre d'individus N.

Nous nous basons plus particulièrement sur une stratégie de seuillage de type SCAD (Antoniadis et Fan 2001), connue pour conduire à de bonnes propriétés de convergence de l'estimateur résultant vers la vraie fonction. Le seuil choisi est alors le seuil universel développé par Donoho et Johnstone (1994) conduisant à des estimateurs atteignant une vitesse de convergence *near-minimax* dans la classe des espaces de Besov, c'est-à-dire minimax pour cette classe à un facteur logarithmique près.

Sous une hypothèse de consistance des estimateurs de variance, nous démontrons

que l'estimateur de l'effet fixe fonctionnel résultant d'un seuillage hétéroscédastique, lorsque les variances sont inconnues, atteint bien la vitesse de convergence near-minimax dans la classe des espaces de Besov. Notons que pour la démonstration de ce résultat, le seuil universel est légèrement modifié (multiplié par une constante) pour des raisons techniques de preuve. Une question ouverte, que nous pensons atteignable, serait de démontrer que ce résultat reste valable dans le cas du seuil universel. Une discussion de ce résultat est donnée vis-à-vis du ratio entre le nombre d'individus N et le nombre de variables M : en effet, intuitivement, l'estimation d'un paramètre de variance à chaque position entraîne la présence d'erreurs paramétriques qui s'additionnent sur l'ensemble des positions, nécessitant de contrôler le rapport M/N afin que celle-ci ne diverge pas. En fait, nous pouvons même montrer que pour un N suffisamment grand devant M, la convergence de l'estimateur de l'effet fixe fonctionnel peut même être accélérée grâce à la présence de répétitions individuelles.

Concernant la problématique de l'estimation des variances, nous proposons une première méthode basée sur une estimation empirique sans biais des variances réalisée grâce à la présence de répétitions individuelles. Cette première méthode présente l'intérêt de conduire à des estimations $\sqrt{N}$-consistantes des paramètres de variances. Dans une optique de sélection des variances associées aux effets aléatoires, nous proposons une deuxième procédure basée sur les techniques de vraisemblance pénalisée au moyen d'une pénalité de type LASSO. L'idée est ici de proposer une méthode permettant de réaliser une sélection des variances tout en conservant le caractère non itératif de la procédure de seuillage développée pour en conserver la rapidité d'exécution. Cela justifie le choix d'une pénalité de type LASSO mais ne nous garantit alors plus de disposer des propriétés de consistance des estimateurs des variances (Donoho et Huo 2002; Meinhausen et Buhlmann 2004).

Dans le Chapitre 8, nous proposons une deuxième approche axée sur la problématique de sélection de variables concernant les effets fixes et les variances des effets aléatoires simultanément. La première approche proposée au Chapitre 7 basée sur les techniques non paramétriques standards, bien que conduisant à un estimateur reconstruit de l'effet fixe fonctionnel optimal en terme de risque quadratique, n'est pas construite dans l'objectif de proposer une sélection des variances des effets aléatoires performante pour des raisons de rapidité numérique. Il est en effet difficile de construire une procédure optimale en terme de sélection des variances des effets aléatoires et d'estimation de l'effet fixe fonctionnel sans passer par une résolution itérative. Notre motivation a donc été de développer proprement une procédure itérative dont l'objectif est de réaliser une sélection des effets fixes et aléatoires simultanément.

Au cours de cette deuxième approche basée sur une vision jointe des modèles mixtes, l'estimation/sélection des paramètres est réalisée par maximum de vraisemblance en optimisant un critère de vraisemblance pénalisée. Plus particulièrement, la vraisemblance du modèle est pénalisée au moyen de deux pénalités de type SCAD concernant les coefficients des effets fixes et les variances des coefficients des effets

aléatoires, induisant une sélection de ces deux types de variables en forçant la mise à zéro d'une partie d'entre elles.

L'optimisation de ce critère doublement pénalisé conduit à des estimateurs des paramètres du modèle possédant la propriété d'oracle, à savoir : le vrai modèle est retrouvé presque sûrement et les estimateurs des paramètres sont asymptotiquement normaux. Ce résultat a été démontré par Bondell et al. (2010) dans un cadre non fonctionnel lorsque le nombre d'individus N tend vers l'infini tandis que le nombre de covariables M est fixé. La contrainte d'un nombre de variables M fixé, correspondant au nombre de points de discrétisations du signal, n'a pas de sens dans un cadre fonctionnel et nous étendons donc ce résultat au cas où M diverge avec N en imposant la contrainte que $M^5/N \to 0$ avec $M < N$. Pour ce faire, nous nous basons sur une preuve proposée par Fan et Peng (2004) concernant les propriétés oraculaires de l'estimateur des effets fixes dans le cadre non mixte sous les mêmes contraintes concernant M et N. L'extension consiste à démontrer que ce résultat reste valable dans le cadre de la "double pénalisation" des effets fixes et des variances des effets aléatoires.

Dans un deuxième temps, nous développons une procédure itérative permettant l'optimisation effective du critère de vraisemblance pénalisée. Le développement de notre procédure nécessite préalablement de considérer une reparamétrisation des effets aléatoires à l'instar de celle proposée par Chen et Dunson (2003). Techniquement, cette reparamétrisation a pour effet de faire passer les paramètres de variances à un statut de "coefficients de régression", constituant une première étape pour le développement d'une procédure d'estimation itérative.

Notre procédure est alors basée sur une variante ECM (Expectation Conditional Maximization) de l'algorithme EM, consistant à remplacer l'étape M par une succession de maximisations conditionnelles. Cet algorithme, développé par Meng et Rubin (1993), possède les mêmes propriétés que l'algorithme EM, garantissant son bon comportement. Les maximisations conditionnelles vis-à-vis des paramètres d'effets fixes et des écarts-types des effets aléatoires se ramènent, dans cette approche, à des seuillages respectifs des données corrigées des effets aléatoires et des données corrigées des effets fixes.

Dans le Chapitre 9, nous proposons en dernier lieu une étude de simulation afin d'étudier les comportements des différentes procédures sur la base de jeux de données synthétiques. Un effort est réalisé quant à la définition d'un cadre de simulation unifié adapté à l'évaluation de l'estimation de l'effet fixe fonctionnel et de la sélection des effets fixes et des variances des effets aléatoires.

Dans un premier temps, les deux approches développées dans cette partie sont évaluées sur la base de jeux de données présentant des configurations particulières permettant de séparer la parcimonie associée aux effets fixes et aléatoires respectivement. Dans ces configurations particulières, les deux approches se révèlent plus performantes lorsque les effets fixes et aléatoires ne s'exercent pas sur les mêmes positions.

Vis-à-vis de l'approche marginale, nous mettons en évidence la supériorité de la procédure de seuillage hétéroscédastique basée sur des estimations empiriques des variances, en terme de reconstruction de l'effet fixe fonctionnel. Cette procédure se distingue des autres quant à l'estimation des variances car c'est la seule à ne pas être basée sur l'estimateur MAD, classiquement utilisé en ondelettes pour l'estimation de la variance du bruit. L'estimateur MAD construit à partir des coefficients du niveau de résolution le plus fin, se révèle être systématiquement biaisé positivement à cause de la présence de signal au niveau de résolution le plus fin. Ce point avait déjà été mentionné initialement par Donoho et Johnstone (1998) et est confirmé par notre étude, particulièrement dans le cas de signaux présentant de fortes discontinuités. De plus, d'après Donoho et Johnstone (1998), on ne peut pas espérer un meilleur comportement de l'estimateur MAD lorsque M augmente car la probabilité que ce dernier soit biaisé positivement augmente avec la taille du signal M.

Concernant les procédures basées sur une estimation/sélection itérative des paramètres, la principale difficulté est d'ordre numérique puisque la présence d'une double pénalisation entraîne la nécessité d'ajuster deux hyperparamètres. Nous proposons de contourner cette difficulté en fixant pour l'hyperparamètre associé à la sélection des effets fixes le seuil universel de Donoho et Johnstone (1994). Cette procédure, outre un temps de calcul réduit, offre de bonnes performances de sélection des effets fixes et des variances d'effets aléatoires.

Dans un dernier temps, nous comparons les approches basées sur des visions marginale et jointe des modèles mixtes sur des données simulées de manière réaliste, c'est-à-dire lorsque les parcimonies des effets fixes et aléatoires sont mélangées et pour des effets aléatoires ayant un sens d'un point de vue fonctionnel. De cette comparaison, ressort principalement le fait que, malgré une grande rapidité d'exécution, la procédure de seuillage hétéroscédastique ne permet pas d'effectuer une sélection satisfaisante des variances associées aux effets aléatoires et ne répond donc pas à l'un des objectifs initialement fixé. D'autre part, les procédures basées sur une approche jointe des modèles mixtes, bien que présentant un temps de calcul largement supérieur, permettent de réaliser de bonnes performances de sélection. Cependant, aucune propriété de convergence de l'effet fixe fonctionnel n'a été démontrée. Sur cette première étude comparative, la qualité de reconstruction de l'estimateur fonctionnel présente des performances similaires pour les deux approches développées, nous incitant à étudier, dans une perspective de ce travail, les propriétés de convergence de l'estimateur de l'effet fixe fonctionnel sous l'approche jointe.

Première partie

Vers le modèle mixte fonctionnel et la classification non supervisée

Introduction

Au cours de cette première partie, nous introduisons les principaux concepts qui constituent les fondements de ce travail de thèse, à savoir : les modèles à effets mixtes et la modélisation fonctionnelle basée sur l'utilisation de bases d'ondelettes. Dans un premier chapitre, nous définissons le cadre usuel des modèles mixtes dans le cas particulier de la modélisation de données longitudinales. Le terme de *données longitudinales* se réfère aux données mesurées pour un individu au cours du temps (ou de la même manière, selon une grille d'espace). Bien que les modèles mixtes soient une classe de modèles beaucoup plus large, nous nous restreignons au cadre longitudinal car il constitue une base naturelle pour une future extension à un cadre fonctionnel. Pour une étude détaillée des modèles mixtes appliqués aux données longitudinales, nous invitons le lecteur à se référer à l'ouvrage de Verbeke et Molenberghs (2000), dont le Chapitre 2 est inspiré.

Le chapitre suivant concerne la modélisation de données fonctionnelles : dans de nombreux domaines scientifiques, le récent développement d'appareils de mesures de plus en plus performants nous donne accès à des données séquencées de manière régulière et à des résolutions de plus en plus conséquentes. Dans ce contexte, Ramsay et Silverman (1997) développent le paradigme des *données fonctionnelles* en désignant par ce terme les données dont l'unité d'observation idéale est la courbe, c'est-à-dire des données mesurées pour chaque individu sur une grille de temps ou d'espace fine et régulière et pour lesquelles on souhaite en exploiter les propriétés de régularité sous-jacentes. Étant donné que ces données sont mesurées sur une grille de temps, elles représentent, en conséquence, une extension naturelle des données longitudinales, la principale différence résidant dans la dimension des données considérées, usuellement largement plus importante pour les données de type fonctionnel. Dans ce cadre, nous présentons les principales problématiques associées à la modélisation de ces données et nous insisterons plus particulièrement sur les modélisations non-paramétriques basées sur l'utilisation de bases d'ondelettes.

Enfin, nous présentons, dans un troisième chapitre, la classe des modèles à variables latentes afin d'exposer le concept de classification non supervisée basée sur une approche probabiliste au sein des modèles fonctionnels ainsi que la prise en compte d'effets aléatoires dans ce contexte. On désigne par le terme de *variables latentes*, ou encore de *données cachées*, la présence de variables non observées au sein d'un modèle. L'introduction des modèles de classification fonctionnelle et mixte

fonctionnel est réalisée au travers de la notion de variables latentes car elle constitue un cadre commun aux problèmes de classification où les variables d'appartenance aux groupes sont non observées et à la présence d'effets aléatoires, représentant eux aussi des variables non observées.

Chapitre 2

Modèle mixte pour données longitudinales

Nous introduisons au sein de ce chapitre la notion de modèles linéaires mixtes dans le cas particulier de l'application aux données longitudinales ainsi que les problématiques usuelles associées à l'étude de tels modèles. Nous présentons deux approches distinctes des modèles mixtes : dans un premier temps, l'approche marginale mettant en avant l'étude des effets fixes du modèle et représentant une base pour l'inférence sur ceux-ci et dans un deuxième temps, l'approche hiérarchique permettant de prendre explicitement en compte la présence d'effets aléatoires et donc de pouvoir faire de l'inférence sur ces derniers. Enfin, nous décrivons brièvement les algorithmes d'optimisation usuels pour l'estimation par maximum de vraisemblance au sein de ces modèles.

2.1 Modèle général

Lors de l'étude de données expérimentales, le statisticien est souvent amené à gérer la présence de variabilité, généralement modélisée sous la forme d'un bruit blanc afin de marquer la présence d'erreurs de mesure. Cependant, il peut aussi être confronté à la présence d'une variabilité spécifique, due à l'individu : on parle alors de variabilité inter-individuelle. Cette variabilité représente la propension des individus à s'écarter du comportement moyen de la population étudiée et la modélisation de celle-ci représente l'essence même des modèles mixtes. Le modèle linéaire mixte peut donc être vu comme une extension du modèle linéaire classique où la présence d'une variabilité individuelle est prise en compte grâce à l'introduction de termes appelés *effets aléatoires* au sein du modèle linéaire. Par opposition, les comportements moyens de la population sont nommés *effets fixes*. Plus particulièrement, un effet est dit fixe si ses différents niveaux représentent entièrement les niveaux existants. On parle *a contrario* d'effets aléatoires lorsque les niveaux de ces effets représentent un échantillon des niveaux possibles, c'est à dire qu'ils présentent une variabilité propre

aux individus.

Laird et Ware (1982) sont les premiers à définir un cadre général pour l'étude des modèles mixtes appliqués aux données longitudinales. Considérons la variable réponse Y_{im} de l'individu i au temps t_m avec $i = 1, \ldots, N$ et $m = 1, \ldots, M$. Dans leur approche, le modèle linéaire mixte pour le vecteur $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iM})$ s'exprime alors, en toute généralité, de la façon suivante :

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\theta}_i + \mathbf{E}_i, \qquad \forall i = 1, \ldots, N, \tag{2.1}$$

où $\boldsymbol{\beta}$ est un vecteur de taille p_1 contenant les effets fixes, $\boldsymbol{\theta}_i$ un vecteur de taille p_2 contenant les effets aléatoires spécifiques à l'individu i. Le vecteur $\mathbf{E}_i$ de taille M contient, quant à lui, les composantes résiduelles ; il est supposé gaussien centré et de matrice de covariance $\sigma_E^2\mathbf{I}_M$, avec $\mathbf{I}_M$ matrice identité de taille $M \times M$. Les effets aléatoires $\boldsymbol{\theta}_i$ sont aussi supposés gaussiens centrés et de matrice de covariance $\mathbf{G}$ de taille $p_2 \times p_2$. Traditionnellement, les composantes $\boldsymbol{\theta}_i$ et $\mathbf{E}_i$ sont supposées indépendantes pour tout $i = 1, \ldots, N$. Enfin, $\mathbf{X}_i$ et $\mathbf{Z}_i$ sont respectivement des matrices de plan d'expérience de taille $M \times p_1$ et $M \times p_2$ contenant les covariables associées aux effets fixes et aléatoires.

Notons que plus généralement, on peut supposer que les pas de temps $\mathbf{t}$ diffèrent en nombre et/ou en espace pour chaque individu. Historiquement, cette possibilité a fortement participé à la popularisation des modélisations mixtes. Nous supposerons, pour notre part, que les temps de mesure sont équirépartis et identiques pour tous les individus afin de simplifier par la suite l'introduction de l'approche fonctionnelle.

2.2 Approche marginale

À l'instar de modèles plus classiques tel que le modèle linéaire, l'objectif principal est d'obtenir des informations sur le comportement moyen des individus vis à vis du phénomène étudié en fonction de variables d'intérêts ou covariables. Cela revient alors à rechercher un estimateur du vecteur de régresseurs $\boldsymbol{\beta}$ et ceci fait l'objet de l'approche dite *marginale* des modèles mixtes. Cette première approche consiste à considérer les effets aléatoires individuels, au vu de leur nature aléatoire, comme une deuxième source de variabilité, ajoutée à la composante résiduelle. La recherche d'une bonne modélisation de la variabilité est uniquement réalisée en vue de l'estimation des effets fixes $\boldsymbol{\beta}$.

La distribution des $\mathbf{Y}_i$ dans le modèle (2.1) est alors définie par :

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i), \qquad \forall i = 1, \ldots, N, \tag{2.2}$$

où $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \sigma_E^2\mathbf{I}_M$ est la matrice de covariance de taille $M \times M$ associée aux observations $\mathbf{Y}_i$. Notons que cette matrice n'est pas supposée posséder de structure particulière et est en général une matrice pleine.

Les paramètres à estimer du modèle (2.2) sont alors constitués du vecteur des effets fixes $\boldsymbol{\beta}$ et des paramètres de variance contenus dans la matrice $\mathbf{V}_i$, c'est-à-dire les $p_2(p_2+1)/2$ paramètres de $\mathbf{G}$ et le paramètre σ_E^2, notés $\mathbf{v}$ dans la suite de ce chapitre.

2.2.1 Estimation des effets fixes par maximum de vraisemblance

Dans un cadre fréquentiste, l'estimation des paramètres dans les modèles linéaires mixtes est généralement réalisée par maximum de vraisemblance. Dans ce cadre, la log-vraisemblance des données observées $(\mathbf{Y}_i)_{i=1,\dots,N}$ est donnée par :

$$-2\log\mathcal{L}(\boldsymbol{\beta},\mathbf{G},\sigma_E^2) = NM\log(2\pi) + \sum_{i=1}^{N}\log|\mathbf{V}_i| + \sum_{i=1}^{N}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}). \tag{2.3}$$

En optimisant (2.3) par rapport au vecteur de paramètres $\boldsymbol{\beta}$ et à $\mathbf{v}$ fixé, on obtient un estimateur des effets fixes donné par :

$$\widehat{\boldsymbol{\beta}}(\mathbf{v}) = \left[\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{Y}_i. \tag{2.4}$$

Cet estimateur dépend des paramètres de variance du vecteur $\mathbf{v}$. Si un estimateur de ces paramètres est disponible, on peut alors estimer $\boldsymbol{\beta}$ en remplaçant $\mathbf{V}_i$ par son estimateur $\widehat{\mathbf{V}}_i = \mathbf{V}(\widehat{\mathbf{v}})$ dans l'expression (2.4). Remarquons que cet estimateur est l'estimateur GLS (Generalized Least Squares), à savoir l'estimateur des moindres carrés pris avec une métrique $\mathbf{V}_i^{-1}$ au lieu de la métrique euclidienne habituelle.

2.2.2 Estimation des paramètres de variance : MLE et REML

L'estimation des paramètres de variances peut être réalisée de deux façons différentes au sein des modèles mixtes : par maximum de vraisemblance (estimateur dit MLE) ou par maximum de vraisemblance restreint (estimateur dit REML). L'approche par maximum de vraisemblance consiste à maximiser la vraisemblance des données (2.3) par rapport aux paramètres de variances $\mathbf{v}$, en ayant au préalable remplacé $\boldsymbol{\beta}$ par son estimateur (2.4).

Cependant, cette approche conduit à l'obtention d'estimateurs biaisés et ceci vient de la non prise en compte de la perte de degré de liberté occasionnée par l'estimation préalable des effets fixes. On peut faire sur ce point l'analogie avec le modèle de régression linéaire usuel où l'on considère un estimateur de la variance non-biaisé en prenant un facteur $\frac{1}{N-1}$ au lieu du facteur $\frac{1}{N}$ dans la variance empirique. Dans le cadre mixte, où la structure de variance est plus complexe, l'obtention d'estimateurs non biaisés est réalisée grâce à l'approche REML développée par Patterson et Thompson (1971).

Cette stratégie consiste, en théorie, à définir une matrice de contraste, notée $\mathbf{A}_i$, de taille $N \times (N - p_1)$ composée de vecteurs orthogonaux aux p_1 colonnes de la matrice de plan d'expérience $\mathbf{X}_i$. L'estimateur REML consiste alors à maximiser la vraisemblance du vecteur des contrastes $\mathbf{A}_i^T \mathbf{Y}_i$ par rapport au vecteur des paramètres de variance $\mathbf{v}$, conduisant ainsi à des estimateurs non biaisés de ces derniers.

Un des avantages de l'utilisation de cette stratégie est qu'en pratique, elle ne nécessite pas de définir explicitement la matrice de contraste $\mathbf{A}$. En effet, Harville (1974) démontre l'égalité suivante :

$$\log \mathcal{L}_{\text{REML}}(\boldsymbol{\beta}, \mathbf{G}, \sigma_E^2) = \left| \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right|^{-\frac{1}{2}} \log \mathcal{L}_{\text{ML}}, (\boldsymbol{\beta}, \mathbf{G}, \sigma_E^2), \tag{2.5}$$

permettant d'accéder aux estimateurs REML à partir des estimateurs MLE.

Les deux stratégies d'estimation MLE et REML conduisent à l'obtention d'estimateurs différents pour les paramètres de variance et par (2.4), à des estimateurs différents des effets fixes. Toutes deux sont basées sur les techniques de maximum de vraisemblance et bénéficient donc des bonnes propriétés de consistance et de normalité asymptotique en découlant. Le choix entre ces stratégies est alors principalement guidé par le nombre d'effets fixes p_1 présents dans le modèle. En effet, pour un nombre p_1 raisonnable (par exemple $p_1 \leq 4$), l'estimateur MLE est relativement peu biaisé et présente l'avantage d'un écart quadratique moyen bien inférieur à celui de l'estimateur REML pour toute valeur de N (Verbeke et Molenberghs 2000). Dans le cas contraire, quand le nombre d'effets fixes est élevé, l'estimateur REML est alors préférable afin de minimiser le biais de l'estimateur. Pour une étude plus complète de ces stratégies, le lecteur pourra se référer à Harville (1977). Dans tous les cas, l'estimation des paramètres au sein de l'approche marginale nécessite une opération d'inversion des matrices de variance $\mathbf{V}_i$ qui peut se révéler coûteuse numériquement en présence d'un nombre élevé de paramètres.

2.2.3 Inférence dans le modèle marginal

Traditionnellement, l'ajustement du modèle n'est pas l'objectif final du praticien mais plutôt un moyen permettant d'inférer par la suite sur la population entière étudiée. En ce sens, la problématique d'inférence occupe une place centrale au sein de l'étude des modèles mixtes. L'inférence n'est pas un sujet explicitement traité dans ce manuscrit et nous développerons peu cette notion mais nous gardons à l'esprit l'importance de cette problématique. Au sein du modèle mixte marginal (2.2), l'inférence concerne en premier lieu les effets fixes et consiste principalement à construire des tests sur les effets fixes basés sur une bonne estimation de la partie résiduelle. Une fois les estimateurs connus, les tests les plus utilisés dans le cadre des modèles mixtes sont, entre autres, le test de Wald permettant les tests de contrastes basés sur les effets fixes ou encore le test du rapport de vraisemblance permettant la comparaison de modèles emboîtés. En revanche, la bonne spécification de la partie résiduelle

du modèle devient alors un enjeu majeur et en nécessite une étude approfondie. Ce point fait l'objet de la prochaine section.

2.3 Approche jointe et prédiction des effets aléatoires

L'approche marginale des modèles mixtes présente quelques lacunes du fait qu'elle ne prend pas explicitement en compte la présence d'effets aléatoires. En effet, d'une part, la qualité d'estimation et d'inférence concernant les effets fixes est directement liée à la bonne modélisation de la variabilité. D'autre part, le praticien peut également être intéressé par la prédiction des effets aléatoires individuels $(\boldsymbol{\theta}_i)_{i=1,\dots N}$ car ceux-ci représentent une source d'information sur les déviations individuelles au profil moyen et permettent ainsi de repérer les comportements spécifiques d'individus ou de groupes d'individus. Pour ce faire, il est nécessaire d'adopter une approche prenant en compte la présence d'effets aléatoires. Nous présentons donc dans cette section l'approche dédiée appelée approche jointe, ou hiérarchique, des modèles mixtes.

Le modèle (2.1) au sein de l'approche jointe est résumé sous la forme suivante :

$$\forall i = 1, \dots, N, \qquad \begin{cases} \mathbf{Y}_i &= \mathbf{Z}_i \mathbf{B}_i + \mathbf{E}_i, \\ \mathbf{B}_i &= \widetilde{\mathbf{X}}_i \boldsymbol{\beta} + \boldsymbol{\theta}_i, \end{cases} \tag{2.6}$$

où $\widetilde{\mathbf{X}}_i$ est une matrice de covariables de taille $(p_2 \times p_1)$ et $\mathbf{B}_i$ un vecteur de régresseurs de taille p_2, spécifique à l'individu i. Les hypothèses sur les autres quantités restent les mêmes que dans le modèle (2.1). Le modèle (2.1) se déduit du modèle (2.6) en posant $\mathbf{X}_i = \mathbf{Z}_i \widetilde{\mathbf{X}}_i$.

Connaissant les effets aléatoires, le modèle suivi par les observations $(\mathbf{Y}_i)_{i=1,\dots N}$ peut alors être décrit par :

$$\mathbf{Y}_i | \boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_E^2 \mathbf{I}_M), \qquad \forall i = 1, \dots, N, \tag{2.7}$$

Cette approche diffère de l'approche marginale principalement par les contraintes concernant l'espace des paramètres de variances du modèle : elles sont, en effet, plus fortes dans l'approche jointe puisque ces contraintes portent à la fois sur les variances associées aux effets aléatoires et sur la variance de l'erreur résiduelle. En pratique, l'ajustement du modèle marginal (2.2) conduit généralement à une convergence vers des paramètres situés en dehors de l'espace des paramètres induit par le modèle hiérarchique (2.6), c'est-à-dire conduisant à des estimations négatives des paramètres de variances.

Au sein du modèle (2.6), les effets aléatoires $(\boldsymbol{\theta}_i)_{i=1,\dots N}$ ne sont pas observés et l'objectif est alors d'en proposer une prédiction basée sur les observations $(\mathbf{Y}_i)_{i=1,\dots N}$. La meilleure prédiction des effets aléatoires, au sens de l'erreur quadratique, est alors

donnée par leur espérance *a posteriori* notée $\mathbb{E}(\boldsymbol{\theta}_i|\mathbf{Y}_i)$. Dans un cadre non gaussien, ces quantités ne sont en général pas calculables et il est alors classique d'en considérer des prédicteurs linéaires en les observations, notés $(\widehat{\boldsymbol{\theta}}_i)_i$. Les prédicteurs ainsi produits sont alors sans biais et de variance minimale parmi les estimateurs linéaires en les observations : ils sont de ce fait usuellement désignés sous la terminologie de BLUP (Best Linear Unbiaised Predictor).

Dans le cadre d'un modèle gaussien, comme en (2.6), la distribution des variables $(\mathbf{Y}_i|\boldsymbol{\theta}_i)_{i=1,\dots,N}$ est encore gaussienne et linéaire par rapport aux observations et on peut alors donner une expression explicite de son espérance :

$$\mathbb{E}(\boldsymbol{\theta}_i|\mathbf{Y}_i) = \widehat{\boldsymbol{\theta}}_i = \mathbf{G}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{v})(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}), \quad \forall i = 1,\dots,N, \tag{2.8}$$

où $\mathbf{v}$ et $\boldsymbol{\beta}$ sont remplacés par leur estimation respective.

Les estimateurs/prédicteurs $\widehat{\boldsymbol{\beta}}$ et $(\widehat{\boldsymbol{\theta}}_i)_{i=1,\dots N}$ des effets fixes et aléatoires peuvent aussi être construits comme solutions d'un système linéaire d'équations. Ces équations ont été proposées par Henderson et al. (1959) et représentent l'équivalent dans le cadre des modèles mixtes des équations normales rencontrées en modèle linéaire classique. Elles sont données par le système d'équations suivant :

$$\begin{bmatrix} \mathbf{X}_i^T\mathbf{X}_i/\sigma_E^2 & \mathbf{X}_i^T\mathbf{Z}_i/\sigma_E^2 \\ \mathbf{Z}_i^T\mathbf{X}_i/\sigma_E^2 & \mathbf{Z}_i^T\mathbf{Z}_i/\sigma_E^2 + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\theta}}_i \end{bmatrix} = \frac{1}{\sigma_E^2} \begin{bmatrix} \mathbf{X}_i^T\mathbf{Y}_i \\ \mathbf{Z}_i^T\mathbf{Y}_i \end{bmatrix}. \tag{2.9}$$

L'intérêt calculatoire d'un tel système est de permettre d'éviter l'étape d'inversion des matrices de covariances $(V_i)_{i=1,\dots,N}$ en se ramenant à l'inversion de matrices de tailles inférieures, ou structurées par blocs pour lesquelles des algorithmes rapides d'inversion existent. Cependant, cet avantage disparaît lorsque l'on est confronté à de grands jeux de données pour lesquels la résolution numérique du système (2.9) peut se révéler coûteuse car les différents blocs deviennent eux-mêmes de taille conséquente. On préférera, dans ce cas, utiliser des algorithmes itératifs d'estimation basés sur les expressions (2.8) et (2.4). Les deux principaux algorithmes utilisés pour l'estimation au sein des modèles mixtes sont présentés dans la section suivante.

2.4 Algorithmes d'estimation

Le problème d'optimisation des vraisemblances (2.3) ou (2.5) n'admet pas de solution explicite en règle générale. Leur maximisation nécessite donc le recours à des algorithmes itératifs. Les deux principaux algorithmes utilisés à cette fin sont l'algorithme EM (Dempster et al. 1977) et l'algorithme de Newton-Raphson adapté à l'estimation dans les modèles mixtes (Lindstrom et Bates 1988).

Pour ces deux méthodes, l'utilisateur fixe des valeurs de départ pour les paramètres du modèle. Ces valeurs sont ensuite mises à jour à chaque itération jusqu'à convergence. L'algorithme EM, basé sur la notion de données incomplètes, présente l'avantage d'assurer l'augmentation de la vraisemblance à chaque itération mais

est aussi connu pour avoir une convergence lente, surtout concernant les estimateurs des paramètres de variance (Laird et Ware 1982). De son coté, l'algorithme de Newton-Raphson est un algorithme rapide de recherche des zéros d'une fonction. Une des principales lacunes de cet algorithme est sa sensibilité particulière aux éventuelles mauvaises spécifications des paramètres de variance, c'est-à-dire à l'introduction d'effets aléatoires non pertinents. Il conduit dans ce cas à des estimations de variances situées sur le bord de l'espace des paramètres, c'est-à-dire, tendant vers zéro. Son application nécessite donc un travail de sélection en amont. Cette problématique de sélection des effets aléatoires sera abordée dans la troisième partie de ce manuscrit dans le cadre des modèles mixtes fonctionnels et la question de stabilisation des algorithmes d'estimation en représente une première motivation. Néanmoins, en pratique, cet algorithme reste aujourd'hui la méthode de résolution la plus populaire pour les modèles mixtes, principalement du fait de sa rapidité de convergence. Pour une revue détaillée des problèmes liés à l'estimation des paramètres au sein des modèles mixtes, le lecteur pourra consulter l'ouvrage de Verbeke et Molenberghs (2000).

Au cours de ce manuscrit, nous utiliserons principalement l'algorithme EM comme algorithme d'estimation au sein des modèles mixtes et celui-ci sera décrit de manière détaillée au Chapitre 4. Cet algorithme présente l'avantage, dans notre contexte, de s'adapter naturellement à la problématique de classification non supervisée au sein des modèles mixtes grâce au paradigme général des modèles à variables latentes (c.f. Chapitre 4).

Chapitre 3

Modélisation fonctionnelle par ondelettes

Dans ce chapitre, nous présentons la notion de modélisation fonctionnelle à partir de projection sur des bases de fonctions. Nous présentons en particulier les bases d'ondelettes et les espaces de Besov, outil de modélisation fonctionnelle privilégié dans ce manuscrit et bien adapté à l'étude de données fonctionnelles irrégulières. Enfin, nous introduisons les principales techniques de régression non paramétrique basée sur les ondelettes, regroupées sous le terme de techniques de seuillage, ainsi que leurs liens avec la classe plus large des régressions pénalisées.

3.1 Modélisation fonctionnelle

De manière formelle, le modèle fonctionnel simple peut être écrit de la façon suivante : nous disposons d'un signal mesuré en M points de temps, notés $\mathbf{t} = (t_1, \ldots, t_M)$. Au point t_m $(m = 1, \ldots, M)$, on a alors :

$$Y(t_m) = \mu(t_m) + E(t_m), \quad \text{avec} \quad E(t_m) \sim \mathcal{N}(0, \sigma_E^2), \tag{3.1}$$

où $Y(t_m)$ est le signal observé, $\mu(t_m)$ le signal fonctionnel moyen et $E(t_m)$ un terme d'erreur de mesure, chacun observé au point t_m. Dans une approche fonctionnelle, ces quantités sont vues comme des discrétisations de courbes sous-jacentes $Y(t), \mu(t), E(t)$. Dans un cadre de régression, le but est alors de donner une estimation de l'effet fixe fonctionnel moyen μ.

Si l'on dispose de connaissances a priori sur les données ou sur le processus générant ces données, on peut alors, dans le cadre du modèle (3.1), se placer dans le cadre de la régression paramétrique et ainsi, spécifier une forme pour la fonction μ. L'objectif est alors d'estimer les paramètres gouvernant le modèle. L'exemple le plus simple d'une telle approche est la régression linéaire où l'ajustement du modèle correspond à l'estimation de la pente et de l'ordonnée à l'origine de la droite

de régression. Cependant, la modélisation paramétrique peut rapidement se révéler trop contraignante pour certaines applications.

Par opposition au cadre paramétrique, une autre stratégie appelée non paramétrique consiste, dans le modèle (3.1), à ne pas spécifier de forme particulière pour la fonction μ et donc de se placer dans un espace de dimension infinie. Le principe est alors de faire peu d'hypothèses sur la fonction de régression μ. Usuellement, on se limite à supposer qu'elle appartient à un certain espace fonctionnel. Dans la suite de ce travail, nous nous intéresserons plus particulièrement aux fonctions d'un sous-espace de $L^2([0,1])$, ensemble des fonctions de carré intégrable à support sur l'intervalle $[0,1]$.

L'objectif dans ce cadre est alors de construire un estimateur de la fonction μ à partir de la connaissance des données. Une technique classique pour atteindre cet objectif consiste à projeter les fonctions du modèle sur une base de fonctions de l'espace fonctionnel considéré.

Ainsi, pour $\{\phi_k\}_k$ une base de Hilbert de l'espace $L^2([0,1])$, toute fonction $f \in L^2([0,1])$ peut être représentée comme suit :

$$f(t) = \sum_{k=0}^{\infty} \rho_k \phi_k(t),$$

où $\rho_k = \langle f, \phi_k \rangle$ est le k-ième coefficient de la projection de f dans la base de fonctions et l'application $\langle \cdot, \cdot \rangle$, le produit scalaire canonique de l'espace $L^2([0,1])$.

Il existe de nombreuses bases de fonctions envisageables pour traiter ce problème. Toutes possèdent des propriétés propres les rendant adaptées ou non à différents types de données et le choix de cette base doit donc se faire en accord avec les hypothèses faites sur les données. Ainsi, les fonctions splines (Wahba 1990) sont connues pour être particulièrement adaptées à l'étude de données mesurées en peu de points de discrétisation et modélisées par des fonctions lisses tandis que les régressions polynomiales sont plus adaptées au traitement des données possédant un design plus dense (Fan et Gijbels 1996).

Tout au long de ce manuscrit, nous nous concentrerons plus particulièrement sur un autre type de base de fonctions : les bases d'ondelettes. Celles-ci possèdent, comme nous le détaillerons plus tard, des propriétés intéressantes en termes de représentation de l'information contenue dans un signal et permettent de modéliser une grande variété de structures fonctionnelles dont des courbes présentant des discontinuités. Ce dernier point est capital pour les signaux que nous cherchons à traiter car la majeure partie de l'information est située précisément dans les irrégularités de ces signaux. Nous allons donc à présent donner une brève introduction des ondelettes et des espaces de Besov.

3.2 Ondelettes et espaces de Besov

L'utilisation des ondelettes en vue d'applications statistiques a connu un fort développement depuis les années 90. Cet outil est à présent très populaire car il permet de modéliser une grande variété de signaux. Ceci est rendu possible principalement grâce aux caractéristiques multi-échelles des ondelettes. En effet, l'utilisation de telles bases de fonctions permet de considérer les signaux étudiés à des niveaux de résolution successifs, et donc d'étudier les fonctions considérées à des fréquences de plus en plus fine. Tout se passe comme si des *zooms* successifs étaient effectués, on parle alors d'une propriété de *zoom in/zoom out*.

3.2.1 Analyse multirésolution

Nous introduisons les ondelettes en prenant comme point de départ la notion d'analyse multirésolution, mettant ainsi en avant cette propriété de *zoom in/zoom out*.

Définition 3.1. *On appelle* analyse multirésolution *de* $L^2(\mathbb{R})$ *toute suite croissante* $\{V_j\}_{j\in\mathbb{Z}}$ *de sous-espaces vectoriels de* $L^2(\mathbb{R})$ *vérifiant :*

(i) ***Complétude :*** $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$ *et* $\bigcup_{j\in\mathbb{Z}} V_j$ *est dense dans* $L^2(\mathbb{R})$,

(ii) ***Auto-similarité en échelle :*** $\forall f \in L^2(\mathbb{R})$, $j \in \mathbb{Z}$, *on a :*

$$f(\cdot) \in V_j \iff f(2\cdot) \in V_{j+1},$$

(iii) ***Auto-similarité en temps :*** $\forall f \in L^2(\mathbb{R})$, $k \in \mathbb{Z}$, *on a :*

$$f(\cdot) \in V_0 \iff f(\cdot - k) \in V_0,$$

(iv) ***Régularité :*** *il existe une fonction* ϕ, *appelée fonction d'échelle, telle que* $\{\phi(\cdot - k)\}_{k\in\mathbb{Z}}$ *soit une base orthonormée de* V_0.

Nous nous limitons ici à l'étude d'analyses multirésolution orthogonales permettant de définir des bases d'ondelettes dites orthogonales. Il est possible d'affaiblir la condition (iv) en se ramenant aux bases de Riesz et ainsi construire des bases bi-orthogonales. Pour plus de détails sur ce sujet, nous renvoyons le lecteur aux ouvrages de Härdle et al. (1998), Daubechies (1992) et Mallat (2008).

De cette définition, on déduit que pour tout $j \in \mathbb{Z}$, la famille de fonctions $\{\phi_{jk}\}_{k\in\mathbb{Z}}$, où ϕ_{jk} est définie par $\phi_{jk} := 2^{j/2}\phi(2^j x - k)$, forme une base orthonormée de l'espace V_j pour la norme L_2.

On peut alors définir W_j, supplémentaire orthogonal de V_j dans V_{j+1}, vérifiant l'égalité :

$$V_{j+1} = V_j \oplus W_j. \tag{3.2}$$

Par rapport à V_j, V_{j+1} est un espace de résolution plus fine car, d'après la définition (3.1), il contient les fonctions contractées de l'espace V_j. D'après la relation (3.2),

les espaces V_j et W_j peuvent être vus, respectivement, comme un espace d'approximation et de détails de V_{j+1}.

De même que pour V_j, on peut construire une base orthonormée $\{\psi_{jk}\}_{k\in\mathbb{Z}}$ de l'espace W_j, pour tout $j \in \mathbb{Z}$, en dilatant et translatant une fonction de base notée ψ. Cette fonction est appelée *ondelette mère* et on définit alors pour tout $k \in \mathbb{Z}$, les fonctions d'ondelettes dilatées et translatées de ψ par :

$$\psi_{jk} = 2^{j/2}\psi(2^j x - k).$$

En combinant la définition (3.1) et la relation (3.2), on en déduit que :

$$L^2([0,1]) = V_{j_0} \oplus \bigoplus_{j=j_0}^{\infty} W_j. \tag{3.3}$$

L'indice $j_0 \in \mathbb{Z}$, représente le premier niveau d'approximation pouvant être choisi arbitrairement.

De la relation (3.3), on déduit alors que la famille de fonctions $\{\phi_{j_0k},\ k \in \mathbb{Z}\ ;\ \psi_{jk}, j \geq j_0, k \in \mathbb{Z}\}$ forme une base orthogonale de l'espace $L^2(\mathbb{R})$ et par conséquent, que toute fonction $f \in L^2(\mathbb{R})$ peut être décomposée dans cette base. Ainsi, on obtient pour f la représentation en série suivante :

$$f(x) = \sum_{k\in\mathbb{Z}} c^*_{j_0k}\phi_{j_0k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} d^*_{jk}\psi_{jk}(x), \tag{3.4}$$

où $\{c^*_{j_0k} = \langle f, \phi_{j_0k}(x)\rangle\}_{k\in\mathbb{Z}}$ sont les coefficients d'approximation théoriques de la décomposition tandis que les $\{d^*_{jk} = \langle f, \psi_{jk}(x)\rangle\}_{j\geq j_0, k\in\mathbb{Z}}$ sont appelés coefficients d'ondelette ou de détail théoriques.

Une propriété intéressante découlant de la représentation en série (3.4) est alors donnée par la formule de conservation de l'énergie, ou identité de Parseval :

$$\|f\|^2_{L_2} = \|(\mathbf{c}^*, \mathbf{d}^*)\|^2_{\ell_2}, \tag{3.5}$$

où $(\mathbf{c}^*, \mathbf{d}^*)$ est la concaténation des vecteurs d'approximation $\mathbf{c}^*$ et de détails $\mathbf{d}^*$.

Par la suite, nous nous restreindrons à l'étude de fonctions à support compact, appartenant à l'espace fonctionnel $L^2([0,1])$ et nous nous placerons de ce fait dans la classe des analyses multirésolution construites sur un intervalle dont la construction est détaillée par Cohen et al. (1993).

Exemple L'exemple le plus simple de telles bases de fonctions est donné par la base de Haar (1910) dont la fonction d'échelle est définie par :

$$\phi(x) = \begin{cases} 1 & \text{si } \ 0 \leq t < 1, \\ 0 & \text{sinon,} \end{cases}$$

tandis que l'ondelette mère est définie par :

$$\psi(x) = \begin{cases} 1 & \text{si } 0 \leq t < \frac{1}{2}, \\ -1 & \text{si } \frac{1}{2} \leq t < 1, \\ 0 & \text{sinon.} \end{cases}$$

Les espaces d'approximation V_j associés à de telles fonctions contiennent les fonctions constantes par morceaux dont la résolution augmente avec j. Toute fonction de $L^2([0,1])$ est donc approchable avec une précision arbitraire, pour un niveau de résolution j suffisamment grand. Cette base d'ondelettes, bien que très simple en apparence, a historiquement permis de développer un cadre pour de nombreuses classes d'ondelettes et beaucoup de propriétés relatives aux ondelettes de Haar restent valables pour toutes les bases d'ondelettes orthogonales.

En pratique cependant, cette base d'ondelettes est très peu utilisée, et ceci est principalement dû à sa mauvaise localisation en échelle et en temps. Ainsi, excepté pour le cas de la modélisation de fonctions constantes par morceaux pour lesquelles la base de Haar est bien adaptée, on préférera utiliser d'autres bases d'ondelettes comme celles de Daubechies dont les fonctions sont orthogonales et à support compact (Daubechies 1992). En Figure 3.1, nous avons représenté des exemples de décomposition du signal `Heavisine` (Donoho et Johnstone 1994) dans deux bases d'ondelettes différentes (Haar et Daubechies à 8 moments nuls), ainsi que les ondelettes mères associées à ces deux bases. On observe que la compression de l'information dépend de la base d'ondelettes utilisée et que la présence de coefficients plus importants est liée aux discontinuités du signal décomposé.

3.2.2 Espaces de Besov

La notion de régularité est au centre de la modélisation fonctionnelle. De par leur construction consistant à successivement affiner la résolution avec laquelle une fonction est observée, on peut modéliser de façon efficace des signaux montrant de fortes irrégularités. De manière à pouvoir en tenir compte dans le cadre d'une modélisation par ondelettes, nous allons à présent introduire la notion d'espaces de Besov noté $B^s_{pq}([0,1])$. Ces espaces fonctionnels permettent de définir très finement la régularité s d'une fonction de $L^p([0,1])$, espace des fonctions p-fois intégrables tout en apportant une correction q à cette régularité. Pour une étude détaillée des espaces de Besov et de leurs propriétés, nous nous référons aux ouvrages de Härdle et al. (1998) et DeVore et Lorentz (1993).

Définition 3.2. *Soient $0 < s < \infty$, $1 \leq p \leq \infty$ et $1 \leq q < \infty$. Pour tout $(x,h) \in \mathbb{R}^2$, on pose $\tau_h f(x) = f(x-h)$. L'espace de Besov $B^s_{pq}(\mathbb{R})$ de paramètre s, p et q est constitué de l'ensemble des fonctions f vérifiant :*

(i) $f \in L^p(\mathbb{R})$,

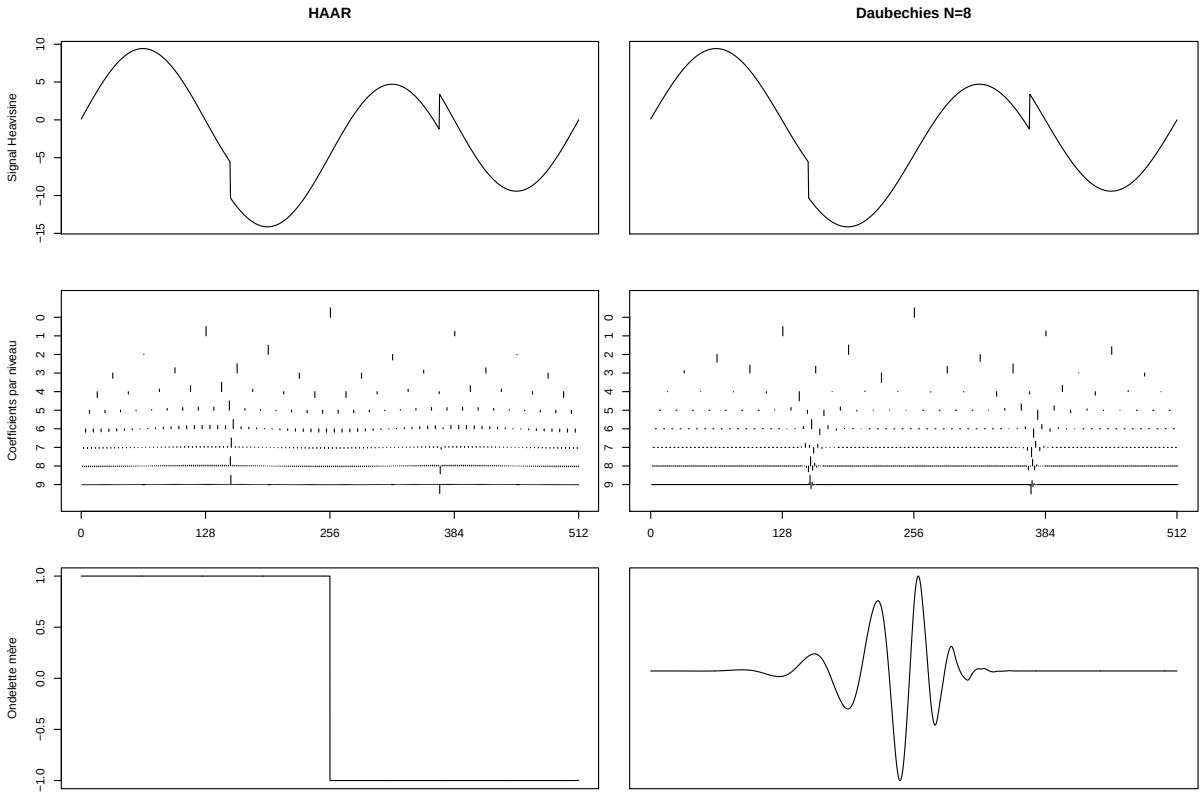


FIGURE 3.1 – *Exemple de décomposition en ondelettes du signal* `Heavisine` *(Donoho et Johnstone 1994). Le signal est représenté sur la première ligne, tandis que sur la deuxième ligne sont représentés les coefficients de détails obtenus avec une base d'ondelettes de Haar ou de Daubechies à 8 moments nuls. La dernière ligne correspond aux représentations des deux ondelettes mère associées respectivement à ces deux bases.*

(ii) $\left(\int_{[0,1]} \left(\frac{\|\tau_h f - f\|_{L^p}}{h^s}\right)^q \frac{dh}{h}\right)^{1/q} < \infty.$

En fait, l'atout majeur des espaces de Besov réside dans leur connexion avec les bases d'ondelettes : en effet, on peut relier l'appartenance d'une fonction à un espace de Besov particulier (c'est-à-dire la régularité de la fonction considérée) à la décroissance de ses coefficients d'ondelettes. Le théorème suivant nous donne une caractérisation des espaces de Besov par la norme des coefficients d'ondelettes d'une fonction $f \in L^p(\mathbb{R})$.

Avant de l'énoncer, nous allons donner une série d'hypothèses sur la fonction d'échelle ϕ ainsi que sur sa transformée de Fourier notée $\widehat{\phi}$.

Hypothèses

1. $\sum_k |\widehat{\phi}(\xi + 2k\pi)|^2 = 1$ *presque partout,*
2. $\widehat{\phi}(\xi) = \widehat{\phi}\left(\frac{\xi}{2}\right) m_0\left(\frac{\xi}{2}\right)$ *presque partout, où* m_0 *est une fonction* 2π*-périodique,*
3. *il existe une fonction* Ψ *bornée et non strictement croissante telle que* $\int \Psi(|u|)du < \infty$, $\int \Psi(|u|)|u|^N du < \infty$ *pour un certain* $N \geq 0$ *et* $|\phi(u)| \leq \Psi(|u|)$ *presque partout,*
4. ϕ *est (N+1) faiblement différentiable et sa dérivée* $\phi^{(N+1)}$ *vérifie :*

$$\operatorname*{ess\,sup}_x \sum_k |\phi(x-k)| < \infty.$$

Le théorème s'exprime alors de la façon suivante (Härdle et al. 1998) :

Théorème 3.1. *Soit* ϕ *une fonction d'échelle suivant les Hypothèses [1-4] pour un* $N \geq 0$ *donné. Alors, pour tout* s *tel que* $0 < s < N + 1$*, pour tout* (p, q) *tels que* $1 \leq p, q \leq \infty$ *et pour toute fonction* $f \in L_p(\mathbb{R})$*, les assertions suivantes sont équivalentes :*

(i) $f \in B^s_{pq}(\mathbb{R})$*,*

(ii) $\|\mathbf{c}^*\|_{\ell_p} < \infty$ *et* $\|\mathbf{d}^*_j\|_{\ell_p} = 2^{-j(s+\frac{1}{2}+\frac{1}{p})}\epsilon_j$*,* $j \in \mathbb{N}$*, où* $\{\epsilon_j\} \in \ell_q$*.*

Cela nous donne ainsi une condition nécessaire et suffisante sur les coefficients d'ondelettes d'une fonction f pour que celle ci appartienne à un certain espace de Besov.

Enfin, notons que le point *(ii)* de la Définition (3.2) définit une norme sur l'espace de Besov $B^s_{pq}(\mathbb{R})$, notée $\|\cdot\|_{spq}$. Par la suite, nous nous restreindrons à la considération de boules de Besov de rayon unitaire, définies comme l'ensemble $\{f \in B^s_{pq}(\mathbb{R})$ telle que $\|f\|_{spq} \leq 1\}$. Afin de simplifier les notations, les boules de Besov de rayon unitaire seront notées par la suite B^s_{pq}.

3.2.3 Transformée en ondelettes rapide et approximation

Un autre point participant à la popularité des ondelettes est l'existence d'algorithmes de décomposition et de reconstruction rapides développés par Mallat (2008). Afin de donner le principe de cet algorithme, on se place à un niveau de résolution j tel que $0<j$. On suppose de plus que les coefficients d'approximation au niveau j, $\{c_{jk}\}_k$, sont connus.

Par les propriétés de l'analyse multirésolution et comme $V_1 \subset V_0$, on sait qu'il existe une suite $\{h_k\}_{k\in\mathbb{Z}}$ de filtres telle que $\phi(x) = \sum_{k\in\mathbb{Z}} h_k \phi_{1k}(x)$. D'où :

$$\begin{aligned}
\phi_{jk}(x) &= 2^{\frac{j}{2}}\phi(2^j x - k) \\
&= 2^{\frac{j}{2}} \sum_{\ell\in\mathbb{Z}} h_\ell \phi_{1\ell}(2^j x - k) \\
&= \sum_{\ell\in\mathbb{Z}} h_\ell \phi_{j+1,\ell+2k}(x) \\
&= \sum_{\ell\in\mathbb{Z}} h_{\ell-2k} \phi_{j+1,\ell}(x).
\end{aligned}$$

On en déduit alors la relation suivante sur les coefficients d'approximation :

$$c^*_{j-1,k} = \sum_{\ell\in\mathbb{Z}} h_{\ell-2k} c^*_{j\ell}. \tag{3.6}$$

De la même manière, une relation permettant d'obtenir les coefficients d'ondelettes à partir des coefficients d'approximation peut être construite. En partant de la relation $\psi(x) = \sum_{k\in\mathbb{Z}} g_k \phi_{1k}(x)$, on obtient alors :

$$d^*_{j-1,k} = \sum_{\ell\in\mathbb{Z}} g_{\ell-2k} c^*_{j\ell}. \tag{3.7}$$

On constate alors que, connaissant les coefficients d'ondelettes d'un certain niveau de résolution j, les coefficients d'approximation et d'ondelettes du niveau inférieur $j-1$ (et donc, récursivement, des niveaux inférieurs) peuvent être déterminés simplement.

Schématiquement, l'algorithme de décomposition peut être représenté de la manière suivante :

$$\begin{array}{ccccccc}
c^*_j & \longrightarrow & c^*_{j-1} & \longrightarrow & c^*_{j-2} & \longrightarrow & \cdots \\
 & \searrow & & \searrow & & \searrow & \\
 & & d^*_{j-1} & & d^*_{j-2} & & d^*_{j-3}
\end{array}$$

Un algorithme de reconstruction, que nous ne détaillerons pas, peut être dérivé de manière similaire en partant de la connaissance des coefficients d'approximation et de détail à un niveau de résolution j.

En pratique, on est en général confronté à l'étude de données discrétisées, c'est-à-dire qu'on ne dispose que d'un nombre fini d'observations de la fonction étudiée.

Dans un souci de simplification des calculs, il est courant de supposer que les observations sont faites de manière équirépartie sur $M = 2^J$ points et donc, que l'on peut disposer au plus des coefficients de niveau J : en effet, on ne pourra de toute façon pas accéder aux détails de résolution plus fine que 2^{-J}. Nous signalons que le cas pour lequel le signal étudié n'est pas de taille 2^J fait l'objet de l'article Todd Ogden (1997). L'algorithme de décomposition sur les coefficients discrétisés se présente alors sous la forme suivante :

$$\begin{pmatrix} \\ \mathbf{c}_J^* \\ \\ \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{c}_{J-1}^* \\ \hline \\ \mathbf{d}_{J-1}^* \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{c}_{J-2}^* \\ \hline \\ \mathbf{d}_{J-2}^* \\ \mathbf{d}_{J-1}^* \end{pmatrix} \rightarrow \cdots \rightarrow \begin{pmatrix} \mathbf{c}_0^* \\ \hline \mathbf{d}_0^* \\ \mathbf{d}_1^* \\ \vdots \\ \mathbf{d}_{J-2}^* \\ \mathbf{d}_{J-1}^* \end{pmatrix}.$$

À chaque étape, on calcule les vecteurs des coefficients d'approximation et de détail de niveau inférieur j. Les vecteurs résultants sont alors de taille 2^j. En concaténant les vecteurs de coefficients obtenus à chaque étape, on obtient alors un vecteur de taille $M = 2^J$ donnant la décomposition discrète du signal en coefficients d'approximation et d'ondelettes. L'algorithme ainsi décrit est un algorithme rapide nécessitant $\mathcal{O}(M)$ opérations. Cependant, l'application de cet algorithme nécessite de disposer des coefficients d'approximation au niveau le plus fin J. Ces derniers peuvent être calculés au moyen d'approximations intégrales mais ce calcul peut alors se révéler coûteux numériquement. Usuellement, les coefficients $\{c_{Jk}^*\}_k$ sont remplacés par les 2^J valeurs observées du signal, permettant ainsi de conserver la rapidité d'exécution de l'algorithme. Pour une base d'ondelettes suffisamment régulière et une grille d'observation fine, ce choix est justifié par la relation (3.8), garantissant une erreur de troncature raisonnable :

$$f\left(\frac{k}{2^J}\right) \approx 2^{J/2}\langle f, \phi_{Jk}\rangle = 2^{J/2} c_{Jk}^*, \tag{3.8}$$

dont la justification peut être trouvée dans l'ouvrage de Daubechies (1992).

Cette approximation, suivie d'une décomposition par l'algorithme décrit ci-dessus, conduit à l'obtention des coefficients d'approximations et d'ondelettes dits *empiriques* que l'on différenciera des coefficients théoriques en utilisant par la suite les notations non étoilées $\mathbf{c}$ et $\mathbf{d}$. En pratique, ce sont les coefficients les plus généralement considérés et par la relation (3.8), on déduit qu'ils sont liés aux coefficients d'approximation et d'ondelette théoriques, respectivement, par les relations : $\mathbf{c}^* \approx M^{-1/2}\mathbf{c}$ et $\mathbf{d}^* \approx M^{-1/2}\mathbf{d}$. L'algorithme basé sur cette approximation porte le nom de transformée en ondelettes discrète ou d'algorithme DWT pour Discrete Wavelet Transform et a été développé par Mallat (2008).

La transformée DWT peut aussi être vue comme un produit matriciel avec une matrice de filtres $\mathbf{W}$ orthogonale. Pour une base d'ondelettes choisie (et donc une matrice de filtres donnée), la décomposition du signal $\mathbf{Y} = (Y(t_1), \ldots, Y(t_M))$ est alors donnée par :

$$\mathbf{WY} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix},$$

où $(\mathbf{c}^T, \mathbf{d}^T)^T$ est le vecteur concaténé des coefficients d'approximations et d'ondelettes empiriques de taille M.

3.2.4 Modélisation statistique par ondelettes

Revenons à présent au modèle fonctionnel initial (3.1). Pour une base d'ondelettes (ϕ, ψ) donnée, on peut décomposer chaque terme du modèle dans cette base. Ainsi, à un niveau d'approximation j_0 fixé, on a :

$$\begin{aligned} Y(t) &= \sum_{k=0}^{2^{j_0}-1} c^*_{j_0k}\phi_{j_0k}(t) + \sum_{j\geq j_0}\sum_{k=0}^{2^j-1} d^*_{jk}\psi_{jk}(t), \\ \mu(t) &= \sum_{k=0}^{2^{j_0}-1} \alpha^*_{j_0k}\phi_{j_0k}(t) + \sum_{j\geq j_0}\sum_{k=0}^{2^j-1} \beta^*_{jk}\psi_{jk}(t). \end{aligned}$$

Par identification, et sachant que la transformée en ondelettes d'un bruit blanc est un bruit blanc, on peut alors donner la modélisation suivante des coefficients d'approximation et d'ondelettes théoriques pour tout $j \geq j_0$, $k = 0, \ldots, 2^j - 1$:

$$\begin{cases} c^*_{j_0k} = \alpha^*_{j_0k} + \varepsilon^*_{j_0k}, \\ d^*_{jk} = \beta^*_{jk} + \varepsilon^*_{jk}, \end{cases}$$

où $\varepsilon^*_{i,jk} \sim \mathcal{N}(0, \sigma^2_{\varepsilon^*})$ pour tout $j \geq j_0, k = 0, \ldots, 2^j - 1$.

Dans le cadre de l'étude de signaux observés de manière discrète sur $M = 2^J$ points, il est courant d'utiliser la transformée en ondelettes rapide DWT. Ainsi, la décomposition du modèle (3.1) par l'algorithme DWT est donné par la relation :

$$\mathbf{WY} = \mathbf{W}\boldsymbol{\mu} + \mathbf{WE},$$

avec $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_M))$ et $\mathbf{E} = (E(t_1), \ldots, E(t_M))$
On obtient alors une représentation du modèle par les coefficients d'ondelettes empiriques :

$$\begin{cases} c_{j_0k} = \alpha_{j_0k} + \varepsilon_{j_0k}, & \forall j \geq j_0, k = 0, \ldots, 2^j - 1, \\ d_{jk} = \beta_{jk} + \varepsilon_{jk}, \end{cases} \tag{3.9}$$

avec $\varepsilon_{jk} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ pour tout $j = j_0, \ldots, J-1$, et $k = 0, \ldots, 2^j - 1$. Par la suite, l'ensemble $\{(j,k) \mid j = j_0, \ldots, J-1, \text{ et } k = 0, \ldots, 2^j - 1\}$ sera noté Λ.

À ce stade, le modèle fonctionnel originel (3.1) a été transformé en un modèle linéaire sur les coefficients empiriques de la décomposition. L'enjeu est alors de trouver une estimation des coefficients d'ondelettes $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ empiriques associés à la fonction de régression μ, observée de manière bruitée, afin d'obtenir un estimateur de celle-ci. Cette problématique fait l'objet de la prochaine section traitant des méthodes de régression non paramétrique basées sur les ondelettes.

3.3 Seuillage et régressions pénalisées

De par leur construction basée sur la notion d'analyse multirésolution, les ondelettes ont la propriété de concentrer ou compresser l'information concernant le signal d'intérêt et on s'attend ainsi à retrouver l'information pertinente sur un petit nombre de coefficients. Inversement, étant donné que la représentation en ondelettes d'un bruit blanc est encore un bruit blanc, l'erreur de mesure est quant à elle répartie uniformément sur tous les coefficients de la décomposition. Une façon d'estimer la fonction μ est alors de retrouver les coefficients contenant l'information sur celle-ci au milieu du bruit de mesure. Ces approches sont regroupées sous le nom de *méthodes de seuillage* et représentent les méthodes privilégiées de la régression non-paramétrique par ondelettes. Ces stratégies seront au cœur des procédures développées au cours de la Partie III de ce manuscrit.

Dans cette section, nous définissons plus formellement l'idée fondatrice du seuillage, développée par Donoho et Johnstone (1994), et basée sur la notion d'adaptativité spatiale. Ensuite, nous introduisons trois méthodes de seuillage parmi les plus populaires qui seront utilisées dans la suite du manuscrit, ainsi que les parallèles existant entre ces stratégies et les méthodes de régressions pénalisées.

3.3.1 Seuillage par ondelettes et risque

Nous reprenons à présent le modèle fonctionnel (3.1)

$$Y_i(t_m) = \mu(t_m) + E_i(t_m), \quad \text{avec} \quad E_i(t_m) \sim \mathcal{N}(0, \sigma_E^2),$$

en supposant, dans le contexte de régression non paramétrique, que la fonction μ appartient à un espace de Besov $B^s_{p,q}[0,1]$.

Ce modèle conduit au modèle (3.9) sur les coefficients d'approximation et d'ondelettes empiriques :

$$\begin{cases} c^i_{j_0k} = \alpha_{j_0k} + \varepsilon^i_{j_0k}, & \forall j \geq j_0, k = 0, \ldots, 2^j - 1, \\ d^i_{jk} = \beta_{jk} + \varepsilon^i_{jk}. \end{cases}$$

Dans ce modèle, notons que la matrice de plan d'expérience implicite est la matrice identité.

Espaces de Besov et risque minimax

La notion de seuillage a initialement été introduite par Donoho et Johnstone (1994). L'idée de départ est basée sur la notion d'adaptativité spatiale : en effet, la classe des espaces de Besov permet de considérer des fonctions possédant une large variété d'irrégularités. Ainsi, par exemple, l'espace *Bump Algebra* décrit par Meyer (1990) est constitué des fonctions se décomposant en une somme de sauts gaussiens dont la hauteur est normalisée à 1 et s'exprimant par :

$$f(x) = \sum_i \alpha_i g_{(t'_i,t_i)}(x), \quad \text{avec } g_{(t',t)}(x) = \exp\left(\frac{-(x-t)^2}{2t'^2}\right)$$

On voit que cet espace peut contenir des fonctions très irrégulières (avec des pics comme la fonction `Bumps` par exemple) et spatialement très inhomogènes (composées de parties avec des pics, opposées à des parties très lisses). Meyer (1990) démontre que cet espace est exactement l'espace de Besov $B^1_{1,1}$.

Cet exemple introductif montre bien la nécessité de développer des méthodes d'estimation fonctionnelle spatialement adaptatives, c'est-à-dire, capables de s'adapter aux spécificités des fonctions étudiées. Sur ce type de fonctions, les estimateurs linéaires ne sont pas adaptés. En effet, ces estimateurs, comme les estimateurs à noyaux à bandes fixes ou basés sur une transformée de Fourier à fenêtre fixe, reposent implicitement sur une notion d'homogénéité spatiale. Or dans ce type de cas, ils tendent à surlisser les parties irrégulières et, au contraire, à ajouter des irrégularités sur les parties initialement lisses.

Des stratégies ont été développées pour remédier à cela et parmi elles, on peut citer les approches de type CART (Breiman et al. 1984)) ou encore les estimateurs à noyaux à bandes variables (Brockmann et al. 1993). Malgré leur caractère adaptatif et non-linéaire, ces méthodes souffrent d'un manque de résultats concernant leurs performances théoriques. La volonté est alors de développer des estimateurs dont le comportement en terme de reconstruction est optimal et contrôlé.

Afin de mesurer ces performances, on peut définir le risque minimax L_2 pour les estimateurs $\widehat{\mu}$ d'une fonction $\mu \in B^s_{pq}$ par :

$$\mathcal{R}(B^s_{pq}) = \inf_{\widehat{\mu}} \sup_{\mu \in B^s_{pq}} \mathbb{E}\|\widehat{\mu} - \mu\|^2_{L^2}. \tag{3.10}$$

Ce risque représente le risque quadratique moyen atteint par le meilleur estimateur dans "la pire des situations" au sein de l'espace $B^s_{pq}[0,1]$. Notons que le risque basé sur une norme L_2 est le plus couramment utilisé mais cette définition peut être étendue aux risques en norme L_r, avec $1 \leq r < \infty$. À partir de cette idée, une classification des taux de convergence atteignables au moyen d'estimateurs linéaires ou non-linéaires peut être établie en fonction du paramètre p associé à l'espace de Besov considéré et de la norme L_r dans laquelle le risque est mesuré. La classification établie par Härdle et al. (1998) est résumée Figure 3.2, où trois zones distinctes possédant des caractéristiques différentes sont représentées :

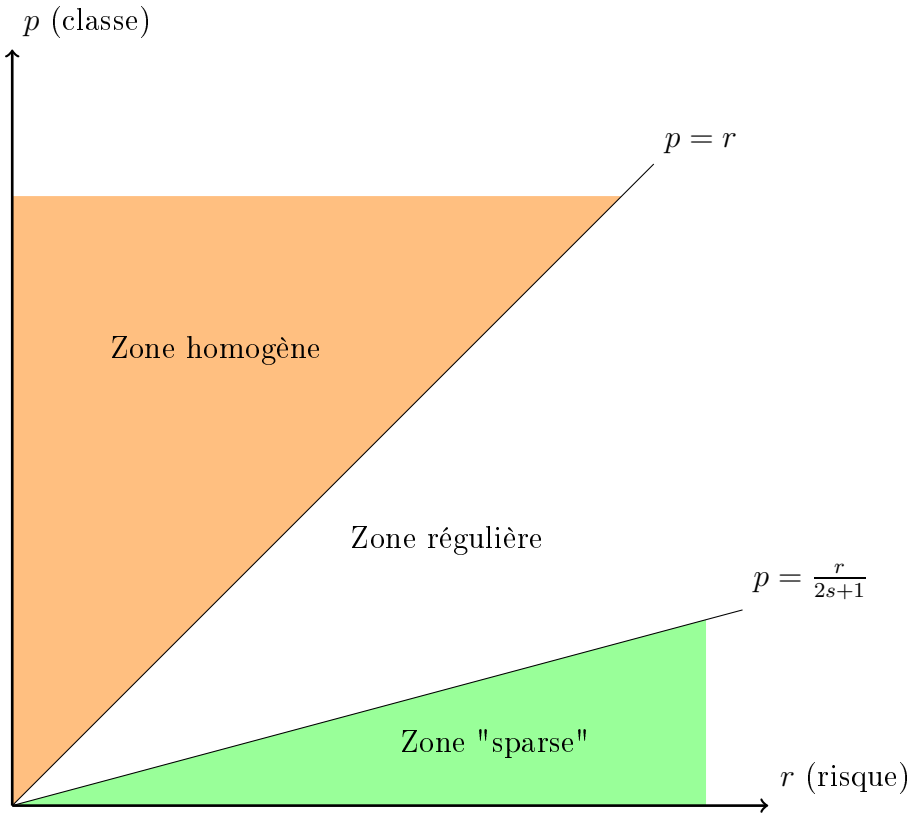


FIGURE 3.2 – *Classification des taux de convergence optimaux pour les estimateurs linéaires et non-linéaires (Härdle et al. 1998).*

- la zone homogène : le taux de convergence optimal est en $\mathcal{O}(M^{-\frac{2s}{2s+1}})$ et ce taux peut être atteint par les estimateurs linéaires.
- la zone régulière : le taux de convergence optimal est en $\mathcal{O}(M^{-\frac{2s}{2s+1}})$ mais ce taux ne peut pas être atteint par les estimateurs linéaires.
- la zone "sparse" : le taux de convergence optimal est en $\mathcal{O}(M^{-\frac{2s'}{2s'+1}})$ avec $s' = s - 1/p - 1/r$ (la convergence est donc plus lente et dépend des valeurs de p et r) et ce taux ne peut pas être atteint par les estimateurs linéaires.

Dans le cas où on s'intéresse au risque quadratique moyen ($r = 2$), la zone homogène recouvre les espaces B^s_{pq} avec $p \geq 2$ et la zone régulière, les espaces tels que $p > \frac{2}{2s+1}$, incluant notamment l'espace $B^1_{1,1}$.

Adaptativité spatiale des ondelettes

Les ondelettes, de par leur bonne localisation temporelle et fréquentielle, répondent naturellement à cette notion d'adaptativité spatiale. Donoho et Johnstone (1994) s'intéressent dans leur article fondateur à l'idée d'adaptation spatiale idéale et plus particulièrement, d'oracle spatial. Pour un estimateur spatialement variable, un oracle spatial nous dit comment cet estimateur peut être adapté au mieux au vu de la vraie fonction. Donoho et Johnstone (1994) démontrent alors que, si l'on dispose d'un oracle spatial pour une fonction μ donnée, alors, on peut aisément construire un estimateur consistant de celle-ci atteignant une vitesse de convergence paramétrique optimale, c'est-à-dire en $\mathcal{O}(1/M)$. Dans le cas d'estimateurs basés sur une décomposition en ondelettes, disposer d'un oracle spatial revient simplement à sélectionner un ensemble de positions $(j,k) \in \Lambda$ sur lesquelles l'information fonctionnelle est concentrée.

Donoho et Johnstone (1994) se basent sur deux propriétés essentielles des ondelettes : d'une part, lorsque l'on s'intéresse à la décomposition d'une fonction régulière sur une base d'ondelettes, l'information concernant la fonction est concentrée sur relativement peu de coefficients (Härdle et al. 1998). D'autre part, pour un signal observé de manière bruitée, le bruit contamine toutes les positions et toutes les échelles de manière égale, car la transformée en ondelettes d'un bruit blanc est aussi un bruit blanc.

À ce stade, la définition d'une "bonne" procédure de seuillage doit vérifier certaines règles. Dans leurs travaux, Fan et Li (2001) définissent un cadre en proposant trois points fondamentaux à vérifier par les estimateurs résultant d'une procédure de seuillage. Ils doivent être :

1. non biaisés,
2. parcimonieux,
3. continus par rapport aux observations $(\mathbf{c}, \mathbf{d})$ afin d'éviter les problèmes d'instabilité.

Plus précisément, les estimateurs de seuillage ne doivent pas introduire de biais, tout en proposant une sélection des coefficients dans un souci de réduction de dimension.

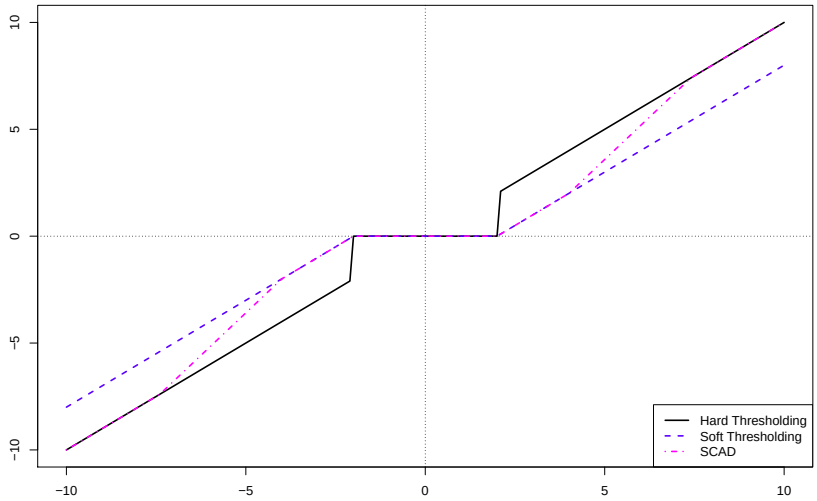


FIGURE 3.3 – *Représentation des seuillages dur et doux développés par Donoho et Johnstone (1994) et du seuillage SCAD développé par Antoniadis et Fan (2001). En abscisse les observations et en ordonnée, les estimations obtenues suivant le seuillage considéré. Les paramètres de régularisation sont fixés à* $\lambda = 2$ *et* $a = 3.7$.

De plus, la continuité des estimateurs est requise afin de garantir une faible sensibilité du modèle aux données de départ.

Nous détaillons à présent trois procédures de seuillages parmi les plus populaires : les seuillages durs et doux de Donoho et Johnstone (1994) ainsi que le seuillage SCAD de Antoniadis et Fan (2001) et Fan et Li (2001). Les fonctions représentatives de ces trois procédures sont représentées en Figure 3.3.

Seuillage dur

Le seuillage dur, plus communément nommé *hard thresholding*, est une règle dite de "keep or kill" appliquée sur les coefficients théoriques de la décomposition et définie par :

$$\delta^{\mathrm{H}}(\mathbf{d}, \lambda) = d_{jk}\mathbf{1}_{\{|d_{jk}|>\lambda\}}, \quad \forall (j,k) \in \Lambda_\delta, \tag{3.11}$$

où λ est un paramètre de régularisation, appelé *seuil*, à déterminer et Λ_δ l'ensemble des positions pour lesquelles le seuillage est appliqué.

Au cours de ce seuillage les coefficients sont soit mis à zéro, soit laissés inchangés s'ils dépassent un certain seuil. Cette procédure, comme on peut le constater en Figure 3.3 est une fonction discontinue des données et, de ce fait, ne respecte donc pas le cadre fixé par Fan et Li (2001). Cela entraîne en effet une instabilité dans

l'estimation des paramètres, c'est-à-dire que pour de faibles variations des données de départ, cette discontinuité peut entraîner de fortes variations sur le modèle estimé.

Seuillage doux

Un autre type de seuillage proposé par Donoho et Johnstone (1994) est le seuillage doux, ou *soft thresholding*, qui est une règle dite de "shrink or kill". Sur les coefficients d'ondelettes, le seuillage est défini par :

$$\delta^{\mathrm{S}}(\mathbf{d}, \lambda) = \mathrm{sign}(d_{jk}) \left(|d_{jk}| - \lambda\right)_+, \quad \forall (j,k) \in \Lambda_\delta. \tag{3.12}$$

On observe ici que les coefficients sont tous réduits d'une amplitude λ et que certains sont, de ce fait, réduits à zéro. Cette procédure, représentée en Figure 3.3, est une procédure continue mais conduit à introduire un biais systématique sur les grands coefficients. Elle ne respecte donc pas non plus dans le cadre fixé par Fan et Li (2001).

Seuillage SCAD

Afin de répondre au cadre fixé pour l'obtention d'une "bonne" procédure, Antoniadis et Fan (2001) ont développé le seuillage SCAD (Smoothly Clipped Absolute Deviation) dans un cadre fonctionnel dans le but de réaliser un compromis entre les seuillages doux et dur. Le seuillage SCAD est défini, pour tout $(j,k) \in \Lambda_\delta$, par :

$$\delta^{\mathrm{SCAD}}(\mathbf{d}, \lambda, a) = \begin{cases} \mathrm{sign}(d_{jk})(|d_{jk}| - \lambda)_+ & \text{si } |d_{jk}| \leq 2\lambda, \\ \frac{(a-1)d_{jk} - a\lambda \mathrm{sign}(d_{jk})}{a-2} & \text{si } 2\lambda < |d_{jk}| \leq a\lambda, \\ d_{jk} & \text{si } |d_{jk}| > a\lambda, \end{cases} \tag{3.13}$$

où $a > 2$ et λ sont des paramètres de régularisation à déterminer. Par une minimisation du risque empirique basée sur des arguments bayésiens, Fan et Li (2001) conseillent de fixer $a = 3.7$ mais ils montrent aussi que l'influence de ce paramètre est faible. Par la suite, nous ne ferons plus référence à ce paramètre dans la définition du seuillage et nous le considérerons fixé à $a = 3.7$.

Dans le cadre fonctionnel, Antoniadis et Fan (2001) démontrent que l'estimateur de seuillage SCAD converge vers la vraie fonction avec une vitesse de convergence minimax dans la classe des espaces de Besov. La Figure 3.3 illustre le fait que le seuillage SCAD se résume à un seuillage doux des données pour les faibles coefficients et à un seuillage dur des grands coefficients reliés par une zone de transition continue.

Choix du paramètre λ et propriétés de reconstruction

Le point commun des trois procédures de seuillages présentées ici est la présence d'un paramètre de seuil λ à déterminer. Le choix de ce paramètre est au cœur de la problématique de seuillage car il contrôle le degré de seuillage. De nombreuses

méthodes de choix du seuil ont été développées et leurs différences résident principalement dans les propriétés attendues sur les estimateurs finaux. Dans la plupart des cas, leur construction vise à avoir de bonnes propriétés de reconstruction fonctionnelle et à atteindre une vitesse de convergence optimale. Cependant, les seuils ainsi construits dépendent généralement de la régularité de la vraie fonction qui n'est pas accessible. Ce paramètre est donc difficile à déterminer en pratique. Une discussion à ce sujet est proposée par Donoho et al. (1995).

Un seuil largement utilisé est le seuil universel de Donoho et Johnstone (1994) défini par :

$$\lambda = \widehat{\sigma}_\varepsilon \sqrt{2 \log M}. \tag{3.14}$$

Cette stratégie de seuillage est très populaire, surtout pour sa simplicité de mise en œuvre (c'est notamment le seuil utilisé par défaut dans la majorité des logiciels). Ce seuil vient de la propriété suivante : on peut démontrer que pour $Z_1, \ldots, Z_n$, variables aléatoires i.i.d suivant une loi normale centrée réduite, alors

$$\mathbb{P}\left\{ \max_{1 \leq j \leq n} |Z_j| > \sqrt{2 \log n} \right\} \sim \frac{1}{\sqrt{\pi \log n}}, \quad \text{quand } n \to \infty.$$

Le point restant à traiter est l'estimation de l'écart-type $\widehat{\sigma}_\varepsilon$ des observations pour calculer la valeur du seuil. L'idée proposée par Donoho et Johnstone (1994) est d'estimer la variance dans le domaine des ondelettes à partir du niveau de résolution le plus fin : on peut en effet raisonnablement penser que la plupart des coefficients à ce niveau de résolution seront constitués en majeure partie de bruit. Cependant, même à cette résolution, les coefficients contiennent également une part de signal, les auteurs proposent alors d'utiliser un estimateur robuste classique de la variance σ_ε basé sur le MAD (Median Absolute Deviation) des coefficients d'ondelettes au niveau de résolution le plus fin :

$$\widehat{\sigma}_\varepsilon^2 = \frac{\widehat{\sigma}_{\text{MAD}}^2}{0.6745}. \tag{3.15}$$

On peut montrer qu'en utilisant ce seuil, les procédures de seuillages dur et doux conduisent à une vitesse de convergence *near-minimax* dans la classe des espaces de Besov pour l'estimateur fonctionnel reconstruit $\widehat{\mu}$. Cela signifie que l'estimateur converge vers la vraie fonction avec un taux optimal à un facteur logarithmique près, soit une convergence en $\mathcal{O}((\log M/M)^{\frac{2s}{2s+1}})$. La preuve de ce résultat peut être trouvée dans Donoho et al. (1995). Un résultat équivalent a été démontré pour la procédure de seuillage SCAD (Antoniadis et Fan 2001).

3.3.2 Lien avec les régressions pénalisées et propriété oracle

Dans cette section, nous nous attachons à décrire les liens existants entre les techniques de seuillage et celles des régressions pénalisées dans un cadre fonctionnel. Nous commençons par introduire le principe des régressions pénalisées et, en

particulier, la régression LASSO (Least Absolute Shrinkage and Selection Operator), méthode populaire d'estimation et de sélection de variables. Nous décrivons ensuite les principales équivalences entre seuillages et régressions pénalisées et nous détaillons alors leurs propriétés respectives du point de vue de la sélection de variables.

Régression pénalisée

Le principe des régressions pénalisées se place dans la large classe des méthodes de sélection de modèle. Dans un contexte scientifique où le développement de nouveaux appareils de mesure à haut-débit permet l'acquisition de données de plus en plus conséquentes, les techniques de sélection classiques, basées sur des procédures pas-à-pas, atteignent leurs limites. En effet, face à de grands nombres de données, les questions de stabilité et de précision des estimateurs, ainsi que la question du choix du bon modèle, deviennent critiques. Cela met en avant la nécessité de développer des approches d'estimation plus globales. C'est principalement dans ce sens qu'ont été développées les approches que l'on peut regrouper sous le nom de régressions pénalisées. Pour illustrer cette notion, nous nous restreindrons par la suite à un cadre fonctionnel qui constitue le contexte de ces travaux. Cependant, signalons que les régressions pénalisées ont été développées dans un cadre de régression plus général et nous invitons le lecteur à se référer aux ouvrages de Hastie, Tibshirani, et Friedman (2009) et Buhlmann et van de Geer (2011) pour plus de détails à ce sujet.

Considérons le modèle de régression sur les coefficients d'ondelettes (3.9) et le problème des moindres carrés classiques :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ \|\mathbf{d} - \boldsymbol{\beta}\|_2^2, \tag{3.16}$$

où $\|\cdot\|_2$ est la norme euclidienne canonique. L'idée des régressions pénalisées est alors d'ajouter une contrainte au problème des moindres carrés classique permettant d'assurer à l'estimateur $\widehat{\boldsymbol{\beta}}$ d'atteindre les propriétés désirées. Ainsi, le critère d'optimisation par moindres carrées pénalisés s'exprime comme :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ \|\mathbf{d} - \boldsymbol{\beta}\|_2^2 + \mathrm{pen}(\boldsymbol{\beta}, \lambda), \tag{3.17}$$

où $\mathrm{pen}(\cdot,\cdot)$ est une fonction de pénalité appliquée au paramètre $\boldsymbol{\beta}$ et dépendant d'un paramètre de régularisation λ. Cette pénalité peut prendre de nombreuses formes mais elle est généralement basée sur une norme ℓ_p des coefficients. Suivant la valeur choisie pour p, l'estimateur pénalisé possède différentes propriétés. Ainsi, pour $p > 1$, les estimateurs construits se situent dans la classe des estimateurs bridges, incluant le cas de la régression ridge pour $p = 2$, dont l'objectif est de régulariser les solutions des moindres carrés. Pour $p \leq 1$, on obtient des méthodes permettant de faire une sélection de variables en forçant certains coefficients à être mis à zéro.

Cette propriété de sélection est due à la présence d'une singularité à l'origine de la fonction de pénalité.

Les performances de sélection des estimateurs pénalisés sont alors évaluées au vu de leur propriétés oraculaires. Notons $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, le vecteur de paramètre tel que $\boldsymbol{\beta}_1$ contient les composantes non nulles du vecteur $\boldsymbol{\beta}$ et $\boldsymbol{\beta}_2$ les composantes nulles. Plus particulièrement, un estimateur est dit posséder la propriété d'*oracle* s'il vérifie asymptotiquement par rapport à N, c'est-à-dire pour un nombre d'individus tendant vers l'infini, les deux propriétés suivantes (Fan et Li 2001) :

1. Identification du bon sous-modèle : $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$.
2. $\sqrt{M}(\mathcal{I}_1(\boldsymbol{\beta}_1) - \mathcal{H}) \left[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 + (\mathcal{I}_1(\boldsymbol{\beta}_1) - \mathcal{H})^{-1}\mathcal{G}\right] \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\boldsymbol{\beta}_1)\Sigma)$,
   où $\mathcal{I}_1(\boldsymbol{\beta}_1)$ est la matrice d'information de Fisher connaissant le bon sous-modèle, $\mathcal{G}$ et $\mathcal{H}$ sont deux quantités associées au gradient et à la dérivée seconde de la pénalité considérée. Ces deux quantités seront détaillées en Section 8.2.3.

Un cas particulier : le LASSO

Un cas particulier de régression pénalisée est la régression LASSO (Least Absolute Shrinkage and Selection Operator) développée initialement par Tibshirani (1996). Notons qu'une approche équivalente, connue sous le nom de *Basis Pursuit Denoising*, a été développée en théorie du signal. Pour plus de détails à ce sujet, le lecteur pourra consulter les travaux et discussions de Chen et Donoho (1995) et Mallat (2008). Cette approche est un cas particulier des régressions pénalisées et met en jeu une pénalité basée sur la norme ℓ_1 des paramètres qui s'exprime comme suit :

$$\text{pen}_{LASSO}(\boldsymbol{\beta}, \lambda) = \lambda \sum_{(j,k)\in\Lambda} |\beta_{jk}|. \tag{3.18}$$

Elle représente une technique de sélection de variables à proprement parler puisqu'elle permet d'obtenir des estimateurs parcimonieux, où certains coefficients sont mis exactement à zéro grâce à la présence d'une singularité à l'origine dans l'expression de la pénalité.

La résolution de ce problème d'optimisation est, de plus, facilitée par l'existence d'un algorithme rapide, le LARS (Least-Angle Regression) développé par Efron et al. (2004). Cet algorithme fournit l'ensemble des solutions du LASSO en fonction des valeurs de λ en utilisant des résultats d'optimisation convexe. L'existence de cet algorithme a grandement participé à la popularité de l'estimateur LASSO.

Dans un contexte général, Fan et Li (2001) démontrent que la régression LASSO ne peut pas être optimale en terme de sélection de variables et de risque simultanément. En effet, sous certaines conditions concernant le modèle sous-jacent (vérifiées dans le cadre fonctionnel) la procédure LASSO est consistante en sélection de variables (Donoho et Huo 2002; Meinhausen et Buhlmann 2004). Cependant, la procédure LASSO produit par ailleurs des estimateurs biaisés pour les grands coefficients, conduisant à une sous-optimalité en terme de risque. Inversement, le choix

d'un paramètre λ optimal en terme de risque conduit à une sélection de variables non consistante (Meinhausen et Buhlmann 2004).

La procédure LASSO reste néanmoins une procédure populaire et de nombreux travaux ont été développés afin de donner des conditions permettant de construire un cadre dans lequel le LASSO possède de bonnes propriétés. Nous ne développons pas explicitement ce point dans ce manuscrit mais nous donnons en référence les travaux de Knight et Fu (2000), Meinhausen et Buhlmann (2006), Zhao et Yu (2006).

Version relaxée du LASSO Meinshausen (2007) propose une procédure LASSO modifiée, appelée LASSO relaxé, permettant de réaliser le compromis d'optimalité en terme de sélection de modèle et de consistance de l'estimateur résultant, de manière simultanée. Cette procédure se déroule en deux étapes dont l'idée sous-jacente est de séparer les deux objectifs principaux de la procédure LASSO, à savoir, la sélection de variables d'une part et l'estimation d'autre part. Une première étape consiste à résoudre le problème d'optimisation (3.17) en retenant uniquement l'information des variables sélectionnées par la procédure tandis que l'estimation des paramètres est réalisée durant une deuxième étape où l'on se place dans le modèle sélectionné au cours de la première étape. Meinshausen (2007) montre qu'en contrôlant ces deux effets séparément, de manière optimale, l'estimateur obtenu possède bien les propriétés attendues. De plus, cette procédure reste de complexité algorithmique raisonnable puisqu'elle est équivalente à celle du LASSO.

Équivalence avec le seuillage

Dans un contexte fonctionnel, un point important est qu'un grand nombre de procédures de seuillage peuvent être reliées aux méthodes de régressions pénalisées par l'intermédiaire d'un choix adapté de fonctions de pénalité. Cela ouvre ainsi la voie à l'exploration d'autres propriétés théoriques concernant les estimateurs issus de seuillages. Dans cette section, nous nous concentrerons principalement sur les propriétés oraculaires des estimateurs de seuillage.

On peut montrer que dans le cadre fonctionnel, c'est-à-dire lorsque le design considéré est orthogonal, l'estimateur obtenu par seuillage doux (3.12) est solution du problème d'optimisation (3.17) combiné à une pénalité de type LASSO (3.18). Cette équivalence a été mise en avant par Tibshirani (1996) dans son article fondateur.

De même, on peut montrer que le seuillage dur (3.11) peut s'exprimer sous la forme d'un problème de régression pénalisée (Antoniadis et al. (1997)). La pénalité est alors définie comme :

$$\text{pen}_{HARD}\left(\boldsymbol{\beta}, \lambda\right) = \lambda^2 - \sum_{(j,k)\in\Lambda} \left(|\beta_{jk}| - \lambda\right)^2 \mathbf{1}_{\{|\beta_{jk}|<\lambda\}}.$$

Comme mentionné précédemment, cette pénalité conduit dans le cadre orthogonal

à une procédure de seuillage discontinue, posant alors des problèmes de stabilité des estimateurs.

De même que pour les seuillages doux et dur, le seuillage SCAD peut être relié à un problème de régression pénalisée en prenant une pénalité de la forme :

$$\text{pen}_{SCAD}(\boldsymbol{\beta}, \lambda, a) = \begin{cases} \lambda|\beta_{jk}| & \text{si } |\beta_{jk}| \leq \lambda, \\ \frac{1}{2(a-1)}\left(|\beta_{jk}|^2 - 2a\lambda|\beta_{jk}| + \lambda^2\right) & \text{si } \lambda < |\beta_{jk}| \leq a\lambda, \\ \frac{1}{2}(a+1)\lambda^2 & \text{si } |\beta_{jk}| > a\lambda. \end{cases} \quad (3.19)$$

Partant de cette pénalité, Fan et Li (2001) démontrent, dans un cadre de régression non fonctionnel, que l'estimateur SCAD (3.13) possède bien la propriété d'oracle. Ce résultat se place dans un cadre asymptotique classique où le nombre d'individus N tend vers l'infini tandis que le nombre de variables M est fixé.

Le cadre d'un nombre de paramètres M fixé peut néanmoins se révéler restrictif à l'heure actuelle où le praticien dispose de plus en plus de variables d'intérêt accessibles, tout particulièrement dans un contexte de données fonctionnelles, mesurées à haut débit. Dans ce contexte, Fan et Peng (2004) démontrent que l'estimateur SCAD possède aussi la propriété d'oracle dans un cadre de double asymptotique, c'est à dire lorsque M et N tendent vers l'infini, avec $M < N$ et en contrôlant le ratio entre M et N en imposant $M^5/N \to 0$. Ce résultat est démontré sous certaines hypothèses propres au cadre de double asymptotique et nous y reviendrons de manière détaillée au cours du Chapitre 8.

Au cours de ce chapitre, nous avons introduit brièvement les bases d'ondelettes et la notion de modélisation fonctionnelle basée sur ces dernières qui constitue l'ingrédient commun à toutes les modélisations développées dans ce travail. Le modèle fonctionnel présenté ici est étendu au cours du prochain chapitre à la notion de modèle de classification fonctionnelle et à la notion de modèle mixte fonctionnel. Dans une deuxième partie, nous avons introduit les principales techniques de seuillage, qui constitueront la base des stratégies développées dans la Partie III de ce manuscrit, ainsi que leur équivalence aux méthodes pénalisées et les problématiques associées en terme de risque des estimateurs et de sélection de variables. Tandis que les méthodes de seuillage par ondelettes sont destinées, par définition, aux problématiques de régression fonctionnelle, les méthodes de régressions pénalisées ont été développées dans un cadre paramétrique. Les propriétés théoriques attendues sont de ce fait différentes : traditionnellement, la qualité d'un seuillage est évalué par rapport au risque sur l'estimateur fonctionnel tandis que les régressions pénalisées le sont sur les propriétés oraculaires, à savoir la capacité à estimer de manière consistante dans le bon sous-modèle. Comme décrit dans ce chapitre, ces deux approches peuvent être mises en correspondance dans un cadre ondelettes en travaillant sur les coefficients empiriques de la décomposition en ondelettes. Ceci est permis grâce aux bonnes propriétés des ondelettes : d'une part, la linéarité de la transformation permettant

un bon comportement sous des hypothèses gaussiennes et d'autre part, la propriété de conservation de l'énergie (3.5).

Chapitre 4

Modèles à variables latentes

Dans ce chapitre, nous nous intéressons à l'étude de modèles à variables latentes, c'est-à-dire à des modèles présentant des variables non-observées. Nous verrons quelles sont alors, dans un cadre d'estimation, les problématiques engendrées par la présence de variables latentes et nous présenterons l'algorithme EM (Expectation-Maximisation), algorithme itératif adapté à l'estimation par maximum de vraisemblance dans ce type de modèle. Dans un cadre fonctionnel, nous nous concentrerons ensuite sur deux modèles à variables latentes distincts : un modèle de classification non-supervisée de courbes sans effet aléatoire, où les variables latentes sont représentées par les classes des individus, ainsi qu'un modèle mixte fonctionnel en présence d'un groupe homogène d'individus dans le cadre duquel les variables non-observées sont les effets aléatoires individuels. Le modèle de classification de courbes dans un contexte mixte, développé dans la Partie II de ce manuscrit, est une association de ces deux modèles à variables latentes particuliers. Étant donné que chacun amène des problématiques qui lui sont propres, nous proposons de les étudier de manière détaillée au cours de ce chapitre introductif.

4.1 Présentation générale

4.2 Estimation dans les modèles à variables latentes

Le terme de modèle à variables latentes désigne la classe des modèles où une partie des données n'est pas observée. On parle alors de la présence de variables latentes ou de données cachées. La principale difficulté rencontrée pour l'étude de tels modèles concerne l'estimation des paramètres par maximum de vraisemblance. En effet, la présence de variables latentes entraîne une incapacité à calculer la vraisemblance du modèle de manière directe. Sous une telle approche, il est donc nécessaire de développer des méthodes d'optimisation itératives, basées sur une étape de prédiction des données non-observées. Dans un premier temps, nous décrivons le contexte général des modèles à variables latentes et nous présentons, dans un second temps,

le principal algorithme conçu pour l'estimation dans ces modèles : l'algorithme EM.

4.2.1 Contexte général

Afin d'introduire en toute généralité les modèles à variables latentes, considérons un jeu de données observées ou données incomplètes, noté $\mathbf{Y} = (Y_1, \ldots, Y_M)$. Les données observées $\mathbf{Y}$ sont supposées dépendre de variables non-observées, appelées variables latentes, que nous noterons $\mathbf{Z} = (Z_1, \ldots, Z_M)$. Le vecteur $(\mathbf{Y}, \mathbf{Z})$ est alors appelé vecteur des données complètes. On suppose que les densités associées aux différentes variables dépendent pour tout ou partie d'un groupe de paramètres noté Υ et on a alors la relation suivante sur la log-vraisemblance des données complètes :

$$\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \Upsilon) = \log \mathcal{L}(\mathbf{Y}; \Upsilon) + \log \mathcal{L}(\mathbf{Z}|\mathbf{Y}; \Upsilon), \tag{4.1}$$

où $\log \mathcal{L}(\mathbf{Z}|\mathbf{Y}; \Upsilon)$ est la log-vraisemblance des variables latentes conditionnellement aux données observées. Afin d'obtenir les estimations des paramètres du modèle par maximum de vraisemblance, notre but est de maximiser la quantité $\log \mathcal{L}(\mathbf{Y}; \Upsilon)$. Étant donné que les données $\mathbf{Y}$ dépendent de variables non-observées, l'optimisation de cette vraisemblance ne peut pas être réalisée de manière directe.

4.2.2 Algorithme EM

Principe de l'algorithme

L'algorithme EM est un algorithme itératif développé par Dempster, Laird, et Rubin (1977) dans un cadre d'estimation au sein des modèles à variables latentes. Cet algorithme permet d'estimer les paramètres par maximum de vraisemblance en utilisant une stratégie d'augmentation des données. L'idée est de se ramener à l'étude des données complètes $(\mathbf{Y}, \mathbf{Z})$ en passant par une étape de prédiction des données cachées $\mathbf{Z}$. Une revue détaillée au sujet cet algorithme se trouve dans l'ouvrage de McLachlan et Krishnan (2008).

L'idée fondamentale de l'algorithme EM repose sur la relation suivante, connue sous le nom d'identité de Fisher (1925) :

$$\frac{\partial \log \mathcal{L}(\mathbf{Y}; \Upsilon)}{\partial \Upsilon} = \mathbb{E}_{\Upsilon}\left[\frac{\partial \log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \Upsilon)}{\partial \Upsilon} \middle| \mathbf{Y} \right]. \tag{4.2}$$

On observe alors que l'optimisation de la vraisemblance des données incomplètes se ramène à l'optimisation de l'espérance des données complètes, conditionnellement aux données observées et aux paramètres du modèle.

L'estimation des paramètres du vecteur Υ par maximum de vraisemblance consiste usuellement à maximiser la quantité $\log \mathcal{L}(\mathbf{Y}; \Upsilon)$ par rapport à Υ, c'est-à-dire, à résoudre l'équation :

$$\frac{\partial \log \mathcal{L}(\mathbf{Y}; \Upsilon)}{\partial \Upsilon} = 0 \tag{4.3}$$

Par l'algorithme EM, cette optimisation est réalisée itérativement. A une itération $[h]$ donnée, supposons que nous disposons d'une solution courante à l'équation (4.3) notée $\Upsilon^{[h]}$. On est ainsi conduit à résoudre le système (4.3) à l'itération $[h+1]$ de manière équivalente en résolvant l'équation suivante par rapport aux données complètes :

$$\mathbb{E}_{\Upsilon^{[h]}}\left[\frac{\partial \log \mathcal{L}\left(\mathbf{Y},\mathbf{Z};\Upsilon\right)}{\partial \Upsilon}\middle|\,\mathbf{Y}\right] = \frac{\partial}{\partial \Upsilon}\Big[\mathbb{E}_{\Upsilon^{[h]}}\left(\log \mathcal{L}\left(\mathbf{Y},\mathbf{Z};\Upsilon\right)|\mathbf{Y}\right)\Big] = 0. \tag{4.4}$$

L'échange de l'espérance et de la dérivation au sein de l'égalité de Fisher (4.2) est rendue possible par le fait qu'on se place à la valeur courante des paramètres $\Upsilon^{[h]}$ conduisant à une espérance ne dépendant plus des paramètres. En ce sens, l'algorithme EM se place dans la classe des méthodes d'augmentation de données (Meng 2000) : l'optimisation est ici réalisée sur les données complètes (non-observées) et non plus sur les données observées seules.
Par la suite, nous adopterons la notation usuelle suivante :

$$Q(\Upsilon,\Upsilon^{[h]}) = \mathbb{E}_{\Upsilon^{[h]}}\left[\log \mathcal{L}\left(\mathbf{Y},\mathbf{Z};\Upsilon\right)|\mathbf{Y}\right]. \tag{4.5}$$

A ce stade, l'algorithme EM se décompose en deux étapes :

- Etape E (Expectation Step) : calcul de $Q(\Upsilon,\Upsilon^{[h]})$,
- Etape M (Maximization Step) : recherche de $\Upsilon^{[h+1]} = \arg\max_{\Upsilon}\ Q(\Upsilon,\Upsilon^{[h]})$.

Le calcul de l'espérance conditionnelle $Q(\Upsilon,\Upsilon^{[h]})$ peut se révéler difficile car on ne dispose pas, dans tous les cas, d'expression explicite pour cette quantité, nécessitant alors d'en proposer une approximation (cf Section 4.2.2). Pour l'ensemble des modèles développés au sein de ce manuscrit, ce calcul se ramène à la prédiction des variables manquantes, données par $\mathbb{E}_{\Upsilon^{[h]}}\left[\mathbf{Z}|\mathbf{Y};\Upsilon\right]$. Ce point est illustré en détail dans le développement de l'algorithme au sein du modèle complet présenté dans la deuxième partie (c.f. Chapitre 5, Section 5.2.2). Ces deux étapes sont répétées de manière itérative jusqu'à convergence de l'algorithme.

Critères d'arrêt

En pratique, la notion de convergence de l'algorithme est donnée en fonction d'un seuil de convergence noté ϵ préalablement fixé. L'algorithme est alors arrêté lorsqu'un critère, défini par l'utilisateur, est inférieur au seuil fixé ϵ.

Le critère d'arrêt le plus populaire est basé sur la différence des normes des paramètres entre deux étapes successives, conduisant ainsi à un critère de la forme :

$$\|\Upsilon^{[h+1]} - \Upsilon^{[h]}\|_2 < \epsilon.$$

Les estimateurs finaux sont alors fixés à la valeur courante des paramètres.

Un autre critère d'arrêt peut également être défini par rapport à la vraisemblance des données :

$$\log \mathcal{L}\left(\mathbf{Y}; \Upsilon^{[h+1]}\right) - \log \mathcal{L}\left(\mathbf{Y}; \Upsilon^{[h]}\right) < \epsilon.$$

Cependant, les critères basés sur l'augmentation de la vraisemblance doivent être considérés avec précaution car l'arrêt de l'algorithme est alors fortement déterminé par la forme générale de la vraisemblance, qui peut contenir des paliers et conduire à l'arrêt de l'algorithme sur un de ces derniers. En conséquence, on privilégie plutôt les critères d'arrêt basés sur les différences des valeurs courantes des paramètres aux étapes successives.

Propriétés théoriques

L'algorithme EM possède différentes propriétés théoriques permettant d'appréhender sa convergence. Nous n'en donnons ici qu'un résumé ; pour une étude plus approfondie, nous renvoyons le lecteur à l'ouvrage de McLachlan et Krishnan (2008, Chap. 3).

La principale propriété de l'algorithme EM, démontrée dans l'article original de Dempster, Laird, et Rubin (1977), est la propriété de monotonie, qui s'exprime comme suit :

$$\mathcal{L}\left(\mathbf{Y}; \Upsilon^{[h+1]}\right) \geq \mathcal{L}\left(\mathbf{Y}; \Upsilon^{[h]}\right) \quad \forall h \in \mathbb{N}. \tag{4.6}$$

Cette propriété est fondamentale car elle garantit à l'utilisateur la croissance de la vraisemblance des données observées à chaque itération. De ce fait, si la vraisemblance des données est bornée, la suite des valeurs $\left[\mathcal{L}(\Upsilon^{[h]})\right]_h$ converge vers une valeur $\mathcal{L}(\Upsilon_0)$.

En revanche, comme c'est le cas pour la plupart des algorithmes d'optimisation itératifs, une des principales lacunes de l'algorithme EM est que rien ne garantit que la valeur $\mathcal{L}(\Upsilon_0)$ corresponde au maximum global de la vraisemblance. La question de la convergence de l'algorithme a fait l'objet de nombreux travaux et nous nous limitons ici à en donner les principales conclusions : on peut démontrer, sous certaines conditions de régularité, que la suite des $\left[\mathcal{L}(\Upsilon^{[h]})\right]_h$ converge vers un point stationnaire de la vraisemblance et que l'ensemble de ces points stationnaires est un ensemble de dimension finie. Donc, excepté dans le cas où cet ensemble est réduit à un singleton, la convergence est garantie uniquement vers un maximum local ou un point selle. En pratique, la convergence vers le maximum global de la vraisemblance est alors fortement conditionnée par l'initialisation de l'algorithme, c'est-à-dire, par le choix des valeurs initiales des paramètres. Les différentes stratégies pouvant être mises en œuvre pour s'assurer de l'atteinte du maximum global sont détaillées dans la prochaine section.

Stratégies d'initialisation

À ce jour, la stratégie d'initialisation la plus utilisée consiste à choisir aléatoirement les valeurs initiales des paramètres. Cependant, ce type de stratégie conduit

en général à une convergence de l'algorithme vers un maximum local de la vraisemblance, mettant alors en avant la nécessité de proposer des techniques d'initialisation mieux adaptées à la recherche du maximum global. Ceci est encore plus important dans un cadre de classification non-supervisée (cf Section 4.3) car d'une part, la vraisemblance est alors multimodale et de plus, la valeur finale de cette dernière sert en général de base au calcul des critères de choix du nombre de groupes. De ce fait, l'initialisation de l'algorithme, et plus particulièrement l'initialisation des labels des individus, représente un point sensible dans ce contexte : en effet, lorsque l'on dispose des labels individuels, il est alors aisé d'en déduire des valeurs initiales des paramètres du modèle par l'utilisation d'estimateurs usuels.

Parmi les stratégies existantes, Biernacki et al. (2003) en étudient trois en particulier, dans un contexte de modèle de mélange gaussien :

Stratégie CEM (EM Classifiant) : l'initialisation des variables d'appartenance aux groupes est réalisée au moyen d'un algorithme des centres mobiles permettant ensuite d'en déduire des estimations initiales des paramètres de moyennes et variances. Lorsque les groupes se démarquent en premier lieu par leur comportement moyen, cette stratégie basée sur l'usage des centres mobiles est alors bien adaptée à l'initialisation des groupes.

Stratégie SEM (EM Stochastique) : pour cette stratégie, l'initialisation des labels individuels est réalisée aléatoirement au moyen de tirages de lois multinomiales. Cette étape, correspondant à l'étape S (Stochastique), est répétée plusieurs fois (période de chauffe) dans l'objectif de perturber les séquences de solutions de l'algorithme afin de prévenir la convergence vers des points selles et s'en écarter. Les valeurs initiales sont ensuite choisies en prenant l'initialisation ayant donné une valeur maximale de vraisemblance durant la période de chauffe.

Stratégie rEM : cette dernière approche consiste à forcer l'algorithme à effectuer un certain nombre d'itérations afin d'initialiser les paramètres. Cette stratégie est celle recommandée par Biernacki et al. (2003) qui observent, sur une vaste étude de simulation, qu'outre une simplicité de mise en oeuvre, elle possède un comportement moins sensible à la présence de bruit pour des performances comparables aux deux autres stratégies en termes de convergence et de temps de calcul.

Dans un contexte différent, lorsque l'espérance de la vraisemblance conditionnellement aux données (étape E) ne possède pas d'expression explicite, une autre stratégie appelée SAEM (Stochastic Approximation EM), développée par Delyon, Lavielle, et Moulines (1999), est aussi largement utilisée en pratique. À une itération $[h]$ donnée, l'espérance calculée au cours de l'étape E est remplacée par une moyenne pondérée d'une approximation stochastique basée sur la simulation de r_h séquences des variables non-observées et du résultat obtenu à l'itération $[h-1]$. Cette stratégie est particulièrement efficace dans le contexte de l'estimation dans les modèles mixtes (Kuhn et Lavielle 2005) et c'est notamment la stratégie implémentée au sein du logiciel MONOLIX dédié à l'analyse de modèles non-linéaires à effets mixtes.

Au cours de ce travail, nous choisissons pour notre part de proposer l'utilisation des deux stratégies SEM et rEM (implémentées au sein du package `curvclust`) : en effet, en pratique, la stratégie rEM, bien que recommandée par Biernacki et al. (2003), est aussi connue pour conduire plus régulièrement à la création de groupes vides, surtout en présence de groupes déséquilibrés, créant alors des difficultés numériques. Dans ce type de cas, l'appel à la stratégie SEM permet alors d'éviter ces difficultés.

4.3 Un modèle de classification de courbes

Dans cette section, nous nous intéressons à la description d'un modèle à variables latentes particulier : le modèle de mélange fonctionnel. Nous nous plaçons ici dans un cadre fonctionnel de classification non-supervisée. Dans un premier temps, nous introduisons les différentes approches existantes de classification non-supervisée dans un cadre classique puis leur extension à un cadre fonctionnel. Nous nous concentrons plus particulièrement sur les approches probabilistes et nous exposons alors les problématiques rencontrées dans ce cadre ainsi que les solutions proposées dans la littérature.

4.3.1 Classification non-supervisée

Nous supposons à présent que nous disposons de plusieurs signaux provenant de N individus distincts, observés de manière discrète sur M points. L'application de l'analyse fonctionnelle à plusieurs individus nous amène naturellement vers une problématique de classification et nous considérons alors que les signaux observés proviennent de plusieurs groupes caractérisés chacun par un comportement fonctionnel moyen différent. On distingue à ce stade deux types de classification : la classification supervisée où les groupes sont connus et où l'objectif est de déterminer une règle de classification optimale et la classification non supervisée pour laquelle on n'a connaissance ni du nombre de groupes recherchés, ni des classes des individus. Cette dernière représente, de fait, une approche plus exploratoire sur laquelle nous nous concentrerons dans ce travail. Notre motivation vient du fait que, bien qu'essentielle pour certaines applications, la classification non supervisée a été relativement peu étudiée, notamment dans le cadre de l'étude de données complexes.

De manière générale, que ce soit dans un cadre fonctionnel ou non, l'objectif premier de la classification non supervisée est la définition de groupes homogènes et compacts. Pour ce faire, trois tâches principales sont à réaliser : estimer le nombre de classes contenus dans les données, affecter chaque individu à une classe et estimer les paramètres de chaque classe en optimisant un certain critère de classification. Le critère usuellement utilisé dans un but d'homogénéité et de compacité consiste à maximiser la variance inter-groupe tout en minimisant la variance intra-groupe,

c'est-à-dire que l'on cherche la partition donnant la plus grande "distance" entre les différents groupes, tout en minimisant leur rayon respectif.

Les méthodes de classification non supervisée dans un contexte non fonctionnel permettant de répondre à un ou plusieurs des enjeux précédemment cités sont nombreuses et se répartissent principalement en 3 catégories. Les deux premières catégories concernent les méthodes de classification multivariées. On peut citer les méthodes de classification hiérarchique (ascendante ou descendante) basées sur des agrégations successives faites de proche en proche (Ward 1963; Duda et Hart 1973). Elles permettent simultanément d'estimer le nombre de groupes et d'affecter les individus aux classes. En revanche, ces méthodes sont difficilement applicables à un grand nombre de données, puisque la complexité de l'algorithme est en $O(N^2)$, et ne conduisent pas en général à un partitionnement optimal du fait de leur caractère hiérarchique. Les algorithmes de réallocation dynamique représentent une deuxième grande catégorie de méthodes multivariées. Ces méthodes consistent à agréger itérativement les individus autour de centres mobiles et nécessitent en général la connaissance *a priori* du nombre de groupes. On peut citer comme exemple les algorithmes de types k-means (MacQueen 1967) ainsi que leurs variantes telles que les nuées dynamiques (Diday 1971) ou l'algorithme PAM (Partitioning Around Medoid) (Kaufman et Rousseeuw 1987). Ces algorithmes, bien que conduisant naturellement à un optimal en terme de critère de classification, sont très sensibles à l'initialisation et ne permettent pas de détecter les individus à cheval sur plusieurs classes (algorithmes dits "classifiants"). Enfin, un dernier type de méthodes est donné par les approches probabilistes de la classification non supervisée pour lesquelles la classification se fait au moyen d'une modélisation préalable des données. L'exemple le plus connu de ce type d'approche concerne les modèles de mélange pour lesquels on utilise classiquement des algorithmes de type EM permettant d'estimer les paramètres du modèle par maximum de vraisemblance (cf Section 4.2.2). La classification finale est alors donnée en terme de probabilités d'appartenance à un groupe.

Le passage au cadre de données fonctionnelles entraîne deux différences majeures : d'une part, la dimension des données considérées devient conséquente et conduit à des difficultés dans leur traitement ; cette caractéristique est communément appelée "fléau de la dimension" (Bellman 1957). D'autre part, le caractère fonctionnel des données induit une notion naturelle d'ordre temporel (ou spatial) qu'il est nécessaire de prendre en compte.

Pour ces deux raisons, les techniques usuelles de classification non supervisée décrites ci-dessus ne sont pas applicables directement dans un cadre fonctionnel et la stratégie la plus courante est de recourir à une étape préalable de réduction de dimension afin de se ramener aux techniques multivariées usuelles. Les méthodes de réduction de dimension sont basées sur le caractère fonctionnel des données et permettent, de ce fait, de prendre en compte l'ordre temporel naturel des données. Deux approches sont majoritairement utilisées : une première approche consistant

en une extension de l'analyse en composantes principales au cadre fonctionnel. Cette technique, notée FPCA pour Functional Principal Component Analysis, est basée sur la décomposition de Karhunen-Loeve de processus continus (Dauxois et al. 1982; Hall et Hosseini-Nasab 2006). Bien que populaire, elle doit néanmoins être considérée avec réserve : en effet, dans un cadre de classification non-supervisée, Chang (1983) démontre que les premières composantes principales ne possèdent pas nécessairement un pouvoir discriminant supérieur aux autres composantes. Une deuxième approche consiste à résumer l'information contenue dans les signaux au moyen d'une décomposition sur une base de fonctions adaptée, comme les bases de splines ou d'ondelettes par exemple (cf. Chapitre 3).

À l'issue de l'étape de réduction de dimension, les techniques de classification non supervisée décrites plus haut peuvent alors être appliquées soit aux coefficients de la décomposition des signaux dans une base fonctionnelle, soit aux scores des projections sur les composantes principales. Les méthodes hiérarchiques ou de centres mobiles nécessitent, dans ce cas, le choix de distances adaptées à l'aspect fonctionnel des données.

Dans ce manuscrit, nous nous concentrons plus particulièrement sur l'extension des approches probabilistes aux données fonctionnelles, basées sur une modélisation des coefficients de la projection dans une base fonctionnelle d'ondelettes. Dans ce cadre de classification non supervisée sans effets mixtes, nous développons plus en détails une approche proposée par Antoniadis et al. (2008). Pour une revue plus générale des stratégies de classification de courbes dans un tel contexte, nous invitons le lecteur à se référer aux revues bibliographiques de Jacques et Preda (2013) et de Bouveyron et Brunet (2013).

4.3.2 Modèle fonctionnel

Dans un cadre fonctionnel et pour fixer les notations, on suppose que l'on a observé N signaux notés $\{\mathbf{Y}_i\}_{i=1,\dots,N}$ sur M points de discrétisation notés $(t_1, \dots, t_M)$ où $M = 2^J$, pour $J \in \mathbb{N}$. On fait l'hypothèse que ces signaux proviennent de L groupes (où L est supposé inconnu) caractérisés par un comportement moyen différent suivant le groupe. Par la suite, on notera $\zeta_{i\ell}$ la variable indicatrice valant 1 si l'individu i est dans la classe ℓ et 0 sinon. Sachant que $\{\zeta_{i\ell} = 1\}$, le modèle fonctionnel de classification proposé par Antoniadis et al. (2008) s'écrit :

$$Y_i(t_m) = \mu_\ell(t_m) + E_i(t_m), \tag{4.7}$$

avec $\mathbf{E}_i \sim \mathcal{N}(0, \sigma_E^2 \mathbf{I}_M)$ et μ_ℓ, effet fixe fonctionnel caractérisant le comportement moyen dans la classe ℓ.

Dans un contexte non-paramétrique, on souhaite alors donner une représentation de ce modèle sur une base de fonctions afin de travailler par la suite sur les coefficients de la décomposition (cf Section 3.1). Dans l'ensemble de ce manuscrit, nous nous concentrerons sur des modélisations non-paramétriques basées sur les ondelettes.

Dans ce cadre, en effectuant la transformée en ondelettes discrètes du modèle (4.7) pour une base d'ondelettes $\{\phi, \psi\}$ choisie et sachant que $\{\zeta_{i\ell} = 1\}$, on obtient, dans le domaine des ondelettes, le modèle suivant sur les coefficients empiriques :

$$\begin{cases} c_{i,j_0k} = \alpha_{\ell,j_0k} + \varepsilon^c_{i,j_0k}, & \forall (j,k) \in \Lambda \\ d_{i,jk} = \beta_{\ell,jk} + \varepsilon^d_{i,jk}, \end{cases} \tag{4.8}$$

avec $(\boldsymbol{\varepsilon}^{c\,T}, \boldsymbol{\varepsilon}_i^{d\,T})^T \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_M)$ pour tout $(j,k) \in \Lambda = \{(j,k) \mid j = j_0, \ldots, J-1, \text{ et } k = 0, \ldots, 2^j - 1\}$.

Par linéarité de la transformée en ondelettes discrète, les coefficients d'approximation c_{i,j_0k} et de détails $d_{i,jk}$ sont alors distribués selon un mélange de lois gaussiennes. En notant π_ℓ la probabilité *a priori* pour un individu d'appartenir à la classe ℓ, on a pour tout $(j,k) \in \Lambda$:

$$\begin{cases} c_{i,j_0k} \sim \displaystyle\sum_{\ell=1}^{L} \pi_\ell \mathcal{N}(\alpha_{\ell,j_0k}, \sigma_\varepsilon^2), \\ d_{i,jk} \sim \displaystyle\sum_{\ell=1}^{L} \pi_\ell \mathcal{N}(\beta_{\ell,jk}, \sigma_\varepsilon^2). \end{cases} \tag{4.9}$$

Finalement, la représentation du modèle de classification fonctionnel (4.7) dans le domaine des ondelettes se ramène à un modèle de mélange gaussien sur les coefficients. La principale différence avec les modèles de mélange classique réside généralement dans la dimension du modèle considéré. En effet, il est courant dans ce type de contexte d'être confronté au cas où $M > N$, rendant les méthodes d'estimations canoniques inapplicables ou coûteuses numériquement. Il est donc souvent nécessaire de passer par une première étape de réduction de dimension.

4.3.3 Réduction de dimension dans les modèles fonctionnels

Par leurs propriétés de compression, l'utilisation d'une stratégie basée sur les ondelettes se révèle être particulièrement adaptée au problème de la réduction de dimension : en effet, il est courant de supposer que les informations concernant les signaux individuels sont concentrées sur une petite partie des coefficients tandis que le bruit affecte l'ensemble des coefficients à des niveaux égaux. L'idée est alors de tirer parti de cette hypothèse en utilisant les techniques décrites sous l'appellation de méthodes de seuillage (cf. Section 3.3).

Dans le cadre d'une démarche de classification non-supervisée, l'objectif principal est de pouvoir attribuer un label à chaque individu, reléguant la question de l'estimation dans ces modèles à une place secondaire : en effet, la présence de positions où les coefficients individuels sont non-discriminants peut être vue comme une source de bruit vis-à-vis de la classification. Pour une position (j,k) donnée, si les coefficients individuels sont tous considérés comme non nuls mais égaux, cette

position n'apportera pas d'information pour la classification. De plus, si de telles positions sont nombreuses par rapport aux positions discriminantes, l'information d'intérêt est alors diluée et conduit à une classification finale moins performante. Le challenge est alors de réussir à identifier ces positions problématiques afin de les retirer de l'analyse. Ceci a été réalisé par Antoniadis, Bigot, et von Sachs (2008) pour la caractérisation de zones homogènes dans des images fonctionnelles et l'idée proposée par les auteurs consiste en trois étapes :

1. Etape de débruitage : Pour chaque signal individuel $\mathbf{Y}_i$, $i = 1, \ldots, N$, on applique un seuillage individuel dur aux coefficients d'ondelettes en utilisant le seuil universel comme défini en Section 3.3.
2. Etape d'union : L'union des positions sélectionnées individuellement durant la première étape est ensuite réalisée. On obtient ainsi un ensemble commun de positions pour lesquelles les coefficients d'ondelettes associés à l'effet fixe fonctionnel sont non-nuls pour l'ensemble des individus. Antoniadis, Bigot, et von Sachs (2008) notent que, formellement, l'estimateur fonctionnel résultant de l'union des coefficients rentre dans la classe des estimateurs dits *d'agrégation* comme défini par Bunea, Tsybakov, et Wegkamp (2007) et de ce fait, sont optimaux en terme de reconstruction.
3. Etape de réduction de dimension : Afin de détecter les positions restantes identiques pour tous les individus, une dernière étape consiste à utiliser le test de Neyman (Fan 1996) sur chaque position afin de déterminer si elle est discriminante ou non. Ainsi, pour une position (j, k) fixée présente dans l'union des coefficients sélectionnés, le test de Neyman est appliqué sur le vecteur des différences $(d^i_{jk} - d^{i-1}_{jk})_{i=2,\ldots,N}$ en prenant comme hypothèse H_0 la nullité de ce vecteur. La multiplicité est ensuite controlée par une approche FDR consistant à contrôler le risque de première espèce par des contraintes sur le nombre de faux positifs (Benjamini et Hochberg 1995).

En pratique, les auteurs ont observé que cette méthode dans son ensemble conduisait en moyenne à une réduction de dimension de l'ordre d'un facteur 2, c'est-à-dire que la taille des vecteurs de coefficients de départ est, en moyenne, réduite de moitié.

On obtient finalement un modèle de taille réduite concentrant l'information utile vis-à-vis de la classification et permettant alors d'appliquer une méthode de classification usuelle sur les positions sélectionnées restantes.

4.3.4 Estimation des paramètres

L'estimation des paramètres dans ce modèle à données cachées est réalisée par maximum de vraisemblance. On utilise pour ce faire l'algorithme EM, où les variables d'appartenance au groupe représentent la structure latente. Ce cadre est un cadre classique d'application de l'algorithme. Afin de faciliter la lecture de ce manuscrit, nous ne détaillerons pas les formulations des estimateurs dans cette section mais

nous en proposerons une formulation dans un contexte plus général de classification non-supervisée au sein de modèles à effets mixtes (Chapitre 5, Section 5.2.2).

4.4 Modèle mixte fonctionnel

Nous présentons dans cette section un autre cas particulier de modèles à variables latentes : le modèle mixte fonctionnel au sein d'un groupe d'individus homogènes. Nous détaillons dans un premier temps la modélisation adoptée, basée sur une approche proposée par Antoniadis et Sapatinas (2007) et plus particulièrement, nous décrivons la modélisation choisie pour les effets aléatoires individuels. Cette modélisation est effectuée dans le domaine des coefficients et nous proposons une étude de simulation visant à visualiser les conséquences de celle-ci sur les effets aléatoires individuels fonctionnels.

4.4.1 Modèle général

Pour définir notre modèle mixte fonctionnel, nous souhaitons étendre la notion de modèle mixte appliqué au cadre des données longitudinales, décrite au Chapitre 2, à un cadre fonctionnel. Les données observées sont supposées fonctionnelles et on suppose de même que les déviations individuelles par rapport au signal moyen sont elles aussi fonctionnelles. A partir du modèle fonctionnel (3.1), cela se traduit par l'ajout d'un effet aléatoire individuel fonctionnel, noté $U_i(t)$, modélisant la variabilité spécifique due à l'individu i. Les effets aléatoires $(U_i(t))_{i=1,\dots,N}$ sont alors modélisés comme des réalisations de processus aléatoires de fonction de covariance commune $K(t,t')$. Dans ce cadre, le modèle fonctionnel s'écrit alors :

$$Y_i(t_m) = \mu(t_m) + U_i(t_m) + E_i(t_m), \tag{4.10}$$

en conservant les hypothèses classiques suivantes,

$$\begin{cases} U_i \text{ processus gaussien centré de covariance } K(t,t') = \mathrm{Cov}(U_i(t), U_i(t')), \\ E_i(t_m) \sim \mathcal{N}(0, \sigma_E^2), \\ \mathbf{U}_i \perp \mathbf{E}_i. \end{cases}$$

En projetant, comme précédemment, le modèle sur une base d'ondelettes $\{\phi, \psi\}$ choisie, on retrouve alors sur les coefficients d'ondelettes, un modèle linéaire mixte canonique donné par :

$$\begin{cases} c_{i,j_0k} = \alpha_{j_0k} + \nu_{i,j_0k} + \varepsilon^c_{i,j_0k}, & \forall (j,k) \in \Lambda, \\ d_{i,jk} = \beta_{jk} + \theta_{i,jk} + \varepsilon^d_{i,jk}, \end{cases} \tag{4.11}$$

avec :

$$\begin{cases} (\boldsymbol{\varepsilon}_i^{c\,T}, \boldsymbol{\varepsilon}_i^{d\,T})^T \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_M), \\ \qquad\qquad \boldsymbol{\nu}_i \sim \mathcal{N}(0, \mathbf{G}_\nu = \mathrm{diag}(\gamma_\nu^2)), \\ \qquad\qquad \boldsymbol{\theta}^i \sim \mathcal{N}(0, \mathbf{G}_\theta). \end{cases}$$

4.4.2 Modélisation de la variabilité individuelle

Spécification de la matrice de covariance $\mathbf{G}_\theta$

Le point difficile ici est de donner une modélisation pour la matrice $\mathbf{G}_\theta$. En effet, suivant la dimension des données considérées, qui peut être conséquente dans un contexte fonctionnel, la taille de la matrice $\mathbf{G}_\theta$ peut rapidement croître. Si de plus, elle est supposée non structurée, le nombre de paramètres à estimer est alors du même ordre et entraîne un problème de sur-ajustement. Il est donc important de proposer une modélisation "simple" permettant de limiter le nombre de paramètres tout en assurant une flexibilité du modèle suffisante.

Dans le cadre de l'étude de données issues des biotechnologies, notre approche consiste à associer les déviations inter-individuelles à des comportements métaboliques propres à l'individu, permettant alors de faciliter l'interprétation des effets aléatoires dans un contexte fonctionnel. Ainsi, pour le cas des données de spectrométrie de masse (cf. Section 6.2.1), les déviations individuelles sont représentées par la présence/absence ou la différence d'expression de certains peptides entre individus et se traduisent par l'observation de la présence/absence de pics et de variabilité dans leur hauteur. Dans une approche fonctionnelle, cela revient à faire l'hypothèse d'effets aléatoires individuels partageant au moins la même régularité que les effets fixes. Nous nous plaçons en ce sens dans le cadre proposé par Antoniadis et Sapatinas (2007) qui proposent une modélisation fonctionnelle des effets fixes et aléatoires dans le contexte des espaces de Besov. L'hypothèse d'effets fixes et aléatoires partageant au moins la même régularité se traduit alors par l'appartenance des fonctions au même espace fonctionnel. Suivant l'idée de Antoniadis et Sapatinas (2007), notre choix est de se placer au sein des espaces de Besov permettant alors une modélisation flexible des deux types d'effets. Dans ce cadre, Antoniadis et Sapatinas (2007) proposent une modélisation des effets aléatoires fonctionnels $U_i(t)$ dans le domaine des ondelettes permettant d'en exploiter pleinement les propriétés ainsi que celles des espaces de Besov. Nous mentionnons à ce propos qu'une telle modélisation n'est pas envisageable à l'aide de splines : en effet, dans le cadre des modèles mixtes, deux approches similaires dans le contexte des modèles mixtes ont été proposées par Huang et Lu (2000) et Angelini et al. (2003) mais les effets fixes et les effets aléatoires sont alors contraints à se situer dans des espaces fonctionnels de régularité différentes d'après un argument soulevé par Green et Silverman (1994)[Chap.3].

Selon notre modélisation, la matrice $\mathbf{G}_\theta$ est, en premier lieu, supposée posséder une structure diagonale, cette hypothèse étant justifiée par la propriété dite *décorrélante* des ondelettes mise en évidence par Zhang et Walter (1994) et Frazier,

Jawerth, et Weiss (1991). Pour une large classe de processus, la matrice de variance covariance associée à la décomposition du processus dans une base d'ondelettes est une matrice "presque diagonale", c'est-à-dire qu'on observe une décroissance rapide des coefficients hors diagonale de la matrice $\mathbf{G}_\theta$.

D'autre part, partant de la propriété (3.1) de caractérisation des espaces de Besov par la norme des coefficients , le théorème (4.1) proposé par Abramovich, Sapatinas, et Silverman (1998) et adapté ensuite par Antoniadis et Sapatinas (2007) dans le cadre des modèles mixtes nous donne une condition nécessaire et suffisante assurant que les effets individuels $U_i(t)$ appartiennent au même espace fonctionnel que l'effet fixe $\mu(t)$.

Théorème 4.1. *Soit ϕ une ondelette de régularité r, où $\max(0, \frac{1}{p} - \frac{1}{2}) < s < r$), $1 \leq p,\ q \leq \infty$, et supposons que les coefficients d'ondelettes $(w_{jk})_{(j,k)\in\Lambda}$ d'une fonction g sont distribués suivant une loi gaussienne centrée de variance $\gamma^2 2^{-j\eta}$. Alors, pour toute valeur fixée des coefficients d'échelle $\boldsymbol{\nu}$, $g \in B^s_{pq}([0,1])$ presque sûrement si et seulement si :*

$$\begin{cases} s + \dfrac{1}{2} - \dfrac{\eta}{2} < 0, \\ s + \dfrac{1}{2} - \dfrac{\eta}{2} = 0 \quad \textit{pour } \ 1 \leq p < \infty \ \textit{et } q = \infty. \end{cases}$$

La preuve de ce théorème est donnée dans l'article de Abramovich, Sapatinas, et Silverman (1998).

Ainsi, on en déduit que le contrôle de la décroissance de la variance des coefficients d'ondelettes permet de contrôler la régularité du processus associé : pour un effet fixe fonctionnel $\mu(t)$ appartenant à l'espace de Besov $B^s_{pq}([0,1])$ et en posant

$$\mathbb{V}(\theta^i_{jk}) = (\mathbf{G}_\theta)_{jk} = \gamma^2 2^{-j\eta}, \quad \forall i = 1, \ldots, N, \ \ \forall (j,k) \in \Lambda,$$

le Théorème (4.1) nous donne une condition nécessaire et suffisante sur la valeur de η assurant que les effets fixes et aléatoires sont bien dans le même espace de Besov.

Morris et Carroll (2006) proposent une procédure d'estimation au sein d'un modèle mixte fonctionnel similaire au modèle (4.10) basée sur l'utilisation d'ondelettes dans un cadre bayésien. Leur modélisation des coefficients dans ce cadre, différente de celle que nous adoptons, est basée sur un mélange entre une masse de Dirac et une loi gaussienne. Comme suggéré par ces auteurs, il peut aussi être nécessaire de permettre plus de flexibilité au modèle en autorisant le paramètre de variance γ^2 à dépendre de la position et du niveau considérés en prenant une modélisation en γ^2_{jk}. Signalons que ce point est illustré de manière visuelle en Figure 1 de leur article.

À ce stade, une première question est de savoir quelle classe de processus peuvent être atteints par ce choix de modélisation. Antoniadis et Sapatinas (2007) affirment que leur modélisation permet d'atteindre une grande diversité de processus par le biais du pouvoir décorrélant des ondelettes mais ce point n'a cependant jamais été

étudié d'un point de vue théorique. L'argument principal est avant tout technique puisqu'une telle modélisation permet de se ramener à la manipulation de matrices diagonales et afin d'en évaluer la richesse, nous proposons dans la section suivante une étude qualitative des processus induits par cette modélisation.

Remarquons enfin que l'approche adoptée ici est différente des approches adoptées plus généralement dans la communauté du traitement du signal : en effet, dans ce contexte, la stratégie largement considérée est de se placer, dans le domaine fonctionnel, dans une classe particulière de processus (par exemple, dans la classe des processus stationnaires, à accroissements stationnaires ou encore autorégressifs) et d'en déduire des propriétés sur les coefficients d'ondelettes associés. On peut citer comme exemple les travaux de Istas (1992) pour les processus gaussiens, de Abry et al. (1995) pour les processus stationnaires ou de Dalhaus et al. (1999) pour les processus autorégressifs. Néanmoins, ce type de modélisation ne permet pas en général d'atteindre les objectifs souhaités dans notre cas, à savoir le contrôle de la régularité des trajectoires mais aussi le contrôle du nombre de paramètres du modèle. En effet, ce type d'approche conduit en général à des matrices de covariance $\mathbf{G}_\theta$ de structure complexe dans le domaine des ondelettes nécessitant, comme Johnstone et Silverman (1997) ou von Sachs et MacGibbon (2000) dans un cadre de seuillage de processus respectivement stationnaires ou localement stationnaires, de se ramener à des matrices diagonales par la propriété décorrélante des ondelettes. Cependant, cette approximation induit alors une restriction quant à la classe de processus concernée qu'il est difficile de maîtriser.

Processus engendrés

Nous plaçons une modélisation particulière sur la structure de covariance associée aux effets aléatoires individuels où les variances des coefficients sont supposées être de la forme $\gamma^2 2^{-j\eta}$. Le choix d'une telle approche nous amène à la question suivante : dans le domaine temporel, quels types de processus peuvent être approchés à l'aide de cette modélisation ? Plus particulièrement, notre objectif dans cette section est d'évaluer la flexibilité d'une telle modélisation afin de ne pas se restreindre à une classe de processus trop restrictive et d'étudier le rôle des paramètres entrant dans la modélisation, comme le paramètre de régularité η ou encore le choix de la base d'ondelettes.

En adoptant une stratégie inverse, on peut montrer que la représentation dans le domaine des ondelettes de processus stationnaires à courte ou longue dépendance conduit à des variances résultantes dépendant uniquement du niveau de résolution j et restant constantes au sein d'un même niveau. Pour des processus non-stationnaires ou sous une hypothèse de stationnarité locale, on obtient dans ce cas, des variances variant avec l'échelle j et la position k. Ces arguments sont notamment utilisés par Johnstone et Silverman (1997) et von Sachs et MacGibbon (2000) pour développer, dans un cadre d'estimation, des stratégies de seuillage pour ces types de processus. Par contre, la décomposition de tels processus dans le domaine des ondelettes ne

résulte pas en général en une modélisation simple de la matrice de covariance.

Par ailleurs, l'approche consistant à modéliser la covariance du processus dans le domaine des ondelettes a fait l'objet de relativement peu de travaux et nous proposons dans cette section une étude qualitative permettant d'appréhender la flexibilité de la modélisation adoptée et l'influence du paramètre de régularité η.

Le calcul de la fonction de covariance $K(s,t)$ à partir de la matrice des variances $\mathbf{G}_\theta$ n'est pas réalisable analytiquement pour la majorité des bases d'ondelettes car elles ne possèdent pas d'expression analytique connue. En revanche, en utilisant la base d'ondelettes de Haar pour laquelle les fonctions d'échelle et de détails s'expriment simplement et possèdent des supports disjoints, il est possible de calculer la fonction de covariance à partir des coefficients d'ondelettes de manière exacte.

On rappelle que les fonctions d'échelle et les ondelettes de la base de Haar sont définies par :

$$\phi_{j_0k}(x) = \begin{cases} 2^{\frac{j_0}{2}} & \text{si } x \in [\frac{k}{2^{j_0}}, \frac{k+1}{2^{j_0}}[, \\ 0 & \text{sinon,} \end{cases}$$

et :

$$\psi_{jk}(x) = \begin{cases} 2^{\frac{j}{2}} & \text{si } x \in [\frac{2k}{2^{j+1}}, \frac{2k+1}{2^{j+1}}[, \\ -2^{\frac{j}{2}} & \text{si } x \in [\frac{2k+1}{2^{j+1}}, \frac{2k+2}{2^{j+1}}[, \\ 0 & \text{sinon.} \end{cases}$$

La décomposition de l'effet aléatoire individuel $U_i(t) \in B^s_{pq}([0,1])$ dans la base de Haar s'écrit alors pour tout $t \in [0,1]$ et tout $i = 1, \ldots, N$:

$$U_i(t) = \sum_{k=0}^{2^{j_0}-1} \nu_{i,j_0k}\phi_{j_0k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \theta_{i,jk}\psi_{jk}(t). \tag{4.12}$$

Soit $(t,s) \in [0,1[$. La fonction de covariance $K(s,t)$ de $U_i(t)$ est donnée par :

$$K(s,t) = \mathrm{Cov}\big(U_i(t)U_i(s)\big) = \mathbb{E}\big(U_i(t)U_i(s)\big) - \mathbb{E}\big(U_i(t)\big)\mathbb{E}\big(U_i(s)\big). \tag{4.13}$$

Par la suite, nous distinguerons les cas $t \neq s$ et $t = s$ et pour simplifier les calculs, nous prendrons $j_0 = 0$.

Cas $\mathbf{t \neq s}$ On peut supposer sans perte de généralité que $s > t$. Les variables $\theta_{i,jk}$ et $\nu_{i,00}$ sont centrées et indépendantes entre elles pour tout $j \geq 0$ et tout $k = 0, \ldots, 2^j - 1$. L'expression (4.13) devient alors :

$$\begin{aligned} K(s,t) &= \mathbb{E}\big(\nu^2_{i,00}\big)\phi_{00}(t)\phi_{00}(s) + \sum_{j \geq 0} \sum_{k=0}^{2^j-1} \mathbb{E}\big(\theta^2_{i,jk}\big)\psi_{jk}(t)\psi_{jk}(s) \\ &= \gamma^2_\nu + \sum_{j \geq 0} \sum_{k=0}^{2^j-1} \mathbb{E}\big(\theta^2_{i,jk}\big)\psi_{jk}(t)\psi_{jk}(s). \end{aligned}$$

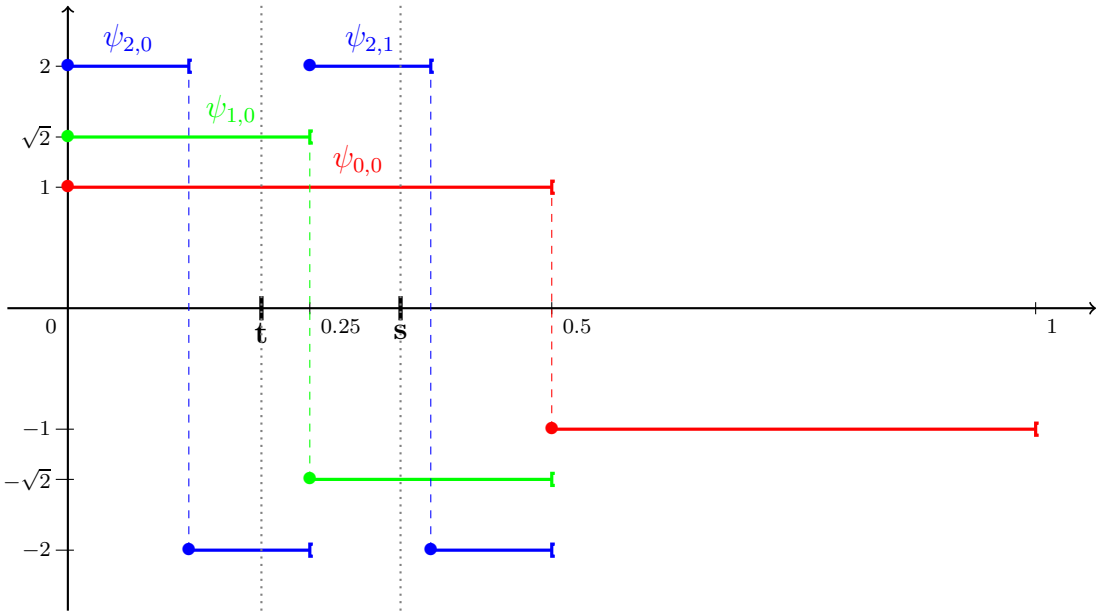


FIGURE 4.1 – *Représentation du couple d'indices* (j_{ts}, k_{ts}) *pour un couple de points* (t, s) *particulier.*

Soit (j_{ts}, k_{ts}) tels que :

$$\begin{cases} t \in [\frac{2k_{ts}}{2^{j_{ts}+1}}, \frac{2k_{ts}+1}{2^{j_{ts}+1}}[\\ s \in [\frac{2k_{ts}+1}{2^{j_{ts}+1}}, \frac{2k_{ts}+2}{2^{j_{ts}+1}}[\end{cases}$$

Le couple (j_{ts}, k_{ts}) existe et est unique. Il dépend des valeurs de t et s et correspond au premier intervalle dyadique sur lequel les points t et s sont à cheval. Ceci est illustré sur la Figure 4.1 pour deux points t et s particuliers : on peut observer qu'à partir du niveau de résolution $j = 2$, les points t et s se trouvent sur deux fonctions d'ondelettes ayant des supports disjoints. Sur cet exemple particulier, on a alors $j_{ts} = 1$ et $k_{ts} = 0$. De manière générale, pour les valeurs (j_{ts}, k_{ts}), on a alors :

$$\begin{aligned} \forall j > j_{ts} \qquad & \text{Pour tout } k = 0, \dots, 2^j - 1, \psi_{jk}(t)\psi_{jk}(s) = 0, \\ \forall j < j_{ts} \qquad & \exists! \ k \text{ tel que } \psi_{jk}(t) \neq 0, \text{ et } \psi_{jk}(t) = \psi_{jk}(s) = 2^{\frac{j}{2}} \neq 0, \\ \text{Pour } j = j_{ts} \qquad & \exists! \ k \text{ tel que } \psi_{jk}(t) \neq 0, \text{ et } \psi_{jk}(t) = -\psi_{jk}(s) = 2^{\frac{j}{2}} \neq 0. \end{aligned}$$

La fonction de covariance s'exprime alors comme :

$$\begin{aligned} K(s,t) &= \gamma_\nu^2 + \sum_{j=0}^{j_{ts}} \sum_{k=0}^{2^j-1} \mathbb{E}\big(\theta_{i,jk}^2\big)\psi_{jk}(t)\psi_{jk}(s) \\ &= \gamma_\nu^2 + \sum_{j=0}^{j_{ts}-1} \mathbb{E}\big(\theta_{i,jk}^2\big)2^j - \mathbb{E}\big(\theta_{i,j_{ts}k}^2\big)2^{j_{ts}}. \end{aligned}$$

En prenant une modélisation telle que proposée par Antoniadis et Sapatinas (2007), à savoir que $\mathbb{E}\big(\theta_{i,jk}^2\big) = \gamma^2 2^{-j\eta}$, pour tout $j \geq 0$ et tout $k = 0, \ldots, 2^j - 1$, on obtient alors :

$$\begin{aligned} K(s,t) &= \gamma_\nu^2 + \sum_{j=0}^{j_{ts}-1} \gamma^2 2^{j(1-\eta)} - \gamma^2 2^{j_{ts}(1-\eta)} \\ &= \gamma_\nu^2 + \gamma^2 \frac{1 - 2^{j_{ts}(1-\eta)}}{1 - 2^{(1-\eta)}} - \gamma^2 2^{j_{ts}(1-\eta)} \\ &= \gamma_\nu^2 + \gamma^2 \frac{1 - 2^{j_{ts}(1-\eta)} \left(2 - 2^{(1-\eta)}\right)}{1 - 2^{(1-\eta)}}. \end{aligned}$$

Cas $t = s$ Soit $t \in [0, 1[$. Dans ce cas, on a :

$$\begin{aligned} K(t,t) &= \mathbb{E}\big(\nu_{i,00}^2\big)\big(\phi_{00}(t)\big)^2 + \sum_{j\geq 0} \sum_{k=0}^{2^{j-1}} \mathbb{E}(\theta_{i,jk}^2)\big(\psi_{jk}(t)\big)^2 \\ &= \gamma_\nu^2 + \sum_{j\geq 0} \sum_{k=0}^{2^{j-1}} \mathbb{E}(\theta_{i,jk}^2)\big(\psi_{jk}(t)\big)^2. \end{aligned}$$

Pour chaque niveau j, il existe un unique indice k tel que $\psi_{jk}(t) = 2^{j/2} \neq 0$, d'où :

$$K(t,t) = \gamma_\nu^2 + \sum_{j\geq 0} \mathbb{E}(\theta_{i,jk}^2) 2^j$$

On constate alors que pour une modélisation constante de la variance, c'est-à-dire avec $\mathbb{E}(\theta_{i,jk}^2) = \gamma^2$ pour tout j, k, la quantité $K(t,t)$ n'est pas définie, car non-sommable. En considérant la modélisation en $\mathbb{E}\big(\theta_{i,jk}^2\big) = \gamma^2 2^{-j\eta}$, pour tout j, k, $K(t,t)$ est définie si et seulement si $2^{(1-\eta)} < 1$, i.e., si et seulement si $\eta > 1$. Dans ce cas, on a :

$$\begin{aligned} K(t,t) &= \gamma_\nu^2 + \sum_{j\geq 0} \gamma^2 2^{-j\eta} 2^j \\ &= \gamma_\nu^2 + \gamma^2 \sum_{j\geq 0} \left[2^{(1-\eta)}\right]^j \\ &= \gamma_\nu^2 + \frac{\gamma^2}{1 - 2^{(1-\eta)}} \end{aligned}$$

Finalement, la fonction $K(s,t)$ s'exprime de la façon suivante :

$$K(s,t) = \begin{cases} \gamma_\nu^2 + \gamma^2 \left(\frac{1 - 2^{j_{ts}(1-\eta)}\left(2 - 2^{(1-\eta)}\right)}{1 - 2^{(1-\eta)}}\right) & \text{si } t \neq s, \\ \gamma_\nu^2 + \frac{\gamma^2}{1 - 2^{(1-\eta)}} & \text{sinon.} \end{cases} \tag{4.14}$$

On constate alors que le processus considéré n'est pas stationnaire car j_{ts} dépend de t et de s et non pas de la différence $(t-s)$. En fait, une telle modélisation donne accès à une large classe de processus, englobant aussi certaines classes de processus non stationnaires. Basé sur le travail d'Abramovich et al. (1998), Antoniadis et Sapatinas (2007) remarquent que la quantité $\mathbb{E}\left(\theta_{i,jk}^2\right)$ représente la variance *a priori* des coefficients d'ondelettes non-nuls de la décomposition de l'effet fonctionnel U_i. Le choix d'une modélisation en $\gamma^2 2^{-j\eta}$ correspond à un *prior* donnant la même probabilité aux coefficients d'un même niveau d'être non-nuls. Ce type de *prior* donne accès entre autres à la modélisation de processus auto-similaires dont le mouvement brownien fait partie.

En Figure 4.2 sont représentées les matrices de variance covariance associées au processus U_i pour différentes valeurs du paramètre η ($\eta \in \{1.001, 1.5, 2, 4\}$). Les valeurs de γ_ν^2 et γ^2 sont choisies égales à 1. Nous pouvons observer la liaison entre le paramètre η et la régularité du processus sous-jacent : en effet, sur la matrice de covariance correspondant à la valeur $\eta = 1.001$, les valeurs de covariances sont très faibles devant les valeurs des variances. Cela correspond à des valeurs successives faiblement corrélées et donc à des processus très irréguliers. A l'inverse, une grande valeur de η induit des dépendances de longue portée conduisant à des processus très lisses.

Ceci est illustré en Figure 4.3 où des exemples de processus sont représentés pour chacune des valeurs du paramètre η données ci-dessus.

Concernant les autres bases d'ondelettes, une manière détournée d'accéder à la fonction $K(s,t)$ consiste à l'approcher en effectuant la transformée en ondelettes inverse de la matrice de variance-covariance $\mathbf{G}_\theta$. Pour des valeurs fixées de γ_ν^2 et $\gamma^2 2^{-j\eta}$ et pour une base d'ondelettes donnée (et donc une matrice de filtres $\mathbf{W}$), la matrice $\mathbf{K} = \mathbf{W}^T \mathbf{G}_\theta \mathbf{W}$ représente une approximation discrétisée de la fonction de covariance $K(s,t)$ dans le domaine fonctionnel. Il est alors possible de simuler des processus dans ce cadre en simulant des réalisations de lois gaussiennes centrées de variance γ_ν^2 pour les coefficients d'approximation et de variance $\gamma^2 2^{-j\eta}$ pour les coefficients de détails. On applique par la suite une transformation discrète inverse aux coefficients ainsi simulés.

En Figure 4.4, nous avons représenté des exemples de processus pour plusieurs valeurs de η (mêmes valeurs que dans le cas précédent) et plusieurs bases d'ondelettes (Haar, Daubechies à 2 moments nuls et Daubechies à 8 moments nuls). On peut observer sur ces graphes que les processus résultant d'une telle approximation sont très dépendants de la base d'ondelettes choisie, traduisant la volonté de modéliser des effets fixes et aléatoires présentant une structure partagée.

Cependant, il est important de garder à l'esprit que cette méthode de simulation conduit à des erreurs d'approximation et dans certains cas à une approximation très imprécise de la fonction $K(s,t)$. La conduite d'une étude de simulation sur l'erreur d'approximation réalisée dans le cas de la base de Haar, où la fonction de covariance est connue de manière exacte, nous a permis de mettre en évidence le fait que la

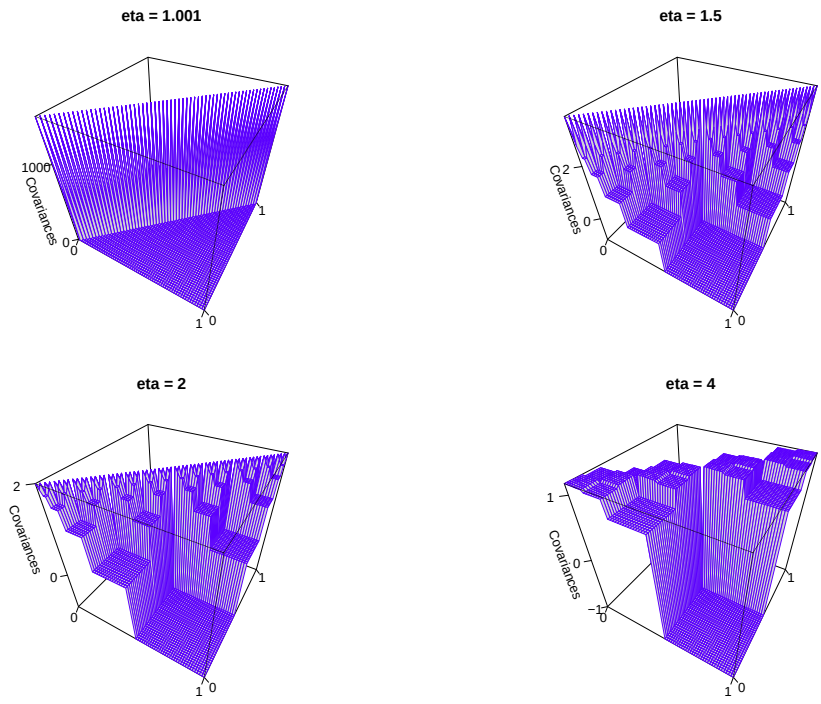


FIGURE 4.2 – *Matrices de variance covariance associées au processus U_i pour différentes valeurs du paramètre η ($\eta \in \{1.001, 1.5, 2, 4\}$). Les matrices représentées sont de taille 64×64. Les paramètres γ_ν^2 et γ^2 sont fixés à 1*

précision de l'approximation est très sensible à la valeur du paramètre η. Ainsi, plus la valeur du paramètre η est faible (c'est à dire, moins le processus est régulier), plus le biais et l'écart quadratique moyen entre la fonction $\mathbf{K}$ et son approximation sont grands. Ce phénomène est représenté en Figure 4.5.

Enfin, comme remarqué par Morris et Carroll (2006), il peut être utile pour modéliser certains types de données de permettre au paramètre γ^2 de dépendre de la position (j, k). En effet, l'hypothèse d'une variance de l'effet aléatoire dépendant uniquement du niveau j (par l'intermédiaire du terme $2^{-j\eta}$) semble être trop restrictive comme nous le verrons avec les exemples d'applications sur données réelles. Cette possibilité est prise en compte dans notre modèle permettant ainsi de modéliser une plus grande variété de processus. Des exemples de tels processus sont représentés en Figure 4.6, en prenant des variances de la forme γ_{jk}^2 pour les coefficients de détails. Cependant, cette modélisation a pour effet d'entraîner une forte augmentation du nombre de paramètres pouvant créer des instabilités d'estimations.

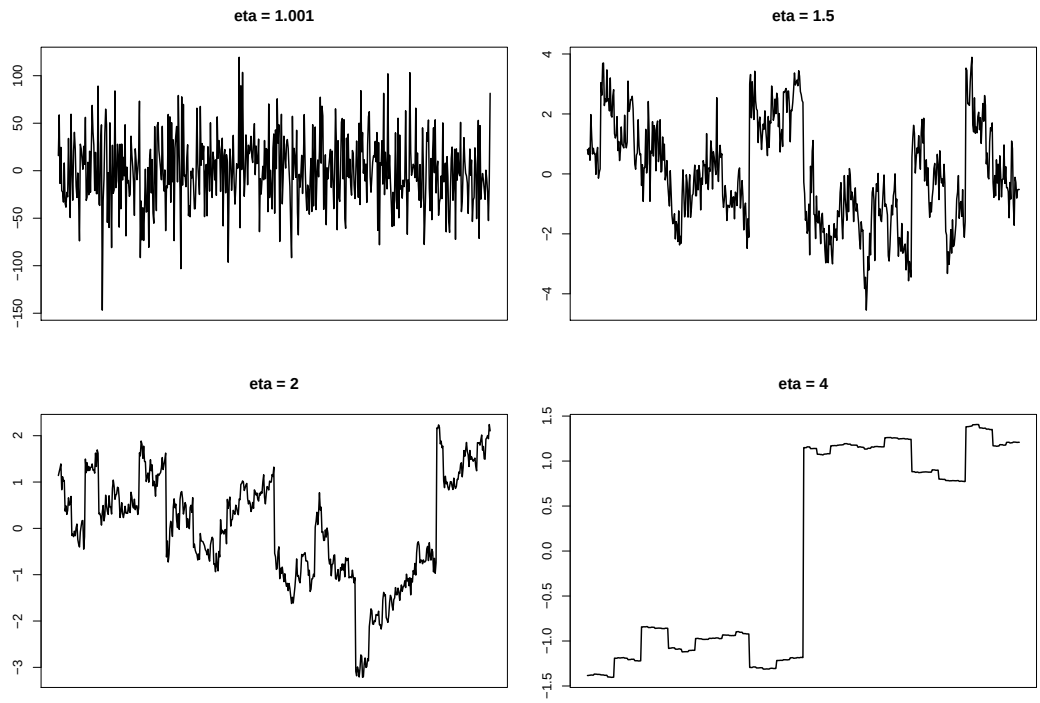


FIGURE 4.3 – *Exemple de processus de fonction de covariance $K(s,t)$ pour différentes valeurs du paramètre η ($\eta \in \{1.001, 1.5, 2, 4\}$). Les signaux représentés sont de taille 512. Les paramètres γ_ν^2 et γ^2 sont fixés à 1*

En résumé, nous avons introduit dans ce chapitre les principaux ingrédients constituant la base de notre modèle de classification non-supervisée dans les modèles mixtes fonctionnels. Le point critique au sein des modèles mixtes fonctionnels est la modélisation des effets aléatoires fonctionnels et plus particulièrement de la covariance associée. Néanmoins, la dimension fonctionnelle des données considérées entraîne, de même que dans les modèles sans effet aléatoire, des difficultés d'ajustement et de ce fait, une nécessité de réduire la dimension du modèle. Ce point n'est pas abordé dans cette partie pour les modèles mixtes fonctionnels car cette problématique est traitée en détails au cours de la deuxième partie de ce manuscrit.

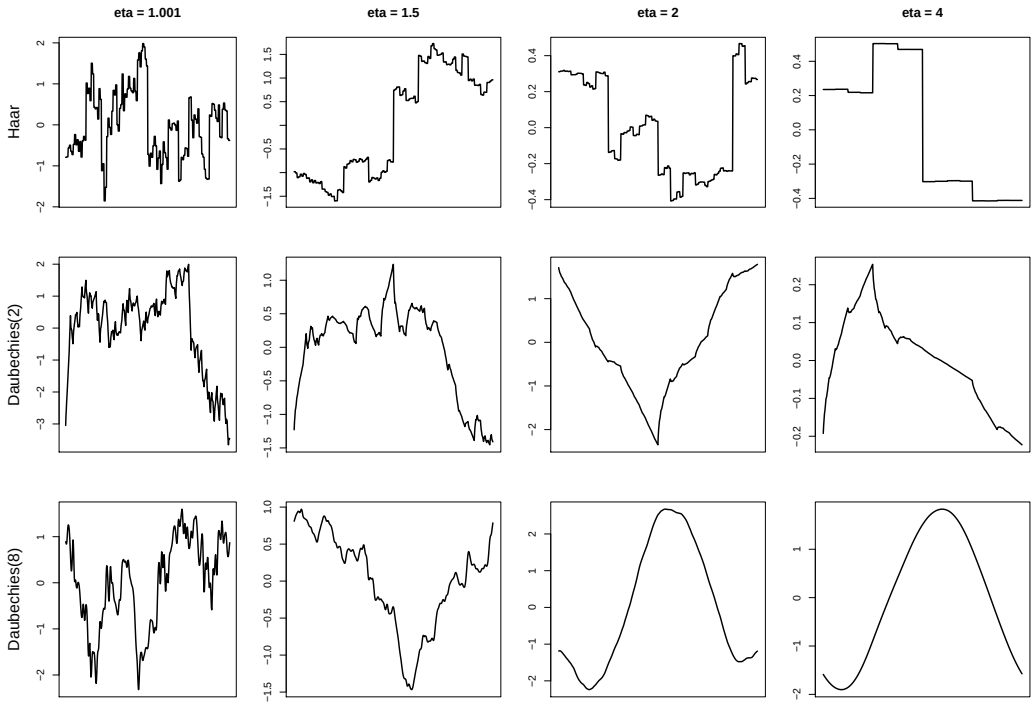


FIGURE 4.4 – *Exemple de processus obtenus pour différentes valeurs du paramètre* η *(*$\eta \in \{1.001, 1.5, 2, 4\}$*) et différentes bases d'ondelettes (Haar, Daubechies à 2 moments nuls et Daubechies à 8 moments nuls). Ces processus ont été obtenus en simulant des coefficients dans le domaine des ondelettes selon notre modèle puis en effectuant une transformation inverse. Les signaux représentés sont de taille 512. Les paramètres* γ_ν^2 *et* γ^2 *sont fixés à 1*

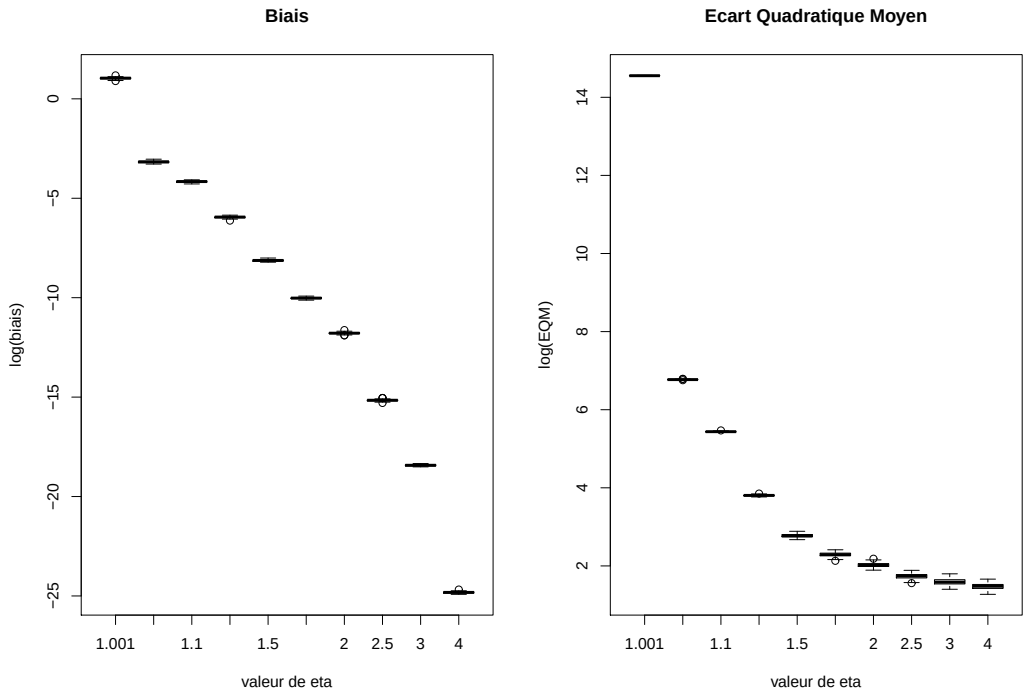


FIGURE 4.5 – *Boxplots des biais et écarts quadratiques moyens (EQM) obtenus entre la fonction de covariance exacte et son approximation basée sur la transformée discrète inverse en utilisant la base d'ondelettes de Haar. Les valeurs observées sont données en fonction du paramètre η associé à la régularité du processus considéré.*

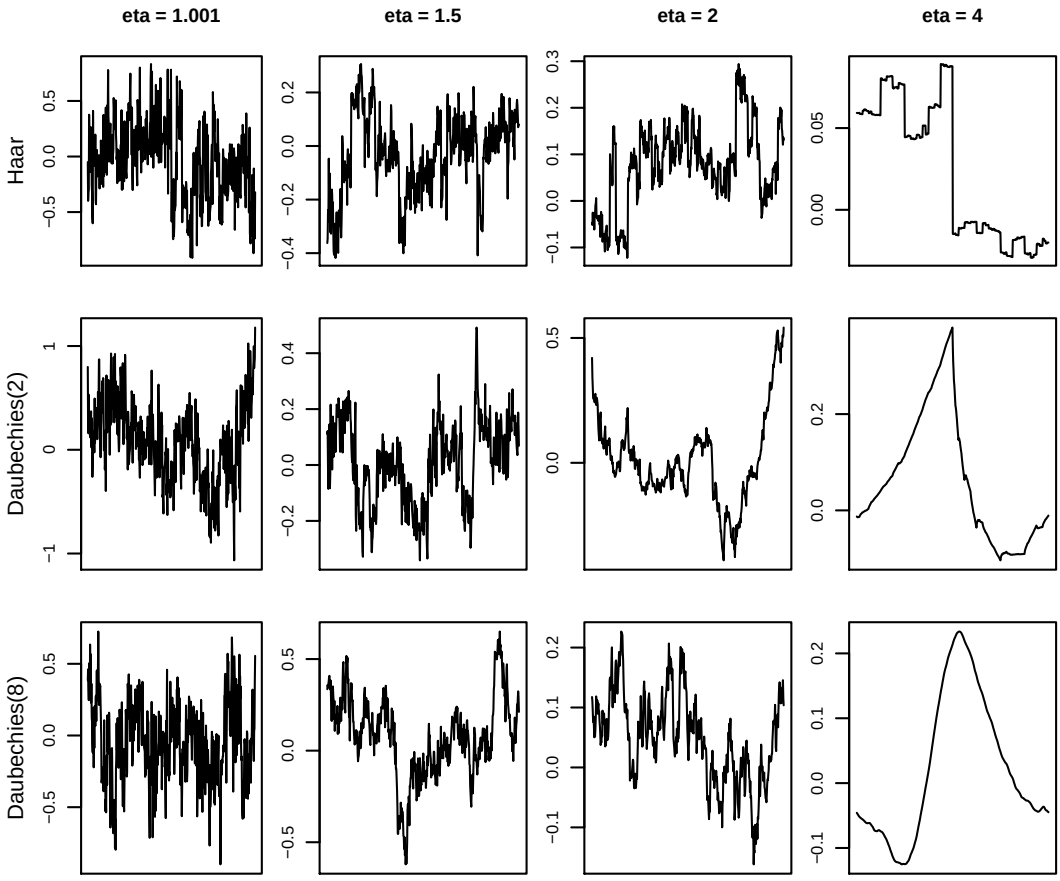


FIGURE 4.6 – *Exemple de processus obtenus pour différentes valeurs du paramètre* η *(*$\eta \in \{1.001, 1.5, 2, 4\}$*) et différentes bases d'ondelettes (Haar, Daubechies à 2 moments nuls et Daubechies à 8 moments nuls). Ces processus ont été obtenus en simulant des coefficients dans le domaine des ondelettes selon notre modèle puis en effectuant une transformation inverse. Les signaux représentés sont de taille 512. Ici, les paramètres de variances* γ_{jk}^2 *varient avec le niveau* (j, k) *et sont simulés selon une loi gamma.*

Deuxième partie

Classification non supervisée dans les modèles mixtes fonctionnels

Introduction

Au cours de cette deuxième partie, nous présentons ce qui représente la première contribution de ce travail de thèse, à savoir, le développement d'une procédure permettant de réaliser une classification non-supervisée de courbes dans le cadre des modèles mixtes fonctionnels.

Dans un premier chapitre, nous introduisons, en premier lieu, le modèle de classification non supervisée au sein des modèles mixtes que nous avons développé. Ce modèle est une extension du modèle de classification fonctionnel et du modèle mixte mixte fonctionnel présentés au Chapitre 4. Puis, nous présentons dans un deuxième temps notre procédure de classification basée sur une approche probabiliste de la classification non supervisée. Notre procédure se déroule en deux étapes : une première étape de réduction de dimension tirant parti des propriétés de parcimonie des ondelettes et une deuxième étape d'estimation des paramètres du modèle, servant à obtenir une classification finale des individus. Le choix du nombre de groupes est réalisé *a posteriori* grâce à un critère de sélection de type BIC (Bayesian Information Criteria).

Dans un deuxième chapitre, nous présentons une vaste étude de simulations permettant d'évaluer le comportement de notre procédure face à différentes configurations de données simulées. Pour ce faire, nous nous sommes attachés à développer un cadre de simulation rigoureux permettant la création systématique de données synthétiques. Nous proposons au cours de cette étude une comparaison à une procédure développée par James et Sugar (2003) reposant sur un modèle non paramétrique similaire mais basée sur l'utilisation de splines. Nous considérons ensuite deux applications à des données issues du domaine des sciences du vivant. Les données considérées sont, pour les premières, des données protéomiques produites à l'aide de la technologie de spectrométrie de masse pour lesquelles l'utilisation d'ondelettes dans un cadre fonctionnel représente un outil d'analyse privilégié. Le deuxième jeu de données considéré est un jeu de données génomiques issu de la technologie des microarray CGH. Concernant ce type de données, la présence d'une variabilité individuelle est une piste de recherche encore peu examinée dans la littérature actuelle et nous proposons ici une première étude qualitative de la classification de ce type de données dans un cadre mixte.

Enfin, nous signalons que les travaux présentés dans cette partie font l'objet d'une publication (Giacofci et al. 2013), et que les méthodes proposées sont implémentées

au sein d'un package `R` appelé `curvclust` disponible librement sur le site du CRAN `http://cran.r-project.org/`.

Chapitre 5

Modèle de mélange mixte fonctionnel

Nous présentons à présent le modèle de classification non supervisée au sein des modèles mixtes utilisé dans la première partie de ce travail. Ce modèle est une extension du modèle de classification fonctionnel et du modèle mixte fonctionnel présentés au Chapitre 4. Au cours de ce chapitre, nous commençons par introduire le modèle complet puis nous décrivons les deux étapes de notre procédure de classification non supervisée consistant en une étape de réduction de dimension et une étape d'ajustement du modèle.

5.1 Présentation du modèle complet

Considérons N individus pour lesquels nous disposons de courbes $(Y_i(t))_{i=1,\dots,N}$ observées au cours du temps en M points équirépartis $\mathbf{t} = (t_1, \dots, t_M)$ et nous supposons que $M = 2^J$ où J est un entier naturel. Sans perte de généralité, nous pouvons supposer que $t_j \in [0,1]$ pour tout $j = 1, \dots, M$. Nous supposons de plus que les individus sont issus de L classes et nous notons $\zeta_{i\ell}$ la variable binaire valant 1 si l'individu i appartient à la classe ℓ et 0 sinon.

Sachant que $\{\zeta_{i\ell} = 1\}$, notre modèle fonctionnel s'écrit alors :

$$Y_i(t_j) = \mu_\ell(t_j) + U_i(t_j) + E_i(t_j), \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, M, \tag{5.1}$$

où μ_ℓ est un effet fixe fonctionnel représentant le comportement moyen des individus au sein de la classe ℓ, U_i est un effet aléatoire fonctionnel propre à l'individu i représentant la déviation de son comportement par rapport au comportement moyen. Dans le domaine fonctionnel, de même qu'au sein du modèle mixte fonctionnel (4.10), U_i est une réalisation d'un processus Gaussien centré de fonction de covariance $K(t,t')$. Enfin, E_i est une réalisation d'un processus Gaussien centré de fonction de covariance $\sigma_E^2 \delta_{tt'}$, avec δ symbole de Kronecker.

Dans un cadre non-paramétrique, James et Sugar (2003) proposent un modèle similaire associé à une représentation de ce dernier dans une base de splines cubiques. Leur modèle est développé dans un contexte de modélisation de courbes régulières

et observées de manière parcimonieuse, c'est-à-dire avec une taille de données raisonnables. Ces deux caractéristiques en font une modélisation non adaptée à notre contexte puisque nous cherchons à modéliser des courbes possédant de fortes irrégularités et pour lesquelles nous disposons d'un grand nombre d'observations. C'est pourquoi notre choix de modélisation s'est dirigé vers l'utilisation d'une représentation du modèle (5.1) dans une base d'ondelettes, particulièrement adaptée à la modélisation de courbes irrégulières en grande dimension.

Ainsi, pour une base d'ondelettes $\{\phi, \psi\}$ choisie, la représentation du modèle (5.1) peut être obtenue dans le domaine des ondelettes sur les coefficients empiriques de la décomposition. Pour tout $i = 1, \ldots, N$, et sachant que $\{\zeta_{i\ell} = 1\}$:

$$\begin{cases} c_{i,00} = \alpha_{\ell,00} + \nu_{i,00} + \varepsilon^c_{i,00}, \\ d_{i,jk} = \beta_{\ell,jk} + \theta_{i,jk} + \varepsilon^d_{i,jk}, \qquad \forall (j,k) \in \Lambda. \end{cases} \tag{5.2}$$

Dans un souci de lisibilité, l'indice 00 concernant les coefficients d'approximation sera omis par la suite. Les hypothèses présentées au Chapitre 4 pour les modèles fonctionnels (4.7) et (4.10) sont conservées pour les coefficients associés aux effets aléatoires fonctionnels et aux erreurs de mesure. À savoir, pour tout $i = 1, \ldots, N$, nous supposons que :

$$\begin{cases} \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2_\varepsilon \mathbf{I}_M), \\ \nu_i \sim \mathcal{N}(0, \gamma^2_\nu), \\ \boldsymbol{\theta}_i \sim \mathcal{N}(0, \mathbf{G}_\theta), \\ (\nu_i^T, \boldsymbol{\theta}_i^T)^T \perp \boldsymbol{\varepsilon}_i, \\ \nu_i \perp \theta_{i,jk}, \qquad \forall (j,k) \in \Lambda, \end{cases}$$

où $\mathbf{G}_\theta$ est une matrice diagonale de terme général $[\mathbf{G}_\theta]_{jk,jk} = 2^{-j\eta}\gamma^2$. Cette hypothèse nous place dans le contexte de modélisation mixte fonctionnelle proposé par Antoniadis et Sapatinas (2007) et détaillé au cours de la Section 4.4.2.

Comme proposé par Morris et Carroll (2006), le modèle peut être étendu aux cas où le paramètre de variance associé aux effets individuels dépend de la position (γ^2_{jk}), de plus nous offrons la possibilité d'une dépendance au groupe (γ^2_ℓ), ou aux deux à la fois $(\gamma^2_{jk\ell})$: en effet, dans un contexte de classification, il peut être utile de permettre des niveaux de variabilité dépendant du groupe. Ceci a pour effet d'élargir la classe de processus atteignables par la modélisation proposée. Nous verrons sur des applications à des données réelles que ce type de modélisation peut conduire à une amélioration des résultats de classification par une meilleure prise en compte de la variabilité inter-individuelle.

Notons tout de même que même si l'on peut définir théoriquement le modèle particulier où la variance est de la forme $\gamma^2_{jk\ell}$, en pratique, cette modélisation conduit en général à des modèles trop riches en paramètres et donc difficiles à ajuster.

5.2 Procédure d'estimation

Nous décrivons à présent la procédure développée pour la classification non supervisée fonctionnelle en présence de variabilité individuelle. La procédure proposée se décompose en deux étapes distinctes.

1. Nous commençons par une étape de réduction de dimension basée sur les techniques de seuillage des ondelettes. La taille conséquente des données considérées tout au long de ce travail rend cette étape nécessaire et notre but est de sélectionner les coefficients les plus informatifs pour la classification.
2. Nous nous intéressons ensuite à l'estimation des paramètres du modèle (4.8) par maximum de vraisemblance. Ceci est réalisé au moyen de l'algorithme EM dans un contexte où deux types de variables latentes sont rencontrées : les variables d'appartenance aux classes et les effets aléatoires individuels.

La description détaillée de ces deux étapes fait l'objet des deux prochaines sections. En dernier lieu, nous présentons le critère utilisé pour le choix du nombre de groupes qui est un critère de type BIC (Schwarz 1978), couramment utilisé dans un contexte de sélection de modèle.

5.2.1 Étape de réduction de dimension

Le problème de la grande dimension des données considérées reste un point critique dans le cas des modèles mixtes fonctionnels, et plus encore dans un cadre de classification non supervisée. Les sources de "bruit" pour la classification sont alors multiples. La présence d'effets aléatoires individuels en est un premier : en effet, du fait de leur structure particulière, des différences dues à une forte variabilité individuelle peuvent être interprétées comme des différences de comportement moyen et introduire des biais dans la classification finale. Par ailleurs, une autre source de bruit vient du fait qu'une partie des positions considérées dans le modèle ne sont pas pertinentes vis-à-vis de l'objectif de classification : en effet, les positions possédant le même niveau pour tous les individus ne permettent pas de discriminer les individus. L'objectif "idéal" serait alors de parvenir à identifier ces positions afin de les retirer de l'analyse comme cela a été réalisé par Antoniadis, Bigot, et von Sachs (2008) dans un cadre de classification fonctionnelle non-supervisée (c.f. Chapitre 4, Section 4.3.3).

Nous proposons d'utiliser la procédure de réduction de dimension suivante, basée sur la stratégie développée par Antoniadis, Bigot, et von Sachs (2008). Notre procédure se décompose en deux étapes distinctes :

1. Effectuer un seuillage individuel dur classique. Pour rappel, cela consiste à mettre à zéro les coefficients d'ondelettes $(d_{i,jk})_{ijk}$ dont la valeur absolue est inférieure au seuil universel $\lambda = \widehat{\sigma}\sqrt{2\log M}$. Pour remplacer $\widehat{\sigma}$, nous choisissons de prendre l'estimateur (3.15) usuellement utilisé dans ce cadre, basé sur le MAD des coefficients au niveau de résolution le plus fin. Cette stratégie est

décrite au Chapitre 3, Section 3.11. Nous obtenons ainsi une estimation robuste de la variance des coefficients au niveau de résolution le plus fin donnée par $\mathbb{V}(d_{i,J-1\,k}) = 2^{-(J-1)\eta}\gamma^2 + \sigma_\varepsilon^2$.

2. Prendre l'union des coefficients sélectionnés durant l'étape de seuillage. Cela a pour effet de retirer les positions dont les coefficients d'ondelettes sont nuls pour tous les individus, ces positions n'apportant, de toute façon, pas d'information pour l'objectif final de classification. De plus, on obtient une sélection de positions qui est commune à tous les individus et comme souligné par Antoniadis, Bigot, et von Sachs (2008), cette stratégie nous assure de disposer de bonnes propriétés concernant l'estimateur résultant.

Par la suite, l'estimation des paramètres du modèle sera réalisée uniquement sur les positions sélectionnées par notre procédure de réduction de dimension. Dans un souci de clarté, nous n'introduisons cependant pas de nouvelles notations.

Signalons en première remarque que notre procédure de réduction de dimension n'atteint pas l'objectif idéal décrit en introduction de cette section : les positions pour lesquelles les niveaux des effets fixes de chaque classe sont égaux ne sont pas identifiées ici. Cette idée correspond à la troisième étape de la stratégie proposée par Antoniadis, Bigot, et von Sachs (2008) basée sur l'utilisation d'un test de Neyman sur les différences individuelles successives à une position donnée. Cependant, ce test n'est pas adapté au type de données considérées dans notre contexte. En effet, ce test est particulièrement adapté à l'étude d'images fonctionnelles car sa puissance est directement liée au caractère creux des vecteurs étudiés. Dans le cas de l'analyse d'images structurées, les individus sont représentés par des pixels et une forte proportion de pixels successifs sont alors similaires, impliquant un vecteur des différences parcimonieux.

Une deuxième remarque concerne le seuillage individuel réalisé durant la première étape. Une stratégie plus adaptée consisterait à prendre à compte la présence d'effets aléatoires, c'est-à-dire, la présence d'un terme de variabilité individuelle, en considérant un seuillage dépendant du niveau de résolution. Cela permettrait d'effectuer une réduction de dimension plus conséquente, grâce à la prise en compte d'une plus grande partie de la variance. Cependant, la mise en place d'une telle stratégie nécessite de disposer d'estimateurs des variances σ_ε^2 et γ^2. Dans un cadre de classification non supervisée, l'estimation de ces deux paramètres revient à un problème d'estimation de variances lorsque les moyennes sont inconnues et différentes. Étant donné que l'on ne dispose d'aucune information sur les labels des individus dans cette première étape, ce problème est difficile à traiter sans sacrifier à la rapidité d'exécution de l'étape de réduction de dimension.

5.2.2 Estimation des paramètres

Les paramètres du modèle (5.2) donnés par $(\boldsymbol{\pi}, \mathbf{G}_\theta, \gamma_\nu^2, \gamma^2, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2)$ sont ensuite estimés par maximum de vraisemblance au moyen de l'algorithme EM (cf. Section

4.2.2). Dans le contexte considéré, nous sommes face à deux types de données non observées : les variables d'appartenance aux groupes ainsi que les effets aléatoires individuels. L'algorithme EM permet naturellement de considérer les deux types de variables latentes et plus particulièrement, de les traiter de manière séparée grâce à la relation suivante sur la log-vraisemblance des données complètes. Cette dernière se décompose en effet en trois termes distincts :

$$\log \mathcal{L}(\mathbf{d}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\zeta}; \boldsymbol{\pi}, \gamma_\nu^2, \gamma^2, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2) = \log \mathcal{L}(\mathbf{d}, \mathbf{c}|\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\zeta}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2) + \log \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\nu}; \gamma^2, \gamma_\nu^2) + \log \mathcal{L}(\boldsymbol{\zeta}; \boldsymbol{\pi}), \quad (5.3)$$

dont les différents termes s'expriment de la manière suivante :

$$\log \mathcal{L}(\boldsymbol{\zeta}; \boldsymbol{\pi}) = \sum_{i=1}^{N} \sum_{\ell=1}^{L} \zeta_{i\ell} \log \pi_\ell, \quad (5.4)$$

$$-2 \log \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\nu}; \gamma_\nu^2, \gamma^2) = \text{Cste} + N \log |\mathbf{G}_\theta| + \sum_{i=1}^{N} \boldsymbol{\theta}_i^T \mathbf{G}_\theta^{-1} \boldsymbol{\theta}_i + N \log \gamma_\nu^2 + \gamma_\nu^2 \sum_{i=1}^{N} \nu_i^2, \quad (5.5)$$

$$-2 \log \mathcal{L}(\mathbf{d}, \mathbf{c}|\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\zeta}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2) = \text{Cste} + NM \log \sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \zeta_{i\ell} \left\{ [\mathbf{d}_i - \boldsymbol{\beta}_\ell - \boldsymbol{\theta}_i]^T [\mathbf{d}_i - \boldsymbol{\beta}_\ell - \boldsymbol{\theta}_i] + [c_i - \alpha_\ell - \nu_i]^2 \right\}. \quad (5.6)$$

L'algorithme EM consiste alors à maximiser l'espérance de ces différentes quantités conditionnellement aux données observées et se décompose en deux étapes distinctes : le calcul des espérances conditionnelles réalisé dans l'étape E et la maximisation de ces dernières dans l'étape M.

Prédiction des variables latentes

L'étape E se ramène à effectuer une prédiction des variables latentes, et donc dans ce cas particulier, au calcul des probabilités *a posteriori* d'appartenances aux groupes et au calcul des prédictions des effets aléatoires ainsi que de leurs variances conditionnellement aux données observées. Les probabilités *a posteriori* d'appartenances aux groupes sont notées :

$$\tau_{i\ell} = \mathbb{E}[\zeta_{i\ell}|\mathbf{d}_i], \qquad \forall i = 1, \ldots, N, \quad \forall \ell = 1, \ldots, L,$$

et ces quantités sont prédites à l'itération $[h]$ par :

$$\tau_{i\ell}^{[h+1]} = \frac{\pi_\ell^{[h]} f\left(c_i, \mathbf{d}_i; \alpha_\ell^{[h]}, \beta_\ell^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2\,[h]} \mathbf{I}_M\right)}{\sum_{p=1}^{L} \pi_p^{[h]} f\left(c_i, \mathbf{d}_i; \alpha_p^{[h]}, \beta_p^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2\,[h]} \mathbf{I}_M\right)}, \tag{5.7}$$

où f est la fonction de densité associée à une distribution Gaussienne et $\mathbf{G} = \mathrm{diag}(\gamma_\nu^2, \mathbf{G}_\theta)$.

A ce stade, en utilisant la formule d'Henderson (1975), les prédictions des effets aléatoires à l'itération $[h]$ sont alors données par :

$$\begin{aligned}
\widehat{\theta}_{i\ell,jk}^{[h+1]} &= \mathbb{E}[\theta_{i,jk} | \zeta_{i\ell} = 1, d_{i,jk}] = \frac{[d_{i,jk} - \beta_{\ell,jk}]}{1 + 2^{j\eta} \sigma_\varepsilon^{2\,[h]} / \gamma^{2\,[h]}}, \qquad \forall (j,k) \in \Lambda, \\
\widehat{\nu}_{i\ell}^{[h+1]} &= \mathbb{E}[\nu_i | \zeta_{i\ell} = 1, c_i] = \frac{[c_i - \alpha_\ell]}{1 + 2^{j\eta} \sigma_\varepsilon^{2\,[h]} / \gamma_\nu^{2\,[h]}}.
\end{aligned} \tag{5.8}$$

Ces quantités sont usuellement appelées BLUP (Best Linear Unbiaised Prediction) dans le contexte des modèles mixtes.

Remarquons que les prédictions des effets aléatoires (5.8) ainsi que de leurs variances conditionnellement aux données observées nécessitent de prendre en compte la dépendance aux groupes. En effet, même si la variable $\boldsymbol{\theta}_i$ est indépendante du groupe auquel appartient l'individu i, le conditionnement des variables $\boldsymbol{\theta}_i$ par les données $\mathbf{d}_i$ fait apparaître une dépendance à la classe de l'individu i. La justification est donnée ci-dessous :

$$\begin{aligned}
\mathbb{E}[\log \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\nu}; \gamma^2, \gamma_\nu^2) | \mathbf{d}] &= \sum_{i=1}^{N} \mathbb{E}[\log \mathcal{L}(\boldsymbol{\theta}_i, \nu_i; \gamma^2, \gamma_\nu^2) | \mathbf{d}_i] \\
&= \sum_{i=1}^{N} \int \log \mathcal{L}(\boldsymbol{\theta}_i, \nu_i; \gamma^2, \gamma_\nu^2) f(\boldsymbol{\theta}_i, \nu_i; \gamma^2, \gamma_\nu^2 | \mathbf{d}_i) \mathrm{d}\theta \\
&= \sum_{i=1}^{N} \sum_{\ell=1}^{L} \mathbb{P}(\zeta_{i\ell} = 1 | \mathbf{d}_i) \mathbb{E}[\log \mathcal{L}(\boldsymbol{\theta}_i, \nu_i) | \mathbf{d}_i, \zeta_{i\ell} = 1].
\end{aligned}$$

Par ailleurs, l'expression de la variance des effets aléatoires conditionnellement aux données observées est alors donnée par :

$$\begin{aligned}
\mathbb{V}[\theta_{i,jk} | d_{i,jk}] &= \frac{\sigma_\varepsilon^{2\,[h]}}{1 + 2^{j\eta} \sigma_\varepsilon^{2\,[h]} / \gamma^{2\,[h]}}, \qquad \forall (j,k) \in \Lambda, \\
\mathbb{V}[\nu_i | c_i] &= \frac{\sigma_\varepsilon^{2\,[h]}}{1 + \sigma_\varepsilon^{2\,[h]} / \gamma_\nu^{2\,[h]}}.
\end{aligned}$$

Nous pouvons remarquer que les variances *a posteriori* ne dépendent, quant à elles, pas des classes individuelles.

Mise à jour des paramètres

Les prédictions calculées au cours de cette première étape (étape E) permettent alors le calcul effectif des espérances conditionnelles (5.4), (5.5) et (5.6) et permettent ainsi la mise à jour des paramètres à l'itération $[h+1]$. Ceci est réalisé par la maximisation de ces espérances conditionnelles au cours de l'étape M. La mise à jour des poids est alors donnée par :

$$\widehat{\pi}_\ell^{[h+1]} = \frac{1}{N}\sum_{i=1}^{N} \tau_{i\ell}^{[h+1]} \qquad \forall \ell \in \{1,\ldots,L\}. \tag{5.9}$$

Vient ensuite la mise à jour des variances γ^2 et γ_ν^2 :

$$\widehat{\gamma}^{2\,[h+1]} = \frac{1}{N(M-1)}\sum_{ijk\ell} 2^{j\eta}\tau_{i\ell}^{[h+1]}\left[\widehat{\theta}_{i\ell,jk}^{2\,[h+1]} + \left(\frac{\sigma_\varepsilon^{2\,[h]}}{1+2^{j\eta}\sigma_\varepsilon^{2\,[h]}/\gamma^{2\,[h]}}\right)\right], \tag{5.10}$$

$$\widehat{\gamma}_\nu^{2\,[h+1]} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\nu}_{i\ell}^{2\,[h+1]} + \frac{\sigma_\varepsilon^{2\,[h]}}{1+\sigma_\varepsilon^{2\,[h]}/\gamma_\nu^{2\,[h]}}\right]. \tag{5.11}$$

Enfin les mises à jour des paramètres de moyennes $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_\ell^{[h+1]} = \left[\sum_{i=1}^{N}\tau_{i\ell}^{[h+1]}\right]^{-1}\sum_{i=1}^{N}\tau_{i\ell}^{[h+1]}\left[\mathbf{d}_i - \widehat{\boldsymbol{\theta}}_{i\ell}^{[h+1]}\right], \qquad \forall \ell = 1,\ldots,L, \tag{5.12}$$

$$\widehat{\alpha}_\ell^{[h+1]} = \left[\sum_{i=1}^{N}\tau_{i\ell}^{[h+1]}\right]^{-1}\sum_{i=1}^{N}\tau_{i\ell}^{[h+1]}\left[c_i - \widehat{\nu}_{i\ell}^{[h+1]}\right], \tag{5.13}$$

et de la variance du bruit :

$$\widehat{\sigma}_\varepsilon^{2\,[h+1]} = \frac{1}{NM}\sum_{i,\ell}\tau_{i\ell}^{[h+1]}\left\{\sum_{j,k}\left[d_{i,jk} - \beta_{\ell,jk}^{[h+1]} - \widehat{\theta}_{i\ell,jk}^{[h+1]}\right]^2 + \sum_{j,k}\frac{\sigma_\varepsilon^{2\,[h]}}{1+2^{j\eta}\sigma_\varepsilon^{2\,[h]}/\gamma^{2\,[h]}}\right.$$
$$\left.\left[c_i - \alpha_\ell^{[h+1]} - \widehat{\nu}_{i\ell}^{[h+1]}\right]^2 + \frac{\sigma_\varepsilon^{2\,[h]}}{1+\sigma_\varepsilon^{2\,[h]}/\gamma_\nu^{2\,[h]}}\right\}. \tag{5.14}$$

Les mises à jour ci-dessus sont données dans le modèle simple pour lequel le paramètre de variance des effets aléatoires est constant (γ^2). On peut aisément dériver les mêmes calculs dans le cas où ce paramètre dépend du groupe (γ_ℓ^2), de la position (γ_{jk}^2) ou des deux à la fois ($\gamma_{\ell jk}^2$). Les formules de mises à jour sont alors données par :

$$\widehat{\gamma}_{\ell}^{2\,[h+1]} = \frac{1}{N(M-1)} \left[\sum_{i=1}^{N} \tau_{i\ell}^{[h+1]}\right]^{-1} \sum_{ijk} 2^{j\eta} \tau_{i\ell}^{[h+1]} \left[\widehat{\theta}_{i\ell,jk}^{2\,[h+1]} + \left(\frac{\sigma_{\varepsilon}^{2\,[h]}}{1 + 2^{j\eta}\sigma_{\varepsilon}^{2\,[h]}/\gamma_{\ell}^{2\,[h]}}\right)\right],$$

$$\widehat{\gamma}_{jk}^{2\,[h+1]} = \frac{1}{N} \sum_{i\ell} 2^{j\eta} \tau_{i\ell}^{[h+1]} \left[\widehat{\theta}_{i\ell,jk}^{2\,[h+1]} + \left(\frac{\sigma_{\varepsilon}^{2\,[h]}}{1 + 2^{j\eta}\sigma_{\varepsilon}^{2\,[h]}/\gamma_{jk}^{2\,[h]}}\right)\right],$$

$$\widehat{\gamma}_{\ell,jk}^{2\,[h+1]} = \left[\sum_{i=1}^{N} \tau_{i\ell}^{[h+1]}\right]^{-1} \sum_{i=1}^{N} 2^{j\eta} \tau_{i\ell}^{[h+1]} \left[\widehat{\theta}_{i\ell,jk}^{2\,[h+1]} + \left(\frac{\sigma_{\varepsilon}^{2\,[h]}}{1 + 2^{j\eta}\sigma_{\varepsilon}^{2\,[h]}/\gamma_{\ell,jk}^{2\,[h]}}\right)\right].$$

Les étapes E et M sont répétées itérativement jusqu'à convergence de l'algorithme, l'arrêt de ce dernier reposant alors sur les différences relatives des valeurs estimées des paramètres entre deux itérations. Les classes des individus sont alors déduites des estimations finales des probabilités d'appartenance $(\tau_{i\ell})_{\ell=1,\ldots,L}^{i=1,\ldots,N}$ par une règle de Maximum A Posteriori (MAP).

L'estimation du paramètre de régularité η se ramène à une problématique d'estimation de régularité d'un processus dont un échantillon de trajectoires est observé. Cette problématique reste à ce jour une question ouverte et non triviale mais il est nécessaire de pouvoir en proposer une estimation afin de définir une procédure complètement fondée sur les données. Nous proposons donc d'estimer ce paramètre par maximisation de la vraisemblance en considérant une grille de valeurs sur laquelle nous effectuons une recherche dichotomique selon un algorithme de *golden search* (Kiefer 1953).

Enfin, signalons que les mises à jour du modèle de classification de courbes (4.8) et du modèle mixte fonctionnel (4.11) présentés au Chapitre 4 peuvent être déduites directement de celles présentées dans cette section en fixant, respectivement, $(\gamma_{\nu}^2, \gamma^2) = (0, 0)$ ou $L = 1$.

Initialisation des paramètres

Comme abordé auparavant, l'algorithme EM est sensible à l'initialisation des paramètres. Les principales techniques d'initialisation ont été décrites en Section 4.2.2. Nous nous concentrons essentiellement sur deux stratégies proposées au sein du package `curvclust` : la procédure SEM permettant une exploration large de l'espace des paramètres mais nécessitant un effort numérique important et la stratégie notée r-EM, moins coûteuse numériquement et connue pour conduire à des niveaux de vraisemblance légèrement plus élevé mais pouvant se montrer inadaptée par sa propension à conduire à l'obtention de groupes vides (Biernacki et al. 2003).

5.2.3 Choix du nombre de groupes - Bayesian Information Criteria

Le choix d'un critère de sélection du nombre de composantes dans un modèle de mélange gaussien est une problématique qui se place dans le contexte plus large de la sélection de modèle et se ramène au dilemme récurrent du compromis entre biais et variance : pour un modèle trop simple, les estimateurs exhibent un fort biais tandis que pour un modèle trop compliqué, on obtient alors des résultats fortement variables.

Nous choisissons d'adopter pour cela le critère de sélection BIC (Bayesian Information Criteria), largement utilisé dans le contexte de classification non-supervisée. Soit $\mathcal{M}$ une famille de modèles représentant l'ensemble des nombres de classes possibles variant dans $[1, L_{\max}]$. Le critère BIC vise à la recherche du modèle le plus vraisemblable parmi la famille $\mathcal{M}$ tout en le pénalisant par un terme visant à limiter le nombre de paramètres du modèle choisi afin d'éviter les problèmes de surajustement.

Ainsi, le critère BIC, défini à partir de la log-vraisemblance des données observées, est donné par :

$$\text{BIC}(\mathbf{m}_L[\gamma^2]) = \log \mathcal{L}\left(\mathbf{c}, \mathbf{d}; \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{G}}, \widehat{\sigma}^2_\varepsilon\right) - \frac{|\mathbf{m}_L[\gamma^2]|}{2} \times \log(N), \tag{5.15}$$

où $\mathbf{m}_L[\gamma^2]$ est le modèle à L classes avec un paramètre de variance γ^2 choisi constant et $|\mathbf{m}_L[\gamma^2]|$ la dimension du modèle, c'est-à-dire, le nombre de paramètres à estimer au sein de ce modèle. On définit de la même manière $\mathbf{m}_L[\gamma^2_{jk}]$, $\mathbf{m}_L[\gamma^2_\ell]$ et $\mathbf{m}_L[\gamma^2_{jk\ell}]$. De manière générale, on a $|\mathbf{m}_L[\gamma^2_\cdot]| = (M+1)L + |G|$ où $|G|$ est le nombre de paramètres libres dans la matrice G, dépendant de la structure de variance choisie pour les effets aléatoires individuels.

Le critère BIC est particulièrement adapté à la problématique de sélection du nombre de groupes dans un contexte de mélange car c'est un critère consistant, c'est-à-dire qu'il converge vers le bon modèle si celui-ci appartient à la famille $\mathcal{M}$ (Keribin 2000). On observe, de plus, un bon comportement de ce critère en pratique dans un cadre non-asymptotique. Cependant, une des limitations de ce critère dans le cadre d'une modélisation mixte est qu'en se basant sur la vraisemblance des données observées, il ne prend pas explicitement en compte la présence d'effets aléatoires. À ce propos, nous signalons qu'un autre critère de choix du nombre de groupes basé sur la vraisemblance des données complétées a été développé au sein de l'article de Giacofci et al. (2013) mais le critère BIC semble tout de même plus performant, en pratique, sur les simulations réalisées.

Finalement, notre procédure globale se décompose comme suit :

1. Projection du modèle fonctionnel (5.1) dans le domaine des ondelettes afin de se ramener à un mélange de modèles linéaires mixtes gaussiens sur les coefficients de la décomposition dans une base d'ondelettes,
2. Réduction de la dimension du modèle par une stratégie basée sur un seuillage des coefficients d'ondelettes,
3. Estimation des paramètres du modèle par maximum de vraisemblance au moyen de l'algorithme EM,
4. Déduction d'une classification finale des individus par une règle du Maximum A Posteriori.

Le choix du nombre de groupes est quant à lui réalisé *a posteriori* sur la base du critère BIC présenté en Section 5.2.3.

Chapitre 6

Applications

Dans ce chapitre, nous présentons d'une part, une étude de simulation destinée à étudier les propriétés de notre modèle sur la base de données simulées réalistes. D'autre part, nous proposons deux applications sur données réelles concernant un jeu de données issu de la spectrométrie de masse et un jeu de données issu de la technologie des biopuces à ADN (données de microarray CGH). L'ensemble des codes développés et des données considérées sont disponibles à l'adresse `http://pbil.univ-lyon1.fr/members/fpicard/software.html`.

6.1 Étude de simulation

6.1.1 Cadre de simulation

Afin de réaliser une étude de simulation approfondie permettant d'évaluer le comportement général de notre modèle, nous nous sommes appliqués à définir un cadre de simulation unifié permettant la construction automatique de jeux de données simulés réalistes.

Contrôle des niveaux de variabilité Notre première tâche a été de définir le rapport signal sur bruit, noté par la suite SNR pour Signal to Noise Ratio, dans un cadre mixte fonctionnel et en présence de plusieurs groupes. Usuellement, cette quantité est définie comme le ratio entre la puissance du signal moyen et la puissance du bruit l'affectant. Une des difficultés dans notre contexte réside dans la présence d'effets aléatoires qui peuvent être vus comme partie intégrante du signal individuel ou comme composante du terme d'erreur globale. Compte tenu de la modélisation adoptée faisant l'hypothèse d'effets aléatoires individuels structurés (cf Section 4.4.2, Chapitre 4), nous considérons qu'ils représentent une part du signal individuel.

Dans ce contexte, la puissance du bruit de mesure est définie d'une part comme la variance fonctionnelle $\sigma_E^2 = M\sigma_\varepsilon^2$. Par ailleurs, la puissance du signal moyen est

définie comme :

$$\lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}\sum_{\ell}\pi_\ell\mathbb{E}\big[|\mu_\ell(t)+U_i(t)|^2\big]\mathrm{d}t=\frac{1}{M}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{\ell,j_0k}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{\ell,jk}^2\right)$$
$$+2^{j_0}\gamma_\nu^2+\frac{2^{j_0(1-\eta)}\gamma_\theta^2}{1-2^{(1-\eta)}}.$$

En effet, en considérant des fonctions à support dans l'intervalle $[0,1]$ et des effets aléatoires fonctionnels $(U_i(t))_{i=1,\dots,N}$ qui sont réalisations d'un processus gaussien centré, la puissance du signal peut être décomposée en :

$$\lim_{T\to\infty}\frac{1}{T}\int_{\frac{T}{2}}^{-\frac{T}{2}}\sum_{\ell}\pi_\ell\mathbb{E}\big[|\mu_\ell(t)+U_i(t)|^2\big]\mathrm{d}t=\sum_{\ell}\pi_\ell\int_0^1|\mu_\ell(t)|^2\mathrm{d}t+\sum_{\ell}\pi_\ell\int_0^1\mathbb{E}\big[U_i(t)^2\big]\mathrm{d}t.$$

Par la formule de conservation de l'énergie (3.5), la puissance associée aux effets fixes fonctionnels est donnée sur les coefficients d'ondelettes par :

$$\sum_{\ell}\pi_\ell\int_0^1|\mu_\ell(t)|^2\mathrm{d}t=\frac{1}{M}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{\ell,j_0k}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{\ell,jk}^2\right).$$

De plus, en utilisant l'orthonormalité des bases d'ondelettes, la puissance associée aux effets aléatoires est donnée par :

$$\begin{aligned}\int_0^1\mathbb{E}[U_i(t)^2]\mathrm{d}t&=\int_0^1\mathbb{E}\left[\left(\sum_{k=0}^{2^{j_0}-1}\nu_{i,j_0k}\phi_{j_0k}(t)+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\theta_{i,jk}\psi_{jk}(t)\right)^2\right]\mathrm{d}t\\&=\sum_{k=0}^{2^{j_0}-1}\gamma_\nu^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}2^{-j\eta}\gamma_\theta^2\\&=2^{j_0}\gamma_\nu^2+\frac{2^{j_0(1-\eta)}\gamma_\theta^2}{1-2^{(1-\eta)}}.\end{aligned}$$

Nous observons alors que la puissance du signal moyen se décompose en deux termes distincts associés respectivement aux effets fixes et aléatoires et nécessite de ce fait des contrôles séparés de ces deux quantités. A cet effet, nous introduisons deux paramètres distincts de contrôle du niveau d'aléa : le rapport SNR associé aux effets fixes fonctionnels

$$\mathrm{SNR}^2=\frac{1}{M\sigma_E^2}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{\ell,j_0k}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{\ell,jk}^2\right),\tag{6.1}$$

ainsi que le paramètre τ_U, associé dans le contexte des modèles mixtes au rapport entre le niveau de bruit et le niveau d'effet aléatoire fonctionnel

$$\tau_{\mathrm{U}} = \sigma_E^2 / \left(\gamma_\nu^2 + \frac{\gamma_\theta^2}{1 - 2^{(1-\eta)}} \right). \tag{6.2}$$

Par rapport aux paramètres de variances σ_ε^2 et γ^2, on obtient alors les relations équivalentes suivantes :

$$\begin{cases} \sigma_\varepsilon^2 = \dfrac{1}{M^2(\mathrm{SNR})^2} \displaystyle\sum_{\ell=1}^{L} \pi_\ell \left(\sum_{k=0}^{2^{j_0}-1} \alpha_{\ell,j_0k}^2 + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \beta_{\ell,jk}^2 \right), \\ \gamma^2 = \dfrac{M\sigma_\varepsilon^2}{\tau_U \left(1 + \frac{1}{1-2^{(1-\eta)}} \right)}. \end{cases} \tag{6.3}$$

Les valeurs classiquement utilisées pour le paramètre SNR varient dans $\{0.1, 1, 3, 5, 7\}$ et dans $\{1/4, 1, 4\}$ pour le paramètre τ_U ; une faible valeur de τ_U indiquant alors un fort niveau d'effets individuels.

Construction des effets fixes fonctionnels Une deuxième tâche est ensuite de construire des effets fixes représentant les comportements fonctionnels moyens au sein des différents groupes. Pour ce faire, nous avons généralisé une approche proposée par Amato et Sapatinas (2005) faisant appel aux fonctions `Blocks`, `Bumps`, `Heavisine` et `Doppler` proposées originellement par Donoho et Johnstone (1994) ; ces fonctions sont représentées en Figure 6.1. Amato et Sapatinas (2005) se place dans une problématique d'estimation d'une courbe de référence au sein de données fonctionnelles se ramenant alors, sans le formaliser explicitement, à une problématique d'estimation de l'effet fixe au sein d'un modèle mixte fonctionnel. Dans ce contexte, les auteurs développent un cadre de simulation unifié permettant de construire des courbes partageant des caractéristiques communes et basées sur les fonctions de Donoho et Johnstone (1994). Nous reprenons cette idée dans le cadre de la définition de L effets fixes fonctionnels en utilisant les formules suivantes, pour $t \in [0, 1]$ et $\ell = 1, \ldots, L$.

$$\begin{aligned} \mu_\ell^{\texttt{Blocks}}(t) &= 10 \sum_{r=1}^{11} \left(1 + \frac{1}{2} h_r^\ell \mathrm{sign}(t - v_r^\ell) \right), \\ \mu_\ell^{\texttt{Bumps}}(t) &= \sum_{r=1}^{11} h_r^\ell / \left(1 + \frac{|t - v_r|}{w_r^\ell} \right)^4, \\ \mu_\ell^{\texttt{Heavisine}}(t) &= 4 \sin(4\pi t) - \mathrm{sign}(t - v_1^\ell) - \mathrm{sign}(v_2^\ell - t), \\ \mu_\ell^{\texttt{Doppler}}(t) &= \sqrt{t(1-t)} \sin\left(2.1\pi / (t - t_0^\ell) \right), \end{aligned}$$

où v_r^ℓ représentent les instants de sauts, choisis aléatoirement dans $[0, 1]$, pour les fonctions `Blocks` et `Bumps`, tandis que h_r^ℓ et w_r^ℓ en représentent respectivement les

hauteurs des sauts et les largeurs des pics. Les scalaires v_1^ℓ, v_2^ℓ sont les places des discontinuités pour la fonction `Heavisine` et t_0^ℓ représente la phase associée à la fonction `Doppler` et ces quantités sont, de même, tirées aléatoirement dans $[0, 1]$.

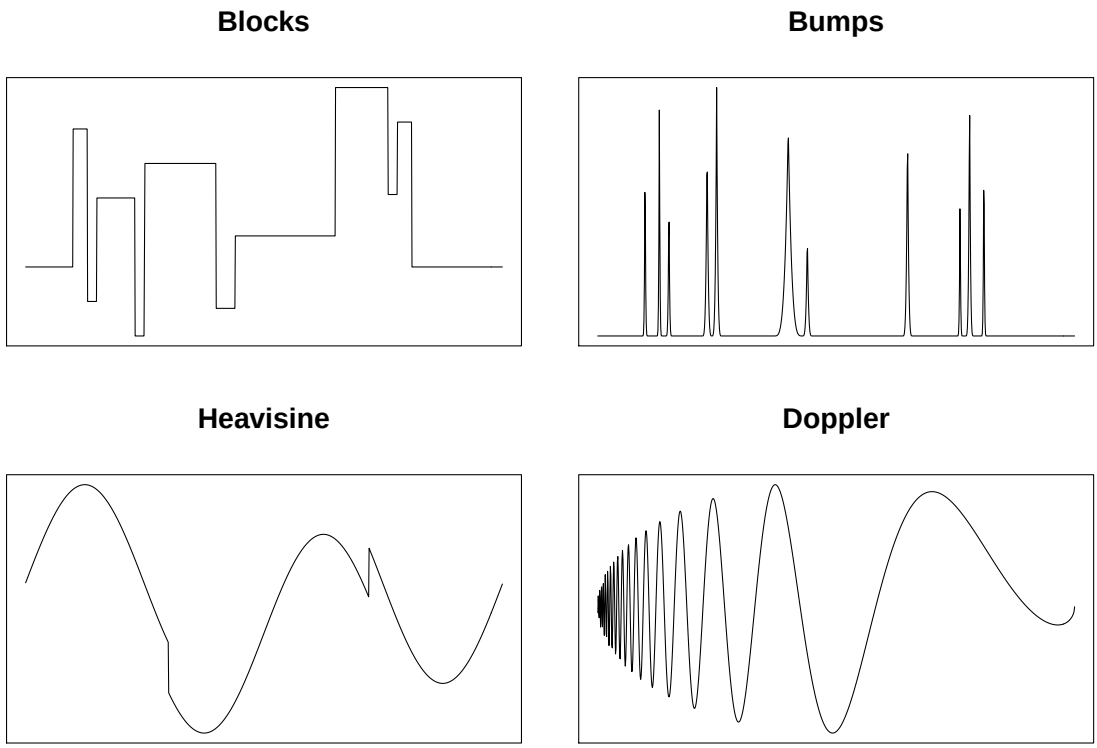


FIGURE 6.1 – *Représentations des fonctions* `Blocks`, `Bumps`, `Heavisine` *et* `Doppler` *proposées par Donoho et Johnstone (1994).*

Construction de données synthétiques et plan de simulation À ce stade, les valeurs des coefficients d'approximation et d'ondelettes associés aux effets fixes fonctionnels $(\alpha_\ell, \boldsymbol{\beta}_\ell)$ peuvent être déduites des L effets fixes fonctionnels construits. De même, les valeurs des paramètres σ_ε^2 et γ^2 peuvent être déduites des relations (6.3) pour des valeurs fixées de SNR et τ_U (dans un souci de simplification, on fixe $\gamma_\nu^2 = \gamma^2$). La simulation des données est alors réalisée dans le domaine des ondelettes en ajoutant aux coefficients $(\alpha_\ell, \boldsymbol{\beta}_\ell)$ des réalisations de lois gaussiennes centrées de variance γ^2 pour représenter les effets aléatoires individuels et de variance σ_ε^2 pour l'erreur de mesure. Les jeux de données fonctionnels peuvent ensuite être obtenus à l'aide de la transformée en ondelettes discrètes inverse. Des exemples de données concernant des effets fixes de type `Blocks` et `Bumps` sont représentés en Figure 6.2 pour deux valeurs de SNR (1 et 5) et deux valeurs de τ_U (0.25 et 4).

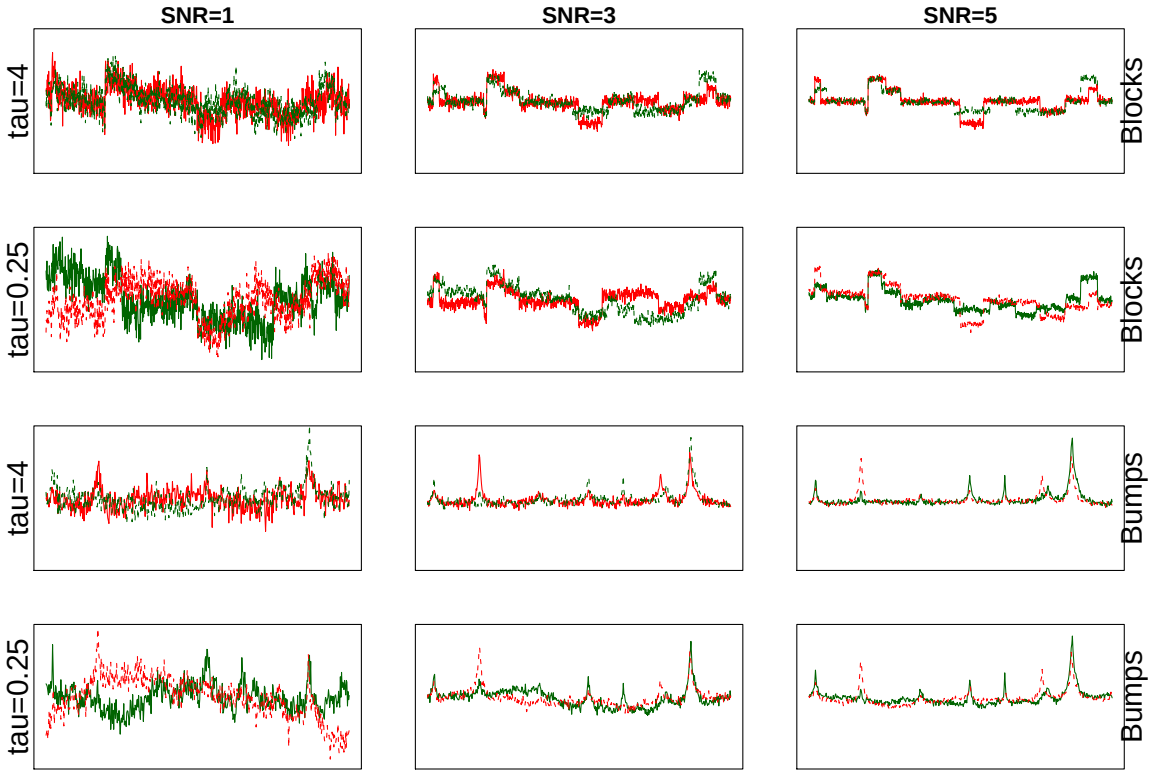


FIGURE 6.2 – *Exemples de courbes simulées pour différentes valeurs de SNR et λ_U (Une seule courbe par groupe est représentée).*

Nous proposons, de plus, de fixer le nombre d'individus à $N = 50$ et le nombre de groupes à $L = 2$ ou 4. Les signaux sont construits avec $M = 512$ points de discrétisations et le paramètre η est fixé à 2. Les jeux de données simulés sont ensuite construits en considérant SNR $\in \{0.1, 1, 3, 5, 7\}$, $\tau_U \in \{0.25, 1, 4\}$ et enfin, $\pi \in \{0.1, 0.25, 0.5\}$, où π est la taille de l'un des deux groupes simulés (dans le cas où $L = 4$, on ne considère que des groupes équilibrés en prenant $\pi = 0.25$). Enfin, nous simulons 50 jeux de données par configuration.

Procédures testées Sur l'ensemble des données simulées, 5 procédures sont mises en compétition : notre procédure de classification de courbes, considérée avec ou sans prise en compte des effets aléatoires individuels (notées respectivement par la suite FCMM/FCM pour Functional Clustering Mixed Model/Functional Clustering Model) afin de souligner le bénéfice à prendre en compte la présence de variabilité individuelle. De plus, ces deux méthodes sont considérées chacune avec ou sans réduction de dimension afin d'en évaluer l'efficacité et l'apport en terme de discrimination. Enfin, ces procédures sont comparées à la procédure de classification de

courbes dans un contexte mixte non-supervisé proposée par James et Sugar (2003) dont le code `R` est disponible librement [1]. L'objectif est ici de mettre en avant l'avantage à employer des techniques basées sur l'utilisation d'ondelettes dans le cadre de l'étude de données irrégulières en grande dimension. En remarque, nous signalons que notre étude est limitée à une taille de signal de $M = 512$ pour des raisons de contraintes mémoires liées à l'utilisation de la procédure Spline.

Critères de comparaison Les performances des procédures testées sont évaluées, au vu de l'objectif principal de classification non supervisée, par rapport au taux d'individus mal classés en se basant sur le critère EER (Empirical Error Rate) défini comme suit :

$$\text{EER} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell}^{L}\mathbb{I}\{\widehat{\zeta}_{i\ell}^{\text{MAP}} \neq \zeta_{i\ell}\}, \tag{6.4}$$

où $\widehat{\zeta}_{i\ell}^{\text{MAP}}$ est la classe prédite pour l'individu i par la règle du Maximum A Posteriori, tandis que $\zeta_{i\ell}$ représente la vraie classe de l'individu i. Le critère EER varie entre 0 et 1, 0 signifiant qu'aucune erreur de classification n'a été faite et 1 signifiant que l'ensemble des individus est mal classé.

Nous considérons, de plus, comme critère de comparaison, les temps d'exécution de chaque procédure (noté TOE pour Time Of Execution).

6.1.2 Résultats de simulation

Nous présentons à présent les résultats de simulations obtenus sur les configurations testées.

Dans une configuration à deux groupes équilibrés

Les résultats vis-à-vis du critère EER dans une configuration à deux groupes équilibrés sont présentés en Figure 6.3 pour les différents types d'effets fixes, de niveaux d'effets aléatoires (une grande valeur de τ_U indiquant un faible niveau d'effets aléatoires) et en fonction de la valeur de SNR. Chaque courbe correspond à une procédure testée et les procédures offrant une étape de réduction de dimension sont différenciées à l'aide de la notation "u". L'observation de ces résultats nous permet de tirer plusieurs conclusions.

La prise en compte de la présence d'effets aléatoires conduit à une amélioration des performances de classification : en effet, dans l'ensemble des configurations étudiées, on peut observer que les méthodes considérant la présence d'effets aléatoires (FCMM/FCMMu), avec ou sans étape de réduction de dimension, donnent lieu à

1. `http://www-bcf.usc.edu/~gareth/`

des taux d'individus mal classés inférieurs ou égaux à ceux obtenus par les autres procédures. On peut alors en conclure que la prise en compte des effets aléatoires au sein de la procédure de classification non-supervisée permet de mieux appréhender la structure discriminante portée par les effets fixes.

L'étape de réduction de dimension permet d'augmenter les performances de classification des procédures prenant en compte les effets aléatoires en permettant l'élimination des coefficients n'apportant pas d'informations vis-à-vis de la classification. Ce n'est pas le cas pour les procédures sans effet aléatoire FCM/FCMu (avec/sans réduction de dimension) : en effet, pour celles-ci, l'étape de réduction de dimension a tendance à détériorer la classification finale, particulièrement dans des configurations avec de forts effets aléatoires. Cette tendance peut s'expliquer par l'important biais d'estimation de la variance résiduelle observé quand le modèle ne prend pas en compte la présence d'effets aléatoires, et ce biais est d'autant plus accentué par l'étape de réduction de dimension. Ce comportement est illustré en Table 6.1 où sont donnés les biais relatifs obtenus sur le paramètre de variance résiduelle σ_ε^2 estimé. On observe que ce biais est bien plus modéré pour les méthodes FCMM/FCMMu.

Notre dernière conclusion porte sur les résultats peu concluants obtenus par la procédure basée sur les splines (notée Spline) : en effet, bien qu'étant une procédure prenant en compte la présence d'effets aléatoires, on observe en Figure 6.3 qu'elle exhibe de mauvais résultats de classification dans l'ensemble des configurations étudiées. Ce résultat, que l'on pouvait attendre, souligne le fait que l'utilisation de splines n'est pas adaptée à la modélisation de données fonctionnelles irrégulières : leurs structures régulières ne permettent de détecter les structures discriminantes au sein des données correspondant aux effets fixes fonctionnels moyens. De plus, on observe un temps d'exécution de la procédure Spline nettement supérieur à ceux des autres procédures mettant en avant la difficulté des splines à gérer les données de grande dimension.

Dans une configuration non équilibrée

Afin d'évaluer le comportement de notre procédure dans le cas d'une étude à deux groupes non équilibrés, nous nous sommes placés dans des configurations où les groupes sont répartis selon les proportions 0.25/0.75 et 0.1/0.9. Les résultats en termes de classification sont présentés en Figure 6.4 et en Figure 6.5 respectivement.

On observe sur ces deux graphiques que les principales conclusions effectuées dans le cas de groupes équilibrés restent valables dans le cas de groupes non équilibrés. Cependant, on observe tout de même une légère dégradation des résultats de classification, principalement visible pour la configuration extrême 0.1/0.9 pour des valeurs élevées de SNR, c'est-à-dire, en présence de faibles erreur de mesure, montrant que la présence de groupes déséquilibrés représente plutôt un désavantage pour notre procédure, comme pour la majorité des procédures de classification non

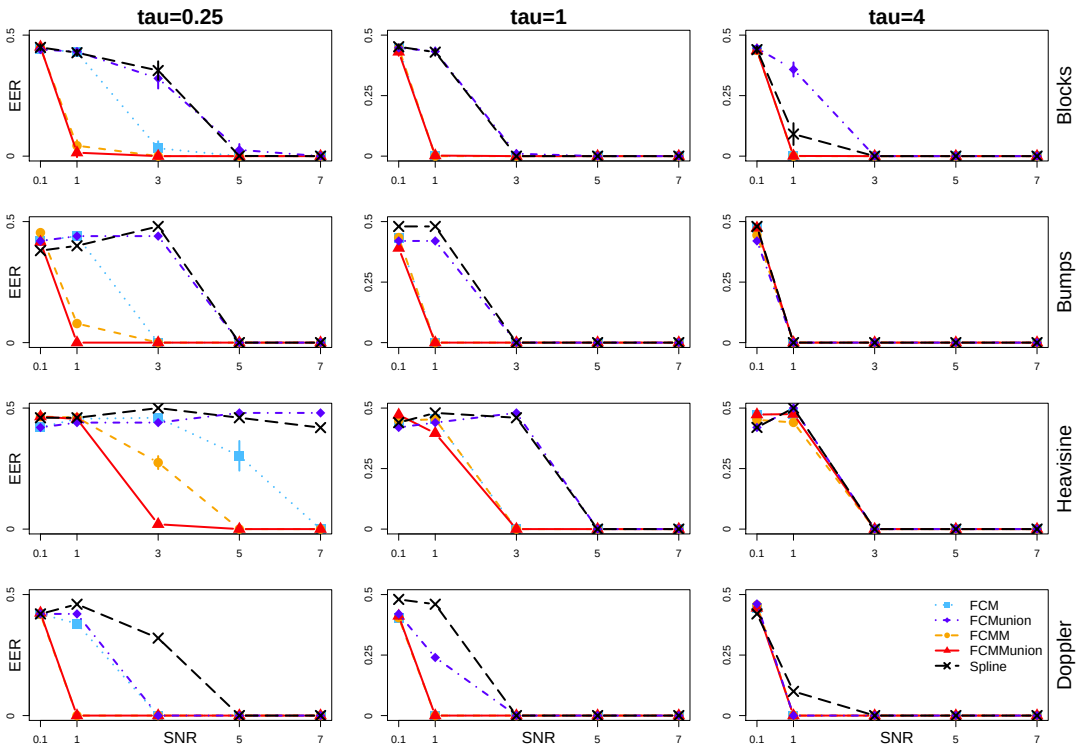


FIGURE 6.3 – *Variations des taux d'individus mal classés (EER) pour les différentes procédures de classification non-supervisée : avec un modèle fonctionnel à effets mixtes avec ou sans réduction de dimension (FCMM/FCMMu), avec un modèle fonctionnel sans effets mixtes avec ou sans réduction de dimension (FCM/FCMu) et la procédure basée sur les splines (Spline) (James et Sugar, 2003). En colonnes sont représentées les différents niveaux d'effets aléatoires ($\tau_U = (0.25, 1, 4)$ pour des effets individuels forts/modérés/faibles) et en lignes, les différents types d'effets fixes considérés (`Blocks`, `Bumps`, `Heavisine`, `Doppler`). Les résultats présentés ici correspondent à une configuration à 2 groupes en présence de groupes équilibrés ($\pi = 0.5$).*

supervisée.

Effet du nombre de groupes et sélection de modèles

De même, nous proposons ici d'évaluer les performances de classification en présence d'un nombre de groupes $L > 2$. Nous nous plaçons dans une configuration à $L = 4$ groupes répartis de manière équilibrée. Le critère EER est encore utilisé comme critère d'évaluation de la classification et les résultats de classification à 4 groupes sont présentés en Figure 6.6. De nouveau, les principales conclusions données dans les cas à deux groupes peuvent être conservées : notre procédure complète de classification non-supervisée en présence d'effets aléatoires présente de meilleurs

		Biais					TOE				
SNR_μ^2		0.1	1	3	5	7	0.1	1	3	5	7
FCM	Blocks	-2.57	-2.66	-2.96	-3.02	-2.99	2.3	2.4	2.3	2.4	2.3
	Bumps	-2.50	-2.69	-2.93	-2.93	-2.93	2.6	2.5	2.6	2.5	2.5
	Heavisine	-2.15	-2.17	-3.22	-4.30	-2.50	2.8	2.7	2.7	2.7	2.8
	Doppler	-2.73	-3.07	-3.32	-3.33	-3.33	2.9	3.2	3.1	3.2	3.2
FCMu	Blocks	-12.93	-11.33	-9.42	-9.38	-8.89	0.4	0.4	0.5	0.5	0.5
	Bumps	-12.98	-11.11	-13.46	-11.98	-11.93	0.5	0.5	0.5	0.5	0.5
	Heavisine	-11.62	-10.20	-10.07	-12.05	-15.68	0.5	0.5	0.5	0.5	0.5
	Doppler	-14.75	-13.14	-11.33	-8.59	-7.87	0.5	0.5	0.5	0.6	0.6
FCMM	Blocks	0.11	0.05	-0.01	-0.01	-0.00	16.0	16.1	15.6	15.8	16.0
	Bumps	0.09	0.04	0.01	0.01	0.01	16.1	16.3	15.2	15.3	15.4
	Heavisine	0.10	0.09	0.08	0.03	0.02	16.4	16.2	16.0	16.4	15.9
	Doppler	0.08	0.01	-0.02	-0.02	-0.01	17.5	17.4	17.5	16.4	17.0
FCMMu	Blocks	-0.11	-0.06	0.03	0.06	0.05	6.9	7.1	7.6	7.6	7.6
	Bumps	-0.10	-0.04	-0.08	-0.08	-0.05	6.7	6.7	6.8	6.7	6.7
	Heavisine	-0.10	-0.10	-0.18	-0.21	-0.19	7.1	7.3	6.8	6.8	6.8
	Doppler	-0.18	-0.06	-0.04	-0.16	-0.11	7.3	7.1	7.3	7.8	7.9
Spline	Blocks	.	.	.	.	.	25.5	26.2	23.0	23.6	22.3
	Bumps	.	.	.	.	.	23.3	26.6	22.0	21.2	21.7
	Heavisine	.	.	.	.	.	24.2	21.6	21.8	22.4	22.3
	Doppler	.	.	.	.	.	33.2	32.4	24.2	24.8	24.2

TABLE 6.1 – *Biais relatif de l'estimateur de la variance résiduelle* $\{(\sigma_\varepsilon^2 - \widehat{\sigma}_\varepsilon^2)/\sigma_\varepsilon^2\}$ *et temps d'exécution moyen (TOE) en minutes sur les données simulées (avec* $N = 50$ *et* $M = 512$*) pour les différentes procédures considérées (FCM/FCMu sans effets aléatoires, FCMM/FCMMu avec effets aléatoires et la procédure 'Spline' de James et Sugar (2003). Les programmes ont été exécutés sur un cluster de calculs de 2 octo-bicore Opteron 2.8Ghz et 2 octo-quadcore Opteron 2.3GHz.*

résultats de classification que les procédures comparées dans une configuration à 4 groupes équilibrés.

Dans un contexte de classification non supervisée, le choix du nombre de groupes représente aussi un point critique de la procédure. Nous présentons en Figure 6.7 les résultats obtenus par le critère BIC de choix du nombre de groupes présentés en Section 5.2.3. On observe que même pour de faibles valeurs de SNR et donc un niveau élevé de variabilité générale, le critère BIC sélectionne en moyenne le bon nombre de groupes. Au sein de l'article de Giacofci et al. (2013), ce critère est comparé à un critère de sélection ICL adapté aux modèles mixtes fonctionnels : en pratique, ce dernier a tendance à sélectionner un nombre plus petit de composantes nous conduisant à préférer l'usage du critère BIC.

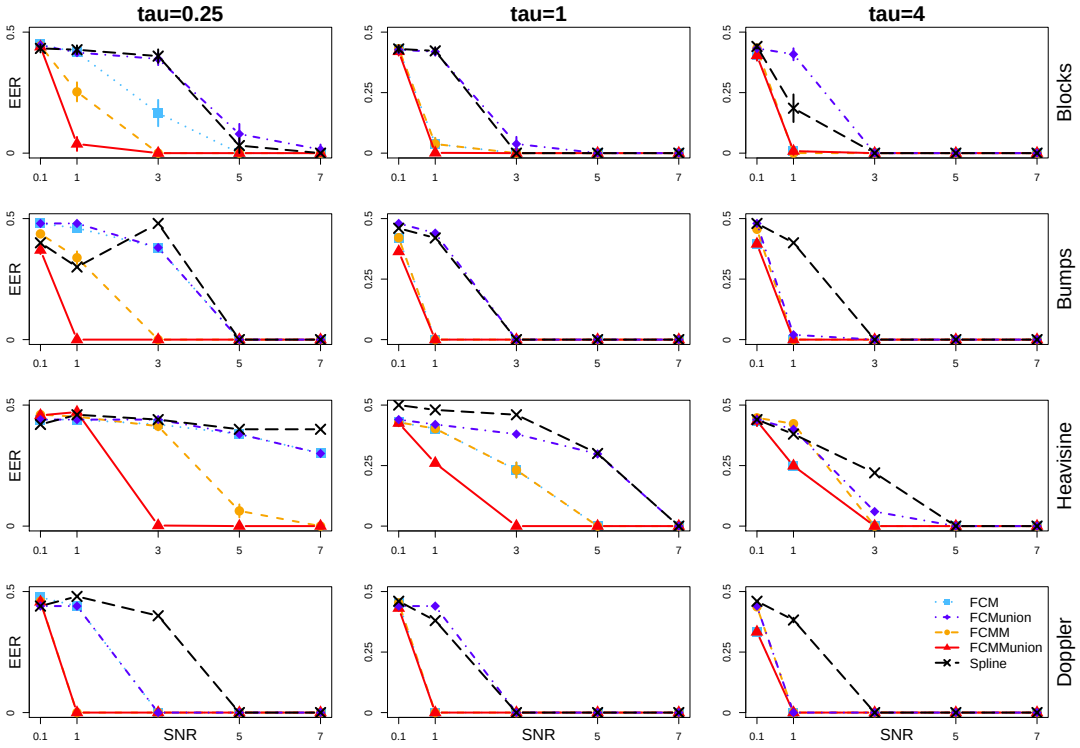


FIGURE 6.4 – *Variations des taux d'individus mal classés (EER) pour les différentes procédures de classification non-supervisée : avec un modèle fonctionnel à effets mixtes avec ou sans réduction de dimension (FCMM/FCMMu), avec un modèle fonctionnel sans effets mixtes avec ou sans réduction de dimension (FCM/FCMu) et la procédure basée sur les splines (Spline) (James et Sugar, 2003). En colonnes sont représentées les différents niveaux d'effets aléatoires ($\tau_U = (0.25, 1, 4)$ pour des effets individuels forts/modérés/faibles) et en lignes, les différents types d'effets fixes considérés (`Blocks`, `Bumps`, `Heavisine`, `Doppler`). Les résultats présentés ici correspondent à une configuration à 2 groupes en présence de groupes non équilibrés ($\pi = 0.25$).*

6.2 Application à des données réelles

Nous présentons à présent deux applications à des données réelles issues du domaine des sciences du vivant. Nous nous intéressons à des types de données regroupées sous le nom de données *omiques*. Ce terme fait référence aux technologies développées depuis le début des années 1990 permettant la mesure relative ou absolue de mécanismes biologiques associés au génome, au protéome ou encore au transcriptome, leur terminaison en "-ome" ayant inspiré le terme de données omiques. Les particularités de ces technologies sont de produire des données de grande, voire très grande, dimension, et à des débits élevés. Nous étudierons dans cette section

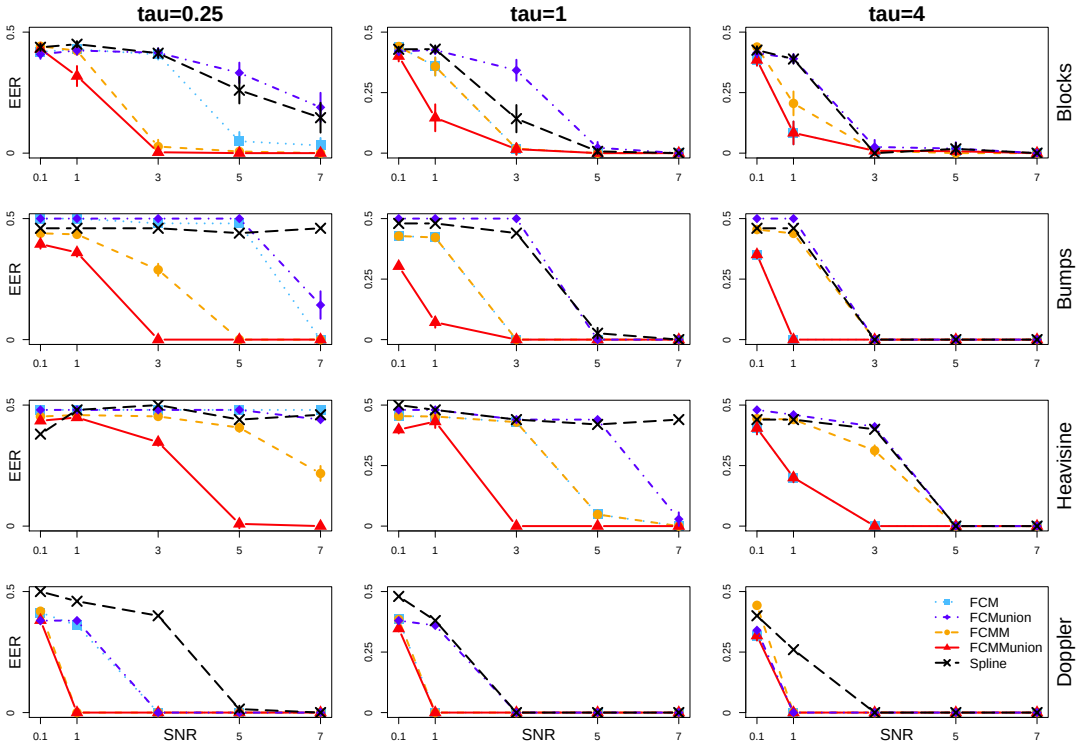


FIGURE 6.5 – *Variations des taux d'individus mal classés (EER) pour les différentes procédures de classification non-supervisée : avec un modèle fonctionnel à effets mixtes avec ou sans réduction de dimension (FCMM/FCMMu), avec un modèle fonctionnel sans effets mixtes avec ou sans réduction de dimension (FCM/FCMu) et la procédure basée sur les splines (Spline) (James et Sugar, 2003). En colonnes sont représentées les différents niveaux d'effets aléatoires ($\tau_U = (0.25, 1, 4)$ pour des effets individuels forts/modérés/faibles) et en lignes, les différents types d'effets fixes considérés (`Blocks`, `Bumps`, `Heavisine`, `Doppler`). Les résultats présentés ici correspondent à une configuration à 2 groupes en présence de groupes non équilibrés ($\pi = 0.1$).*

plus particulièrement un jeu de données issu de la spectrométrie de masse (données protéomiques) et un jeu de données issu de la technologie des microarray CGH (données génomiques).

6.2.1 Données de spectrométrie de masse

Technologie

Les protéines occupent un rôle central dans la vie d'une cellule dans la mesure où l'immense majorité des fonctions cellulaires est assurée par des protéines. Elles sont de ce fait souvent reliées au développement de certaines maladies ou de processus

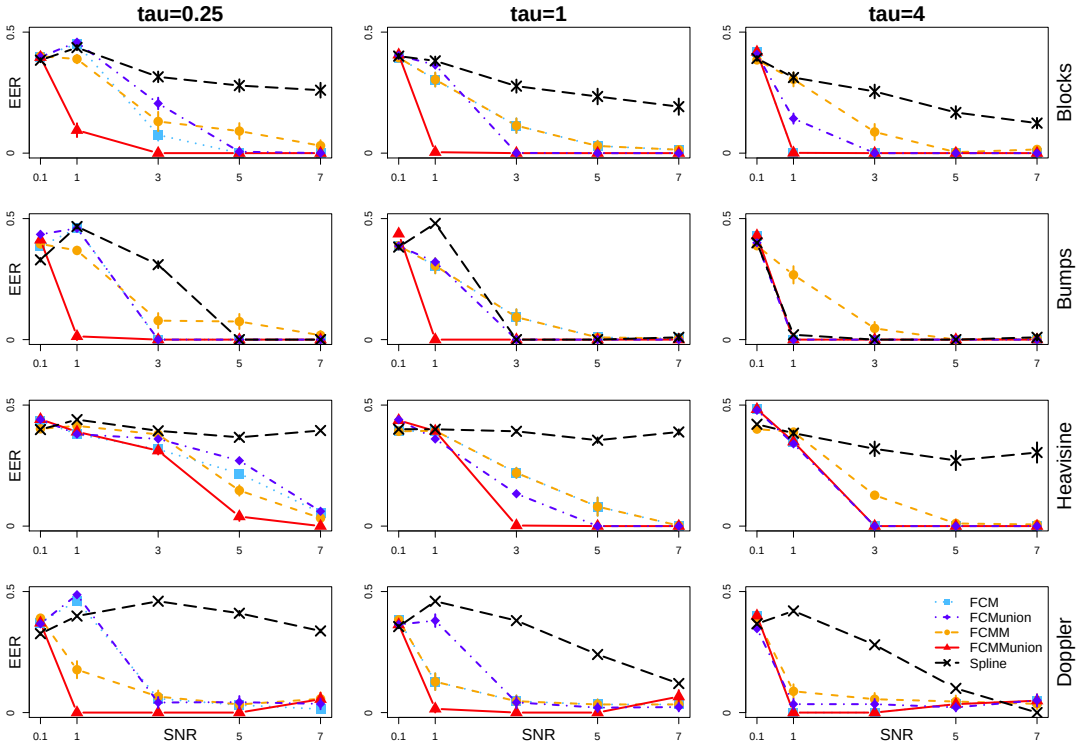


FIGURE 6.6 – *Variations des taux d'individus mal classés (EER) pour les différentes procédures de classification non-supervisée : avec un modèle fonctionnel à effets mixtes avec ou sans réduction de dimension (FCMM/FCMMu), avec un modèle fonctionnel sans effets mixtes avec ou sans réduction de dimension (FCM/FCMu) et la procédure basée sur les splines (Spline) (James et Sugar, 2003). En colonnes sont représentées les différents niveaux d'effets aléatoires ($\tau_U = (0.25, 1, 4)$ pour des effets individuels forts/modérés/faibles) et en lignes, les différents types d'effets fixes considérés (`Blocks`, `Bumps`, `Heavisine`, `Doppler`). Les résultats présentés ici correspondent à une configuration à 4 groupes en présence de groupes équilibrés ($\pi = 0.25$).*

métaboliques particuliers. L'intérêt porté par la communauté scientifique à l'étude du protéome est postérieur aux travaux de séquençage et d'étude du génome humain et représente actuellement un espoir pour une meilleure compréhension de l'étiologie de certaines pathologies : en effet, bien que les protéines soient directement codées par le génome, l'inventaire des protéines présentes au sein d'un échantillon biologique n'est pas la simple traduction des gènes présents sur le génome. Ainsi, l'étude du protéome peut représenter une manière d'étudier certaines pathologies et peut conduire au développement de traitements mieux ciblés.

Une technologie largement utilisée à cet effet est la spectrométrie de masse permettant d'étudier la composition protéomique d'échantillons biologiques aisément accessibles tels que le plasma ou le sérum. Elle permet d'une part l'identification

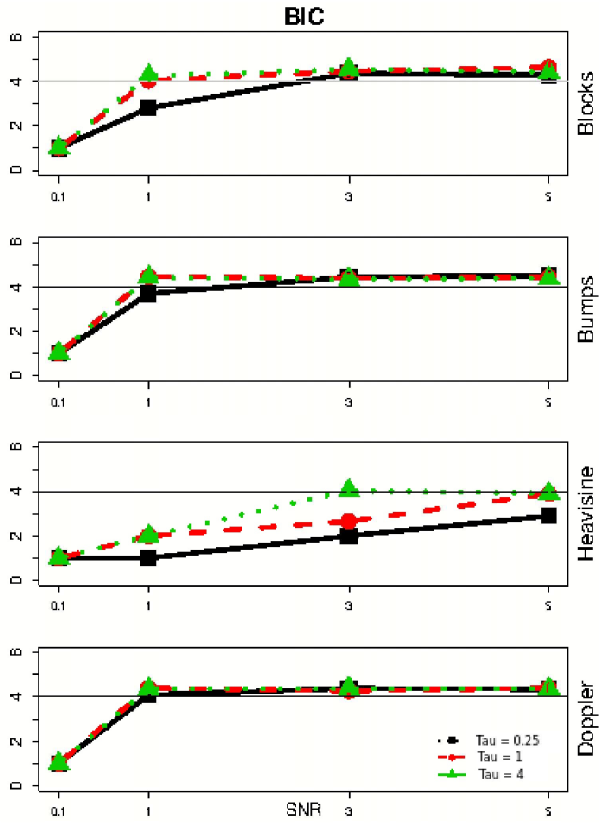


FIGURE 6.7 – *Nombres moyens de classes sélectionnées par le critère BIC dans une configuration à 4 groupes équilibrés au moyen de la procédure FCMM avec réduction de dimension. Le nombre moyen de classes est donné en fonction des valeurs de SNR tandis que chaque courbes correspond à une valeur particulière de* τ_U.

des protéines ou polypeptides présentes dans un échantillon mais aussi de mesurer l'intensité d'expression de ces dernières au sein de l'échantillon. Le principe consiste à différencier les protéines présentes au vu de leur ratio masse sur charge (noté m/z), ce ratio permettant de caractériser chaque protéine de manière unique.

Un schéma présentant la technologie est présenté en Figure 6.8. Deux principaux spectromètres de masse sont communément utilisés à ce jour, l'instrument MALDI-TOF (Matrix-Assisted Laser Desorption and Ionisation- Time Of Flight) et l'instrument SELDI-TOF (Surface-Enhanced Laser Desorption and Ionisation-Time Of Flight). Les principes de mesures pour ces deux outils sont sensiblement les mêmes, à savoir : en premier lieu, l'échantillon biologique est mixé à un composant organique qui agit comme une matrice afin de faciliter les étapes suivantes. Les molécules sont ensuite ionisées et désorbées par le biais d'une exposition à un

faisceau laser pulsé. Les ions sont ensuite accélérés au travers d'un tube de vol par un champ électrostatique jusqu'à atteindre une énergie cinétique commune. Ils sont à ce stade séparés par leur temps de vol au travers du tube qui est déterminé par leur ratio m/z. Les données recueillies correspondent alors aux nombres d'ions atteignant le détecteur placé à la sortie du tube à des temps donnés. Les spectres ainsi créés contiennent de nombreux pics caractérisés en abscisse par le ratio m/z associé tandis que l'ordonnée représente l'abondance de la protéine correspondante dans l'échantillon analysé. Des exemples de tels spectres sont présentés en Figure 6.9.

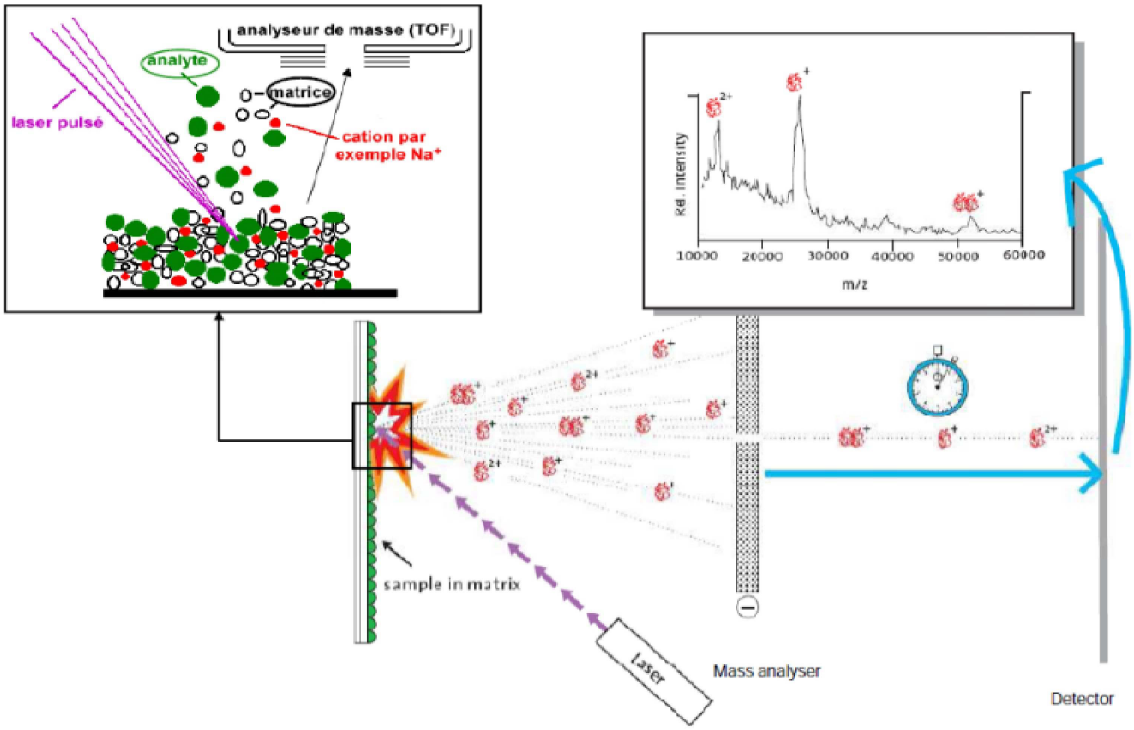


FIGURE 6.8 – *Schéma simplifié de l'acquisition de données de spectrométrie de masse.*

La nécessité d'analyse de telles données a entraîné le développement de nombreuses méthodes statistiques. Parmi celles-ci, les stratégies basées sur une approche fonctionnelle se révèlent particulièrement adaptées à l'étude de telles données, caractérisées par leur dimension élevée et leur haute résolution. Un point critique réside

alors dans la prise en compte de la variabilité individuelle pour l'analyse de ces données car elle a été identifiée comme étant la principale source de variabilité pour les données de spectrométrie de masse (Eckel-Passow et al. 2009). Dans un contexte de classification non-supervisée, les approches existantes concernant les données de spectrométrie de masse sont basées sur la détection préalable des pics contenus dans le signal, utilisés ensuite comme base pour la classification. Ces stratégies présentent néanmoins le désavantage de dépendre fortement des techniques de détection de pics utilisées et de ne pas prendre en compte la présence d'une forte variabilité individuelle. Ceci justifie le développement d'approches fonctionnelles dans un cadre mixte pour l'étude des données de spectrométrie de masse.

Données de cancer de l'ovaire

Nous présentons ici une application à un jeu de données issu d'une étude portant sur le cancer de l'ovaire, publié originalement par Petricoin et al. (2002). Ces données de spectrométrie de masse acquises au moyen d'un spectromètre SELDI-TOF ont été produites grâce à la plateforme protéomique Ciphergen WCX2 et sont disponibles librement dans la banque de données cliniques protéomiques de la FDA/NCI (Food and Drug Administration/National Cancer Institute) [2] .

Le jeu de données étudié est composé de profils protéomiques de 253 sujets féminins dont 162 atteints d'un cancer de l'ovaire et 91 sujets de contrôle sains. Chaque profil est composé de 15154 mesures correspondant à des valeurs m/z allant de 0.0000786 à 19995.513. Ces données sont analysées sur un sous-ensemble de 8192 ratio m/z dans l'intervalle $[1500, 14000]$, les valeurs en dessous 1500 étant retirées pour éviter les effets de la matrice.

L'étude de ce type de données nécessite au préalable une importante étape de pré-traitement : en effet, dû à des effets de la matrice, les données brutes exhibent, d'une part, une ligne de base représentant un bruit de fond qu'il est souhaitable de corriger. D'autre part, les spectres bruts peuvent présenter des problèmes d'alignement ce qui dans un cadre de classification peut conduire à la découverte de structures discriminatoires artificielles. Cette étape de pré-traitement est réalisée grâce à une procédure développée par Antoniadis et al. (2007) qui proposent de corriger le bruit de fond à l'aide d'une régression quantile basée sur l'utilisation de splines tandis que les spectres sont ensuite alignés à l'aide d'une procédure basée sur une technique de détection des zero-crossings, au moyen des ondelettes.

Des exemples de spectres pré-traitées sont représentés en Figure 6.9. Les données sont ici représentées par groupe (cancer/contrôle) mais nous soulignons que les procédures testées sont appliquées dans un cadre non-supervisé, sans connaissance préalable des groupes, le but étant d'évaluer les procédures au vu de leur aptitude à retrouver classes.

Sur ces données, nous avons appliqué notre procédure de classification fonction-

2. `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`, ovarian dataset 8-7-02

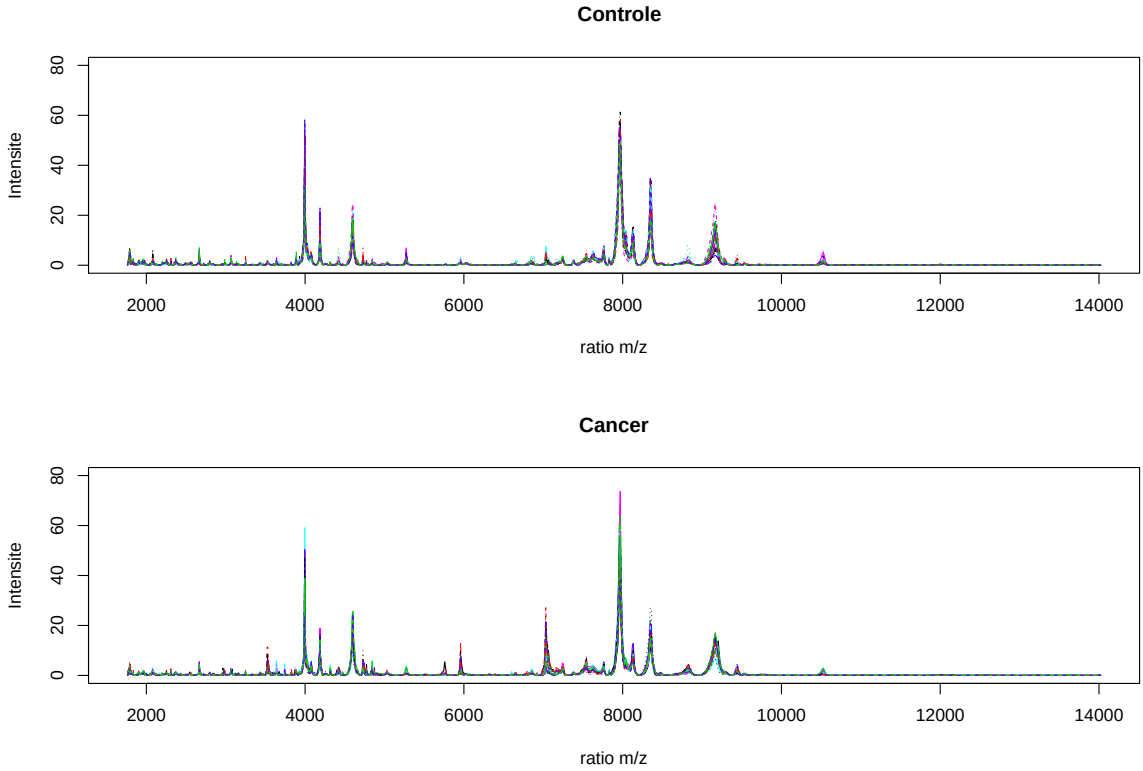


FIGURE 6.9 – *Données de spectrométrie de masse issu d'un jeu de données concernant le cancer de l'ovaire (Petricoin et al.,2002). Les données ont été pré-traitées et alignées et sont représentées par groupe (cancer/contrôle) sur 8192 valeurs pour des ratio m/z allant de 1500 à 14000.*

nelle non supervisée sans effet aléatoire (notée $\mathbf{m}_2$) et avec prise en compte des effets aléatoires pour différentes structures de variances : constante ($\mathbf{m}_2[\gamma^2]$), dépendant du groupe ($\mathbf{m}_2[\gamma^2_\ell]$), de la position ($\mathbf{m}_2[\gamma^2_{jk}]$), ou des deux à la fois ($\mathbf{m}_2[\gamma^2_{\ell,jk}]$). Nous avons, pour ce jeu de données, fixé préalablement le nombre de groupes à deux car ce paramètre ne présentait pas de réel enjeu pour l'étude de ces données. Les résultats sont évalués au vu du critère EER, taux d'individus mal classés, et les résultats sont présentés en Table 6.2. Deux types de pré-traitement sont appliqués : un premier sur les données brutes globales, sans prise en compte des groupes correspondant au cadre non supervisé général et un deuxième en réalisant l'alignement des spectres avec connaissance des groupes.

Signalons que la procédure développée par James et Sugar (2003) n'est pas évaluée sur ce jeu de données à cause de sa dimension trop importante que leur procédure n'est pas en mesure de gérer.

	$\mathbf{m}_2$	$\mathbf{m}_2[\gamma^2]$	$\mathbf{m}_2[\gamma^2_\ell]$	$\mathbf{m}_2[\gamma^2_{jk}]$	$\mathbf{m}_2[\gamma^2_{\ell,jk}]$
alignement global	38	24	24	23	23
alignement par groupe	20	21	22	0.4	36

TABLE 6.2 – *EER (en pourcentage) obtenus sur les données pour différents modèles : classification non supervisée fonctionnelle sans effets aléatoires à 2 groupes ($\mathbf{m}_2$) et le modèle de classification fonctionnelle en présence d'effets aléatoires avec différentes structures de variances : constante $\mathbf{m}_2[\gamma^2]$, dépendant du groupe $\mathbf{m}_2[\gamma^2_\ell]$, de la position $\mathbf{m}_2[\gamma^2_{jk}]$ ou des deux à la fois $\mathbf{m}_2[\gamma^2_{\ell,jk}]$.*

Résultats

En premier lieu, nous nous intéressons aux résultats correspondant à un prétraitement appliqué aux données globales (première ligne de la Table 6.2). Nous pouvons alors observer que, dans ce cas, la prise en compte des effets aléatoires permet une réduction du nombre d'individus mal classés et cela pour toutes les structures de variances considérées : le EER décroît en effet de 38% à environ 24% entre les modèles sans ou avec effets aléatoires respectivement.

Les procédures ont également été appliquées à la suite d'un pré-traitement où l'alignement des spectres est réalisé avec la connaissance des groupes. Dans un contexte de classification non-supervisée, cette approche n'est pas réaliste puisque les labels individuels sont supposés inconnus, cependant, l'étape d'alignement est une étape de pré-traitement des données et ne rentre pas, à proprement parlé, dans la procédure proposée dans cette première contribution. L'observation des résultats nous montre néanmoins une réelle amélioration des performances de classification, plus particulièrement dans le cas d'une structure de variances des effets aléatoires dépendant de la position γ^2_{jk} où seul un individu est mal classés à l'issue de la procédure. Cela nous amène à tirer deux conclusions principales concernant ce jeu de données.

Les performances de classification sont fortement dépendantes de l'étape d'alignement des spectres : en effet, la procédure d'alignement utilisée est basée sur l'idée d'alignement des spectres par rapport à un spectre moyen calculé à partir des données. Les résultats observés induisent qu'il existe une différence d'alignement par groupe et qu'il est alors nettement préférable de réaliser l'alignement par rapport à des spectres moyens calculés par groupe. Ceci met en avant la nécessité de développer une méthode permettant de réaliser l'étape d'alignement et de classification simultanément. Cependant, en l'état, une procédure itérative mêlant alignement et classification est difficilement imaginable au moyen de la procédure de Antoniadis et al. (2007) car elle possède un coût numérique important dépendant de la dimension des données. Cette idée représente néanmoins une perspective de ce travail pour le contexte particulier des données de spectrométrie de masse.

Dans un contexte d'alignement par groupe, la meilleure performance observée est celle obtenue par la procédure basée sur une modélisation mixte avec des variances associées aux effets aléatoires dépendant de la position. Ce résultat nous conduit à penser que la variabilité inter-individuelle n'est pas homogène au long du spectre. De plus, les résultats montrent qu'une grande partie des estimations des variances γ^2_{jk} obtenues sont proches de zéro, accréditant l'idée que la représentation des effets aléatoires dans le domaine des ondelettes est parcimonieuse. Ce point est étudié plus en détails au cours de la deuxième contribution de ce travail de thèse où des stratégies de seuillage des effets fixes et aléatoires sont étudiés dans le contexte des modèles mixtes fonctionnels au sein d'un groupe d'individus homogène.

6.2.2 Données de microarray CGH

Technologie

Les données de microarray CGH (Comparative Genomic Hybridization) représentent une technique permettant l'étude du génome dans sa globalité. Développée au début des années 90, cette technologie permet la détection d'anomalies touchant une petite portion du génome entre un génome d'intérêt et un génome de référence. Le principe de cette technologie est basé sur le caractère diploïde de la majorité des êtres vivants (la plupart des êtres vivants possèdent, normalement, deux copies de chaque chromosome et donc, deux copies de chaque "gène" ou portion de génome) et sur les propriétés d'hybridation des deux brins composant l'ADN. Ainsi, l'objectif est de détecter et de cartographier, tout au long du génome en une seule expérience, les mutations conduisant à la présence d'un nombre de copies de gènes plus élevé que le nombre normal, ce genre de mutation ayant été relié au développement de certaines pathologies comme les cancers par exemple. La compréhension de ces mutations et des parties du génome mises en causes peuvent à terme apporter une meilleure compréhension des mécanismes mis en jeu dans le développement de certaines pathologies ou même pour un meilleur diagnostic des pathologies.

Les profils génomiques individuels ainsi acquis correspondent aux log-ratio du nombre de copies des gènes dans l'ADN testé par rapport à l'ADN de référence. Ils peuvent soit montrer une amplification, le nombre de copies est plus important dans l'échantillon testé, soit une délétion, le nombre de copies est moins élevé dans l'échantillon testé. Une représentation du type de signaux obtenus est donné en Figure 6.11. Par définition, les signaux attendus doivent être de forme constante par morceaux. On observe néanmoins en Figure 6.11, que les données brutes présentent une forte variabilité autour des segments attendus provenant principalement d'erreurs de l'appareil de mesure mais aussi d'une la variabilité génomique naturelle dépendant de la nature des génomes étudiés. L'analyse de ce type de données nécessitent en premier lieu le développement d'outils statistiques permettant d'extraire l'information biologique sous-jacente aux signaux étudiés. Cette problématique est couramment abordée sous l'angle des *méthodes de segmentation* visant à la détec-

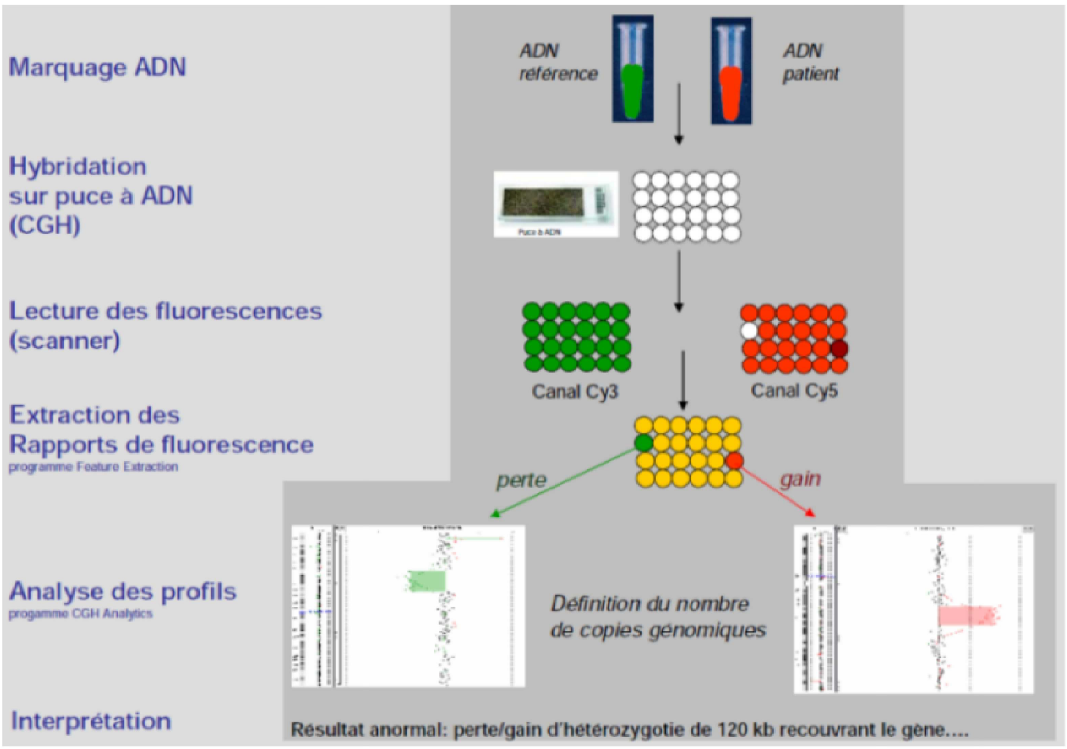


FIGURE 6.10 – *Schéma explicatif de la technologie des microarray CGH*

tion de changement abrupts d'un signal (Picard et al. 2005). Pour une revue plus large des techniques d'analyse de données de microarray CGH, le lecteur pourra se référer à l'article de van de Wiel et al. (2011). Les efforts se sont jusqu'à présent concentrés sur la compréhension de la variabilité induite par des fortes variabilités biologiques et des imprécisions des mesures, laissant les questions de la quantification et de l'éventuelle prise en compte de la variabilité inter-individuelle ouvertes. De même que pour les données de spectrométrie de masse, la nature même de ces données, issues du vivant, nous amène à penser que la variabilité inter-individuelle représente une forte source de variabilité.

Données de cancer du sein

Nous nous intéressons pour ce deuxième type d'applications à un jeu de données portant sur le cancer du sein, publié originellement par Fridlyand et al. (2006). Ce jeu de données contient les profils CGH de 55 individus mesurés en 2464 points correspondant à autant de positions au long du génome. En outre, des informations cliniques concernant les patientes telles que la taille de la tumeur, son stade, son grade, la durée de suivi du patient, la survenue du décès ou non et d'autres encore.

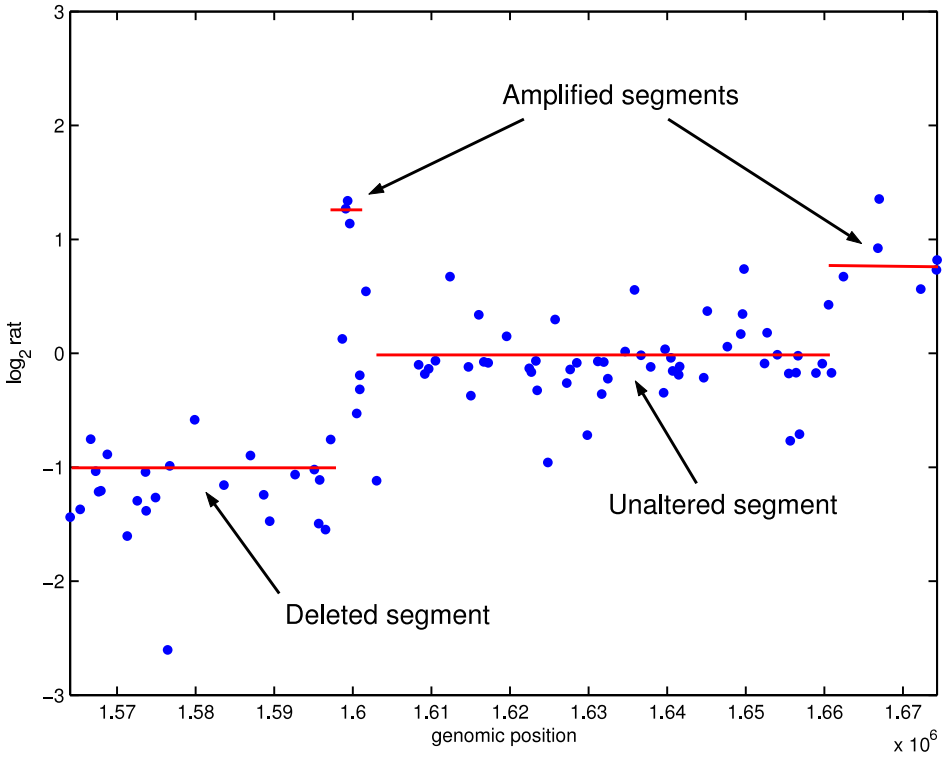


FIGURE 6.11 – *Exemple de données de microarray CGH*

Les données sont téléchargeables librement en tant que documentation supplémentaire de l'article principal. Fridlyand et al. (2006) identifient dans leur étude trois groupes principaux ("1q16q", "amplifier", "complex") qui se différencient principalement par le degré d'instabilité induit sur le génome et que les auteurs relient à des espérances de survie différentes, le groupe 1q16q étant associé à la plus grande espérance de survie. Il est intéressant de noter que ce jeu de données a été analysé de nouveau par Van Wieringen et al. (2008) qui proposent une classification relativement différente pour laquelle ils distinguent 5 groupes et suggèrent que ce jeu de données puisse même être considéré comme plus hétérogène et induire la présence d'un nombre plus élevé de groupes distincts.

Nous analysons ces données au moyen de notre procédure de classification au sein des modèles mixtes dans un contexte entièrement non-supervisé. Notre contribution par rapport aux analyses précédentes est de proposer une classification prenant en compte les effets aléatoires, ce qui n'a jamais été réalisé sur ce type de données.

Résultats

Notre procédure conclut aussi à la présence de 5 groupes et donc à des données plus hétérogènes que ce que propose l'analyse initiale. Notre procédure rejoint néan-

moins l'analyse initiale concernant le sous-groupe 1q16q associé au meilleur taux de survie puisque celui-ci est retrouvé, avec seulement une erreur. Nous signalons que ce groupe n'avait pas été distingué par les analyses suivant celle de Fridlyand et al. (2006), bien qu'il soit clairement associé à de meilleures chances de survie. Comme deux des trois groupes originalement identifiés sont associés à la présence de cellules ER+ (c'est-à-dire, de cellules présentant des récepteurs à oestrogènes), nous avons analysé les 36 individus présentant cette configuration à part, conduisant cette fois-ci à l'identification sans erreur du sous-groupe 1q16q. De plus, notre procédure nous permet d'identifier 3 autres tumeurs (S0041, S1519 et S0303) présentant des comportements génétiques similaires et identifiées comme telles dans l'article original.

En dernier lieu, nous avons souhaité estimer *a posteriori*, à l'aide des paramètres du modèle ajusté par notre procédure, les valeurs des quantités SNR et τ_U obtenues sur ce jeu de données. Bien que documenté dans le cadre des données de spectrométrie de masse, ces paramètres, et plus particulièrement le paramètre τ_U, ont peu été regardés dans le contexte des données CGH. Les résultats sont présentés en Table 6.3. On observe que les valeurs estimées sont très éloignées des valeurs prises en compte lors de notre étude de simulation et correspondent à des niveaux de bruit et d'effets aléatoires très élevés. Ce dernier point nous amène à conclure que la découverte de groupes homogènes ayant une véritable signification biologique en présence de tels niveaux de variabilité nécessiterait de disposer de beaucoup plus d'individus (de l'ordre du millier). En ce sens, l'étude que nous proposons sur ces données se place sur un plan plus qualitatif que quantitatif.

	Données complètes			Données ER+	
cluster ID	$\widehat{\mathrm{SNR}}_\mu^2$	$\widehat{\tau}_U$	cluster ID	$\widehat{\mathrm{SNR}}_\mu^2$	$\widehat{\tau}_U$
1	2.1e-4	3.9e-04	1	2.1e-3	2.2e-04
2	2.3e-3	3.8e-05	2	7.8e-3	1.9e-05
3	1.3e-3	6.4e-04	3	1.1e-2	3.8e-05
4 (1q/16p)	1.5e-3	1.3e-04	4 (1q/16p)	4.4e-3	4.4e-04
5	9.3e-4	4.3e-05			

TABLE 6.3 – *SNR_μ^2 et τ_U estimé pour le jeu de données de cancer du sein (Fridlyand et al., 2006).*

Troisième partie

Réduction de dimension dans les modèles mixtes fonctionnels

Introduction

Dans la première partie de ce manuscrit, nous nous sommes attachés à décrire notre procédure de classification non supervisée au sein des modèles mixtes fonctionnels. Une fois les groupes formés arrive naturellement la question de l'estimation dans les modèles mixtes fonctionnels pour un groupe d'individus homogène. C'est l'objet de cette troisième partie qui représente la deuxième contribution de ce travail de thèse.

En faisant un parallèle avec les modèles mixtes dans un contexte non fonctionnel (cf Chapitre 2), l'estimation au sein des modèles mixtes fonctionnels soulève deux problématiques principales : l'estimation des effets fixes du modèle et celle des effets aléatoires. Dans le cadre de l'étude de données issues de la biologie moléculaire, l'estimation de l'effet fixe fonctionnel se traduit par une meilleure compréhension des mécanismes biologiques du phénomène étudié. L'estimation des effets aléatoires revêt aussi une importance particulière dans la mesure où elle peut permettre, d'une part, une meilleure caractérisation de la variabilité des réponses physiologiques individuelles mais aussi une plus grande précision de l'estimation de l'effet fixe fonctionnel.

À partir d'une modélisation basée sur les ondelettes, la représentation de l'effet fixe fonctionnel dans le domaine des coefficients est naturellement creuse. Les méthodes d'estimation dédiées sont alors les méthodes de seuillage telles que décrites au Chapitre 3. Par ailleurs, dans le domaine des ondelettes, les représentations des effets aléatoires fonctionnels, partageant la même régularité que l'effet fixe fonctionnel sous la modélisation décrite en Section 4.4, sont aussi naturellement parcimonieuses. Comme les effets aléatoires fonctionnels sont supposés être des réalisations de processus gaussiens centrés, cette parcimonie se traduit dans le domaine des coefficients, par des variances nulles à certaines positions : en effet, pour des coefficients distribués à une position (j, k) donnée selon une loi gaussienne centrée, une variance nulle entraîne des coefficients d'effets aléatoires nuls à cette même position. À ce propos, la nécessité de la prise en compte de la parcimonie associée aux effets aléatoires a été illustrée lors de l'étude des données de spectrométrie de masse en Section 6.2.1 où les résultats de classification non supervisée laissent supposer que certains paramètres de variance sont nuls.

Au cours de cette partie, nous abordons les problématiques d'estimation de l'effet fixe et de sélection des effets fixes et aléatoires de manière distincte en proposant

deux approches répondant aux deux approches marginale et jointe classiquement adoptées pour l'étude des modèles mixtes. Dans le Chapitre 7, nous nous intéressons au problème de reconstruction de l'effet fixe fonctionnel et nous nous basons sur une approche marginale des modèles mixtes fonctionnels. Le problème ainsi posé se ramène à une problématique de seuillage dans un contexte de régression non paramétrique hétéroscédastique en présence de répétitions. Nous démontrons que l'estimateur fonctionnel résultant, dans un contexte où les variances sont inconnues, atteint bien une vitesse *near-minimax* dans l'espace fonctionnel considéré. Nous étudions plusieurs stratégies d'estimation des paramètres de variance dont une, basée sur les techniques de vraisemblance pénalisée, permettant de réaliser une sélection des paramètres de variances tout en conservant la rapidité d'exécution caractérisant les méthodes de seuillage non paramétrique. Les estimations des variances sont ensuite utilisées pour le seuillage avec une stratégie de type plug-in.

Dans le Chapitre 8, nous nous concentrons plus particulièrement sur la problématique de sélection des effets fixes et aléatoires. À partir d'une approche jointe du modèle mixte fonctionnel, nous proposons une stratégie basée sur une double pénalisation de la vraisemblance du modèle vis-à-vis des effets fixes et des variances des effets aléatoires grâce à des pénalités de type SCAD. Nous démontrons que l'optimisation du critère de vraisemblance pénalisée ainsi défini conduit à des estimateurs possédant des propriétés oraculaires dans un contexte de double asymptotique où le nombre d'individus N et le nombre de variables M tendent vers l'infini avec $M < N$. Nous développons, parallèlement, une procédure d'optimisation basée sur l'algorithme EM, dédiée à l'estimation par maximum de vraisemblance en présence de données non observées.

Enfin, les performances de ces différentes approches sont étudiées au cours d'une étude de simulation présentée au Chapitre 9, réalisée, dans un premier temps sur des jeux de données présentant des configurations particulières où les parcimonies associées aux effets fixes et aléatoires sont traitées de manière séparée. Cette première étude présente l'avantage de permettre l'évaluation du comportement d'estimation et de sélection des procédures tout en gardant un contrôle sur la configuration considérée. Dans un deuxième temps, nous comparons les comportements des procédures proposées sur des jeux de données simulés de manière réaliste avec comme objectif de se rapprocher de l'étude de données réelles. Nous nous comparons, de plus, pour l'ensemble des simulations à la procédure de seuillage SCAD homoscédastique, couramment utilisée pour ce type de problématique.

Chapitre 7

Seuillage pour le modèle hétéroscedastique

Nous proposons dans ce chapitre une première stratégie d'estimation basée sur une approche marginale des modèles mixtes fonctionnels, telle que décrite au Chapitre 2. Notre principal objectif est de conserver la simplicité de mise en œuvre caractérisant les procédures de seuillage. Notre stratégie représente une extension de ces procédures à un cadre de régression non-paramétrique hétéroscédastique en présence de répétitions et nous démontrons une propriété de convergence de l'estimateur fonctionnel de l'effet fixe. Plus particulièrement, nous verrons que sa convergence est directement liée au ratio entre le nombre de points de discrétisation et le nombre d'individus. Enfin, nous proposons une procédure d'estimation de variances permettant d'effectuer une sélection des positions où s'exerce la variabilité individuelle.

7.1 Modèle marginal et problématique

Dans une approche marginale, seule la loi des observations $(\mathbf{d}_i)_{i=1,\dots,N}$ est considérée. On ne tient en effet pas compte de la loi des observations conditionnellement aux effets individuels $(\mathbf{d}_i|\boldsymbol{\theta}_i)_{i=1,\dots,N}$. Dans le domaine fonctionnel, cela revient à modéliser les signaux de la façon suivante :

$$Y(t) = \mu(t) + \widetilde{E}(t), \tag{7.1}$$

où $\mu(t)$ est un effet fixe fonctionnel supposé appartenir à un espace de Besov B^s_{pq} tandis que le terme d'erreur $\widetilde{E}(t)$ est supposé être une réalisation d'un processus aléatoire dépendant dont la spécification est réalisée dans le domaine des coefficients. Sur les coefficients empiriques de la décomposition, le modèle (7.1) s'écrit alors :

$$\forall i = 1, \dots, N, \;\; \forall (j,k) \in \Lambda, \qquad \begin{cases} c_i = \alpha + \tilde{\varepsilon}^c_i \\ d_{ijk} = \beta_{jk} + \tilde{\varepsilon}^d_{ijk}, \end{cases} \tag{7.2}$$

où le vecteur $\left[(\tilde{\varepsilon}^c)^T, (\tilde{\boldsymbol{\varepsilon}}^d)^T\right]^T \sim \mathcal{N}(\mathbf{0}_M, \mathbf{R})$ où $\mathbf{R}$ est une matrice diagonale dont le terme diagonal à l'indice (j,k) est égal à $\sigma^2_{jk} = \sigma^2_{\theta,jk} + \sigma^2_\varepsilon = 2^{-j\eta}\gamma^2_{jk} + \sigma^2_\varepsilon$, représentant variance totale associée à la position (j,k). Le vecteur $\left[(\alpha)^T, (\boldsymbol{\beta})^T\right]^T$ contient les coefficients empiriques de la décomposition en ondelettes de l'effet fixe fonctionnel μ.

Dans un cadre usuel de régression fonctionnelle, c'est-à-dire pour $N = 1$ et $\sigma^2_{jk} = \sigma^2$ pour tout $(j,k) \in \Lambda$, il est classique, pour une modélisation basée sur les ondelettes, d'utiliser des techniques de seuillage afin d'estimer l'effet fixe fonctionnel inconnu μ, comme décrit au Chapitre 3.

Le modèle (7.2) se distingue de ce cadre classique principalement sur deux points :

- d'une part, le bruit affectant les observations dans le domaine des ondelettes n'étant plus homoscédastique, on fait l'hypothèse que les variances σ^2_{jk} varient avec la position,
- d'autre part, nous nous plaçons dans un contexte de répétitions individuelles, c'est-à-dire pour $N > 1$.

Nous exposons, dans un premier temps, les problématiques associées à ces deux points et présentons les travaux existants sur ces sujets.

7.2 Procédures de seuillage pour modèles hétéroscédastiques sans répétition

La problématique du seuillage dans un contexte hétéroscédastique a déjà été largement abordée dans la littérature mais sous un angle différent. Les approches existantes se placent sous le modèle (7.1), dans le cas où $N = 1$ (sans répétition individuelle), et cherchent à modéliser le processus engendrant le terme d'erreur $\widetilde{E}(t)$. Ce processus est supposé être soit de Wiener, correspondant alors au seuillage classique de Donoho et Johnstone (1994), soit stationnaire (Johnstone et Silverman 1997), soit non-stationnaire (Gao (1997), von Sachs et MacGibbon (2000)).

Le point commun de ces approches est de se ramener, dans le domaine des ondelettes, à un modèle similaire au modèle (7.2) en négligeant les dépendances entre observations par la propriété de décorrélation des ondelettes (Frazier et al. 1991). La principale différence avec le modèle (7.2) concerne la modélisation des variances. Plus précisément, le cadre stationnaire se ramène à une hypothèse de variances dépendant uniquement du niveau de résolution j, c'est-à-dire $\sigma^2_{jk} = \sigma^2_j$, tandis que le cas plus général des processus non-stationnaires se ramène à une modélisation en σ^2_{jk}, où les variances associées aux coefficients varient selon la position (j,k) (pour $(j,k) \in \Lambda$). Dans notre contexte, nous nous plaçons, de même, dans un cadre de variances de la forme σ^2_{jk} mais possédant une structure particulière en $(2^{-j\eta}\gamma^2_{jk} + \sigma^2_\varepsilon)$ pour tout (j,k) d'après les considérations fonctionnelles développées en Section 4.4.2.

En se plaçant à variance connue, en utilisant les techniques de seuillage usuelles telles que décrites par Donoho et Johnstone (1994) avec le seuil universel, les estima-

teurs construits dans le cadre des différentes modélisations atteignent une vitesse de convergence near-minimax dans la classe des espaces de Besov en $\mathcal{O}\left((\log M/M)^{\frac{2s}{2s+1}}\right)$. Cette vitesse correspond à la vitesse minimax non-paramétrique à un facteur logarithmique près, qui est aussi atteinte dans le cadre homoscédastique.

En pratique, la plupart des seuils couramment utilisés, tel que le seuil universel, dépendent des paramètres de variance et l'application d'un seuillage nécessite donc de disposer d'estimateurs des variances. Dans le cadre stationnaire, les $\left(\sigma_j^2\right)_{j=1,\ldots,J}$ sont estimées niveau par niveau, en utilisant un estimateur robuste de type MAD. Cela est rendu possible grâce à la présence de répétitions au sein d'un même niveau de résolution.

Dans un cadre non-stationnaire, la problématique de l'estimation des variances σ_{jk}^2 est un problème plus difficile puisque qu'étant donné que le nombre d'individus est fixé à $N = 1$, la notion de répétitions au sein d'un niveau est alors absente. Cette problématique a été étudiée dans de nombreuses configurations et concernant la régression basée sur les ondelettes, une approche largement adoptée consiste à se placer dans un cadre non-paramétrique et à considérer le vecteur des variances $\left(\sigma_{jk}^2\right)_{(j,k)\in\Lambda}$ comme les coefficients de la décomposition d'une fonction de variance, réalisation d'un processus non stationnaire, dans une base d'ondelettes. Les variances sont alors estimées au sein de chaque niveau de résolution j par une procédure basée sur les différences d'ordre v ($v \in \mathbb{N}$). Ces estimateurs peuvent être vus comme des contrastes locaux d'ordre v sur les observations. Cette procédure peut ensuite être couplée à une méthode de régularisation. Concernant ce type d'approches, nous citons (non-exhaustivement) les travaux de Gasser et al. (1989), Antoniadis et Lavergne (1995), Gao (1997), et Fryzlewicz (2008).

Contrairement aux approches décrites au cours de ce paragraphe, notre modélisation est réalisée directement dans le domaine des ondelettes en considérant une forme particulière des variances associées aux coefficients des effets individuels. Néanmoins, notre modélisation peut être placée dans le contexte décrit ci-dessus en remarquant qu'elle conduit à considérer dans le domaine fonctionnel des processus non-stationnaires, dont la covariance est diagonalisable par la transformée en ondelettes discrète.

7.3 Procédures de seuillage pour modèles hétéroscédastiques avec répétitions

L'étude de données présentant des répétitions individuelles est un sujet relativement peu abordé dans la littérature non-paramétrique bien que ce type de données devienne actuellement de plus en plus courant grâce aux progrès rapides des appareils et techniques de mesures. La majorité des approches non-paramétriques peuvent être étendues trivialement à un cadre de répétitions cependant leurs propriétés théoriques sont rarement explorées. Considérons à présent que nous disposons

de $N > 1$ signaux individuels. La question est alors de savoir comment tirer parti du supplément d'informations apporté par les répétitions individuelles dans le cadre de l'estimation des effets fixes en présence de variabilité inter-individuelle.

Amato et Sapatinas (2005) proposent une étude empirique dans un cadre d'estimation d'une courbe de référence en présence de répétitions. Le modèle considéré est, pour tout $i = 1, \ldots, N$ et tout $m = 1, \ldots, M$:

$$Y_i(t_m) = f_i(t_m) + E_i(t_m) \qquad \text{où} \quad E_i(t_m) \sim \mathcal{N}(0, \sigma_E^2),$$

La particularité de leur approche est de considérer que les f_i sont aléatoires sans le formaliser complètement sous l'angle des modèles mixtes fonctionnels.

Amato et Sapatinas (2005) cherchent alors à estimer une courbe de référence correspondant à $\mathbb{E}\left(f_i(t)\right)$. De plus, ils supposent que $\mathbb{E}\left\{f_i(t) - \mathbb{E}\left(f_i(t)\right)\right\}$ est constant, ce qui revient à considérer un modèle tel que :

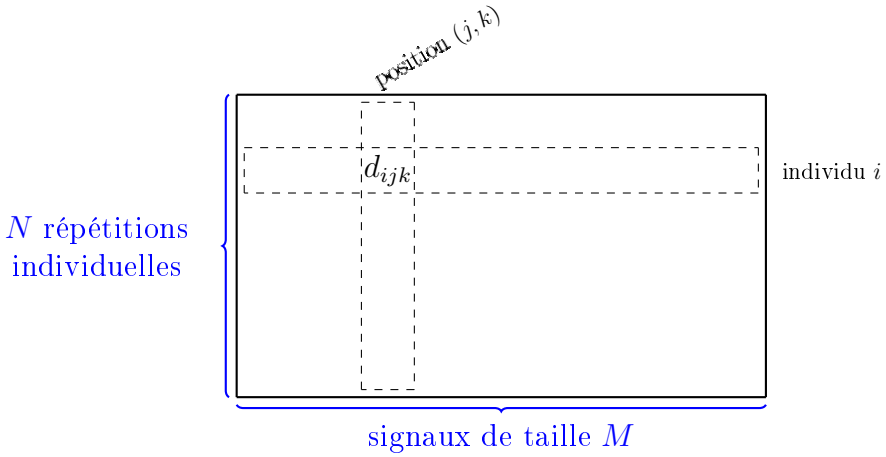
$$\begin{cases} f_i(t_m) = \mu(t_m) + \xi_i(t_m) & \text{où} \quad \xi_i(t_m) \sim \mathcal{N}(0, \sigma_\xi^2), \\ Y_i(t_m) = f_i(t_m) + E_i(t_m). \end{cases}$$

On constate donc que notre approche, en considérant le modèle hétéroscédastique, est plus générale et fondée sur une écriture de type "mixte" du modèle.

Malgré ces différences, leur approche est une première étape : ils mettent en avant, au travers d'une étude de simulation approfondie, que la stratégie consistant à seuiller la moyenne des coefficients est préférable à celle consistant à seuiller la moyenne des signaux seuillés individuellement. Cette stratégie se ramène alors à seuillage non-linéaire homoscédastique usuel et profite donc des propriétés de convergence associées, à savoir, l'estimateur résultant atteint le taux de convergence minimax attendu. Notre objectif est alors de considérer cette procédure dans un cadre hétéroscédastique et d'en étudier les propriétés de convergence, en fonction de la taille des signaux M et du nombre de répétitions N.

7.4 Considérations asymptotiques

Notre contexte est le suivant : nous disposons de N signaux individuels de taille M, représentés chacun par leur M coefficients d'ondelettes.



L'objectif principal est de retrouver le signal moyen commun aux N individus en présence d'un bruit hétéroscédastique, dans le domaine des ondelettes. L'atteinte de cet objectif passe nécessairement par la détermination d'estimations des variances σ^2_{jk}, agissant sur chaque position $(j,k) \in \Lambda$. Nous sommes alors face à deux types d'estimations, entraînant deux notions d'asymptotiques distinctes :

- d'une part, en se plaçant à variance connue, la qualité d'estimation fonctionnelle est dépendante de la taille M des signaux et de l'espace fonctionnel dans lequel ils se situent. Dans la classe des boules de Besov, la vitesse de convergence minimax vers la vraie fonction pouvant être atteinte par les estimateurs de seuillage est en $M^{-2s/(2s+1)}$.
- d'autre part, la présence de N répétitions pour chaque position (j,k) nous permet d'estimer paramétriquement le paramètre de variance σ^2_{jk}. L'erreur attendue pour l'estimation des variances à chaque position est alors paramétrique en N, c'est-à-dire en $1/\sqrt{N}$, dés lors que nous disposons d'un estimateur consistant des paramètres de variance.

La principale problématique associée à cette configuration est alors en rapport avec le ratio M/N : sur chacune des M positions, une erreur est commise, due à l'estimation de σ^2_{jk}. Ces erreurs se cumulent sur les M positions et nécessitent de ce fait un contrôle du ratio entre le nombre d'individus N et le nombre de pas de discrétisation M, de manière à obtenir un estimateur minimax de l'effet fixe fonctionnel. Cette problématique est traitée dans la Section 7.5 et nous verrons que la convergence de l'estimateur de l'effet fixe est asymptotiquement liée à ce ratio.

La problématique de l'estimation des paramètres de variances $(\sigma_{jk})_{(j,k)\in\Lambda}$ est traitée dans un deuxième temps. Dans ce cadre, nous proposons deux stratégies dont une basée sur les techniques de vraisemblance pénalisée permettant de faire une sélection des variables.

7.5 Estimation de l'effet fixe fonctionnel et risque quadratique

Dans cette section, nous présentons une procédure de seuillage hétéroscédastique et nous montrons que l'estimateur proposé atteint bien la vitesse de convergence non paramétrique "near-minimax" dans la classe des espaces de Besov. Nous rappelons que le terme "near-minimax" est employé pour des estimateurs convergeant vers la vraie valeur avec une vitesse minimax dans la classe des Besov à un facteur logarithmique près. Notre résultat se place dans un contexte fonctionnel où les paramètres de variances sont inconnus et en présence de répétitions individuelles.

Partant du modèle marginal (7.2), nous nous ramenons à une procédure de seuillage hétéroscédastique en présence de répétitions, appliquée sur la moyenne des coefficients d'ondelettes associés aux observations, dont la loi est donnée pour tout $(j,k) \in \Lambda$ par :

$$\overline{d_{jk}} \sim \mathcal{N}\left(\beta_{jk}, \frac{\sigma_{jk}^2}{N}\right) \qquad \text{et} \qquad \overline{c} \sim \mathcal{N}\left(\alpha, \frac{\sigma_\nu^2}{N}\right). \tag{7.3}$$

L'estimation des paramètres associés aux coefficients d'approximations $(c_i)_{i=1,\dots,N}$ est traitée séparément, à l'aide d'estimateurs usuels définis par :

$$\widehat{\alpha} = \overline{c} \qquad \text{et} \qquad \widehat{\sigma_\nu^2} = \frac{1}{N-1}\sum_i (c_i - \overline{c})^2. \tag{7.4}$$

Par la suite, ces estimateurs seront appelés estimateurs "de type moment".

Concernant l'estimation des paramètres associés aux coefficients de détails, considérons une fonction de seuillage $\delta(.,.)$. Par la suite, nous nous concentrons principalement sur les fonctions de seuillage continues, à savoir le seuillage doux de Donoho et Johnstone (1994) et le seuillage de type SCAD proposé par Antoniadis et Fan (2001), notés respectivement δ^S et δ^{SCAD}. Pour ces deux types de seuillage, nous choisissons d'utiliser le seuil universel proposé par Donoho et Johnstone (1994) égal à $\lambda\sigma$ où λ est défini par $\lambda = \sqrt{2\log M}$.

Nous étendons cette définition au cas hétéroscédastique en considérant le seuil $\lambda\sigma_{jk}$ dépendant de la position (j,k) considérée pour tout $(j,k) \in \Lambda$. De plus, nous nous plaçons dans un cadre où les variances $(\sigma_{jk})_{(j,k)\in\Lambda}$ sont inconnues mais dont des estimateurs, notés $\left(\widehat{\sigma}_{jk}^2\right)_{(j,k)\in\Lambda}$ sont disponibles et que nous utilisons grâce à une stratégie de type *plug-in*.

Notre estimateur de seuillage est alors défini de la manière suivante :

$$\widehat{\beta}_{jk}(\widehat{\sigma}_{jk}) = \begin{cases} \overline{d_{jk}} & \text{si } 0 \le j \le j_0, \\ \delta(\overline{d_{jk}}, \lambda\widehat{\sigma}_{jk}) & \text{si } j_0 < j \le j_1, \\ 0 & \text{si } j > j_1, \end{cases} \tag{7.5}$$

avec δ fonction de seuillage choisie parmi $\{\delta^S, \delta^{SCAD}\}$. Les niveaux de résolution j_0 et j_1, représentant respectivement les premiers et derniers niveaux de seuillage, vérifient :

$$\begin{cases} 2^{j_0} = \mathcal{O}\left[\left(\log M\right)^{\frac{1-p/2}{1+s'p}} (MN)^{\frac{p/2}{1+s'p}}\right], \\ \dfrac{M}{\log M} \leq 2^{j_1} \leq \dfrac{2M}{\log M}, \end{cases} \tag{7.6}$$

avec $s' = s - \frac{1}{p} + \frac{1}{2}$. L'entier j_0 est fixé en fonction des paramètres de régularité et de la taille des signaux afin d'atteindre une vitesse de convergence optimale tandis que l'entier j_1 est fixé comme dans Juditsky et Delyon (1996). La justification de ces choix est donnée au cours de la démonstration du Théorème (7.1).

La fonction estimée reconstruite à partir de l'estimateur $\widehat{\boldsymbol{\beta}}$ et à l'aide d'une transformation inverse sera notée par la suite $\widehat{\boldsymbol{\mu}}$ et est définie comme $\widehat{\boldsymbol{\mu}} = \mathbf{W}^T(\widehat{\alpha}^T, \widehat{\boldsymbol{\beta}}^T)^T$, avec $\mathbf{W}$ matrice de filtres associée à la base d'ondelettes considérée (cf. Chapitre 3).

Notre premier résultat est donné par le Théorème (7.1). Nous montrons que l'estimateur $\widehat{\mu}$ de l'effet fixe μ atteint bien une vitesse de convergence near-minimax pour le risque L^2. La démonstration de ce résultat est présentée en Annexe A.

Théorème 7.1. *Nous nous plaçons sous le modèle marginal (7.2). Soit $\mu \in B^s_{pq}$ avec $s' = s - \frac{1}{p} + \frac{1}{2} > 0$, $p \geq 1$, $q \geq 1$ et $s \geq 1/p$. Soit $\widehat{\mu}$ l'estimateur d'ondelettes de la fonction μ résultant du seuillage défini par (7.5) et appliqué avec un seuil égal à $2\lambda = 2\sqrt{2\log M}$. On suppose de plus que les variances σ^2_ν et $\left(\sigma^2_{jk}\right)_{(j,k)\in\Lambda}$ sont bornées par une constante σ^2_{max} et que l'on dispose d'estimateurs $\sqrt{N}$-consistants de ces paramètres notés $\widehat{\sigma}^2_\nu$ et $\left(\widehat{\sigma}^2_{jk}\right)_{(j,k)\in\Lambda}$. On a alors :*

$$\mathbb{E}\left(\|\widehat{\mu} - \mu\|^2_{L^2}\right) \leq \max\left\{ \underbrace{\mathcal{O}\left[\left(\frac{\log M}{MN}\right)^{\frac{2s}{2s+1}}\right]}_{T_1} + \underbrace{\left[\mathcal{O}\left(\frac{\log M}{M}\right)^{2s'}\right]}_{T_2} \right\}. \tag{7.7}$$

Le seuil adopté dans ce théorème est une version modifiée du seuil universel, il est multiplié par une constante égale à deux. Cette modification est effectuée pour des raisons techniques de preuve et pour être plus précis, pour obtenir une convergence suffisamment rapide du terme $T_{4.2}$ dans la démonstration donnée en Annexe A. Asymptotiquement, cette modification n'a pas d'impact mais la question de la démonstration du même résultat pour le seuil universel $\lambda = \sqrt{2\log M}$ reste ouverte. Comme cette limitation est due à des raisons techniques et que nous pensons que ce résultat reste vrai pour le seuil universel, les simulations effectuées au Chapitre 9 sont basées sur l'utilisation du seuil universel standard.

Nous nous intéressons dans un premier temps à la vitesse near-minimax non-paramétrique usuelle en $\mathcal{O}\left[\left[\frac{\log M}{M}\right]^{\frac{2s}{2s+1}}\right]$. L'atteinte de cette vitesse par notre estimateur fonctionnel est alors vérifiée sous les conditions énoncées concernant les

valeurs de s et p. En effet, le terme T_2 converge plus rapidement que la vitesse near-minimax non-paramétrique usuelle, c'est-à-dire que nous avons :

$$\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}} = o\left[\left(\frac{\log M}{M}\right)^{2s'}\right].$$

La preuve de ce résultat peut être consultée dans l'article de Juditsky et Delyon (1996) (Lemme 5).

Dans un deuxième temps, nous pouvons remarquer qu'idéalement, cette vitesse devrait être comparée à la borne inférieure du risque L_2 que l'on obtiendrait dans un modèle de régression en présence de N répétitions. La présence du terme T_2 dans la majoration du risque quadratique donnée par l'expression (7.7) correspond à un terme d'erreur d'approximation correspondant au fait que nous n'avons accès qu'à une discrétisation de la fonction μ. La présence de répétitions ne permettra donc pas de diminuer ce risque et pour cette raison, nous nous attendons à une borne inférieure du risque de la forme :

$$\max\left\{\mathcal{O}\left[\left(\frac{\log M}{MN}\right)^{\frac{2s}{2s+1}}\right] + \left[\mathcal{O}\left(\frac{\log M}{M}\right)^{2s'}\right]\right\}.$$

Ceci n'a pas été démontré de manière formelle au cours de ce travail de thèse mais en représente une perspective à court terme.

7.6 Estimation des variances

Nous nous concentrons à présent sur l'estimation des variances $\left(\sigma^2_{jk}\right)_{(j,k)\in\Lambda}$ afin de permettre l'obtention d'une estimation de l'effet fixe fonctionnel à partir d'une stratégie de type *plug-in*. Dans un contexte de seuillage hétéroscédastique, cette étape représente le cœur du problème. Les approches existantes dans un contexte non-paramétrique concernant le problème d'estimation des variances se placent généralement en présence d'une unique courbe ($N = 1$). Nous nous plaçons pour notre part dans un contexte de répétitions individuelles, permettant de disposer d'une plus grande liberté concernant l'estimation des paramètres $\left(\sigma^2_{jk}\right)_{(j,k)\in\Lambda}$. Nous proposons principalement deux approches : la première est basée sur des estimations empiriques usuelles tandis que la deuxième est une procédure d'estimation permettant de faire une sélection des paramètres de variances associés aux effets aléatoires.

7.6.1 Estimation de type moment

Notre première approche consiste à proposer des estimateurs des paramètres $\left(\sigma^2_{jk}\right)_{(j,k)\in\Lambda}$ basés sur une stratégie de type moment. Rappelons que dans une approche marginale, les coefficients d'ondelettes associés aux données observées sont

indépendants et identiquement distribués pour tout $i = 1, \dots, N$, selon :

$$d_{ijk} \sim \mathcal{N}(\beta_{jk}, \sigma^2_{jk}), \quad \forall j, k \in \Lambda.$$

À une position (j, k) donnée, nous disposons de N observations, correspondant aux N répétitions individuelles. L'estimateur considéré, défini par :

$$\widehat{\sigma}^2_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(d_{ijk} - \overline{d_{jk}}\right)^2, \tag{7.8}$$

est alors un estimateur sans biais et $\sqrt{N}$-consistant du paramètre σ^2_{jk}. Cette stratégie représente une première approche simple à mettre en œuvre. Cependant, un des intérêts d'une modélisation mixte est de pouvoir distinguer la variabilité due à une erreur de mesure de celle due à l'individu ce qu'une telle stratégie ne permet pas puisque la variabilité est estimée de manière globale à chaque position (j, k).

7.6.2 Estimation pénalisée

Au sein du modèle marginal (7.2), la variabilité globale des coefficients à une position (j, k) donnée peut être décomposée en :

$$\forall (j, k) \in \Lambda, \qquad \sigma^2_{jk} = \sigma^2_{\theta,jk} + \sigma^2_{\varepsilon}, \tag{7.9}$$

avec $\sigma^2_{\theta,jk} = 2^{-j\eta}\gamma^2_{jk}$. Notre but est alors d'obtenir des estimateurs séparés de σ^2_{ε} et $\sigma^2_{\theta,jk}$ pour tout $(j, k) \in \Lambda$. De plus, notre approche s'appuie sur une hypothèse de parcimonie, à savoir que certaines composantes du vecteur $\left(\sigma^2_{\theta,jk}\right)_{(j,k)\in\Lambda}$ sont nulles comme expliqué dans l'introduction de la Partie III. Ainsi, nous nous ramenons à un seuillage des effets fixes effectué avec une variance σ^2_{ε} à certaines positions et σ^2_{jk} aux positions les plus variables.

Afin d'obtenir un estimateur parcimonieux du vecteur $\left(\sigma^2_{\theta,jk}\right)_{(j,k)\in\Lambda}$, nous nous appuyons sur les techniques de vraisemblance pénalisée, calquées sur les techniques de régression pénalisées décrites en Section 3.3.2, qui, par l'ajout d'une pénalité de type ℓ_1 sur les paramètres $\left(\sigma^2_{\theta,jk}\right)_{(j,k)\in\Lambda}$, "force" la mise à zéro de certains d'entre eux. Le critère que l'on cherche à optimiser est alors le suivant :

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}^2, \sigma^2_{\varepsilon}) = -2 \log \mathcal{L}(\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\gamma}^2, \sigma^2_{\varepsilon}) - \lambda \sum_{jk} |\sigma^2_{\theta,jk}|, \tag{7.10}$$

où λ est un paramètre de régularisation à déterminer. L'optimisation du critère (7.10) par rapport aux paramètres $(\beta_{jk})_{jk}$ conduit à prendre $\widehat{\beta}_{jk} = \overline{d_{jk}}$. Par contre, le critère (7.10) n'est pas optimisable simplement par rapport aux paramètres $\sigma^2_{\theta,jk}$ et σ^2_{ε} : en effet, les optimisations selon ces deux paramètres ne sont pas séparables et nécessitent alors le recours à des algorithmes itératifs. Un de nos objectifs étant

de conserver le rapidité d'exécution des techniques de seuillage non-paramétrique, nous proposons donc de fixer $\widehat{\sigma}^2_\varepsilon$ au moyen de l'estimateur usuellement utilisé en ondelettes basé sur le MAD des coefficients au niveau de résolution le plus fin, noté $\widehat{\sigma}^2_{\text{MAD}}$.
La solution à l'optimisation du critère (7.10) en $\sigma^2_{\theta,jk}$ est alors donnée par :

$$\widehat{\sigma}^2_{\theta,jk} = \left[\frac{-(N + 2\lambda\widehat{\sigma}^2_{\text{MAD}}) + \sqrt{N^2 + 4\lambda \sum_i (d_{ijk} - \overline{d_{jk}})^2}}{2\lambda} \right]_+, \tag{7.11}$$

pour tout $(j,k) \in \Lambda$.
Et, on a alors que :

$$\widehat{\sigma}^2_{\theta,jk} \geq 0 \quad \Longleftrightarrow \quad v^2_{jk} - \widehat{\sigma}^2_{\text{MAD}} \geq \frac{\lambda \widehat{\sigma}^4_{\text{MAD}}}{N} \tag{7.12}$$

où les quantités $v^2_{jk} = \frac{1}{N}\sum_i (d_{ijk} - \overline{d_{jk}})^2$ sont les moments centrés d'ordre 2.

Cette condition se ramène à un seuillage : en effet, si la différence entre la variance empirique des observations et la variance associée au bruit de mesure est en deçà d'un certain seuil, la variance associée aux effets individuels $\sigma^2_{\theta,jk}$ est mise à zéro.

Cette condition peut aussi s'interpréter en terme d'optimisation : en effet, si la condition (7.12) est respectée, alors, le critère Q est convexe. Dans ce cas, il existe une unique solution au problème de minimisation de ce critère et elle est donnée par (7.11). Dans le cas où la condition (7.12) n'est pas respectée, la dérivée du critère objectif Q est strictement positive sur le domaine d'optimisation. Le minimum du critère objectif Q est alors situé sur le bord du domaine d'optimisation, à savoir en $\widehat{\sigma}^2_{\theta,jk} = 0$.

Enfin, nous utilisons un critère de type BIC pour la détermination du paramètre de régularisation λ défini par :

$$\text{BIC}(\lambda) = Q(\boldsymbol{\beta}, \boldsymbol{\gamma}^2, \sigma^2_\varepsilon) + \text{df}_\lambda \times \log(M),$$

où df_λ est le nombre de paramètres $\left(\sigma^2_{\theta,jk}\right)_{\{(j,k)\in\Lambda\}}$ non nuls pour une valeur λ donnée.

Remarques

- Dans l'approche présentée ci-dessus nous avons choisi de définir une pénalité sur les paramètres $\left(\sigma^2_{\theta,jk}\right)_{(j,k)\in\Lambda}$. Un autre choix possible est de définir une pénalité ℓ_1 sur les paramètres de variances $\left(\gamma^2_{jk}\right)_{(j,k)\in\Lambda}$. Cela nous conduit, par une optimisation similaire, à des estimateurs définis pour tout $(j,k) \in \Lambda$:

$$\widehat{\gamma}^2_{jk} = \left[\frac{-(N2^{-j\eta} + 2\lambda\widehat{\sigma}^2_{\text{MAD}}) + \sqrt{N^2 2^{-2j\eta} + 4 \times 2^{-j\eta}\lambda \sum_i (d_{ijk} - \overline{d_{jk}})^2}}{2 \times 2^{-j\eta}\lambda} \right], \tag{7.13}$$

si $2^{-j\eta}(v_{jk}^2 - \widehat{\sigma}_{\text{MAD}}^2) \geq \frac{\lambda \widehat{\sigma}_{\text{MAD}}^4}{N}$ et 0 sinon.
Sur un plan théorique, pour des choix appropriés du paramètre de régularisation λ, ces deux stratégies mènent à des résultats équivalents. Cependant, d'un point de vue numérique, des problèmes de stabilité peuvent intervenir. Sur simulations, on observe néanmoins qu'une pénalisation ℓ_1 sur les paramètres $\sigma_{\theta,jk}^2$ conduit à des résultats plus stables d'un point de vue numérique.

- Pour l'approche sélection de variances, les estimateurs obtenus pour les variances $\sigma_{jk}^2 = \widehat{\sigma}_{\theta,jk}^2 + \widehat{\sigma}_{\varepsilon}^2$ sont des estimateurs biaisés. En effet, d'une part, l'estimateur $\widehat{\sigma}_{\varepsilon}^2$ est basé sur le MAD, qui est un estimateur biaisé. D'autre part, les variances associées aux effets aléatoires $\left(\sigma_{\theta,jk}^2\right)_{(j,k)\in\Lambda}$ sont estimées grâce à une procédure de type LASSO et leurs estimations sont, de ce fait, aussi biaisés (cf. Chapitre 4, Section 3.3.2). Deux alternatives usuelles pourraient être adoptées pour "débiaiser" les estimateurs de type LASSO : soit l'utilisation d'une version relaxée du LASSO (Section 3.3.2), soit l'emploi d'une pénalité de type SCAD. L'utilisation d'une version relaxée du LASSO est envisageable mais nécessite alors l'ajustement d'un modèle mixte par un algorithme itératif. Nous considérons qu'un tel surcoût numérique n'est pas souhaitable dans une approche non-paramétrique basée sur les techniques de seuillages car l'étape de réestimation serait alors plus coûteuse que la procédure complète. De même, l'utilisation d'une pénalité de type SCAD conduit à une optimisation non-convexe ne possédant pas de solution explicite, induisant un coût numérique de résolution important. Ces stratégies seront considérées lors de l'approche jointe axée sur la problématique de sélection des effets fixes et aléatoires au cours du Chapitre 8.
- Signalons qu'avec une telle approche, il est possible de donner une prédiction des effets aléatoires à chaque position $(j,k) \in \Lambda$ par l'expression classique :

$$\widehat{\theta}_{i,jk} = \mathbb{E}[\theta_{i,jk}|\mathbf{d}_i] = \frac{2^{-j\eta}\widehat{\gamma}_{jk}^2}{2^{-j\eta}\widehat{\gamma}_{jk}^2 + \widehat{\sigma}_{\varepsilon}^2}[d_{i,jk} - \widehat{\beta}_{jk}]. \tag{7.14}$$

  Cependant, de telles prédictions ne possèdent alors plus la propriété de BLUP. En effet, ce prédicteur dépend de l'estimateur $\widehat{\boldsymbol{\beta}}$ résultant d'un seuillage non-linéaire et qui n'est, de ce fait, plus linéaire en les données. Ainsi, on ne peut pas espérer obtenir simultanément une bonne vitesse de convergence pour l'estimateur de l'effet fixe et un prédicteur linéaire pour les effets aléatoires individuels.

Finalement, un des principaux attraits de cette première approche réside dans sa rapidité d'exécution et sa simplicité de mise en œuvre. Néanmoins, la principale difficulté de l'estimation au sein des modèles linéaires mixtes classiques est la bonne spécification des effets aléatoires, car elle conditionne la qualité d'estimation de l'effet fixe fonctionnel. Les procédures d'estimation des variances proposées dans

ce chapitre possèdent chacune leurs limitations : la stratégie de type moment ne propose pas de sélection des paramètres de variances tandis que les approches pénalisées, basées sur des pénalités de type LASSO, ne garantissent pas de disposer de la propriété d'oracle pour les estimateurs. En ce sens, l'approche marginale peut donc se révéler limitée pour l'estimation des paramètres du modèle mixte fonctionnel et ceci est dû à la difficulté d'estimation des paramètres de variances. Nous proposons dans le chapitre suivant une procédure basée sur une représentation mixte du modèle, pour lesquels les paramètres sont estimés à l'aide d'un algorithme EM. Cette solution itérative peut être interprétée comme une extension de l'approche marginale (qui représenterait la première étape de cet algorithme itératif) et peut permettre d'améliorer la qualité des estimateurs de variances en termes de précision et de sélection de variables.

Chapitre 8

Sélection de variables dans les modèles mixtes

L'application de notre procédure de classification à des données réelles au cours de la Partie II a mis en avant la problématique de la sélection des effets fixes et aléatoires au sein des modèles mixtes fonctionnels et plus particulièrement celle de la sélection des effets aléatoires lors de l'étude des données de spectrométrie de masse (cf. Section 6.2.1). Au cours du Chapitre 7, nous nous sommes intéressés plus particulièrement à l'estimation dans les modèles mixtes fonctionnels et, dans ce contexte, les résultats de simulations, présentés au Chapitre 9, montrent que l'usage d'une stratégie non itérative basée sur une approche non paramétrique du problème d'estimation ne permet pas de proposer une sélection performante des effets aléatoires.

L'approche présentée dans ce chapitre vise à atteindre deux objectifs connexes : d'une part, proposer une meilleure reconstruction de l'effet fixe fonctionnel μ et d'autre part, sélectionner les positions correspondant à une variance non-nulle des effets aléatoires dans le domaine des ondelettes. Cette problématique est abordée, à ce stade, en adoptant une approche jointe des modèles mixtes. Nous présentons dans un premier temps le modèle mixte fonctionnel dans une approche jointe ainsi que la stratégie d'estimation/sélection adoptée, basée sur les techniques de vraisemblance pénalisée. Nous montrons qu'une telle stratégie conduit à des estimateurs possédant la propriété d'oracle dans un contexte de double asymptotique, c'est-à-dire lorsque le nombre d'individus N et le nombre de variables M tendent vers l'infini (en supposant $M < N$). Nous présentons ensuite notre procédure d'estimation basée sur une reparamétrisation du modèle initial et une utilisation de l'algorithme ECM (Expectation Conditional Maximization), variante de l'algorithme EM, pour la maximisation de la vraisemblance pénalisée.

8.1 Modèle et vraisemblance pénalisée

Reprenons à présent le modèle mixte comme défini en (4.10). Ainsi, dans le domaine des ondelettes, les coefficients empiriques de la décomposition sont modélisés, pour tout $i = 1, \ldots, N$ et pour tout $(j, k) \in \Lambda$, par :

$$\begin{cases} c_i = \alpha + \nu_i + \varepsilon_i^c \\ d_{i,jk} = \beta_{jk} + \theta_{i,jk} + \varepsilon_{i,jk}^d. \end{cases}$$

On suppose de plus que pour tout $i = 1, \ldots, N$:

$$\begin{cases} \begin{bmatrix} \nu_i \\ \boldsymbol{\theta}_i \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G} = \mathrm{diag}(\gamma_\nu^2, \mathbf{G}_\theta)\right) \\ \begin{bmatrix} \varepsilon_i^c \\ \boldsymbol{\varepsilon}_i^d \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M). \end{cases}$$

où γ_ν^2 est la variance associée au terme d'approximation ν_i et $\mathbf{G}_\theta$ est la matrice de covariance associée au vecteur de coefficients $\boldsymbol{\theta}_i$. De plus, la matrice $\mathbf{G}_\theta$ est supposée diagonale et de terme général $[\mathbf{G}_\theta]_{jk} = 2^{-j\eta}\gamma_{jk}^2$, pour tout $(j, k) \in \Lambda$.

Par la suite, dans un souci de simplification des notations et des calculs, nous nous intéresserons exclusivement aux coefficients de détails de la décomposition du modèle, soit les coefficients $\mathbf{d}_i$ pour $i = 1, \ldots, N$. Les paramètres associés aux coefficients c_i sont estimés de façon distincte, comme en (7.4).

Sur les coefficients de détails et en omettant les termes constants, la vraisemblance du modèle est alors donnée par :

$$-2\log\mathcal{L}(\mathbf{d}, \boldsymbol{\theta}; \mathbf{G}_\theta, \boldsymbol{\beta}, \mathbf{R}) = N\log|\sigma_\varepsilon^2\mathbf{I}_M| + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N}\|\mathbf{d}_i - \boldsymbol{\beta} - \boldsymbol{\theta}_i\|^2 + N\log|\mathbf{G}_\theta| + \sum_{i=1}^{N}\boldsymbol{\theta}_i^T\mathbf{G}_\theta^{-1}\boldsymbol{\theta}_i. \quad (8.1)$$

Dans un objectif de sélection des effets fixes et aléatoires et partant de l'idée des régressions pénalisées (cf. Section 3.3.2), nous proposons de pénaliser la vraisemblance (8.1) par rapport aux effets fixes $\boldsymbol{\beta}$ et par rapport aux variances associées aux effets aléatoires $\boldsymbol{\gamma}$. Ainsi, le critère de vraisemblance pénalisée, noté $\boldsymbol{\ell}$, est donné par :

$$\boldsymbol{\ell}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) = -\log\mathcal{L}(\mathbf{d}, \boldsymbol{\theta}; \mathbf{G}_\theta, \boldsymbol{\beta}, \mathbf{R}) + \mathrm{pen}(\boldsymbol{\beta}, \lambda_1) + \mathrm{pen}(\boldsymbol{\gamma}, \lambda_2), \quad (8.2)$$

où $\mathrm{pen}(\cdot, \cdot)$ est une fonction de pénalité, positive, croissante et dérivable sur $(0, \infty)$. Cette fonction dépend, suivant les paramètres pénalisés, de λ_1 et λ_2, deux paramètres de régularisation permettant de contrôler le degré de pénalisation. Il serait plus juste de noter ces paramètres $\lambda_1^{(N)}$ et $\lambda_2^{(N)}$ car leur valeur dépend de la taille d'échantillon

N. Cependant, pour des questions de lisibilité, nous ne mettons pas de référence à la dimension N.

Les estimations des paramètres du modèle correspondent alors aux quantités vérifiant :

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\sigma_\varepsilon^2}) \;=\; \underset{(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma_\varepsilon^2)}{\arg\max}\; \boldsymbol{\ell}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2). \tag{8.3}$$

Face à tous les choix de pénalité possibles, nous nous focalisons sur l'utilisation de pénalités de type SCAD comme définies en (3.13) pour le vecteur des paramètres d'effets fixes $\boldsymbol{\beta}$ et le vecteur des variances associées aux effets aléatoires $\boldsymbol{\gamma}$: en effet, cette pénalité est connue pour disposer de bonnes propriétés de sélection dans un cadre classique de régression pénalisée (Fan et Li 2001) et se révèle être performante pour le seuillage par ondelettes dans un cadre non paramétrique (Antoniadis et Fan 2001). Pour rappel, on a alors, pour tout $(j,k) \in \Lambda$:

$$\mathrm{pen}(\beta_{jk}, \lambda_1) = \begin{cases} \lambda_1 |\beta_{jk}| & \text{si } |\beta_{jk}| \leq \lambda_1, \\ \frac{1}{2(a-1)} \left(|\beta_{jk}|^2 - 2a\lambda_1|\beta_{jk}| + \lambda_1^2\right) & \text{si } \lambda_1 < |\beta_{jk}| \leq a\lambda_1, \\ \frac{1}{2}(a+1)\lambda_1^2 & \text{si } |\beta_{jk}| > a\lambda_1, \end{cases} \tag{8.4}$$

et

$$\mathrm{pen}(\gamma_{jk}, \lambda_2) = \begin{cases} \lambda_2 \gamma_{jk} & \text{si } \gamma_{jk} \leq \lambda_2, \\ \frac{1}{2(a-1)} \left(\gamma_{jk}^2 - 2a\lambda_2\gamma_{jk} + \lambda_2^2\right) & \text{si } \lambda_2 < \gamma_{jk} \leq a\lambda_2, \\ \frac{1}{2}(a+1)\lambda_2^2 & \text{si } \gamma_{jk} > a\lambda_2, \end{cases} \tag{8.5}$$

avec

$$\mathrm{pen}(\boldsymbol{\beta}, \lambda_1) = \sum_{jk \in \Lambda} \mathrm{pen}(\beta_{jk}, \lambda_1),$$
$$\mathrm{pen}(\boldsymbol{\gamma}, \lambda_2) = \sum_{jk \in \Lambda} \mathrm{pen}(\gamma_{jk}, \lambda_2).$$

Par la suite, nous désignerons par $\Upsilon = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ le vecteur des paramètres pénalisés du modèle de taille $2M$ et par $\Upsilon_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$ les valeurs des vrais paramètres. Sans perte de généralité, on peut alors réécrire le vecteur des paramètres $\Upsilon_0 = (\Upsilon_0^1, \Upsilon_0^2)$ avec $\Upsilon_0^1 = (\Upsilon_{01}^1, \ldots, \Upsilon_{0m_N}^1)$, vecteur contenant les paramètres non nuls et $\Upsilon_0^2 = (\Upsilon_{0m_N+1}^2, \ldots, \Upsilon_{02M}^2)$ vecteur contenant les paramètres nuls.

8.2 Propriétés asymptotiques des estimateurs

Dans cette partie, nous nous intéressons aux propriétés de sélection des estimateurs (8.3) proposés. Notre objectif dans ce chapitre est de proposer des estimateurs des paramètres du modèle permettant d'effectuer une sélection performante des effets fixes et des variances des effets aléatoires. Nous nous interrogeons de ce fait sur les

propriétés de sélection des estimateurs par vraisemblance pénalisée (8.3). Plus particulièrement, nous démontrons que ces estimateurs jouissent de la propriété d'oracle (cf. Section 3.3.2), à savoir, on a :

1. Le bon modèle est atteint presque sûrement, c'est-à-dire, $\widehat{\Upsilon}^2 = 0$ presque sûrement,
2. L'estimateur $\widehat{\Upsilon}^1$ est asymptotiquement normal.

Dans la littérature, deux résultats principaux ayant trait à notre problématique peuvent être mis en avant. Un premier résultat proposé par Bondell et al. (2010) se place dans le cadre de la sélection des effets fixes et aléatoires au sein des modèles mixtes. Les estimateurs proposés sont solutions d'un problème de moindres carrés pénalisés au moyen de pénalités de type LASSO adaptatif sur les effets fixes et sur les termes de covariances associés aux effets aléatoires. Bondell et al. (2010) démontrent que leurs estimateurs possèdent bien la propriété d'oracle. Cependant, ce résultat se place dans un cadre non fonctionnel pour lequel le nombre de covariables est fixé par rapport au nombre d'individus N qui tend vers l'infini. Dans notre cadre, le nombre de points de discrétisations M joue le rôle du nombre de covariables, traditionnellement noté p, néanmoins, considérer un nombre de variables M fixé n'a pas de sens lorsque l'on se place dans un contexte fonctionnel.

Un deuxième résultat remarquable est donné par Fan et Peng (2004). Les auteurs se placent dans ce travail dans un cadre non fonctionnel, sans effets aléatoires, de moindres carrés pénalisés à l'aide d'une pénalité non concave (dont font partie les pénalités de type SCAD). Lorsque le nombre de variables M est fixé, Fan et Li (2001) ont démontré que l'estimateur pénalisé possédait bien la propriété d'oracle. Fan et Peng (2004) étendent ce résultat à un cadre de double asymptotique, c'est-à-dire, lorsque le nombre de variables M tend vers l'infini, de même que le nombre d'individus N. Les auteurs font tout de même l'hypothèse que $M < N$ et que le ratio entre ces deux dimensions vérifie $M^5/N \to 0$. Sous ces hypothèses, Fan et Peng (2004) démontrent que les estimateurs résultant du problème de moindres carrés pénalisés possèdent bien la propriété d'oracle. Cette approche considérant un nombre divergent de variables se rapproche du cadre fonctionnel qui est le nôtre et notre objectif est d'étendre ce résultat au cadre des modèles mixtes fonctionnels lorsqu'une sélection des variances des effets aléatoires est réalisée parallèlement à une sélection des effets fixes, c'est-à-dire, en présence de deux termes de pénalités concernant les effets fixes et les variances des effets aléatoires.

Afin de démontrer les propriétés oraculaires de nos estimateurs, il est nécessaire de faire des hypothèses sur la fonction de pénalité et sur la vraisemblance des données. Nous reprenons ici les hypothèses énoncées par Fan et Peng (2004) que nous étendons aux paramètres de variances des effets aléatoires du vecteur $\boldsymbol{\gamma}$. Pour les deux types de paramètres, la pénalité commune utilisée dans notre travail est une pénalité de type SCAD vérifiant les hypothèses concernant la fonction de pénalité. Concernant les hypothèses se référant à la vraisemblance des observations, nous

consacrons l'Annexe (B.1) à la vérification de ces hypothèses pour notre modèle mixte fonctionnel.

8.2.1 Hypothèses sur les pénalités

On commence par définir les quantités :

$$a_N^{\boldsymbol{\beta}} = \max_{1\leq m\leq M} \big\{ \frac{\partial}{\partial \beta_m} \text{pen}(|\beta_{0m}|, \lambda_1), \beta_{0m} \neq 0 \big\},$$
$$a_N^{\boldsymbol{\gamma}} = \max_{1\leq m\leq M} \big\{ \frac{\partial}{\partial \gamma_m} \text{pen}(\gamma_{0m}, \lambda_2), \gamma_{0m} \neq 0 \big\},$$

ainsi que

$$b_N^{\boldsymbol{\beta}} = \max_{1\leq m\leq M} \big\{ \frac{\partial^2}{\partial \beta_m^2} \text{pen}(|\beta_{0m}|, \lambda_1), \beta_{0m} \neq 0 \big\},$$
$$b_N^{\boldsymbol{\gamma}} = \max_{1\leq m\leq M} \big\{ \frac{\partial^2}{\partial \gamma_m^2} \text{pen}(\gamma_{0m}, \lambda_2), \gamma_{0m} \neq 0 \big\}.$$

La fonction de pénalité utilisée doit vérifier les hypothèses suivantes :

(H1) $\liminf_{N\to\infty} \liminf_{\beta\to 0^+} \frac{\frac{\partial}{\partial\beta}\text{pen}(\beta,\lambda_1)}{\lambda_1} > 0$
et $\liminf_{N\to\infty} \liminf_{\gamma\to 0^+} \frac{\frac{\partial}{\partial\gamma}\text{pen}(\gamma,\lambda_2)}{\lambda_2} > 0,$

(H2) $a_N^{\boldsymbol{\beta}} = \mathcal{O}(N^{-\frac{1}{2}})$ et $a_N^{\boldsymbol{\gamma}} = \mathcal{O}(N^{-\frac{1}{2}})$,

(H2') $a_N^{\boldsymbol{\beta}} = o(\frac{1}{\sqrt{NM}})$ et $a_N^{\boldsymbol{\gamma}} = o(\frac{1}{\sqrt{NM}})$,

(H3) $\lim_{N\to\infty} b_N^{\boldsymbol{\beta}} = 0$ et $\lim_{N\to\infty} b_N^{\boldsymbol{\gamma}} = 0$,

(H3') $b_N^{\boldsymbol{\beta}} = o_P(\frac{1}{\sqrt{M}})$ et $b_N^{\boldsymbol{\gamma}} = o_P(\frac{1}{\sqrt{M}})$,

(H4) Il existe $C^{\boldsymbol{\beta}}$ et $D^{\boldsymbol{\beta}}$ constantes telles que, pour $\beta_1, \beta_2 > C^{\boldsymbol{\beta}}\lambda_1$:

$$|\frac{\partial^2}{\partial\beta^2}\text{pen}(|\beta_1|,\lambda_1) - \frac{\partial^2}{\partial\beta^2}\text{pen}(|\beta_2|,\lambda_1)| \leq D^{\boldsymbol{\beta}}|\beta_1 - \beta_2|,$$

Et, il existe $C^{\boldsymbol{\gamma}}$ et $D^{\boldsymbol{\gamma}}$ constantes telles que, pour $\gamma_1, \gamma_2 > C^{\boldsymbol{\gamma}}\lambda_2$:

$$|\frac{\partial^2}{\partial\gamma^2}\text{pen}(\gamma_1,\lambda_2) - \frac{\partial^2}{\partial\gamma^2}\text{pen}(\gamma_2,\lambda_2)| \leq D^{\boldsymbol{\gamma}}|\gamma_1 - \gamma_2|,$$

L'ensemble de ces propriétés sont vérifiées en particulier par la pénalité de type SCAD. Cela nous montre néanmoins que les théorèmes (8.1) et (8.2) à suivre restent valables dans le cadre plus général des pénalités vérifiant ces différents points.

L'hypothèse (H1) se réfère à la propriété de parcimonie de l'estimateur final en obligeant la fonction de pénalité à être "pincée" en 0. Les hypothèses (H2) et (H2') assurent que les estimateurs seront non-biaisés tandis que les hypothèses (H3) et (H3') assurent que les valeurs des pénalités resteront raisonnables face à celles de vraisemblance. Enfin, l'hypothèse (H4) se réfère à la continuité de la fonction de pénalité et donc assure la continuité des estimateurs en les données afin d'éviter les problèmes d'instabilités des estimations (Fan et Li 2001).

8.2.2 Hypothèses sur la vraisemblance

De la même manière, des hypothèses concernant la vraisemblance des données sont nécessaires. Par la suite, on notera $\log \mathcal{L}^i$, la vraisemblance associée à chaque individu i, $i = 1, \ldots, N$, issue d'une loi de probabilité commune et qui s'exprime par :

$$-2 \log \mathcal{L}^i(\Upsilon) = \sum_{jk} \log(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2) + \sum_{jk} \frac{1}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2}(d_{jk}^i - \beta_{jk})^2.$$

(H5) Les dérivées première et seconde de la vraisemblance vérifient pour tout $i = 1, \ldots, N$:

$$\mathbb{E}_\Upsilon \left[\frac{\partial \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_m} \right] = 0 \quad \forall m = 1, \ldots, 2M,$$

et

$$\mathbb{E}_\Upsilon \left[\frac{\partial \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1}} \frac{\partial \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_2}} \right] = -\mathbb{E}_\Upsilon \left[\frac{\partial^2 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2}} \right] \quad \forall m_1, m_2 = 1, \ldots, 2M.$$

(H6) La matrice d'information de Fisher définie pour tout $i = 1, \ldots, N$ par :

$$\mathcal{I}(\Upsilon) = \mathbb{E}\big[\nabla \log \mathcal{L}^i(\Upsilon) \;\; \nabla^T \log \mathcal{L}^i(\Upsilon)\big],$$

vérifie :

$$0 < C_1 < \lambda_{vp}^{\min}(\mathcal{I}(\Upsilon)) \leq \lambda_{vp}^{\max}(\mathcal{I}(\Upsilon)) < C_2 < \infty.$$

où $\lambda_{vp}^{\min}(\mathcal{I}(\Upsilon))$ et $\lambda_{vp}^{\max}(\mathcal{I}(\Upsilon))$ sont respectivement la plus petite et la plus grande valeur propre de la matrice d'information de Fisher. De plus, pour tout $m_1, m_2 = 1, \ldots, 2M$:

$$\mathbb{E}_\Upsilon \left[\frac{\partial^2 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1}} \frac{\partial^2 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_2}} \right]^2 < C_3 < \infty,$$

et

$$\mathbb{E}_\Upsilon \left[\left[\frac{\partial^2 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2}} \right]^2 \right] < C_4 < \infty.$$

(H7) Il existe des fonctions B_{m_1,m_2,m_3} telles que pour tout $m_1, m_2, m_3 = 1, \ldots, 2M$:

$$\left| \frac{\partial^3 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2} \partial \Upsilon_{m_3}} \right| \leq B_{m_1,m_2,m_3}(\mathbf{d}_i) \quad \forall i = 1, \ldots, N,$$

et telles que, pour tout m_1, m_2, m_3 et pour tout i :

$$\mathbb{E}_\Upsilon \big[B_{m_1,m_2,m_3}^2(\mathbf{d}_i) \big] < C_5 < \infty.$$

On cherche ici à majorer les dérivées partielles d'ordre 3 de la log-vraisemblance des données par une fonction dépendant uniquement des coefficients $\mathbf{d}_i$ mais non des paramètres.

(H8) Sans perte de généralité, en réécrivant le vecteur $\Upsilon_0 = (\Upsilon_0^1, \Upsilon_0^2)$ avec $\Upsilon_0^1 = (\Upsilon_{01}^1, \ldots, \Upsilon_{0m_N}^1)$, vecteur contenant les paramètres non nuls et $\Upsilon_0^2 = (\Upsilon_{0,m_N+1}^2, \ldots, \Upsilon_{02M}^2)$ vecteur contenant les paramètres nuls, on a la condition suivante :

$$\min_{1 \leq m \leq m_N} \frac{|\Upsilon_{0m}^1|}{\lambda_\omega} \to 0 \quad \text{quand } n \to \infty,$$

où ω prend les valeurs 1 ou 2 suivant le paramètre considéré (β_m ou γ_m).

La dernière hypothèse (H8) se réfère au cadre d'estimation par maximum de vraisemblance pénalisée dans un contexte de double asymptotique, c'est-à-dire, quand le nombre de variables et d'individus tendent vers l'infini. Elle sert, en effet, à contrôler la capacité à distinguer les vrais paramètres non nuls des autres suivant la grandeur de N. Dans un cadre d'asymptotique simple, c'est à dire lorsque le nombre de paramètres M ne diverge pas, cette hypothèse est artificielle car elle est alors implicite. Dans le cas divergent, elle devient essentielle à la démonstration de propriétés oraculaires.

Les hypothèses (H5), (H6) et (H7) sont classiques dans le cadre d'estimation par maximum de vraisemblance pénalisée. Dans le cas de l'estimation de paramètres de deux types différents, d'effet fixe et de variance, la vérification de ces hypothèses semble moins évidente et, de ce fait, nous donnons en Annexe B.1 le détail de la vérification effective de ces hypothèses dans notre cadre.

8.2.3 Propriétés oraculaires

Sous ces hypothèses, nous allons à présent énoncer les propriétés vérifiées par les estimateurs maximisant le critère pénalisé (8.2), conduisant au fait qu'ils possèdent des propriétés oraculaires. Cela se déroule en plusieurs étapes. Le premier théorème énoncé ci-dessous montre l'existence d'un maximum local du critère de vraisemblance pénalisée asymptotiquement proche du vrai paramètre.

Théorème 8.1. *Supposons que les fonctions de pénalité vérifient les hypothèses (H2)-(H4) et que la vraisemblance des données vérifie les hypothèses (H5)-(H7). Si $M^4/N \to 0$ quand $N \to \infty$, alors, il existe un maximum local $\widehat{\Upsilon}_N$ du critère $\boldsymbol{\ell}(\Upsilon)$ donné en (8.2) tel que :*

$$\|\widehat{\Upsilon}_N - \Upsilon_0\| = \mathcal{O}_P\left(\sqrt{M}\left(\tfrac{1}{\sqrt{N}} + a_N\right)\right),$$

avec a_N défini par $a_N = \max\left\{a_N^{\boldsymbol{\beta}}, a_N^{\boldsymbol{\gamma}}\right\}$.

La preuve de ce résultat, basée sur un développement de Taylor dans un voisinage du vrai paramètre, est donnée en Annexe B.2.1.

Une fois l'existence d'un maximum local démontrée, on peut alors montrer que les estimateurs par maximum de vraisemblance pénalisée possèdent la propriété

d'oracle. Pour cela, notons :

$$\mathcal{H}_N = \mathrm{diag}\Big[\frac{\partial^2}{\partial \Upsilon_1^2}\mathrm{pen}(\Upsilon_{01},\lambda),\ldots,\frac{\partial^2}{\partial \Upsilon_{m_N}^2}\mathrm{pen}(\Upsilon_{0m_N},\lambda)\Big] = \frac{1}{N}\nabla^2\mathrm{pen}(\Upsilon_0,\lambda),$$

et :

$$\mathcal{G}_N = \Big[\frac{\partial}{\partial \Upsilon_1}\mathrm{pen}(|\Upsilon_{01}|,\lambda)\mathrm{sign}(\Upsilon_0^1),\ldots,\frac{\partial}{\partial \Upsilon_{m_N}}\mathrm{pen}(|\Upsilon_{0m_N}|,\lambda)\mathrm{sign}(\Upsilon_{0m_N})\Big] = \frac{1}{N}\nabla\mathrm{pen}(|\Upsilon_0|,\lambda).$$

On a alors le théorème suivant :

Théorème 8.2. *On suppose les hypothèses (H1)-(H7) vérifiées. De plus, on suppose que* $\sqrt{N/M}\ \lambda \to \infty$ *si* $\lambda \to 0$ *et* $M^5/N \to 0$ *quand* $N \to \infty$*. Alors, l'estimateur* $\sqrt{N/M}$*-consistant* $\widehat{\Upsilon} = \begin{bmatrix}\widehat{\Upsilon}^1\\ \widehat{\Upsilon}^2\end{bmatrix}$ *du Théorème (8.1) vérifie :*

i. (Parcimonie) $\widehat{\Upsilon}_2 = 0$ *presque sûrement,*

ii. (Normalité Asymptotique)

$$\sqrt{N}\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\left[\mathcal{I}(\Upsilon_0^1)+\mathcal{H}_N\right]\left[\widehat{\Upsilon}^1-\Upsilon_0^1+\left(\mathcal{I}(\Upsilon_0^1)+\mathcal{H}_N\right)^{-1}\mathcal{G}_N\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,\mathbf{H}),$$

où $\mathbf{A}_N$ *est une matrice de taille* $q \times m_N$ *telle que* $\mathbf{A}_N\mathbf{A}_N^T \to \mathbf{H}$*, avec* $\mathbf{H}$ *matrice symétrique positive.*

La preuve de ce théorème, donnée en Annexe B.2.2, se décompose en deux parties distinctes. On cherche en premier lieu à démontrer les propriétés de parcimonie (*i*) des estimateurs. Dans un deuxième temps, on montre que, de plus, ces estimateurs sont asymptotiquement gaussiens (*ii*).

8.3 Procédure de sélection des effets fixes et aléatoires

Après avoir démontré les propriétés théoriques des estimateurs du maximum de vraisemblance pénalisée, nous nous intéressons à présent aux aspects computationnels du calcul de ces estimateurs en pratique.

8.3.1 Reparamétrisation du modèle mixte fonctionnel

Le modèle mixte fonctionnel fait partie de la classe des modèles à variables latentes (cf. Chapitre 4) : en effet, les effets aléatoires $\boldsymbol{\theta}_i$ sont des variables non

observées. A l'instar du modèle de classification mixte fonctionnel (5.2), la log-vraisemblance des observations peut-être décomposée de la manière suivante :

$$\log \mathcal{L}(\mathbf{d}, \boldsymbol{\theta}; \boldsymbol{\beta}, \gamma^2, \sigma_\varepsilon^2) = \log \mathcal{L}(\mathbf{d}|\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\varepsilon^2) + \log \mathcal{L}(\boldsymbol{\theta}; \gamma^2), \tag{8.6}$$

conduisant au critère pénalisé à optimiser :

$$\boldsymbol{\ell}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) = \underbrace{-\log \mathcal{L}(\mathbf{d}|\boldsymbol{\theta}; \boldsymbol{\beta}, \sigma_\varepsilon^2) + \text{pen}(\boldsymbol{\beta}, \lambda_1)}_{(I)} \underbrace{-\log \mathcal{L}(\boldsymbol{\theta}; \gamma^2) + \text{pen}(\boldsymbol{\gamma}, \lambda_2)}_{(II)}. \tag{8.7}$$

L'optimisation de ce critère peut donc être réalisée séparément vis-à-vis des paramètres $(\boldsymbol{\beta}, \sigma_\varepsilon^2)$ d'une part, en optimisant le terme (I) par rapport à ces derniers et du vecteur des paramètres $\boldsymbol{\gamma}$ d'autre part en optimisant le terme (II). De manière explicite, le terme (II) s'exprime de la manière suivante :

$$\frac{N}{2} \log |\mathbf{G}_\theta| + \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{\theta}_i^T \mathbf{G}_\theta^{-1} \boldsymbol{\theta}_i + \text{pen}(\boldsymbol{\gamma}, \lambda_2).$$

Ce terme, et plus particulièrement le terme de perte, se révèle être non convexe en $\boldsymbol{\gamma}$, rendant alors son optimisation difficile : en effet, en présence d'une fonction objectif non-convexe, l'existence d'un minimum global n'est plus assuré. Comme proposé par Chen et Dunson (2003), nous utilisons à ce stade une reparamétrisation du modèle (5.2) en remplaçant les effets aléatoires individuels $\boldsymbol{\theta}_i$ par l'équivalent $\mathbf{G}_\theta^{1/2} \boldsymbol{\vartheta}_i$ où $\boldsymbol{\vartheta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$. Le modèle (5.2) sur les coefficients de détails de la décomposition du modèle se réécrit alors pour tout $i = 1, \ldots, N$:

$$\mathbf{d}_i = \boldsymbol{\beta} + \mathbf{G}_\theta^{1/2} \boldsymbol{\vartheta}_i + \boldsymbol{\varepsilon}_i,$$

dont la vraisemblance, en omettant les termes constants, s'exprime comme :

$$-2 \log \mathcal{L}(\mathbf{d}, \boldsymbol{\vartheta}_i; \mathbf{G}_\theta, \boldsymbol{\beta}, \mathbf{R}) = NM \log \sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \|\mathbf{d}_i - \boldsymbol{\beta} - \mathbf{G}_\theta^{1/2} \boldsymbol{\vartheta}_i\|_2^2 + \sum_{i=1}^{N} \boldsymbol{\vartheta}_i^T \boldsymbol{\vartheta}_i. \tag{8.8}$$

Cette reparamétrisation n'affecte pas les propriétés des estimateurs par maximum de vraisemblance puisque le modèle reparamétrisé reste équivalent au modèle de départ (5.2). L'effet de cette dernière se concentre sur le "statut" des paramètres du vecteur $\boldsymbol{\gamma}$: en effet, cette reparamétrisation conduit les paramètres du vecteur $\boldsymbol{\gamma}$ à passer d'un "statut de variances" à un "statut de coefficients de régression". Cela se traduit, techniquement, par une vraisemblance associée aux effets aléatoires, correspondant au terme (II) dans l'expression (8.7), ne dépendant plus des paramètres $\boldsymbol{\gamma}$, et permettant ainsi de simplifier l'algorithme EM d'optimisation développé dans la suite.

En pénalisant cette vraisemblance de la même manière qu'en (8.2), le critère pénalisé que l'on souhaite optimiser est alors donné par :

$$\boldsymbol{\ell}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) = \frac{NM}{2} \log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} \|\mathbf{d}_i - \boldsymbol{\beta} - \mathbf{G}_\theta^{1/2} \boldsymbol{\vartheta}_i\|_2^2 + \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{\vartheta}_i^T \boldsymbol{\vartheta}_i + \text{pen}(\boldsymbol{\beta}, \lambda_1) + \text{pen}(\boldsymbol{\gamma}, \lambda_2). \quad (8.9)$$

L'objectif par la suite est alors de proposer une méthode de résolution adaptée à l'optimisation du critère pénalisé reparamétrisé (8.9). Étant en présence d'un modèle à variables latentes, on pense naturellement à l'utilisation d'un algorithme d'optimisation itératif de type EM.

8.3.2 Algorithme EM pour la sélection

Le principe de l'algorithme EM dans le cadre de l'optimisation d'un problème de vraisemblance pénalisé reste le même que décrit au Chapitre 4, à savoir que l'optimisation du critère (8.9) est réalisée vis-à-vis de l'espérance de ce dernier conditionnellement aux données pour palier à l'impossibilité du calcul de la vraisemblance des données complètes.

L'espérance du critère pénalisé (8.9) conditionnellement aux observations est alors donnée par :

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) = \mathbb{E}\left[\boldsymbol{\ell}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2) | \mathbf{d}\right] &= \frac{NM}{2} \log \sigma_\varepsilon^2 \\ &+ \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \left[\|\mathbf{d} - \boldsymbol{\beta} - \mathbf{G}_\theta^{1/2} \widehat{\boldsymbol{\vartheta}}_i\|_2^2 + \text{tr}\left((\mathbf{G}_\theta^{1/2})^T \mathbb{V}(\widehat{\boldsymbol{\vartheta}}_i | \mathbf{d}_i) \mathbf{G}_\theta^{1/2} \right) \right] \\ &+ \sum_{i=1}^{N} \widehat{\boldsymbol{\vartheta}}_i^{\,T} \widehat{\boldsymbol{\vartheta}}_i + \sum_{i=1}^{N} \text{tr}(\mathbb{V}(\widehat{\boldsymbol{\vartheta}}_i | \mathbf{d}_i)) + \text{pen}(\boldsymbol{\beta}, \lambda_1) + \text{pen}(\boldsymbol{\gamma}, \lambda_2), \quad (8.10) \end{aligned}$$

où $\widehat{\boldsymbol{\vartheta}}_i = \mathbb{E}(\boldsymbol{\vartheta}_i | \mathbf{d}_i)$. On constate à ce stade que le calcul effectif de l'expression (8.10) dépend uniquement des quantités $\widehat{\boldsymbol{\vartheta}}_i$ et $\mathbb{V}(\widehat{\boldsymbol{\vartheta}}_i | \mathbf{d}_i)$, qui représentent les prédictions des variables cachées du modèle. Ce calcul correspond à l'étape E de l'algorithme.

On retrouve, à l'itération $[h+1]$, les expressions classiques dans le cadre des modèles mixtes, données par :

$$\begin{cases} \widehat{\vartheta}_{i,jk}^{[h+1]} = \dfrac{\gamma_{jk}^{2\,[h]} 2^{-j\eta}}{\gamma_{jk}^{2\,[h]} 2^{-j\eta} + \sigma_\varepsilon^{2\,[h]}} [d_{i,jk} - \beta_{jk}^{[h]}], \\ \mathbb{V}(\widehat{\boldsymbol{\vartheta}}_{i,jk}^{[h+1]} | \mathbf{d}_i) = \dfrac{\sigma_\varepsilon^{2\,[h]}}{\gamma_{jk}^{2\,[h]} 2^{-j\eta} + \sigma_\varepsilon^{2\,[h]}}, \qquad \forall i = 1, \ldots, N, \;\; \forall (j,k) \in \Lambda. \end{cases} \quad (8.11)$$

Nous constatons sur les expressions (8.11) que pour une variance γ_{jk}^2 mise à zéro à une itération donnée, la prédiction des effets aléatoires à la position (j,k) à l'itération suivante est alors nulle aussi.

L'étape M de l'algorithme consiste alors, à l'itération $[h+1]$ en la maximisation du critère $Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\varepsilon^2)$ par rapport aux paramètres $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ et σ_ε^2, connaissant les valeurs courantes de ces derniers $\boldsymbol{\beta}^{[h]}$, $\boldsymbol{\gamma}^{[h]}$ et $\sigma_\varepsilon^{2\,[h]}$.

Afin d'éviter des problèmes de non-convexité, soulevés plus loin dans ce chapitre, nous choisissons d'optimiser le critère (8.10) de manière successive par rapport aux paramètres $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ et σ_ε^2, les autres paramètres étant fixés à leur valeur courante à l'itération $[h]$. Cela revient à adopter une stratégie d'optimisation de type ECM (Expectation Conditional Maximization), variante de l'algorithme EM développée par Meng et Rubin (1993) où la maximisation selon un paramètre est réalisée conditionnellement aux autres paramètres. Meng et Rubin (1993) démontrent que cette variante de l'algorithme en conserve les propriétés de convergence et de monotonie dans l'accroissement de la vraisemblance au cours des étapes successives. Du point de vue de la sélection de modèles, nous signalons cependant que cette approche revient à explorer moins de modèles puisqu'au lieu d'en explorer 2^{M+M}, on se restreint à $2^M + 2^M$ modèles, car la sélection des paramètres du vecteur des variances $\boldsymbol{\gamma}$ est réalisée à $\boldsymbol{\beta}$ fixé et inversement.

Mise à jour de l'effet fixe

A l'itération $[h]$, la minimisation du critère (8.10) par rapport au paramètre $\boldsymbol{\beta}$ est équivalente à minimiser la fonction objectif suivante :

$$Q_{\boldsymbol{\beta}}(\Upsilon^{[h]}, \Upsilon) = \frac{1}{2\sigma_\varepsilon^{2\,[h]}} \sum_{i=1}^{N} \|\mathbf{d}_i - \mathbf{G}_\theta^{1/2} \widehat{\boldsymbol{\vartheta}_i^{[h]}} - \boldsymbol{\beta}\|_2^2 + \mathrm{pen}(\boldsymbol{\beta}, \lambda_1). \tag{8.12}$$

Ce problème se présente comme un problème de seuillage classique appliqué aux données corrigées de la prédiction des effets aléatoires. La fonction de pénalité choisie dans ce chapitre est une pénalité de type SCAD, comme définie en (8.4). En notant $d_{i,jk}^\beta = d_{i,jk} - \sqrt{2^{-j\eta}}\gamma_{jk}\widehat{\vartheta}_{i,jk}$ pour tout $i = 1, \ldots, N$ et tout $(j,k) \in \Lambda$, les données observées corrigées des prédictions des effets aléatoires, la mise à jour des coefficients $(\beta_{jk})_{jk\in\Lambda}$ associés à l'effet fixe fonctionnel à l'itération $[h+1]$ est donnée par :

$$\widehat{\beta}_{jk}^{[h+1]} = \begin{cases} \mathrm{sign}\left(\overline{d_{jk}^\beta}\right)\left[|\overline{d_{jk}^\beta}| - \lambda_1 \sigma_\varepsilon^{2\,[h]}/N\right]_+ & \text{si } |\overline{d_{jk}^\beta}| \leq \lambda_1(1 + \sigma_\varepsilon^{2\,[h]}/N), \\ \dfrac{(a-1)\overline{d_{jk}^\beta} - \sigma_\varepsilon^{2\,[h]} a\lambda_1 \mathrm{sign}\left(\overline{d_{jk}^\beta}\right)/N}{a-(1+\sigma_\varepsilon^{2\,[h]}/N)} & \text{si } \lambda_1(1+\sigma_\varepsilon^{2\,[h]}/N) < |\overline{d_{jk}^\beta}| \leq a\lambda_1, \\ \overline{d_{jk}^\beta} & \text{si } |\overline{d_{jk}^\beta}| > a\lambda_1, \end{cases} \tag{8.13}$$

où $\overline{d_{jk}^\beta} = N^{-1}\sum_{i=1}^N d_{i,jk}^\beta$. Le détail de l'optimisation du critère (8.12) est donné en Annexe C.

La principale différence avec le seuillage SCAD classique réside dans l'apparition du terme en $\left(1+\sigma_\varepsilon^{2\,[h]}/N\right)$ qui remplace à présent le facteur 2 présent dans le seuillage usuel. Ceci est dû au fait que l'on travaille à présent dans un cadre de maximum de vraisemblance en cherchant à prendre en compte les paramètres de variances. Cela a plusieurs conséquences, notamment celle de contraindre différemment le paramètre a. En effet, dans le cadre connu, la paramètre a est contraint de vérifier $a > 2$ pour une bonne définition des estimateurs. Dans notre cas, il nous faut imposer que $a > (1+\sigma_\varepsilon^{2\,[h]}/N)$ mais on ne possède alors plus la garantie que la valeur classiquement utilisée de $a = 3.7$ reste valable. Cette contrainte est peu gênante car d'une part, dans leur article fondateur, Fan et Li (2001) montrent que les performances du seuillage en terme de risque sont peu dépendantes de la valeur du paramètre a. De plus, dans un cadre asymptotique, on s'attend naturellement à ce que la quantité $(1+\sigma_\varepsilon^{2\,[h]}/N)$ devienne négligeable. Cependant, c'est un point qu'il faudra contrôler d'un point de vue numérique.

Mise à jour des paramètres de variances

Après la mise à jour du vecteur des paramètres d'effet fixe $\boldsymbol{\beta}$, nous cherchons à présent à mettre à jour les paramètres de variances, constitués du vecteur des variances associées aux effets aléatoires individuels $\boldsymbol{\gamma}$ et du paramètre σ_ε^2. Ces mises à jour sont réalisées en fixant les autres paramètres à leur valeur courante. Le problème d'optimisation s'exprime ici comme :

$$\begin{aligned}\underset{\boldsymbol{\gamma}}{\arg\min} \quad & \frac{1}{2\sigma_\varepsilon^{2\,[h]}}\sum_{i=1}^{N}\left[\|\mathbf{d}-\boldsymbol{\beta}^{[h]}-\mathbf{G}_\theta^{1/2}\widehat{\boldsymbol{\vartheta}}_i\|^2+\operatorname{tr}\left(\left(\mathbf{G}_\theta^{1/2}\right)^T\mathbb{V}(\widehat{\boldsymbol{\vartheta}}_i|\mathbf{d}_i)\mathbf{G}_\theta^{1/2}\right)\right] \\ & \qquad\qquad + \operatorname{pen}(\boldsymbol{\gamma},\lambda_2), \\ \text{t.q.} \quad & \gamma_{jk}>0 \quad \forall (j,k)\in\Lambda.\end{aligned} \tag{8.14}$$

Une contrainte supplémentaire a été ajoutée pour les paramètres de variance du vecteur $\boldsymbol{\gamma}$ qui, de par leur nature, sont contraints à être positifs. La pénalité choisie est une pénalité de type SCAD définie par (8.5)

La mise à jour des paramètres fait apparaître deux quantités centrales pour lesquelles nous adopterons les notations suivantes, pour tout $(j,k)\in\Lambda$:

$$\begin{aligned} d_{+jk}^{\gamma} &= \sum_i \sqrt{2^{-j\eta}}\widehat{\vartheta}_{i,jk}\left(d_{i,jk}-\widehat{\beta}_{jk}\right), \\ \mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h]})^2|\mathbf{d}\right] &= \sum_i\left\{(\vartheta_{ijk}^{[h]})^2+\mathbb{V}(\widehat{\vartheta}_{jk}|\mathbf{d}_i)\right\}. \end{aligned} \tag{8.15}$$

La quantité d_{+jk}^{γ} représente la somme des coefficients observés, centrés et normalisés par les prédictions des effets aléatoires tandis que la quantité $\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h]})^2|\mathbf{d}\right]$

représente la contribution totale des moments d'ordre 2 des prédictions des effets aléatoires.

A l'itération $[h+1]$, les paramètres de variance $\left(\gamma_{jk}^{2\,[h+1]}\right)_{(j,k)\in\Lambda}$ sont mis à jour par :

$$\widehat{\gamma}_{jk}^{[h+1]} = \begin{cases} \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]\right\}^{-1}\left[d_{+jk}^{\gamma} - \lambda_2\sigma_{\varepsilon}^{2\,[h]}\right]_+ \\ \qquad\qquad \text{si } d_{+jk}^{\gamma} \leq \lambda_2\left[2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] + \sigma_{\varepsilon}^{2\,[h]}\right], \\ \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] - \frac{\sigma_{\varepsilon}^{2\,[h]}}{a-1}\right\}^{-1}\left[d_{+jk}^{\gamma} - \frac{\sigma_{\varepsilon}^{2\,[h]}a\lambda_2}{a-1}\right] \\ \qquad\qquad \text{si } \left[2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] - \sigma_{\varepsilon}^{2\,[h]}\right]\lambda_2 < d_{+jk}^{\gamma} \leq a_j\lambda_2 2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right], \\ \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]\right\}^{-1} d_{+jk}^{\gamma} \\ \qquad\qquad \text{si } d_{+jk}^{\gamma} > a_j\lambda_2 2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]. \end{cases} \tag{8.16}$$

Le détail de l'optimisation du critère (8.14) est donnée en Annexe C.

De même que pour les paramètres d'effets fixes du vecteur $\boldsymbol{\beta}$, des contraintes supplémentaires, dépendant du niveau j apparaissent sur le paramètre a qui est de ce fait noté a_j : pour pouvoir définir proprement la mise à jour des paramètres de variances, il nous faut imposer la condition :

$$a_j > 1 + \frac{\sigma_{\varepsilon}^{2\,[h]}}{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]}, \quad \forall (j,k) \in \Lambda.$$

Étant donné que la quantité $\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h]})^2|\mathbf{d}\right]$ augmente avec le nombre N d'individus, la quantité $\sigma_{\varepsilon}^{2\,[h]}/2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]$ devient faible dans un cadre asymptotique, rejoignant ainsi le cadre classique de définition du seuillage SCAD.

Enfin, la mise à jour du paramètre σ_{ε}^2 est réalisée à $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ fixés et elle est donnée par :

$$\widehat{\sigma}_{\varepsilon}^{2\,[h+1]} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{(j,k)\in\Lambda}\left[\left(d_{i,jk} - \beta_{jk}^{[h+1]} - \sqrt{2^{-j\eta}}\gamma_{jk}^{[h]}\widehat{\vartheta}_{i,jk}\right)^2 + 2^{-j\eta}\gamma_{jk}^{2\,[h+1]}\mathbb{V}(\widehat{\theta}_{,jk}|\mathbf{d})\right],$$

Choix des paramètres λ_1 et λ_2

Pour le choix des paramètres d'ajustement λ_1 et λ_2, nous avons choisi d'utiliser un critère BIC (cf. Section 5.2.3).

Dans notre cadre, le critère que l'on souhaite optimiser est la vraisemblance pénalisée (8.9) et la dimension de notre modèle est le nombre de variables sélectionnées, soit le nombre de coefficients non nuls donné par l'algorithme EM pour les vecteurs $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$.
Ainsi, le critère BIC que nous utilisons est défini par :

$$\mathrm{BIC}_{(\lambda_1,\lambda_2)} = -2\boldsymbol{\ell}(\Upsilon) + \log(M) \times \mathrm{df}_{(\lambda_1,\lambda_2)},$$

On cherche à minimiser ce critère en fonction des valeurs choisies pour λ_1 et λ_2. En toute rigueur, cela demande alors de parcourir une grille à deux dimensions pour les 2 paramètres d'ajustement λ_1 et λ_2 et de calculer le critère BIC pour chaque couple de valeurs. Suivant la précision des grilles fixées pour ces paramètres, cela conduit rapidement à un nombre très important de couples et donc à une procédure très coûteuse. Plusieurs solutions s'offrent à nous pour limiter le coût numérique de telles procédures : on peut penser en premier lieu à l'utilisation de grilles de valeurs logarithmiques réduites permettant un maillage fin pour les petites valeurs et plus large pour les grandes valeurs. Une deuxième possibilité est d'implémenter un seuillage basé sur le seuil universel, comme développé dans le Chapitre 7, pour les paramètres d'effet fixe $\boldsymbol{\beta}$ en fixant λ_1 à la valeur du seuil universel. Ainsi, le paramètre de régularisation λ_1 est défini d'avance et l'optimisation ne se fait alors que sur le paramètre λ_2. Cette approche purement fonctionnelle semble assez naturelle et peut se justifier par le fait que le vecteur $\boldsymbol{\beta}$ est effectivement un vecteur de coefficients associé à un effet fixe fonctionnel pour lequel cela a un sens de considérer un seuillage basé sur une hypothèse de normalité des coefficients.

Ces deux stratégies seront implémentées lors d'une étude de simulation afin d'évaluer leur compromis temps de calcul/performances sur des données simulées.

8.3.3 Comparaison avec l'approche de Bondell et al. (2010)

Dans un même objectif, Bondell et al. (2010) proposent une procédure pour la sélection d'effets fixes et aléatoires au sein des modèles mixtes mais dans un cadre non fonctionnel. Leur procédure est basée sur l'utilisation de l'algorithme EM pour l'estimation des paramètres par maximum de vraisemblance pénalisée pour les termes d'effets fixes et les termes d'effets aléatoires, en utilisant des pénalités de type LASSO adaptatif pour les paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$. Dans un cadre de modèle mixte classique, les auteurs considèrent des structures de covariances diverses et non nécessairement diagonales. Notre approche est semblable à celle proposée par Bondell et al. (2010), considérée dans un cadre plus restreint car notre modélisation mixte fonctionnelle (4.10) conduit à des structures de variances diagonales dans le domaine des ondelettes. Néanmoins, notre approche est différente sur certains points pour lesquels nous émettons des réserves.

Dans un premier temps, le calcul de l'espérance conditionnellement aux données dérivé par Bondell et al. (2010) ne prend pas en compte le terme de trace associé à la

variance des effets aléatoires conditionnellement aux données. Ce terme correspond au terme de trace dans l'expression (8.14) qui, contrairement à ce qui est affirmé par les auteurs en page 1071 de leur article, dépend explicitement des paramètres du vecteur $\boldsymbol{\gamma}$, situés sur la diagonale de la matrice $\mathbf{G}_\theta$.

Dans un deuxième temps, au cours de l'étape M de l'algorithme, Bondell et al. (2010) proposent d'adopter une approche globale pour l'optimisation de l'espérance conditionnelle de la vraisemblance par rapport aux paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$, c'est-à-dire, d'optimiser de manière simultanée par rapport à ces deux types de paramètres. Cette stratégie a l'avantage d'explorer l'ensemble des modèles possibles de manière exhaustive en optimisant selon deux "directions" simultanément. Ainsi, dans l'optique de se ramener à un problème d'optimisation connu, Bondell et al. (2010) transforment leur problème d'optimisation, par une réécriture matricielle de leur critère d'optimisation, en un problème de la forme :

$$\begin{aligned} \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \quad & -\begin{bmatrix}\boldsymbol{\beta}^+\\ \boldsymbol{\beta}^-\\ \boldsymbol{\gamma}\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\boldsymbol{\beta}^+\\ \boldsymbol{\beta}^-\\ \boldsymbol{\gamma}\end{bmatrix} + \mathbf{B}\begin{bmatrix}\boldsymbol{\beta}^+\\ \boldsymbol{\beta}^-\\ \boldsymbol{\gamma}\end{bmatrix} \\ \text{t.q.} \quad & \boldsymbol{\beta}^+ \geq 0 \; ; \; \boldsymbol{\beta}^- \geq 0 \; ; \; \boldsymbol{\gamma} \geq 0, \end{aligned} \tag{8.17}$$

où $\boldsymbol{\beta}^+$ et $\boldsymbol{\beta}^-$ sont respectivement les parties positives et négatives du vecteur de paramètres $\boldsymbol{\beta}$ et la matrice $\mathbf{A}$ est une matrice carrée de taille $3M \times 3M$. Nous ne détaillons pas explicitement les expressions de la matrice $\mathbf{A}$ et du vecteur $\mathbf{B}$ dans un souci de lisibilité mais le lecteur peut en trouver une description détaillée dans les ressources complémentaires de l'article, accessibles librement en ligne. Pour que ce problème d'optimisation entre dans la classe des problèmes convexes, il faut et il suffit que la matrice $\mathbf{A}$ soit définie positive. On peut vérifier que la matrice $\mathbf{A}$, telle que définie par Bondell et al. (2010), ne l'est pas en général et cette question n'est pas abordée par les auteurs. Par contre, ceux-ci utilisent, au sein de leur code de calcul [1], une méthode de régularisation (de Tikhonov) en ajoutant un terme positif sur la diagonale de la matrice $\mathbf{A}$ dans une optique de régularisation numérique des solutions. Nous affirmons pour notre part que ce phénomène est dû à la non convexité de la fonction objectif considérée. Sur l'exemple proposé par les auteurs dans la documentation de leur code de calcul, en ôtant le terme de régularisation au sein de leur optimisation, leur procédure n'est plus à même de fonctionner en raison de la présence de valeurs propres négatives.

La stratégie proposée par Bondell et al. (2010) présente néanmoins l'avantage d'être développée pour des structures de covariances des effets aléatoires diverses et potentiellement non-diagonales dans le cadre des modèles mixtes non fonctionnels. Une perspective intéressante serait d'adapter notre stratégie du cadre fonctionnel au cadre des modèles mixtes non fonctionnels. Cette problématique ne rentre pas

1. disponible librement à l'adresse : `http://www4.stat.ncsu.edu/~bondell/Software/PenLME/`

dans le contexte de ce travail de thèse mais peut être intéressante d'un point de vue pratique, pour la sélection d'effets fixes et aléatoires au sein des modèles mixtes.

Au cours de ce chapitre, nous nous sommes concentrés sur la problématique de sélection des effets fixes et des variances des effets aléatoires dans les modèles mixtes fonctionnels et avons démontré que la résolution du critère de vraisemblance pénalisé (8.2) conduit à des estimateurs possédant la propriété d'oracle dans un cadre de double asymptotique quand M et N tendent vers l'infini avec $M < N$ et $M^5/N \to 0$. Étant donné que nous nous plaçons dans un contexte fonctionnel, la prochaine étape serait alors de s'intéresser aux propriétés de reconstruction de l'estimateur fonctionnel de l'effet fixe et plus particulièrement à sa vitesse de convergence vers la vraie fonction. Ce point représente une perspective de ce travail, la difficulté résidant ici dans le fait que nous ne disposons pas d'expression explicite de l'estimateur $\widehat{\mu}$, celui-ci étant uniquement défini comme solution d'un problème de vraisemblance pénalisée.

Remarquons enfin que, de la même manière que dans le Chapitre 7, un prédicteur des effets aléatoires aléatoires peut être défini par (7.14), possédant les mêmes propriétés de non linéarité en les données, nécessaire à une potentielle bonne vitesse de convergence de l'estimateur de l'effet fixe fonctionnel.

Chapitre 9

Simulations

Nous proposons dans ce chapitre une étude de simulation permettant d'évaluer les comportements et limites des approches marginale et mixte (présentées respectivement aux Chapitres 7 et 8) vis-à-vis des problématiques d'estimation et de sélection au sein des modèles mixtes fonctionnels. Les deux approches sont étudiées dans un premier temps sur des jeux de données présentant des configurations de parcimonie spécifiques dans le but d'évaluer leur comportement vis-à-vis des effets fixes et aléatoires en distinguant les influences de ces derniers. Les deux approches, ainsi que le seuillage homoscédastique usuel sont ensuite comparés sur des données simulées de manière réaliste afin de mettre en évidence l'apport des méthodes itératives sur la sélection et l'estimation au sein des modèles mixtes fonctionnels.

Nous cherchons, tout au long de ce chapitre, à produire des jeux de données simulés possédant des propriétés spécifiques aux modèles étudiés dans cette partie du manuscrit, à savoir :

- une représentation parcimonieuse de l'effet fixe fonctionnel dans le domaine des ondelettes,
- une représentation parcimonieuse des effets aléatoires par l'intermédiaire d'un vecteur des variances associées aux effets aléatoires parcimonieux.

9.1 Approche marginale et seuillage hétéroscédastique

Notre objectif pour cette approche est de comparer, dans un premier temps, les performances du seuillage hétéroscédastique proposé au Chapitre 7 à celle du seuillage homoscédastique usuel, vis-à-vis de la reconstruction de l'effet fixe fonctionnel.

9.1.1 Construction de jeux de données simulées

La construction de jeux de données répondant aux objectifs de parcimonie recherchés passe en premier lieu par la définition d'effets fixes : nous choisissons les fonctions `Blocks`, `Bumps`, `Heavisine` et `Doppler`, introduites initialement par Donoho et Johnstone (1994), afin de considérer plusieurs régularités d'effet fixe. Ces effets fixes ont une représentation naturellement parcimonieuse dans le domaine des ondelettes.

Les effets aléatoires sont ensuite simulés dans le domaine des ondelettes et nécessitent de définir au préalable une structure d'effets aléatoires, c'est-à-dire, un ensemble de positions $(j,k) \in \Lambda$ auxquelles correspond une variance non-nulle des effets aléatoires ; cet ensemble sera noté Λ_1^γ. Nous choisissons de nous concentrer sur deux structures particulières :

Configuration A - les effets aléatoires s'exercent sur les coefficients nuls de l'effet fixe,

Configuration B - les effets aléatoires s'exercent sur les coefficients non-nuls de l'effet fixe,

De telles structures ne correspondent pas à des jeux de données réalistes, pour lesquels ces deux configurations particulières sont mélangées, mais permettent de traiter séparément la parcimonie associée aux coefficients de l'effet fixe $(\beta_{jk})_{(j,k)\in\Lambda}$ de celle associée aux effets aléatoires $(\sigma_{\theta,jk})_{(j,k)\in\Lambda}$.

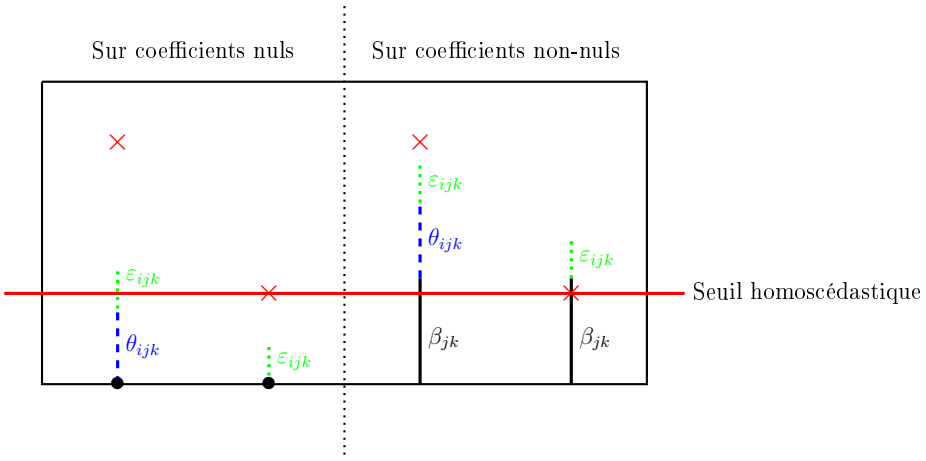


FIGURE 9.1 – *Schéma simplifié comparant les seuillages homoscédastique et hétéroscédastique dans les configurations A et B et en présence ou non d'effets aléatoires.*

Ainsi, comme illustré en Figure 9.1 et sans tenir compte du problème d'estimation des variances, si le coefficient associé à l'effet fixe est nul, le fait d'utiliser un

seuillage hétéroscédastique, qui est, par définition, plus grand qu'un seuillage homoscédastique, aura pour effet de diminuer le nombre de coefficients sélectionnés à tort (faux positifs). *A contrario*, pour un coefficient β_{jk} non-nul, notre procédure peut conduire à l'obtention d'un nombre plus élevé de coefficients seuillés à tort (faux négatifs)

Les niveaux de bruit et de variabilité individuelle sont ensuite contrôlés par l'intermédiaire du rapport signal sur bruit (SNR) et du paramètre τ_U, contrôlant le ratio entre l'intensité du bruit de mesure et de la variabilité individuelle. Le SNR est défini par (6.1), de même qu'en Partie I, tandis que la définition du paramètre τ_U est adaptée afin de prendre en compte la parcimonie et l'hétérogénéité des effets aléatoires. Nous définissons pour cela la quantité γ^2_{REF} représentant un niveau référent de variabilité individuelle et la contribution de la variance des effets aléatoires est alors définie en sommant uniquement sur les positions correspondant à des variabilités non nulles des effets aléatoires. Ainsi, τ_U est défini par :

$$\tau_U = \frac{M\sigma^2_E}{\gamma^2_{\text{REF}} \sum_{(j,k)\in\Lambda_1^\gamma} 2^{-j\eta}}. \tag{9.1}$$

Les variances γ^2_{jk} sont alors simulées à partir de γ^2_{REF} selon une loi Gamma comme suit :

$$\gamma^2_{jk} \sim \Gamma(\gamma^2_{\text{REF}}, 2) \quad \text{si } (j,k) \in \Lambda_1^\gamma.$$

Les valeurs des quantités $(\mu, \Lambda_1^\gamma, \tau_U, \text{SNR})$ nous permettent alors d'en déduire des valeurs pour les paramètres $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2_\varepsilon)$. Les jeux de données synthétiques sont ensuite construits dans le domaine des coefficients en ajoutant au coefficient de l'effet fixe β_{jk}, selon la configuration (A ou B) adoptée, une réalisation de loi gaussienne de variance $2^{-j\eta}\gamma^2_{jk}$ pour la contribution de l'effet individuel et de variance σ^2_ε pour la contribution de l'erreur, pour tout $(j,k) \in \Lambda$.

Dans cette étude, le nombre d'individus est fixé à $N = 100$ et le nombre de pas de discrétisations à $M = 64$. Le paramètre η est quant à lui fixé à 1.5. Quatre types d'effets fixes sont considérés, basés sur les fonctions `Blocks`, `Bumps`, `Heavisine` et `Doppler`. Le paramètre SNR varie dans une plage de valeurs $(1, 3, 5, 7)$ et le paramètre τ_U dans $(0.1, 0.25, 1, 4)$. Enfin, pour chaque configuration, 50 jeux de données sont simulés.

9.1.2 Procédures comparées et indicateurs de performance

Nous comparons sur ces jeux de données simulés les comportements des procédures de seuillage distinguées par les procédures d'estimation des variances. Plus particulièrement, nous comparons les performances de la procédure de seuillage homoscédastique classique, notée "Homoscédastique" par la suite, basée sur un estimateur de type MAD de la variabilité, à celle de procédures hétéroscédastiques. Les

différentes procédures hétéroscédastiques correspondent aux différentes stratégies d'estimation de variances proposées en Section 7.6 : une procédure basée sur des estimations empiriques, notée "Type Moment", et deux procédures basées sur les techniques de vraisemblance pénalisée, avec une pénalité sur les paramètres $(\sigma_{\theta,jk})_{(j,k)\in\Lambda}$, notée "PEN.Var", ou sur les paramètres $\left(\gamma^2_{jk}\right)_{(j,k)\in\Lambda}$, notée "PEN.Gamma".

Dans un cadre d'estimation non-paramétrique par ondelettes, nous nous concentrons principalement sur la précision d'estimation de l'effet fixe fonctionnel reconstruit et des paramètres de variances $\boldsymbol{\sigma}$. Nous utilisons comme critère de comparaison les écarts quadratiques moyens (EQM) relatifs définis par :

$$\mathrm{EQM}_{\boldsymbol{\beta}} = \sqrt{\frac{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|_2^2}{\|\boldsymbol{\mu}_0\|_2^2}}, \quad \text{et} \quad \mathrm{EQM}_{\boldsymbol{\sigma}} = \sqrt{\frac{\|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}_0\|_2^2}{\|\boldsymbol{\sigma}_0\|_2^2}}, \tag{9.2}$$

où $\boldsymbol{\mu}_0$, $\boldsymbol{\sigma}_0$ et sont les vecteurs contenant les vrais paramètres d'effet fixe fonctionnel et des variances globales associées à chaque position. Le vecteur $\widehat{\boldsymbol{\mu}}$ est l'effet fixe fonctionnel reconstruit à l'aide de la transformée en ondelettes inverse appliquée au vecteur estimé $\widehat{\boldsymbol{\beta}}$.

Afin de préciser les résultats d'estimation des variances, des résultats de sélection de variables sont présentés pour les paramètres de variances. Rappelons néanmoins qu'on ne s'attend pas à ce que les procédures pénalisées proposées soient consistantes en terme de sélection de variables puisqu'elles sont basées sur une pénalité de type LASSO. Nous considérons comme vrai positif (TP - True Positive) une position sélectionnée par la procédure et dont la vraie valeur est non-nulle, comme faux positif (FP - False Positive), une position sélectionnée alors que le vrai paramètre est nul. De même, les vrais négatifs (TN - True Negative) et les faux négatifs (FN - False Negative) représentent les positions non sélectionnées par la procédure et dont les vraies valeurs sont respectivement vraiment nulles ou au contraire non-nulles. Les critères de sélection de variances considérés sont alors :

- La sensibilité, définie comme le ratio $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, qui représente la capacité de la procédure à retrouver les vrais non-nuls.
- La spécificité, définie par le ratio $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$, qui représente la capacité de la procédure à retrouver les vrais positions nulles
- Le critère PPV (Positive Predictive Value), défini par le ratio $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$, représente le nombre de variables correctement sélectionnées parmi l'ensemble des positions sélectionnées.
- Le critère NPV (Negative Predictive Value), défini comme le ratio $\frac{\mathrm{TN}}{\mathrm{TN+FN}}$, est le nombre de variables non sélectionnées à raison parmi l'ensemble des positions non sélectionnées.

Ces critères varient tous entre 0 et 1 où 1 représente la valeur à atteindre pour une sélection parfaite.

9.1.3 Résultats

Usage de l'estimateur MAD en présence de répétitions

Nous faisons à ce stade une première conclusion préliminaire concernant les stratégies de seuillage homoscédastique en présence de répétitions. En effet, les travaux d'Amato et Sapatinas (2005) nous apprennent que la manière la plus adaptée de retrouver l'effet fixe dans un cadre de répétitions individuelles est de réaliser un seuillage sur la moyenne des coefficients $\left(\overline{\mathbf{d}_i}\right)_{i=1,\ldots,N}$ avec une variance divisée par un facteur N^{-1}. En revanche, le problème de l'estimation du paramètre de variance, nécessaire au seuillage n'est pas traité par Amato et Sapatinas (2005). Or, dés l'instant où l'on utilise une stratégie de type plug-in, les solutions sont multiples. Nous avons choisi de comparer deux stratégies concurrentes : une première où le MAD est calculé sur le signal moyen, correspondant à $\widehat{\sigma}^2_{\mathrm{MAD}}(\overline{\mathbf{d}}_i)$ et une deuxième, où les MAD sont déterminés individu par individu, l'estimateur global étant ensuite la moyenne de ces estimateurs individuels, soit $\overline{\widehat{\sigma}^2_{\mathrm{MAD}}(\mathbf{d}_i)}$. *A priori*, aucune stratégie ne semble être à privilégier.

Sur simulations réalisées dans un contexte d'absence d'effets aléatoires, nous avons pu constater qu'un estimateur de la forme $\overline{\widehat{\sigma}^2_{\mathrm{MAD}}(\mathbf{d}_i)}$ conduit à l'obtention d'estimateurs de la variance moins biaisés en moyenne et plus précis. De plus, on peut observer que cet estimateur permet une nette diminution de l'erreur de reconstruction de l'effet fixe, particulièrement pour les fonctions `Blocks` et `Bumps`, comme représenté sur la Figure 9.2. Nous avons donc adopté cette stratégie d'estimation au cours des simulations effectuées dans cette partie.

Résultats de reconstruction pour le seuillage hétéroscédastique

Les résultats concernant les précisions de reconstruction de l'effet fixe fonctionnel et d'estimation des variances sont présentés respectivement en Figure 9.3 et 9.4. Des exemples de reconstruction de l'effet fixe fonctionnel sont aussi représentés en Figure 9.5 pour les 4 types d'effets fixes étudiés et pour les deux configurations d'effets aléatoires proposées. L'observation de ces graphes nous conduit à plusieurs conclusions.

- Le seuillage hétéroscédastique conduit à une meilleure reconstruction de l'effet fixe quand les effets aléatoires s'exercent sur les coefficients nuls de l'effet fixe (configuration A). Cette conclusion concerne plus particulièrement le seuillage hétéroscédastique basé sur des estimateurs des variances de type moment et elle est illustrée en Figure 9.3. Les méthodes pénalisées présentent, quant à elles, un comportement similaire à celui du seuillage homoscédastique classique en termes de reconstruction de l'effet fixe dans la configuration A. À l'inverse, lorsque les effets aléatoires s'exercent sur les coefficients non-nuls de l'effet fixe (configuration B), la méthode homoscédastique classique montre alors de meilleures performances pour la reconstruction de l'effet fixe, ce qui

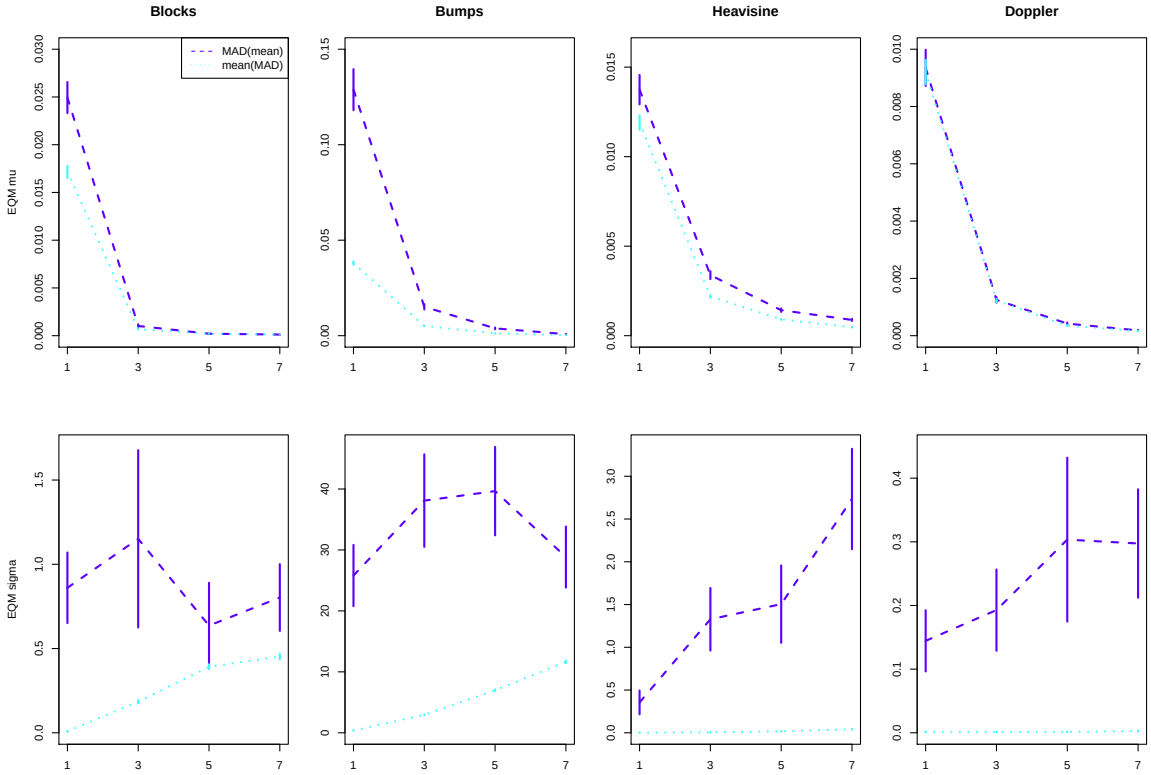


FIGURE 9.2 – *Écarts quadratiques moyens obtenus pour l'estimation de la variance basée sur le MAD des coefficients au niveau le plus fin dans un contexte de régression sans effets aléatoires. Deux stratégies sont comparées : une première où le MAD est calculé sur le signal moyen, noté MAD(mean), et une deuxième consistant à prendre la moyenne des MAD individuels, notée mean(MAD). Sur chaque graphe, l'axe des abscisses correspond aux différentes valeur de SNR et l'ordonnée donne la valeur de l'EQM.*

est cohérent avec le fait que la présence de bruit sur les coefficients non-nuls entraîne une augmentation du nombre de faux négatifs pour les procédures hétéroscédastiques.

- Parmi les stratégies de seuillage hétéroscédastique, le choix de la méthode d'estimation des paramètres de variances est important car elle influence les résultats de reconstruction de l'effet fixe. Nous pouvons constater en Figure 9.4-(a) que la méthode à privilégier dans ce contexte est la méthode de type moment conduisant à une estimation robuste des variances. Le terme de robustesse est employé ici dans le sens où cette méthode conduit à une précision des estimateurs des variances peu dépendantes du niveau de variabilité général. A l'inverse, la précision des stratégies pénalisées dépend fortement de leur ca-

pacité à détecter les positions où s'exercent les effets aléatoires. L'observation des Figures 9.4-(a) et 9.4-(b) nous montre que les méthodes pénalisées sont peu performantes dans la sélection des positions non-nulles, particulièrement lorsque la variabilité des effets aléatoires diminue, entraînant de ce fait, des écarts quadratiques plus importants. Ce comportement en terme de sélection est plutôt attendu car lorsque le niveau de variabilité diminue, les paramètres deviennent proches de zéro numériquement rendant la sélection plus difficile. Néanmoins, nous constaterons au cours des simulations sur l'approche jointe que la sélection peut être améliorée en considérant une stratégie itérative. Nous en déduisons que ce manque d'adaptation à la parcimonie des paramètres est en grande partie lié à la contrainte d'une procédure se déroulant en une seule étape et à la non-consistance des procédures étudiées. Enfin, notons que comme déclaré en Section 7.6.2, la procédure basée sur une pénalisation des variances $\sigma^2_{\theta,jk}$ conduit, numériquement, à des résultats présentant de meilleurs EQM.

- Enfin, une conclusion importante ici concerne le seuillage hétéroscédastique basé sur une estimation empirique des variances à partir des répétitions individuelles. Nous pouvons observer en Figure 9.3 et en Figure 9.5 que cette procédure conduit à une réelle amélioration de l'effet fixe reconstruit dans le cas d'effets aléatoires placés sur des positions nulles (configuration A). Contrairement aux autres procédures basées sur un estimateur MAD du bruit de mesure, cette procédure se base sur une estimation de la variance globale calculée à partir des répétitions individuelles. Cette stratégie a pour effet de conduire à des seuils mieux adaptés au signal sous-jacent en utilisant sur certaines positions un seuil inférieur (contrairement aux procédures basée sur l'estimateur MAD). Ce phénomène est illustré en Figure 9.6 où un exemple comparatif entre les seuillages hétéroscédastiques et le seuillage usuel est représenté pour un effet fixe de type `Blocks`. Ceci est dû à la nature même de l'estimateur MAD classiquement utilisé qui est calculé à partir des coefficients du niveau de décomposition le plus fin d'un même signal : en effet, l'hypothèse sous-jacente est alors que le dernier niveau de résolution est essentiellement constitué de bruit. Or, comme avancé par Donoho et Johnstone (1998), certains coefficients contiennent une part non négligeable de signal à ce niveau de résolution, donnant un estimateur de la variance du bruit systématiquement biaisé. Ceci est particulièrement vrai pour les signaux exhibant des discontinuités car ces dernières entraînent la présence de grands coefficients à tous niveaux de résolution localisés autour des discontinuités. Ce phénomène est vrai dans notre contexte et est illustré en Table 9.1, ou nous comparons le biais de l'estimateur MAD pour les signaux de type `Blocks`, contenant de nombreuses discontinuités, et les signaux `Doppler` qui en présentent peu. On constate que pour l'effet fixe `Blocks`, le biais moyen observé pour l'estimateur MAD, calculé par signal, est nettement supérieur à celui de l'estimateur empirique, calculé à partir des répétitions individuelles. Pour un effet fixe de type `Doppler` ne montrant pas

		Estimateur MAD	Estimateur de type moment
Blocks	SNR = 1	0.0730 (1.5740)	0.0019 (0.1063)
	SNR = 5	0.6313 (3.5078)	0.0118 (0.0074)
Doppler	SNR = 1	-0.0032 (0.0133)	0.0082 (0.0045)
	SNR = 5	0.0070 (3.8×10^{-4})	0.0018 (1.4×10^{-4})

TABLE 9.1 – *Biais moyens observés relatifs pour l'estimateur de type MAD usuel et l'estimateur des variances empiriques (7.8) sur 50 répétitions. Le paramètre τ est fixé à 4, le paramètre SNR prend les valeurs 1 et 5 et deux effets fixes sont considérés* `Blocks` *et* `Doppler`*. Les résultats présentés concernent la configuration A d'effets aléatoires.*

de discontinuités, on observe que la différence entre les deux estimateurs est bien moins significative. Donoho et Johnstone (1998) vont même plus loin en affirmant que la probabilité pour l'estimateur MAD de surestimer le vrai paramètre de variances augmente avec la taille de signal M, signifiant qu'on ne peut *a priori* pas espérer un meilleur comportement en augmentant la taille des signaux. De ce fait, en présence de répétitions individuelles et tout particulièrement pour des signaux présentant de fortes discontinuités, l'utilisation d'estimateurs des variances tirant parti des répétitions semble plus adaptée et permet d'améliorer les performances de reconstruction.

Finalement, nous avons exploré les performances comparées des méthodes homoscédastique et hétéroscédastique en fonction de la configuration des effets aléatoires et au vu de la reconstruction de l'effet fixe fonctionnel. Au sein des méthodes hétéroscédastiques et en présence de répétitions, la méthode basée sur une estimation empirique des variances, comparable numériquement au seuillage homoscédastique classique, montre un comportement relativement stable et meilleur que les stratégies pénalisées dans un cadre de procédures non-itératives.

9.2 Approche jointe et sélection de variables

Nous nous consacrons à présent à l'évaluation des performances de l'approche jointe décrite au Chapitre 8. Cette approche est regardée principalement vis-à-vis de ses performances de sélection des effets fixes et aléatoires, objectif principal dans lequel cette stratégie a été développée.

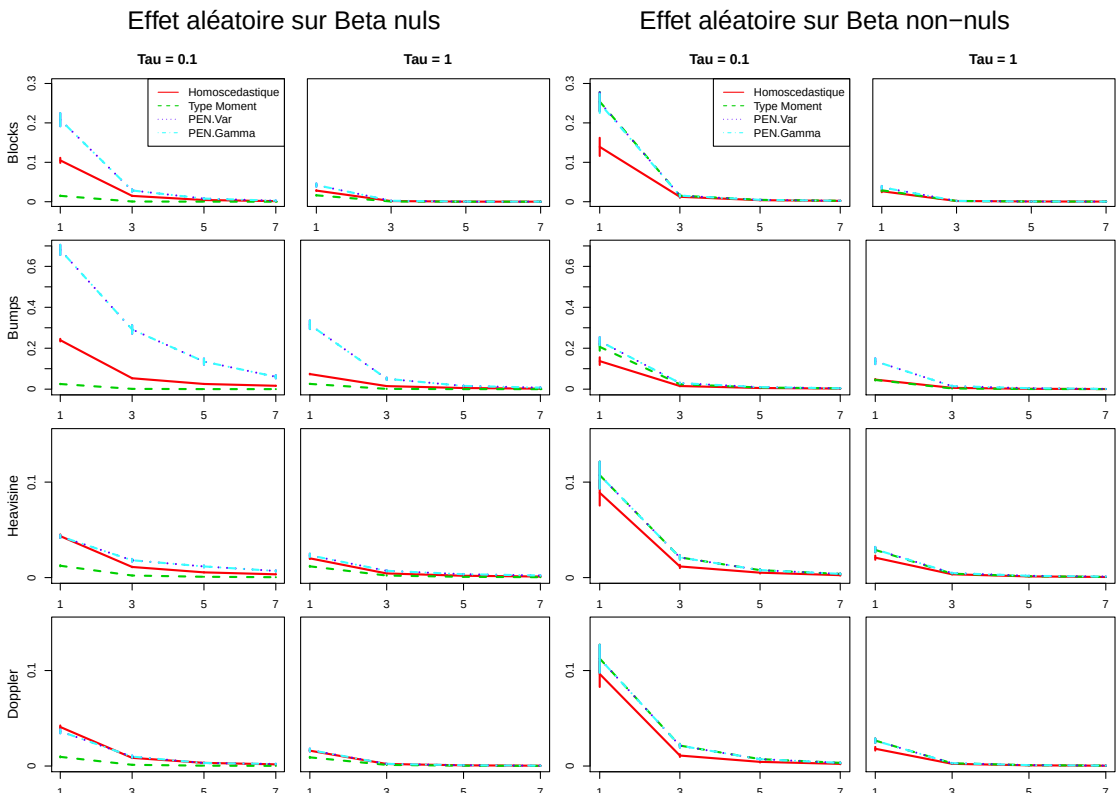
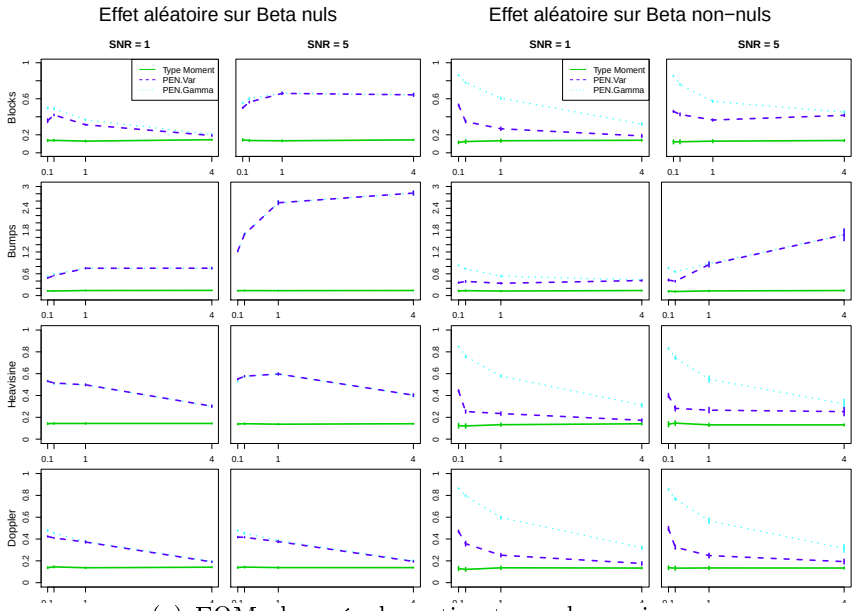
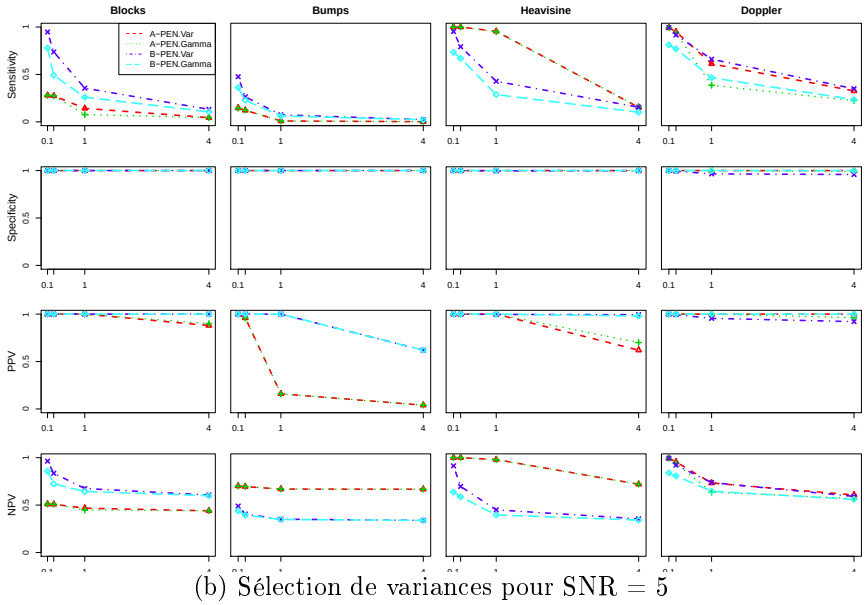


FIGURE 9.3 – *EQM - Résultats de simulations concernant la précision de la procédure de seuillage de l'effet fixe en fonction de la procédure d'estimation de variances considérée, pour des effets aléatoires placés sur les coefficients nuls ou non-nuls et pour les 4 types de fonctions étudiées :* `Blocks`, `Bumps`, `Heavisine` *et* `Doppler`. *Les résultats sont représentés pour* $\tau_U = 0.1$ *et* $\tau_U = 1$. *Au sein de chaque graphe, chaque courbe représente une méthode et varie en abscisse par rapport au paramètre SNR.*



(a) EQM observés des estimateurs des variances



(b) Sélection de variances pour SNR = 5

FIGURE 9.4 – *(a) - Résultats de simulations concernant la précision des procédures d'estimation de variances, pour des effets aléatoires placés sur les coefficients nuls ou non-nuls et pour les 4 types de fonctions étudiées :* `Blocks`, `Bumps`, `Heavisine` *et* `Doppler`. *Les résultats sont représentés pour deux valeurs de SNR (1 et 5). Au sein de chaque graphe, chaque courbe représente une méthode et varie en abscisse par rapport au paramètre* τ_U, *(b) - Résultats de sélection de variances pour SNR = 5 (les comportements en terme de sélection de variables restent sensiblement les mêmes pour d'autres valeurs de SNR).*

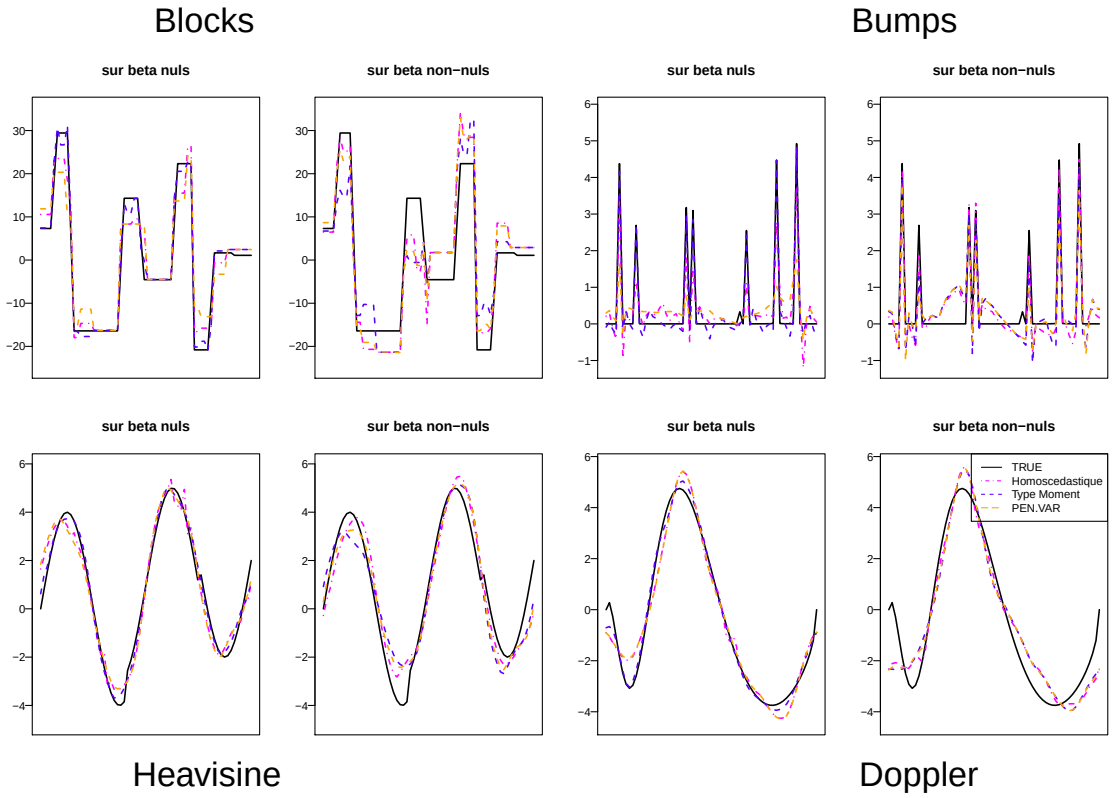


FIGURE 9.5 – *Exemple de reconstruction de l'effet fixe seuillé pour SNR=1 et $\tau = 0.1$ et pour les deux configurations d'effets aléatoires. L'effet fixe original est représenté en noir, la reconstruction de l'effet fixe seuillé grâce à la procédure Pen.Var en orange, celle de l'effet fixe seuillé grâce à une procédure SCAD basée sur l'estimateur MAD en rose et celle reconstruite grâce à des estimations empiriques des variances en bleu. Les exemples présentés correspondent, pour chaque configuration, au jeu de données associé à l'EQM médian pour la reconstruction de l'effet fixe par la procédure notée "Type Moment".*

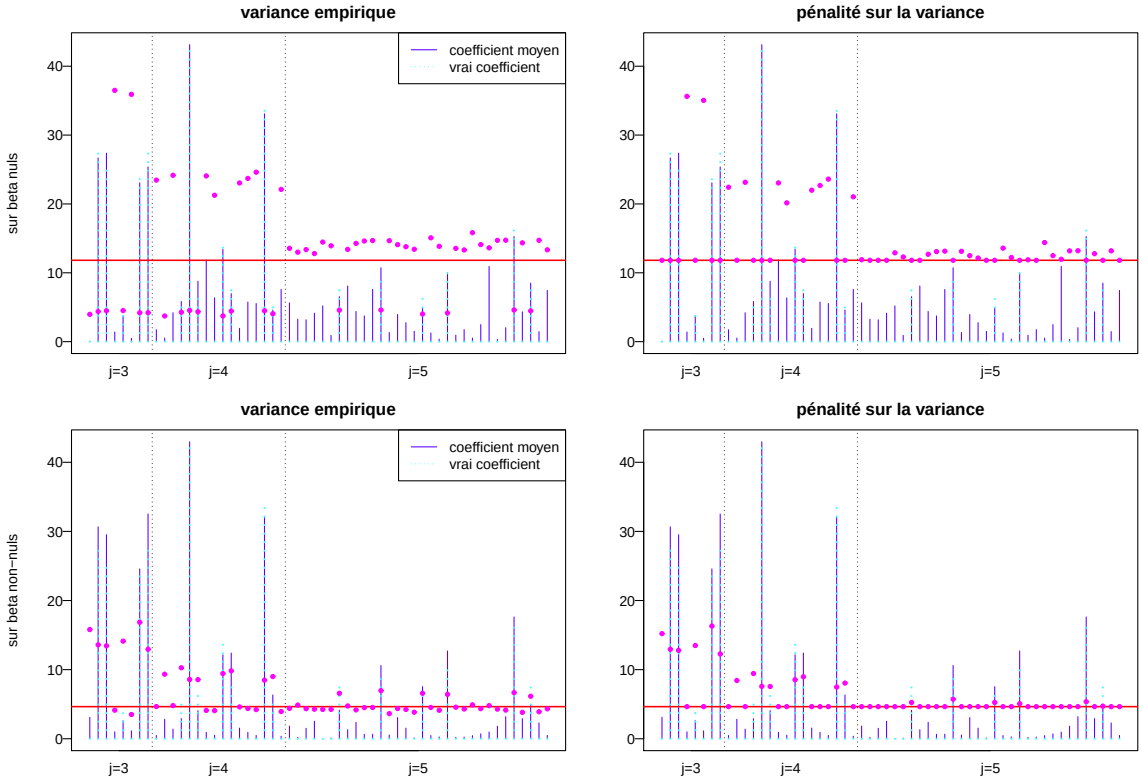


FIGURE 9.6 – *Exemple de seuillage homoscédastique et hétéroscédastique pour SNR=1, $\tau = 0.1$, pour les deux configurations d'effets aléatoires et deux types de seuillage hétéroscédastique (variance empirique et avec pénalité sur la variance). Sur chaque graphe, le seuil homoscédastique usuel est représenté en rouge tandis que les seuils à chaque position sont représentés par des points roses. Les coefficients associés au signal moyen sont en bleu et sont seuillés par les seuillages homoscédastique ou hétéroscédastique s'ils se situent respectivement en dessous du seuil constant en rouge ou du point représenté en rose.*

9.2.1 Données considérées et procédures testées

Données synthétiques pour la sélection

Nous nous basons dans un premier temps sur les jeux de données synthétiques décrits en Section 9.1.1 car ceux-ci possède l'avantage de séparer la parcimonie des effets fixes de celle des effets aléatoires, nous permettant dans ce cadre de mieux appréhender le comportement de notre procédure basée sur une approche jointe des modèles mixtes, vis-à-vis de la sélection des effets fixes et aléatoires séparément. Ici encore, la configuration A est destinée à conduire à de meilleures performances en terme de sélection des effets fixes et aléatoires. En effet, l'estimateur (8.13) des effets fixes dans l'approche jointe s'apparente à un seuillage SCAD usuel des coefficients corrigés des prédictions des effets aléatoires. Ainsi, dans la configuration A et en faisant abstraction de la problématique de sélection des variances des effets aléatoires, le seuillage est plus fort sur les coefficients nuls conduisant à l'obtention de moins de coefficients sélectionnés à tort (faux positifs) tandis que dans la configuration B, le contraire est valable, à savoir, un seuillage moins important sur les coefficients effectivement nuls, qui ne sont pas touchés par les effets aléatoires, pouvant conduire alors à l'obtention de plus de faux positifs. Par ailleurs, l'estimateur (8.16) des variances des effets aléatoires se ramène aussi à un seuillage des données, corrigées des prédictions des effets fixes, et devrait conduire de ce fait à un comportement, là encore, meilleur dans la configuration A par les mêmes arguments.

Stratégies étudiées

Nous appliquons à ces jeux de données notre procédure de sélection basée sur une double pénalisation de la vraisemblance du modèle par rapport aux effets fixes et aléatoires. Un des points central de l'ajustement du modèle réside dans le choix des paramètres de régularisation λ_1 et λ_2. Ainsi, nous proposons une première stratégie durant laquelle, ces deux paramètres sont fixés au moyen d'un critère de type BIC comme défini en Section 8.3.2. Les paramètres (λ_1, λ_2) parcourent une double grille suivant une échelle logarithmique, le couple étant ensuite fixé à la valeur minimisant le critère BIC. Cette procédure peut se révéler coûteuse d'un point de vue numérique car elle nécessite un grand nombre d'appel à l'algorithme EM (de l'ordre de la centaine suivant le pas de grille choisi), dont un seul appel est déjà relativement coûteux. Afin de limiter ce coût, nous en proposons une variante consistant à remplacer le paramètre λ_1 par le seuil universel homoscédastique égal à $\sigma_\varepsilon^{[h]}\sqrt{2\log M/N}$, dépendant de l'itération courante $[h]$ de l'algorithme. Cela peut se justifier par le fait que le paramètre λ_1 correspond à la sélection d'effets fixes et est donc lié à un problème d'estimation d'une fonction μ observée dans un bruit. On est donc, pour ce paramètre, dans un cadre similaire à celui développé par Donoho et Johnstone (1994). Pour le paramètre λ_2, lié à une problématique de sélection de variances, nous choisissons de conserver la grille logarithmique telle que définie précédemment. Cette stratégie sera par la suite appelée "Minimax" tandis que la procédure basée sur une

optimisation selon une double grille sera notée "BIC". Nous proposons en outre une version relaxée de ces deux procédures, comme décrit en Section 3.3.2 (Meinshausen 2007), dans un but de stabilisation de l'estimation des paramètres du modèle. Ces deux dernières seront notées respectivement "BIC.relaxe" et "Minimax.relaxe".

Enfin, nous utilisons comme critère d'évaluation ceux utilisés pour l'approche marginale (cf. Section 9.1.2) concernant les performances de sélection et de précision des estimateurs.

9.2.2 Résultats

Nous donnons à présent les résultats obtenus sur les données simulées. Les résultats principaux sont donnés au vu de l'objectif de sélection des effets fixes d'une part et des effets aléatoires d'autre part.

Sélection des effets fixes

Les résultats de sélection concernant la sélection des effets fixes sont présentés en Figure 9.7 où sont représentés pour les 4 types d'effets fixes, les critères de sensibilité, de spécificité ainsi que les critères PPV et NPV. Les graphiques correspondent à une valeur τ_U fixée à 0.1 et nous signalons que les conclusions restent les mêmes dans le cas d'effets aléatoires plus modérés. En Figure 9.8, nous avons représenté les écarts quadratiques moyens obtenus pour la reconstruction de l'effet fixe fonctionnel pour les 4 types d'effets fixes considérés et plusieurs intensités d'effets aléatoires.

- La configuration A, où les effets aléatoires sont placés sur les positions nulles des effets fixes, favorise bien la sélection puisqu'on observe de meilleures performances générales de sélection pour l'ensemble des procédures considérées (cf Figure 9.7). Une exception est à faire concernant la procédure BIC qui présente un comportement moins bon en terme de spécificité, c'est-à-dire dans sa capacité à retrouver les positions nulles. Cependant, cette procédure est la plus instable car elle nécessite l'ajustement de deux hyperparamètres et on peut observer que la version relaxée de cette dernière permet de stabiliser la sélection.
- Dans la configuration B, désavantageuse quant à la sélection des effets fixes, l'utilisation d'un seuil basé sur le seuil minimax pour le choix du paramètre λ_1 conduit les procédures "Minimax" et "Minimax.relaxe" à ne pas distinguer la présence d'effets fixes sur les coefficients non nuls dans le cas d'une variabilité élevée. En effet, dans ce contexte, on observe pour ces deux procédures des sensibilité et des valeurs de NPV proches de zéro signifiant que très peu de coefficients sont sélectionnés. Ce phénomène s'observe notamment du point de vue de la sélection des variances des effets aléatoires qui, au contraire, se révèle performante dans ce cas particulier (détaillée au paragraphe sur la sélection des variances). De plus, dans la configuration B, on observe aussi que la procédure BIC a tendance *a contrario* à sélectionner beaucoup de coefficients conduisant

à de faibles valeurs de spécificité et de PPV, confortant l'idée d'instabilité de la procédure BIC.

- Bien que n'étant pas l'objectif premier des procédures présentées au Chapitre 8, nous pouvons observer en Figure 9.8 que les écarts quadratiques moyens associés à la reconstruction de l'effet fixe fonctionnel restent dans des gammes de valeurs faibles, comparables à celle obtenues pour le seuillage hétéroscédastique associé à une estimation des variances de type moment, présenté au Chapitre 7 et ceci reste valable pour les deux configurations particulières étudiées et l'ensemble des procédures testées. Ce résultat est encourageant et nous incite à nous intéresser, dans une perspective de ce travail, aux propriétés de reconstruction de l'estimateur de l'effet fixe fonctionnel en terme de risque quadratique.

Sélection des variances des effets aléatoires

Les résultats de sélection concernant la sélection des variances des effets aléatoires sont présentés de manière similaire. En Figure 9.9, les critères de sensibilité, de spécificité ainsi que les critères PPV et NPV sont représentés pour les 4 types d'effets fixes. Les graphiques correspondent à une valeur de SNR fixée à 1 mais les conclusions restent aussi les mêmes dans le cas de paramètres SNR plus élevés.

- On observe en premier lieu en Figure 9.9 que, dans les deux configurations, les performances de sélection de variances sont influencées par le niveau d'effets aléatoires présents et deviennent meilleures lorsque le niveau d'effets individuels est fort : en effet, dans ce cas, la discrimination entre les positions affectées par un effet aléatoire et celles qui ne le sont pas est alors facilitée.
- De plus, entre les différentes procédures, la différence de comportement se situe ici entre les méthodes relaxées et les non relaxées. Cela nous conduit à conclure que l'étape de réestimation revêt une importance particulière pour la sélection des variances associées aux effets aléatoires. Ceci est particulièrement vrai dans la configuration B, désavantageuse pour la sélection, où pour des forts niveaux d'effets aléatoires, les méthodes non relaxées réalisent peu de sélection (c'est-à-dire, sélectionnent plus de positions que nécessaire), se traduisant alors par des valeurs de spécificité et de PPV proches de zéro. Ceci est illustré en Figure 9.10 présentant un exemple d'estimation des variances pour la configuration B avec un fort niveau d'effets individuels entre les méthodes "Minimax" et "Minimax.relaxe". La présence de forts effets aléatoires est pourtant sensée conduire à une sélection facilitée et ce phénomène nous conduit donc à préférer l'utilisation de méthodes relaxées pour la sélection/estimation des variances associées aux effets aléatoires.

Temps de calcul associés aux procédures de sélection

Les temps de calcul associés aux procédures basées sur l'algorithme EM sont aussi à prendre en compte : en effet, la procédure de seuillage hétéroscédastique basée sur une estimation des variances de type moment ne nécessite pas de puissance de calcul particulière et le résultat est obtenu de manière immédiate. Ce n'est pas le cas pour les procédures décrites dans cette section, qui sont, d'une part, basée sur une optimisation au moyen d'un algorithme itératif et d'autre part, nécessitent le réglage d'hyperparamètres, notamment pour la procédure BIC où l'on doit fixer les deux paramètres de régularisation λ_1 et λ_2. Nous présentons en Table 9.2 les temps de calcul moyens obtenus pour les procédures de sélection BIC et Minimax pour les deux configurations particulières étudiées et différentes valeurs de SNR, τ_U et d'effet fixe. Les temps de calcul associés aux méthodes relaxées sont comparables à leur version non relaxées car l'étape de réestimation possède un coût numérique négligeable devant celui des procédures itératives complètes.

Nous constatons sur cette table que les procédures de sélection Minimax où le paramètre λ_1 est préalablement fixé permettant un vrai gain en terme de temps de calcul. Un deuxième comportement particulier peut être mis en avant : dans la configuration A, la diminution du niveau de variabilité globale par le biais de SNR et τ_U conduit à des temps de calcul plus importants, tandis que dans la configuration B, la baisse du niveau d'effets aléatoires entraîne une diminution du temps de calcul. Ce type de phénomène est cohérent avec les conclusions réalisées quant à la sélection des variances associées aux effets aléatoires, on peut donc penser que la difficulté de sélection/estimation des variances associées aux effets aléatoires influence fortement le coût numérique global de la procédure.

Au terme de cette première étude concernant les procédures de sélection des effets fixes et aléatoires basées sur une approche jointe des modèles mixtes, deux conclusions principales peuvent être mises en avant : d'une part, l'utilisation de méthodes relaxées telles que proposées par Meinshausen (2007) permettent de stabiliser le processus de sélection et de ce fait, d'améliorer la précision des estimateurs résultants. De plus, en offrant un temps de calcul bien inférieur, nous privilégierons l'usage de la procédure appelée "Minimax.relaxe“ où l'hyperparamètre associé à l'estimation des effets fixes est fixé au seuil universel.

Néanmoins, les configurations simulées dans cette première étude sont peu réalistes car elles séparent les parcimonies des effets fixes et aléatoires. Nous comparerons dans une deuxième partie les procédures étudiées selon les approches marginale et jointe sur des données synthétiques simulées de manière plus réaliste.

9.3 Comparaison des approches sur données réalistes

Nous avons jusqu'à présent développé deux types d'approches concernant l'estimation au sein des modèles mixtes et cela dans des objectifs différents. La première approche décrite au Chapitre 7 se résume à un seuillage hétéroscédastique des données combiné à une étape d'estimation des variances aux différentes positions utilisées par une stratégie de type plug-in. L'étude de simulation réalisée sur cette dernière fait apparaître de bonnes propriétés de reconstruction de l'effet fixe fonctionnel lorsque l'estimation des variances est réalisée par une méthode de type moment, avec l'idée sous-jacente que les stratégies de sélection développées dans ce cadre sont peu performantes et dégradent de ce fait la reconstruction de l'effet fixe fonctionnel. *A contrario*, l'approche développée au Chapitre 8 présentant des procédures basées sur une double pénalisation de la vraisemblance vis-à-vis des effets fixes et des variances des effets aléatoires par des pénalités de type SCAD s'attache plutôt aux propriétés de sélection de ces variables. Les simulations réalisées dans ce sens montrent effectivement de bons comportements de ces procédures en terme de sélection des effets fixes et aléatoires mais aussi de bons résultats de reconstruction de l'effet fixe fonctionnel ouvrant la voie à la recherche de propriété de convergence de l'estimateur fonctionnel associé.

Notre but dans cette section est double : d'une part, les procédures développées dans cette partie ont été pour l'instant évaluées sur la base de jeux de données synthétiques présentant des structures particulières, où les parcimonies associées aux effets fixes et aléatoires sont considérées de manière séparée par le biais des configurations A et B. Ces deux configurations sont mélangées lors de l'étude de données réelles et notre volonté est alors d'en étudier les implications sur les performances respectives des procédures développées. Nous souhaitons pour cela construire des jeux de données synthétiques *réalistes* et nous entendons par le terme réaliste, des jeux de données pour lesquels les configurations A et B sont mélangées et pour lesquels la variabilité des effets aléatoires dans le domaine des coefficients a un sens d'un point de vue fonctionnel. De plus, nous nous attacherons sur ces nouvelles données synthétiques à la comparaison des procédures correspondant à l'approche marginale d'une part et l'approche jointe d'autre part, en terme de sélection de variables et plus particulièrement en terme de reconstruction de l'effet fixe fonctionnel.

9.3.1 Simulation de données réalistes

Nous adoptons ici un plan de simulation similaire à celui décrit en Section 9.1.1. L'unique différence réside dans la définition des structures des effets aléatoires dans le domaine des coefficients, c'est-à-dire, dans la définition des positions (j, k) auxquelles correspondent une variabilité non nulle des effets aléatoires dans le domaine de ondelettes.

Pour ce faire, nous nous inspirons d'un travail réalisé par Amato et Sapatinas (2005), évoqué en Section 7.3, lors duquel les auteurs, dans un contexte de seuillage des effets fixes au sein de modèles mixtes fonctionnels, proposent une définition analytique de jeux de données simulant la présence d'effets individuels autour des fonctions `Blocks`, `Bumps`, `Heavisine` et `Doppler` de Donoho et Johnstone (1994). Dans leur approche, Amato et Sapatinas (2005) postulent que l'influence des effets aléatoires pour des fonctions de type `Blocks` s'exprime sur la variabilité de la place et de la hauteur des discontinuités. Pour les fonctions de type `Bumps`, elle se traduit par une variabilité de la largeur et de la hauteur des pics, par une variabilité de la localisation des discontinuités pour les fonctions de type `Heavisine` et par une variabilité de la fréquence pour les fonctions de type `Doppler`.

À partir de la simulations d'effets aléatoires fonctionnels, nous en déduisons une structure d'effets aléatoires individuels en projetant dans le domaine des ondelettes les effets aléatoires individuels fonctionnels qui possèdent alors une représentation parcimonieuse. Cela nous permet de repérer pour chaque individu, les positions auxquelles correspondent des valeurs non nulles d'effets aléatoires. Par une stratégie d'union sur ces positions, on en déduit alors une structure pour les effets aléatoires individuels, c'est-à-dire, une liste de couples (j, k) qui correspondront, dans le domaine des coefficients, aux positions associées à des variances d'effets aléatoires non-nulles sur les données synthétiques.

En Figure 9.11, nous avons représenté les différents effets fixes ainsi que les écart-types empiriques des données simulées, estimées à chaque position, pour plusieurs valeurs de SNR et τ_U. De plus, en Figure 9.12, nous avons représenté pour chaque type d'effet fixe, les enveloppes de données pour des exemples de données simulées, et cela pour plusieurs valeurs de SNR et τ_U.

9.3.2 Comparaison des approches marginale et jointe

Sur ces jeux de données simulés de manière réalistes, nous nous attachons à comparer les performances des procédures basées sur des approches marginales et jointe des Chapitres 7 et 8. Nous nous limitons pour cela à la comparaison de 4 procédures particulières : la procédure de seuillage classique notée "Homoscédastique", la procédure de seuillage hétéroscédastique basée sur une estimation des variances de type moment, qui présente les performances les plus intéressantes après les premières simulations (cf. Section 9.1). Cette dernière sera notée "Type Moment". Enfin, nous considérons en outre deux procédures issues de l'approche jointe : les procédures "BIC.relaxe" et "Minimax.relaxe". Nous choisissons d'utiliser les versions relaxées des procédures développées car au vu des simulations précédentes (cf Section 9.2), l'étape de relaxation apporte un véritable bénéfice en terme de stabilisation des estimations.

Les 4 procédures considérées dans cette Section sont comparées essentiellement sur les écarts quadratiques moyens associés à la reconstruction de l'effet fixe fonc-

tionnel et aux estimations de la variabilité globale à chaque position. Nous nous intéresserons aussi aux résultats de sélection concernant les effets fixes et aléatoires basés sur les critères de sensibilité, de spécificité, PPV et NPV. Les définitions des critères utilisés sont donnés en Section 9.1.2. Enfin, nous comparerons les temps de calcul des diverses procédures mises en compétition.

Résultats

- En Figure 9.13-(b) sont représentées les performances de sélection des variables du vecteur $\boldsymbol{\beta}$ associées aux effets fixes pour les quatre procédures considérées. On observe en premier lieu que la sélection des effets fixes de type `Heavisine` et `Doppler` montre une faible sensibilité pour l'ensemble des procédures considérées. Cela peut être imputé à la régularité de ces fonctions par rapport aux signaux `Blocks` et `Bumps`, entraînant de ce fait des coefficients s'écrasant rapidement lorsque le niveau de résolution j augmente. On observe alors en Figure 9.13-(a), où sont représentés les écarts quadratiques moyens associés à l'estimation des effets fixes fonctionnels reconstruits, que les performances d'estimation sont alors dégradées pour les procédures itératives basées sur une approche jointe dans le cas des effets fixes `Heavisine` et `Doppler`. Cela montre que les performances de sélection ont une influence sur la reconstruction de l'effet fixe fonctionnel alors que celle-ci est moindre dans le cas des procédures de seuillage homoscédastique ou hétéroscédastique.
  Pour des fonctions régulières, dans un objectif d'estimation de l'effet fixe fonctionnel, on est alors conduit à privilégier le seuillage usuel homoscédastique, basé sur une estimation robuste de la variance, qui présente des performances équivalentes ou meilleures que les procédures hétéroscédastiques.
- Pour des effets fixes présentant des discontinuités (`Blocks` et `Bumps`) et malgré des spécificités légèrement inférieures, les méthodes itératives "BIC.relaxe" et "Minimax.relaxe" conduisent à de meilleures performances de reconstruction de l'effet fixe fonctionnel, les rendant préférables dans ce cadre.
- En Figure 9.14-(b) sont représentées les performances de sélection pour les procédures "BIC.relaxe" et "Minimax.relaxe" offrant une sélection des paramètres de variances associés aux effets aléatoires. On observe que les sélections sont équivalentes pour ces deux procédures tandis que la Table 9.3 montre un réel gain de temps de calcul pour la procédure "Minimax.relaxe" où un seul hyperparamètre est à déterminer. Cela nous conduit alors à privilégier cette dernière par rapport à la procédure "BIC.relaxe".

En Figure 9.15, nous avons représenté des exemples de reconstruction des effets fixes fonctionnels `Blocks`, `Bumps`, `Heavisine` et `Doppler` pour $\tau_U = 0.1$ et pour des SNR de 1 ou 5.

En résumé, nous avons étudié au cours de ce chapitre les comportements des procédures selon des approches marginale ou jointe vis-à-vis de la problématique d'estimation au sein des modèles mixtes fonctionnels. L'approche marginale, basée sur les techniques de seuillage non paramétrique couramment utilisées dans un cadre ondelettes présentent des points forts : elle exhibe de bonnes performances de reconstruction de l'effet fixe fonctionnel associées à une grande rapidité d'exécution pour des variances estimées à chaque position par les estimateurs empiriques usuels. Cette approche se révèle néanmoins peu adaptée au deuxième objectif de cette partie, à savoir, la sélection des effets aléatoires. Dans cet objectif, les procédures basées sur une approche jointe des modèles mixtes, prenant explicitement en compte la présence d'effets aléatoires, se révèlent performantes et permettent d'effectuer une sélection simultanée des effets fixes et aléatoires. La meilleure stratégie pour ce type d'approche est la procédure pour laquelle seul l'hyperparamètre λ_2 associé aux effets aléatoires est à déterminer tandis que la valeur de λ_1 est fixée à la valeur du seuil universel. De plus, l'étude de simulation réalisée dans ce chapitre fait apparaître un gain de précision et une meilleure sélection grâce à l'utilisation d'une version relaxée de la procédure de sélection. L'approche jointe n'est étudiée ici que d'un point de vue de la sélection des variables d'effets fixes et aléatoires pour lesquelles des propriétés oraculaires ont été démontrées au Chapitre 8, on peut cependant se demander quelles sont les propriétés de l'estimateur reconstruit de l'effet fixe fonctionnel. Les premiers résultats de simulations sont encourageants et nous incitent à poursuivre en ce sens dans la mesure où les gammes de risques quadratiques obtenus sont comparables à ceux obtenus pour l'approche marginale.

Une prochaine étape nécessaire est de réaliser une étude de simulation en considérant des tailles de signaux M qui augmente afin de se placer dans un cadre asymptotique et comparer les approches marginale et jointe dans ce contexte. Ceci représente une perspective directe de ce travail et n'a pas encore été réalisé car une telle étude nécessite, en particulier pour l'approche jointe basée sur l'algorithme EM, des ressources numériques importantes. Notons néanmoins qu'on ne peut pas espérer un meilleur comportement de l'estimateur MAD au vu des propriétés de cet estimateur mentionnées par Donoho et Johnstone (1998) (p.23) et les résultats d'une étude de simulation restreinte, non présentée ici, sur signaux plus conséquents (M=2048) vont dans le même sens. Nous espérons de ce fait retrouver nos conclusions avec une taille de signaux plus grande.

Enfin, après une étude de simulation sur données simulées de manière réaliste, une autre perspective importante de ce travail est d'étudier le comportement de ces procédures sur des données réelles issues du domaine de la biologie moléculaire. Pour ce type de données, une meilleure compréhension de la variabilité inter-individuelle ainsi qu'une estimation performante du comportement moyen au sein d'une population homogène représente à l'heure actuelle un défi et un espoir pour la compréhension des mécanismes biologiques mis en jeux.

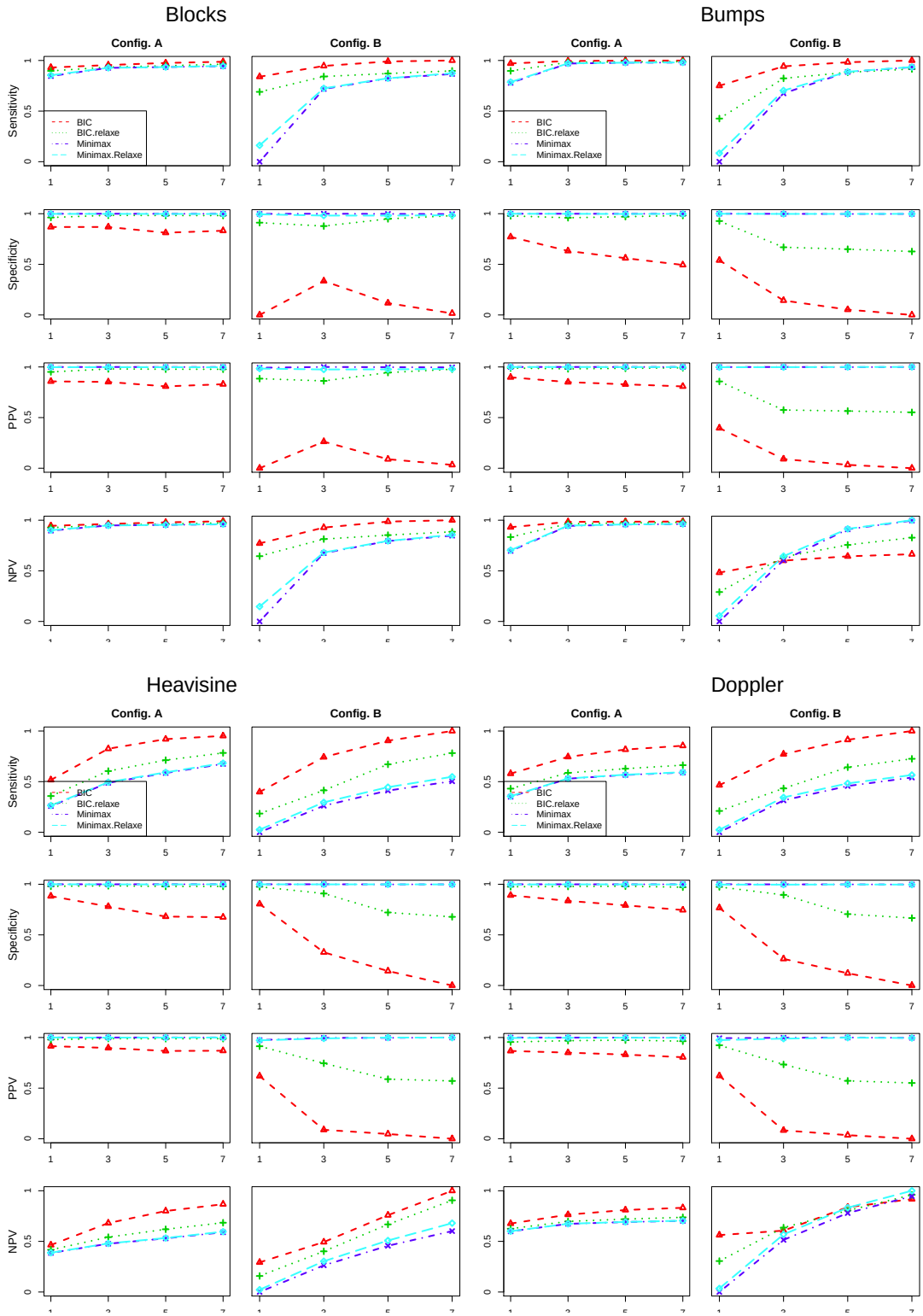


FIGURE 9.7 – *Résultats de simulations concernant les performances de sélection pour les effets fixes et pour 4 procédures d'estimation/sélection basées sur l'algorithme EM. Les graphes correspondent aux 4 types de fonction* `Blocks`, `Bumps`, `Heavisine` *et* `Doppler`*; les résultats sont représentés pour* $\tau_U = 0.1$ *pour les configurations A et B. Chaque ligne correspond à un critère de sélection, parmi la sensibilité, la spécificité, le PPV ou le NPV. Au sein de chaque graphe individuel, chaque courbe représente une procédure et varie en abscisse par rapport au paramètre SNR.*

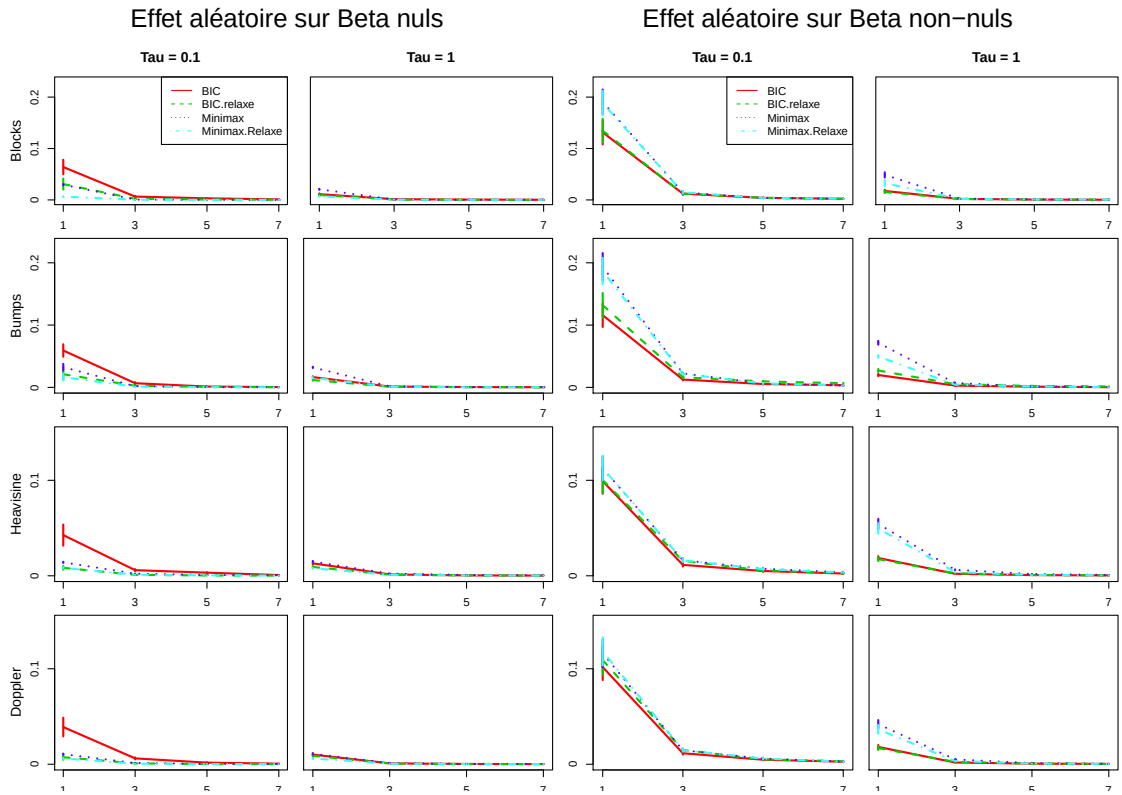


FIGURE 9.8 – *EQM observés par rapport à la reconstruction de l'effet fixe fonctionnel pour 4 procédures d'estimation/sélection basées sur l'algorithme EM et pour les 4 types de fonctions étudiées. Les résultats sont représentés pour pour* $\tau_U = 0.1$ *et* $\tau_U = 1$ *et les deux configurations A et B. Au sein de chaque graphe individuel, chaque courbe représente une procédure et varie en abscisse par rapport au paramètre SNR.*

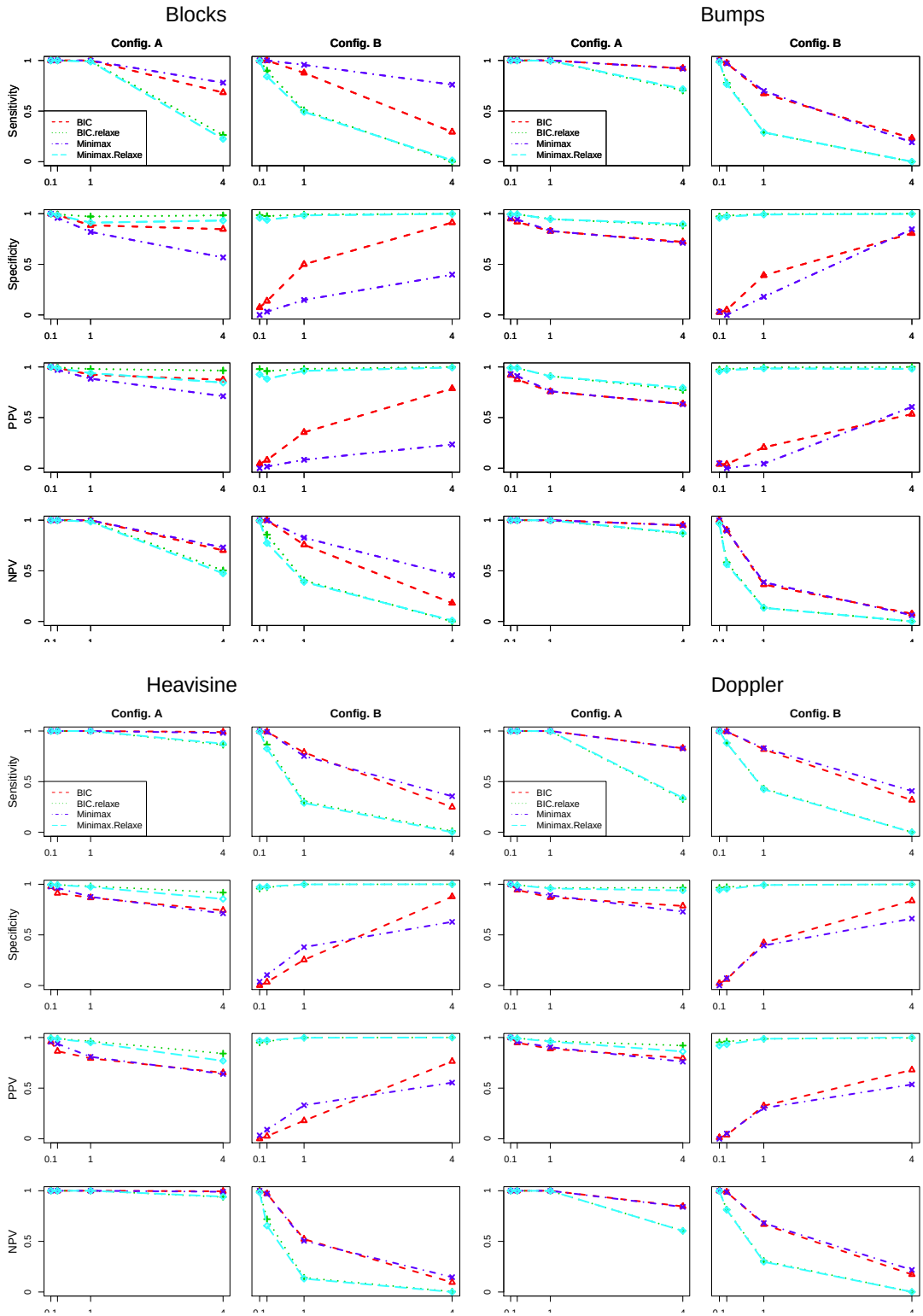


FIGURE 9.9 – *Résultats de simulations concernant les performances de sélection des variances associées aux effets aléatoires pour 4 procédures d'estimation/sélection basées sur l'algorithme EM. Les graphes correspondent aux 4 types de fonction. Les résultats sont représentés pour* $SNR = 1$ *pour les configurations A et B. Chaque ligne correspond à un critère de sélection, parmi la sensibilité, la spécificité, le PPV ou le NPV. Au sein de chaque graphe individuel, chaque courbe représente une procédure et varie en abscisse par rapport au paramètre SNR.*

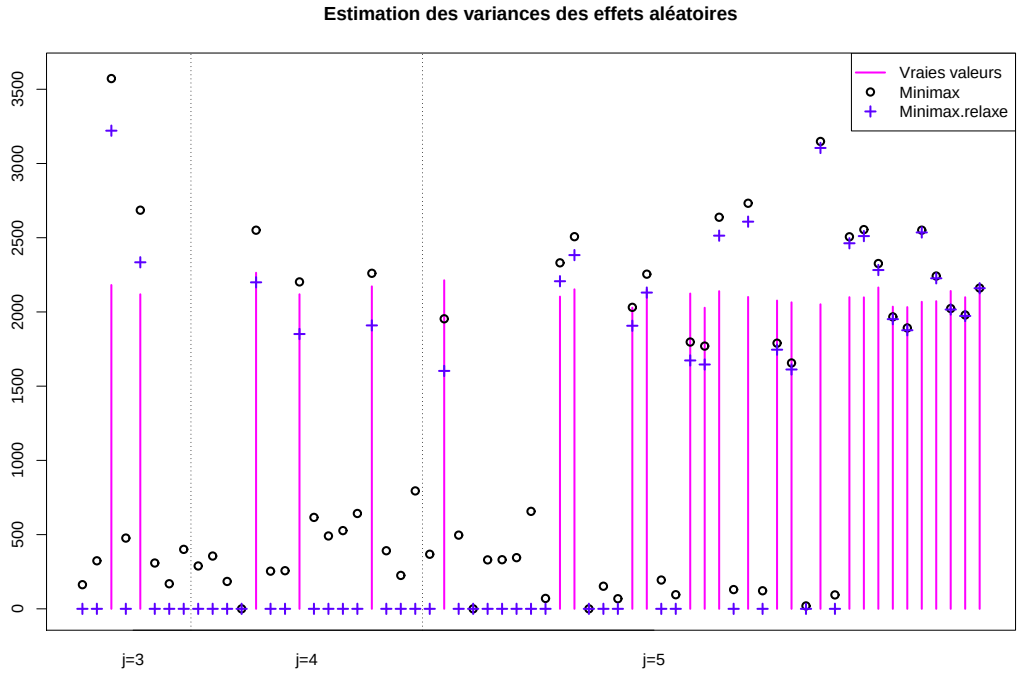


FIGURE 9.10 – *Exemple d'estimation de variances dans la configuration B. Les vraies valeurs de variances sont représentées en rose tandis que les variances estimées par les procédures Minimax et Minimax.relaxe sont représentées respectivement en noir et bleu. Les valeurs de SNR et τ_U sont fixées respectivement à 5 et 0.1.*

			Configuration A		Configuration B	
	τ_U	SNR	1	7	1	7
Blocks	0.1	BIC	45.3 (1.4)	164.5 (20.52)	1086.5 (519.8)	1519.3 (559.6)
		Minimax	19.5 (1.8)	30.2 (2.3)	101.6 (52.5)	148.8 (63.1)
		Type Moment	0.009 (0.002)	0.009 (0.002)	0.01 (0.005)	0.009 (0.002)
	4	BIC	158.6 (12.5)	467.4 (54.35)	342.0 (35.7)	573.04 (85.5)
		Minimax	31.3 (2.0)	60.1 (6.5)	57.5 (7.5)	67.8 (6.6)
		Type Moment	0.01 (0.002)	0.01 (0.003)	0.009 (0.002)	0.009 (0.002)
Bumps	0.1	BIC	196.1 (17.9)	760.5 (77.7)	1342.5 (559.1)	1805.6 (457.1)
		Minimax	26.4 (2.1)	66.7 (5.1)	100.0 (58.0)	139.5 (49.8)
		Type Moment	0.009 (0.002)	0.01 (0.005)	0.009 (0.002)	0.009 (0.002)
	4	BIC	357.8 (37.4)	1082.8 (143.7)	805.8 (110.0)	1225.6 (231.6)
		Minimax	40.2 (3.9)	81.6 (10.8)	88.2 (11.6)	98.1 (25.9)
		Type Moment	0.009 (0.002)	0.009 (0.002)	0.009 (0.002)	0.008 (0.001)
Heavisine	0.1	BIC	150.0 (12.6)	598.6 (69.5)	1460.7 (496.2)	1723.0 (413.5)
		Minimax	23.0 (1.5)	52.8 (4.5)	117.3 (49.3)	165.4 (55.1)
		Type Moment	0.01 (0.005)	0.011 (0.005)	0.009 (0.002)	0.009 (0.002)
	4	BIC	301.8 (29.2)	1035.4 (111.9)	841.1 (137.5)	1260.3 (170.1)
		Minimax	37.0 (3.6)	83.6 (9.4)	82.2 (12.7)	97.5 (14.3)
		Type Moment	0.01 (0.005)	0.009 (0.002)	0.008 (0.002)	0.009 (0.002)
Doppler	0.1	BIC	99.6 (14.7)	449.6 (52.0)	1684.8 (682.4)	1801.8 (539.8)
		Minimax	20.2 (2.2)	45.7 (4.0)	139.4 (54.3)	157.8 (65.7)
		Type Moment	0.009 (0.002)	0.011 (0.006)	0.01 (0.005)	0.009 (0.002)
	4	BIC	278.4 (22.9)	1044.4 (101.0)	710.2 (86.7)	1326.8 (162.9)
		Minimax	34.7 (2.7)	81.2 (9.5)	71.4 (9.4)	96.5 (14.7)
		Type Moment	0.01 (0.003)	0.009 (0.003)	0.009 (0.002)	0.009 (0.005)

TABLE 9.2 – *Temps de calculs moyens et écarts-types (entre parenthèse) en secondes associés aux procédures "BIC", "Minimax" et "Type Moment" pour les configurations A et B et pour différentes valeurs de* (μ, SNR, τ_U). *Le coût numérique de réestimation pour les procédures relaxées est négligé devant le coût total des procédures itératives.*

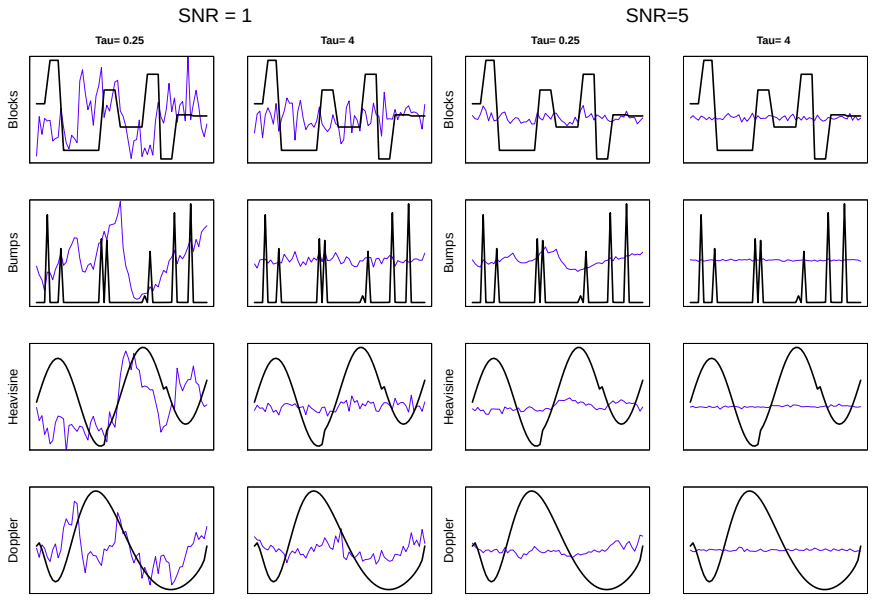


FIGURE 9.11 – *Effets fixes et écart-types empiriques représentés pour plusieurs valeurs de SNR et τ_U. Les écart-types estimés sont représentés en bleu sur les graphes et on peut observer l'effet des variations du couple (SNR,τ_U) sur la variabilité des données simulées.*

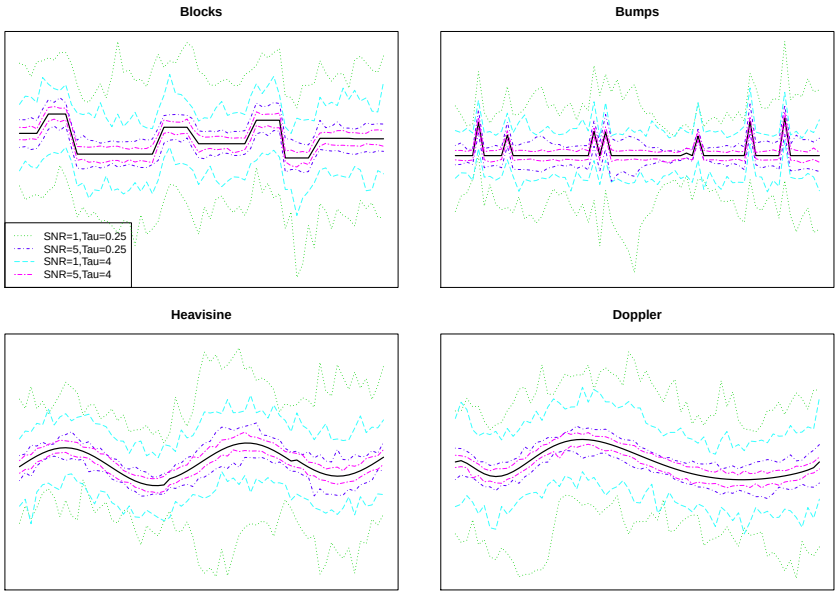
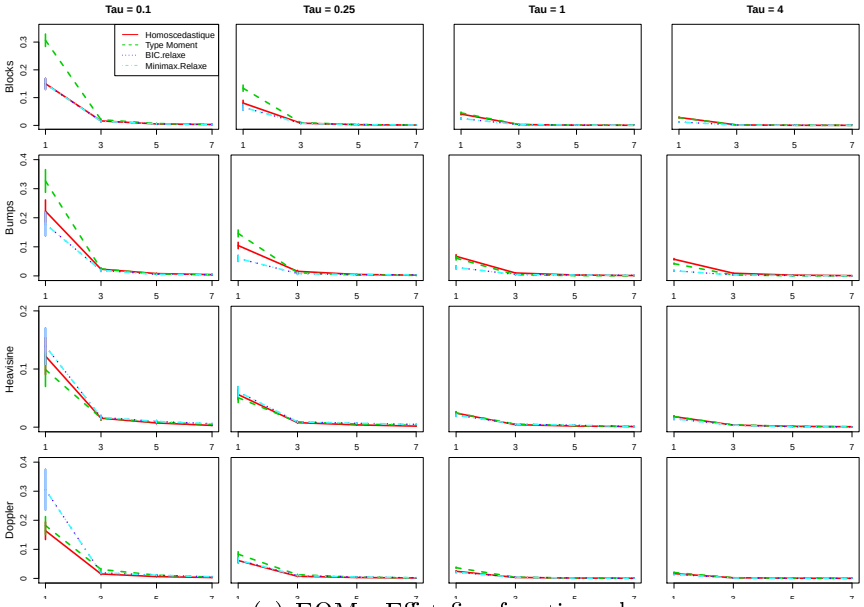
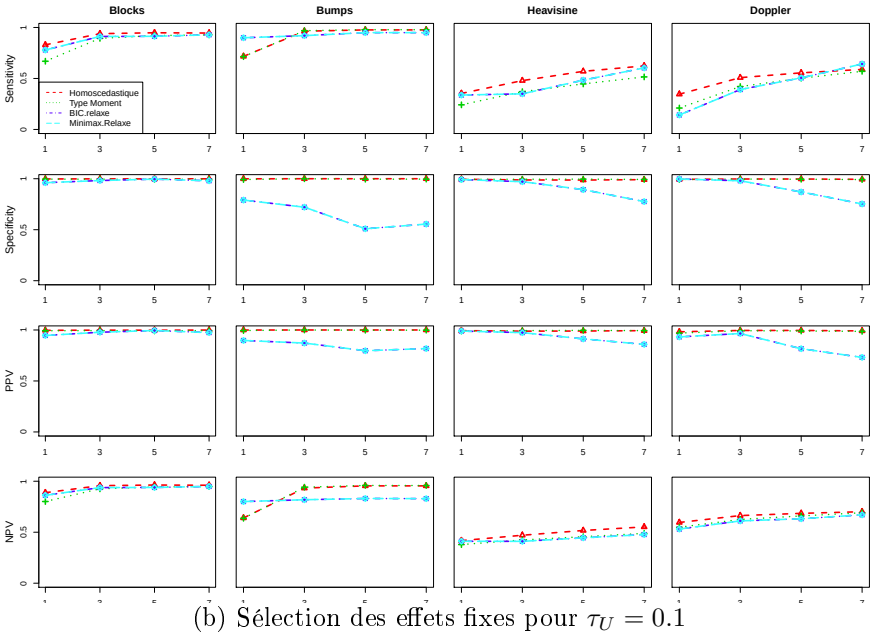


FIGURE 9.12 – *Exemples d'enveloppes de données simulées pour les 4 types d'effets fixes. L'effet fixe est représenté en noir tandis que les enveloppes des données simulées sont représentées en couleur pour plusieurs couples (SNR,τ_U).*

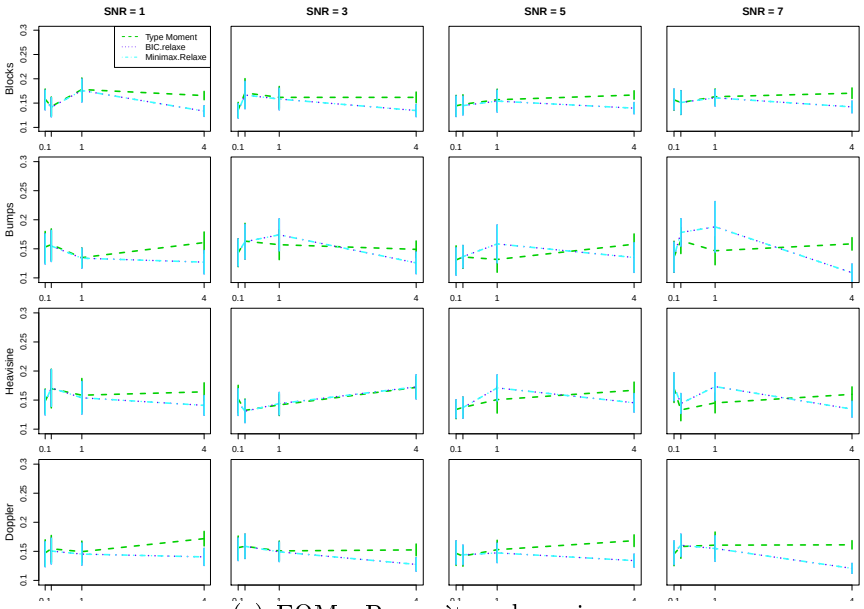


(a) EQM - Effet fixe fonctionnel

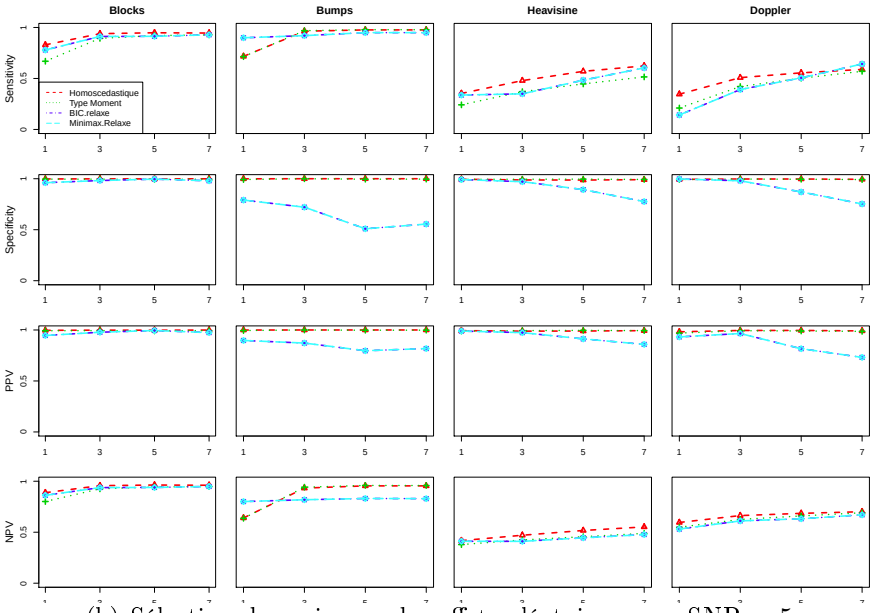


(b) Sélection des effets fixes pour $\tau_U = 0.1$

FIGURE 9.13 – *(a) - Résultats de simulations concernant la précision des procédures d'estimation de l'effet fixe fonctionnel pour 4 procédures d'estimation/sélection : un seuillage SCAD usuel basé sur le seuil universel, un seuillage hétéroscédastique basé sur une estimation des variances de type moment et deux procédures basées sur des techniques de vraisemblance pénalisée. Les résultats sont représentés pour les 4 types de fonctions étudiées (`Blocks`, `Bumps`, `Heavisine` et `Doppler`) et pour* $\tau_U = \{0.1, 0.25, 1, 4\}$. *Au sein de chaque graphe individuel, chaque courbe représente une procédure et varie en abscisse par rapport au paramètre SNR. (b) - Résultats de sélection de variances pour* $\tau_U = 0.1$ *(les comportements en terme de sélection de variables restent sensiblement les mêmes pour d'autres valeurs de* τ_U*).*



(a) EQM - Paramètres de variances

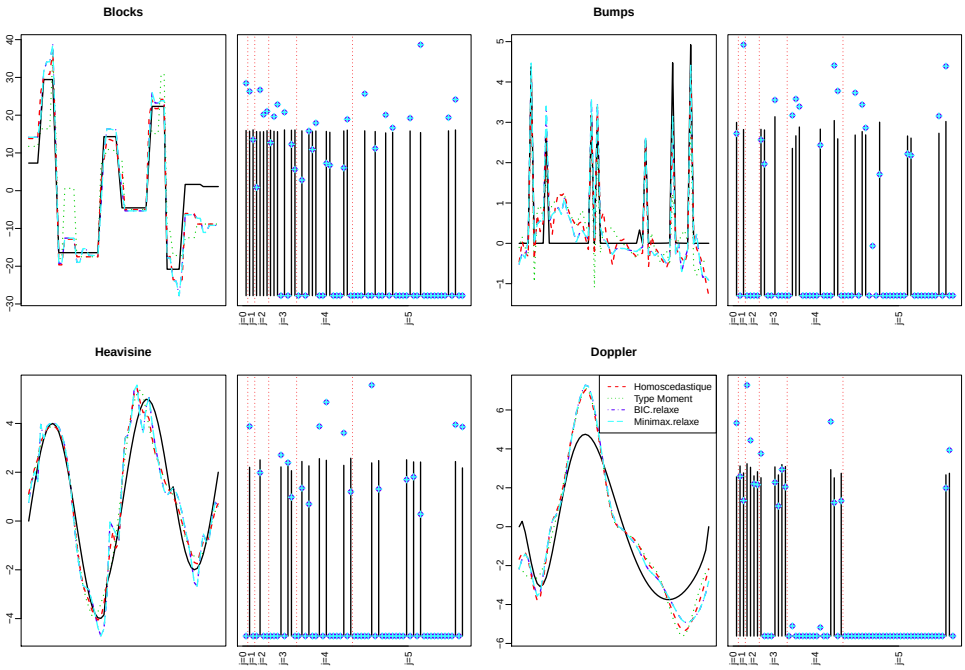


(b) Sélection de variances des effets aléatoires pour SNR = 5

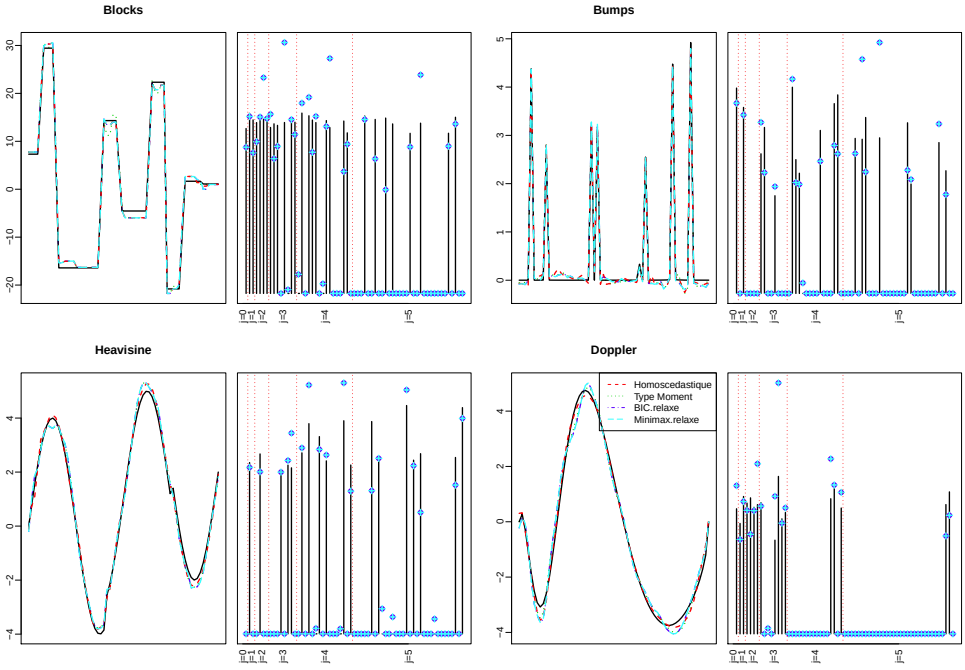
FIGURE 9.14 – *(a) - Résultats de simulations concernant la précision des estimateurs des paramètres de variances pour 4 procédures d'estimation/sélection : un seuillage SCAD usuel basé sur le seuil universel, un seuillage hétéroscédastique basé sur une estimation des variances de type moment et deux procédures basées sur des techniques de vraisemblance pénalisée. Les résultats sont représentés pour les 4 types de fonctions étudiées (`Blocks`, `Bumps`, `Heavisine` et `Doppler`) et pour SNR* $= \{1, 3, 5, 7\}$. *Au sein de chaque graphe individuel, chaque courbe représente une procédure et varie en abscisse par rapport au paramètre SNR. (b) - Résultats de sélection de variances pour SNR* $= 5$ *pour les procédures BIC.relaxe et Minimax.relaxe proposant une sélection des paramètres de variances.*

			SNR	
	τ_U		1	7
Blocks	0.1	BIC	729.9 (191.9)	905.1 (274.3)
		Minimax	54.3 (20.2)	104.3 (39.6)
		Type Moment	0.03 (0.008)	0.02 (0.009)
	4	BIC	220.0 (15.2)	397.1 (65.2)
		Minimax	39.6 (3.2)	45.5 (3.8)
		Type Moment	0.02 (0.016)	0.02 (0.01)
Bumps	0.1	BIC	1154.0 (412.8)	1521.4 (510.1)
		Minimax	108.1 (56.2)	151.6 (60.2)
		Type Moment	0.01 (0.008)	0.02 (0.007)
	4	BIC	462.0 (58.0)	669.2 (108.7)
		Minimax	52.6 (3.3)	56.6 (8.6)
		Type Moment	0.01 (0.007)	0.02 (0.009)
Heavisine	0.1	BIC	887.8 (474.5)	1516.6 (524.2)
		Minimax	82.9 (39.3)	118.1 (32.4)
		Type Moment	0.02 (0.007)	0.02 (0.011)
	4	BIC	426.5 (54.7)	620.21 (60.0)
		Minimax	47.4 (5.5)	57.9 (5.4)
		Type Moment	0.03 (0.008)	0.02 (0.009)
Doppler	0.1	BIC	1199.4 (442.0)	1532.6 (415.0)
		Minimax	106.3 (46.9)	121.5 (40.0)
		Type Moment	0.02 (0.008)	0.02 (0.008)
	4	BIC	431.4 (55.9)	632.73 (71.5)
		Minimax	43.7 (4.0)	55.3 (3.5)
		Type Moment	0.02 (0.008)	0.02 (0.008)

TABLE 9.3 – *Temps de calculs moyens et écarts-types (entre parenthèse) en secondes associés aux procédures "BIC", "Minimax" et "Type Moment" pour différentes valeurs de* (μ, SNR, τ_U)*. Le coût numérique de réestimation pour les procédures relaxées est négligé devant le coût total des procédures itératives.*



(a) Pour SNR $= 1$ et $\tau_U = 0.1$



(b) Pour SNR $= 5$ et $\tau_U = 0.1$

FIGURE 9.15 – *Exemple de reconstruction de l'effet fixe seuillé. L'effet fixe original est représenté en noir. Dans chaque cas sont représentées les estimations des paramètres de variance $\boldsymbol{\gamma}$ dans le domaine des ondelettes obtenues par les procédures BIC.relaxe et Minimax.relaxe, respectivement en bleu foncé et bleu clair. Les vraies valeurs des variances associées aux effets aléatoires sont représentées en noir. Les reconstructions présentées correspondent aux données conduisant à l'EQM médian pour la reconstruction de l'effet fixe fonctionnel avec la procédure "Minimax.relaxe".*

Chapitre 10

Conclusion et perspectives

Dans ce travail de thèse, nous nous sommes intéressés à la problématique de la classification non supervisée et de la sélection de variables dans les modèles mixtes fonctionnels en adoptant une stratégie non paramétrique basée sur les ondelettes.

10.1 Classification non supervisée dans les modèles mixtes fonctionnels

Pour la problématique de la classification non supervisée dans les modèles mixtes fonctionnels, nous avons proposé une modélisation permettant de prendre en compte la présence d'effets aléatoires dans un cadre fonctionnel de classification. Nous avons développé une procédure basée sur une représentation du modèle dans le domaine des ondelettes permettant l'estimation des paramètres du modèle. Notre procédure se déroule en deux étapes : une première de réduction de dimension des données, nécessaire au vu de la dimension des données dans notre cadre fonctionnel, basée sur les techniques de seuillage par ondelettes. La deuxième étape est basée sur l'utilisation de l'algorithme EM pour l'estimation des paramètres par maximum de vraisemblance.

Une étude de simulation approfondie montre de réels gains de performance de classification en présence de variabilité inter-individuelle et l'application de notre procédure à des jeux de données réelles apporte de nouvelles pistes intéressantes vis-à-vis de leur compréhension. Toutefois, certains aspects de notre approche peuvent être améliorés.

- La première étape de réduction de dimension est réalisée *a minima* à l'heure actuelle : en effet, seules les positions nulles pour l'ensemble des individus sont retirées de l'étude, ces coefficients n'étant pas informatifs pour la classification. Cependant, notre objectif idéal serait de pouvoir retirer les positions ayant le même niveau pour tous les individus. Cette problématique est difficile car nous sommes dans un cadre non supervisé en présence de variabilité individuelle où

les labels individuels sont inconnus et elle se ramène en fait à une problématique de seuillage dans les modèles de mélanges. À ce titre, nous pensons que les travaux de Pan et Shen (2007) et la thèse de Meynet (2012) pourraient représenter une base de réflexion intéressante.

- Dans le cadre particulier de la classification non supervisée appliquée à des données de spectrométrie de masse, nous avons pu constater que les résultats de classification dépendent fortement de l'étape de pré-traitement (alignement) des signaux. Ainsi, les résultats sont grandement meilleurs lorsque cette étape est réalisée en connaissance des labels individuels. L'idée serait alors d'intégrer cette étape dans l'algorithme itératif de classification. Pour des raisons de coût numérique, ceci n'est pas envisageable avec la procédure de Antoniadis et al. (2007) mais une piste intéressante pourrait être de s'intéresser aux récents travaux de Bigot (2011), proposant une procédure moins coûteuse de recalage de signaux basée sur la notion de moyenne de Fréchet.

10.2 Sélection de variables et estimation dans les modèles mixtes fonctionnels

Dans la seconde partie de cette thèse, nous nous sommes intéressés à l'estimation dans les modèles mixtes fonctionnels. Nous avons abordé cette problématique de deux façons différentes répondant aux deux approches marginale et jointe des modèles mixtes classiques. Notre procédure, basée sur une représentation marginale des modèles mixtes fonctionnels, se ramène pour l'estimation des effets fixes à un seuillage hétéroscédastique des données. Sous une hypothèse de consistance des estimateurs des variances, nous montrons que l'estimateur de l'effet fixe fonctionnel converge vers la vraie fonction dans la classe des espaces de Besov avec une vitesse near-minimax vis-à-vis du risque quadratique.

Notre deuxième stratégie basée sur une approche jointe des modèles mixtes est développée dans un objectif de sélection des effets fixes et des variances des effets aléatoires. Dans ce but, nous nous sommes basés sur la définition d'un critère de vraisemblance doublement pénalisé par rapport aux effets fixes et aux variances des effets aléatoires. Nous avons montré que l'optimisation de ce critère conduit à des estimateurs possédant la propriété d'oracle dans un cadre de double asymptotique quand la taille des signaux M et le nombre d'individus N divergent, avec $M < N$. Nous avons proposé une procédure basée sur l'algorithme EM permettant l'optimisation itérative de ce critère pénalisé.

Cependant, certains aspects théoriques et applicatifs restent encore à étudier.

- Dans la stratégie jointe, les propriétés de convergence de l'estimateur de l'effet fixe fonctionnel dans la classe des espaces de Besov n'ont pas été étudiées. La difficulté de ce point réside dans le fait que nous ne disposons pas d'estimateurs explicites des paramètres du modèle. Pour traiter ce problème, l'approche ori-

ginelle de la notion de seuillage de Donoho et Johnstone (1994) nous semble être une bonne piste de départ.

- De plus, dans le cas de l'étude de données issues de la biologie moléculaire, il n'est pas rare d'être confronté à des données de très grande dimension (avec $M > N$) justifiant notre approche fonctionnelle. Il serait intéressant de pouvoir étendre les propriétés oraculaires précédemment développées au cadre où $M > N$ en se basant sur les travaux récents de Kim et al. (2008).
- D'un point de vue applicatif, notre étude doit être complétée en réalisant une étude de simulations dans un cadre plus réaliste d'un point de vue fonctionnel en choisissant des tailles de signaux plus conséquentes. Cela nécessitera d'importantes ressources numériques pour s'adapter au nombre de configurations fixées. Enfin, l'application des procédures à des jeux de données réelles issus de la spectrométrie de masse et des microarray CGH représente une perspective nécessaire et intéressante de ce travail.

Annexe A

Vitesse de convergence de l'estimateur de seuillage hétéroscédastique

Notre but est de majorer le risque L^2 de l'estimateur de l'effet fixe fonctionnel $\widehat{\mu}$ pour une fonction μ appartenant à une boule de Besov $B^s_{pq}[0,1]$. Nous nous plaçons dans une configuration où les paramètres de variances sont inconnus et nous distinguons par la suite les estimateurs $\widehat{\mu}_\sigma$ et $\widehat{\mu}_{\widehat{\sigma}}$, correspondant respectivement aux contextes où les variances sont connues ou inconnues mais pour lesquelles on dispose d'estimateurs $\sqrt{N}$-consistants, notés $\left[\widehat{\sigma}^2_{jk}\right]_{(j,k)\in\Lambda}$. Nous cherchons alors à majorer le risque $\mathbb{E}\left(\|\widehat{\mu}_{\widehat{\sigma}} - \mu\|^2_{L^2}\right)$

Nous considérons au cours de cette démonstration, les coefficients *théoriques* de la décomposition en ondelettes afin de prendre en compte la décomposition entière du signal. Ces coefficients sont différenciés des coefficients empiriques par une notation étoilée (*) et sont liés aux coefficients empiriques par un facteur $\sqrt{M}$ (cf. Section 3.2.3). La loi des coefficients théoriques associés aux observations est alors donnée, pour tout $i = 1, \ldots, N$ et tout $(j,k) \in \Lambda$, par :

$$d^*_{ijk} \sim \mathcal{N}\left[\beta^*_{jk}, \frac{\sigma^2_{jk}}{M}\right] \quad \text{et} \quad c^*_i \sim \mathcal{N}\left[\alpha^*, \frac{\sigma^2_\nu}{M}\right] \tag{A.1}$$

Soit $\widehat{\boldsymbol{\beta}}^*$ estimateur de seuillage défini comme en Section 7.5.

On a alors :

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{\mu}_{\widehat{\sigma}}-\mu\|_{L^2}^2\right) &\leq \mathbb{E}\left(\|\widehat{\mu}_{\widehat{\sigma}}-\widehat{\mu}_{\sigma}\|_{L^2}^2\right)+\mathbb{E}\left(\|\widehat{\mu}_{\sigma}-\mu\|_{L^2}^2\right)\\
&\leq \mathbb{E}\left[\sum_{j=j_0+1}^{j_1}\sum_k |\widehat{\beta}^*_{jk}(\widehat{\sigma}_{jk})-\widehat{\beta}^*_{jk}(\sigma_{jk})|^2\right]\\
&\quad+\mathbb{E}(|\widehat{\alpha}^*(\sigma_\nu)-\alpha^*|^2)+\mathbb{E}\left[\sum_{j=0}^{j_0}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk})-\beta^*_{jk}|^2\right]\\
&\quad+\mathbb{E}\left[\sum_{j=j_0+1}^{j_1}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk})-\beta^*_{jk}|^2\right]+\mathbb{E}\left[\sum_{j=j_1+1}^{\infty}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk})-\beta^*_{jk}|^2\right]\\
&= T_1+T_2+T_3+T_4+T_5. \qquad \text{(A.2)}
\end{aligned}$$

On cherche ensuite à majorer chaque terme de la décomposition (A.2).

En considérant des estimateurs empiriques des paramètres de variance définis par (7.8) pour toute position (j,k) et en utilisant l'approximation de la *delta méthode*, basée sur un développement limité de la fonction de seuillage, on a alors :

$$\begin{aligned}
T_1 &\leq \sum_{j=j_0+1}^{j_1}\sum_k C_1\,\frac{2N-1}{(MN)^2}\,\sigma^4_{jk}\\
&\leq \sum_{j=j_0+1}^{j_1} C_1\,\frac{2^j}{M^2N}\,\sigma^4_{jk}\\
&\leq C_1\sigma^4_{\max}\,\frac{2^{j_1}}{M^2N} \leq C_1\sigma^4_{\max}\,\frac{(\log M)^{-1}}{MN} \qquad \text{(A.3)}
\end{aligned}$$

en prenant j_1 comme défini dans Juditsky et Delyon (1996), c'est-à-dire, vérifiant $\frac{M}{\log M} \leq 2^{j_1} \leq \frac{2M}{\log M}$.

Ensuite :

$$T_2 = \mathbb{E}(|\widehat{\alpha}^*(\sigma_\nu)-\alpha^*|^2) \leq \frac{\sigma^2_{\max}}{MN}.$$

De la même manière, pour le terme T_3 correspondant à des niveaux de résolution pour lesquels on ne fait pas de seuillage, on a :

$$\begin{aligned}
T_3 \;=\; \mathbb{E}\left[\sum_{j=0}^{j_0}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk})-\beta^*_{jk}|^2\right] &= \sum_{j=0}^{j_0}\sum_k \mathbb{E}\left(|d^*_{ijk}-\beta^*_{jk}|^2\right)\\
&= \sum_{j=0}^{j_0} 2^j\,\frac{\sigma^2_{\max}}{M} \leq C_3\,\frac{\sigma^2_{\max}}{N}\,\frac{2^{j_0}}{M}. \qquad \text{(A.4)}
\end{aligned}$$

La majoration du terme T_5 nous donne :

$$T_5 \;=\; \mathbb{E}\left[\sum_{j=j_1+1}^{\infty}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk})-\beta^*_{jk}|^2\right] = \sum_{j=j_1+1}^{\infty}\sum_k |\beta^*_{jk}|^2 \leq C_5 2^{-2j_1 s'}, \qquad \text{(A.5)}$$

où $s' = \begin{cases} s - \frac{1}{p} + \frac{1}{2} & \text{si } p < 2, \\ s & \text{sinon.} \end{cases}$

La justification de cette majoration est faite à partir du livre de Härdle et al. (1998). Ainsi, par le Théorème 9.5 du livre, on sait que pour toute fonction $f \in B^s_{pq}$, et pour P_j opérateur de projection sur l'espace d'approximation V_j, on a :

$$\|P_j f - f\|_p = 2^{-js} h_j \quad \text{où } \{h_j\} \in \ell_q.$$

De plus :

- si $p \geq 2$, $L_p(L) \subset L_2(L)$ et donc $\|P_j f - f\|_2 \leq \|P_j f - f\|_p \leq C 2^{-js}$,
- si $p < 2$, alors (Corollaire 9.2) $B^s_{pq} \subset B^{s'}_{2q}$ pour $s' = s - \frac{1}{p} + \frac{1}{2}$ et donc $\|P_j f - f\|_2 \leq C' 2^{-js'}$,

avec C et C' constantes réelles. On en déduit alors l'inégalité (A.5).

Concentrons nous à présent sur le terme T_4. En utilisant le Lemme 2 de Juditsky et Delyon (1996), on a :

$$\mathbb{E}\left[\sum_{j=j_0+1}^{j_1}\sum_k |\widehat{\beta}^*_{jk}(\sigma_{jk}) - \beta^*_{jk}|^2\right] \leq \underbrace{\mathbb{E}\left[\sum_{j=j_0+1}^{j_1}\sum_k \min\left(|\beta^*_{jk}|, \frac{3}{2}\frac{2\lambda\sigma_{jk}}{\sqrt{MN}}\right)^2\right]}_{\text{Terme } T_{4.1}}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{j=j_0+1}^{j_1}\sum_k 3^2 |\varepsilon^*_{ijk}|^2 \mathbf{1}_{|\varepsilon^*_{ijk}| > \frac{2\lambda\sigma_{jk}/\sqrt{MN}}{2}}\right]}_{\text{Terme } T_{4.2}}. \quad \text{(A.6)}$$

Commençons par le terme $T_{4.2}$. On a :

$$\begin{aligned}
T_{4.2} &= \sum_{j=j_0+1}^{j_1}\sum_k 9\mathbb{E}\left(|\varepsilon^*_{ijk}|^2 \mathbf{1}_{|\varepsilon^*_{ijk}| > \lambda\sigma_{jk}/\sqrt{MN}}\right) \\
&\leq \sum_{j=j_0+1}^{j_1}\sum_k 9\mathbb{E}\left(|\varepsilon^*_{ijk}|^4\right)^{\frac{1}{2}} \mathbb{E}\left[\left(\mathbf{1}_{|\varepsilon^*_{ijk}| > \lambda\sigma_{jk}/\sqrt{MN}}\right)^2\right]^{\frac{1}{2}} \quad \text{par Cauchy-Schwartz} \\
&\leq \sum_{j=j_0+1}^{j_1} C_{4.2} \times \frac{\sigma^2_{\max}}{MN} \times \exp\left[\frac{-\left(\lambda\sigma_{jk}/\sqrt{MN}\right)^2}{\frac{2\sigma^2_{jk}}{MN}}\right]^{\frac{1}{2}} \\
&\leq \sum_{j=j_0+1}^{j_1} C_{4.2} \times \frac{\sigma^2_{\max}}{MN} \times M^{-1} \\
&\leq C_{4.2}\, \frac{\sigma^2_{\max}}{N} M^{-2} \times 2^{j_1}. \qquad \text{(A.7)}
\end{aligned}$$

On en déduit alors :

$$T_{4.2} \leq C_{4.2}\, \frac{\sigma^2_{\max}}{MN} (\log M)^{-1}.$$

Pour le dernier terme $T_{4.1}$, on a :

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=j_0+1}^{j_1} \sum_k \min\left(|\beta^*_{jk}|, 3\lambda \frac{\sigma_{jk}}{MN}\right)^2\right] &\leq \mathbb{E}\left[\sum_{j=j_0+1}^{j_1} \sum_k \left(3\lambda \frac{\sigma_{jk}}{MN}\right)^{2-p} |\beta^*_{jk}|^p\right] \\
&\leq \sum_{j=j_0+1}^{j_1} \sum_k \left(3\lambda \frac{\sigma_{jk}}{MN}\right)^{2-p} |\beta^*_{jk}|^p \\
&\leq 3^{2-p} \left(\frac{2\log M}{M}\right)^{1-\frac{p}{2}} \left(\frac{\sigma^2_{\max}}{N}\right)^{1-\frac{p}{2}} \\
&\qquad \times \underbrace{\sum_{j=j_0+1}^{j_1} \sum_k |\beta^*_{jk}|^p}_{=\mathcal{O}(C 2^{-s' p j_0})} \\
&\leq C_{4.1} \left(\frac{\log M}{M}\right)^{1-\frac{p}{2}} \left(\frac{\sigma^2_{\max}}{N}\right)^{1-\frac{p}{2}} 2^{-s' p j_0}.
\end{aligned} \tag{A.8}$$

Finalement, en regroupant tous les termes, on obtient l'inégalité suivante :

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{\mu}_{\widehat{\sigma}} - \mu\|^2_{L^2}\right) \leq C_1\, \frac{(\log M)^{-1}}{MN} + \frac{\sigma^2_{\max}}{MN} + C_{4.1} \left(\frac{\log M}{M}\right)^{1-\frac{p}{2}} \left(\frac{\sigma^2_{\max}}{N}\right)^{1-\frac{p}{2}} 2^{-s' p j_0} \\
+ C_3 \frac{\sigma^2_{\max}}{N} \frac{2^{j_0}}{M} + C_{4.2}\, \sigma^2_{\max} \frac{M^{-1}}{N \log M} + C_5 2^{-2 j_1 s'}.
\end{aligned} \tag{A.9}$$

Afin de trouver la forme optimale du niveau j_0, niveau à partir duquel le seuillage commence, on cherche à balancer, en fonction de M, les deux termes :

$$\underbrace{\mathcal{O}\left(\frac{2^{j_0}}{MN}\right)}_{(\star)} \qquad \text{et} \qquad \underbrace{\mathcal{O}\left(\left[\frac{\log M}{MN}\right]^{1-p/2} 2^{-s' p j_0}\right)}_{(\star\star)}$$

Ainsi, on en déduit que le paramètre j_0 doit être de la forme :

$$2^{j_0} = \mathcal{O}\left[(\log M)^{\frac{1-p/2}{1+s'p}} \; (MN)^{\frac{p/2}{1+s'p}}\right].$$

En remplaçant dans $(\star)$ et dans $(\star\star)$, on obtient alors que ces deux termes sont de la forme :

$$\mathcal{O}\left[\left[\frac{\log M}{MN}\right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}}\right].$$

Et donc, l'inégalité (A.9) devient :

$$\begin{aligned}\mathbb{E}\left(\|\widehat{\mu}_{\widehat{\sigma}}-\mu\|_{L^2}^2\right) \leq C_1\ \frac{(\log M)^{-1}}{MN} + \frac{\sigma_{\max}^2}{MN} + C_3\ \sigma_{\max}^2 \left[\frac{\log M}{MN}\right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}} \\ + C_{4.1}\ \sigma_{\max}^{2-p} \left[\frac{\log M}{MN}\right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}} \\ + C_{4.2}\ \sigma_{\max}^2 \frac{M^{-\frac{1}{8}}}{N\log M} + C_5 \left[\frac{\log M}{M}\right]^{2s'} \\ = T_1' + T_2' + T_3' + T_{4.1}' + T_{4.2}' + T_5'.\end{aligned} \tag{A.10}$$

Pour $p > \frac{2}{2s+1}$, parmi les termes limitant de l'inégalité (A.10), c'est-à-dire les termes ayant la convergence la plus lente, intéressons nous aux termes T_3' et $T_{4.1}'$ en $\mathcal{O}\left[\left[\frac{\log M}{M}\right]^{\frac{2s}{2s+1}} (\log M)^{\frac{-2s'}{2s+1}}\right]$.

À ce stade, on peut donc distinguer deux cas :

- Si $\frac{2}{2s+1} < p \leq 2$, on a :

  $$(\log M)^{-\frac{2s'}{2s+1}} \underset{M\to\infty}{\longrightarrow} 0.$$

  La vitesse de convergence des termes T_3' et $T_{4.1}'$ est alors en $\mathcal{O}\left[\left[\frac{\log M}{MN}\right]^{\frac{2s}{2s+1}}\right]$.
- Si $p > 2$, alors $0 < s < s'$ et donc :

  $$(\log M)^{\frac{2s}{2s+1}-\frac{2s'}{2s+1}} \underset{M\to\infty}{\longrightarrow} 0.$$

  Dans ce cas, la vitesse de convergence des termes T_3' et $T_{4.1}'$ est alors en $\mathcal{O}\left[\left[\frac{1}{MN}\right]^{-\frac{2s}{2s+1}}\right]$.

Remarque : Dans cette preuve, on utilise en particulier un résultat donné par Juditsky et Delyon (1996) (Lemme 2). Ce résultat concerne initialement l'estimateur obtenu avec un seuillage dur δ^H et s'exprime comme suit :

Lemme A.1. *Soit $\beta^* \in \mathbb{R}$ et ε^* une variable aléatoire réelle. Soit $d^* = \beta^* + \varepsilon^*$. Alors, l'estimateur $\widehat{\beta^*}^{HARD} = d^*\mathbf{1}_{|d^*|>\lambda}$ vérifie :*

$$|\widehat{\beta^*}^{HARD} - \beta^*| \leq \min(|\beta^*|, \frac{3}{2}\lambda) + 3|\varepsilon^*|\mathbf{1}_{|\varepsilon^*|>\frac{\lambda}{2}}.$$

Or, dans le cas d'un seuillage de type SCAD et en notant $\widehat{\beta}^{\mathrm{SCAD}}$ l'estimateur résultant, cette inégalité reste encore valable, étant donné que ce dernier vérifie la relation :

$$\left|\widehat{\beta^*}^{\mathrm{SCAD}}\right| \leq \left|\widehat{\beta^*}^{\mathrm{HARD}}\right|,$$

donnant alors la relation :

$$|\widehat{\beta^*}^{\mathrm{SCAD}} - \beta^*| \leq |\widehat{\beta^*}^{\mathrm{HARD}} - \beta^*| \leq \min(|\beta^*|, \frac{3}{2}\lambda) + 3|\varepsilon^*|\mathbf{1}_{|\varepsilon|>\frac{\lambda}{2}}.$$

Annexe B

Propriétés oraculaires pour la sélection des effets fixes et aléatoires

B.1 Vérification des hypothèses sur la vraisemblance

Le résultat présenté en Section (8.2.3) suppose préalablement la vérification des hypothèses présentées dans la même section. Nous nous intéresserons plus particulièrement à la vérification des hypothèses concernant la vraisemblance et pour lesquelles nous nous plaçons dans un cadre particulier puisque nous travaillons avec des modèles mixtes fonctionnels. Notre but est alors de sélectionner aussi bien les effets fixes que les effets aléatoires. Il nous faut donc vérifier ces hypothèses aussi bien sur les paramètres d'effet fixe du vecteur $\boldsymbol{\beta}$ que les paramètres des variances des effets aléatoires du vecteur $\boldsymbol{\gamma}$.

Pour rappel, la vraisemblance des données pour l'individu i est donnée par :

$$\log \mathcal{L}^i(\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\gamma}^2, \sigma_\varepsilon^2) = -\frac{1}{2}\sum_{jk} \log\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right) - \frac{1}{2}\sum_{jk} \frac{1}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2}\big(d_{i,jk} - \beta_{jk}\big).$$

- Hypothèse (H5)

  Pour tout $i = 1, \ldots, N$ et tout $(j,k) \in \Lambda$, on vérifie bien, d'une part, que :

$$\mathbb{E}\left[\frac{\partial \log \mathcal{L}^i(\mathbf{d})}{\partial \beta_{jk}}\right] = \frac{1}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2}\mathbb{E}\big(d_{i,jk} - \beta_{jk}\big) = 0.$$

Et d'autre part :

$$\mathbb{E}\left[\frac{\partial \log \mathcal{L}^i(\mathbf{d})}{\partial \gamma_{jk}^2}\right] = \frac{2^{-j\eta}}{2} \times \frac{1}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2} + \frac{2^{-j\eta}}{2} \times \frac{1}{\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right)^2} \underbrace{\mathbb{E}\left((d_{i,jk} - \beta_{jk})^2\right)}_{=\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2}$$
$$= 0.$$

De plus, on vérifie facilement que, dans tous les cas possibles, on a :

$$\mathbb{E}_\Upsilon\left[\frac{\partial \log \mathcal{L}^i(\mathbf{d})}{\partial \Upsilon_{m_1}} \frac{\partial \log \mathcal{L}^i(\mathbf{d})}{\partial \Upsilon_{m_2}}\right] = -\mathbb{E}_\Upsilon\left[\frac{\partial^2 \log \mathcal{L}^i(\mathbf{d})}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2}}\right] \quad \forall\, m_1, m_2 = 1, \ldots, 2M.$$

- Hypothèse (H6)

  La matrice d'information de Fisher est définie par :

  $$\mathcal{I}(\Upsilon) = \mathbb{E}\left[\nabla \log \mathcal{L}^i(\Upsilon) \;\; \nabla^T \log \mathcal{L}^i(\Upsilon)\right].$$

  Dans notre cadre, la calcul de cette quantité conduit à l'obtention d'une matrice diagonale définie positive de taille $2M \times 2M$ dont l'expression est donnée par :

  $$\left[\begin{array}{c|c} \ddots & \\ & \frac{1}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2} \\ & \ddots \end{array} \begin{array}{c} \\ 0 \\ \\ \end{array} \right.$$
  $$\left. \begin{array}{c} \\ 0 \\ \\ \end{array} \begin{array}{c|c} \ddots & \\ & \frac{1}{2} \times \frac{2^{-j\eta}}{\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right)^2} \\ & \ddots \end{array} \right].$$

  On peut alors remarquer qu'en imposant la condition $\sigma_\varepsilon^2 > 0$, alors, les valeurs propres de cette matrice sont bien bornées, ainsi que leur carrés, cette condition restant peu contraignante.

  Cela entraîne que, sous la même condition, les quantités $\mathbb{E}_\Upsilon\left[\left[\frac{\partial^2 \log \mathcal{L}^i(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2}}\right]^2\right]$ sont bornées elles aussi.

- Hypothèse (H7)

Cette hypothèse concerne les moments d'ordre 3 de la fonction de vraisemblance. Ces moments ont les expressions suivantes pour tout $i = 1, \ldots, N$ et tout $(j,k) \in \Lambda$:

$$\begin{aligned}
\frac{\partial^3 \log \mathcal{L}^i(\Upsilon)}{\partial^3 \beta_{jk}} &= 0, \\
\frac{\partial^3 \log \mathcal{L}^i(\Upsilon)}{\partial^2 \beta_{jk} \partial \gamma_{jk}^2} &= \frac{2^{-j\eta}}{\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2}, \\
\frac{\partial^3 \log \mathcal{L}^i(\Upsilon)}{\partial^2 \gamma_{jk}^2 \partial \beta_{jk}} &= \frac{-2 \times \left(2^{-j\eta}\right)^2}{\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right)^3}(d_{i,jk} - \beta_{jk}), \\
\frac{\partial^3 \log \mathcal{L}^i(\Upsilon)}{\partial^3 \gamma_{jk}^2} &= \left(2^{-j\eta}\right)^3 \left[\frac{3(d_{jk}^i - \beta_{jk})^2}{\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right)^4} - \frac{1}{\left(\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2\right)^3}\right].
\end{aligned}$$

En supposant que le paramètre σ_ε^2 est strictement positif, on remarque alors que tous les moments d'ordre 3 peuvent être majorés par une fonction (notée B dans la démonstration) dépendant de $(d_{i,jk} - \beta_{jk})$. Or, de par leur appartenance à un espace de Besov, on sait que la norme du vecteur $\boldsymbol{\beta}$ est bornée, impliquant ainsi que les coefficients β_{jk} le sont.
En supposant de plus que les variances $\left(\gamma_{jk}^2\right)_{\{(j,k)\in\Lambda\}}$ sont bornées par une quantité notée $\gamma_{\max}^2$, on montre alors facilement que l'espérance du carré de cette même fonction est aussi bornée car elle dépend uniquement de la quantité $\sigma_\varepsilon^2 + 2^{-j\eta}\gamma_{jk}^2$, cette hypothèse sur les variances restant une hypothèse relativement courante.

B.2 Propriétés oraculaires des estimateurs

B.2.1 Preuve du théorème 8.1

Soit $\widetilde{a}_N = \sqrt{M}(\frac{1}{\sqrt{N}} + a_N)$. Pour démontrer le théorème (8.1), on cherche à montrer que pour tout $\epsilon > 0$, il existe une constante $C > 0$ telle que pour N assez grand, on ait :

$$\mathbb{P}\left\{ \sup_{\|u\|=C} \boldsymbol{\ell}(\Upsilon_0 + \widetilde{a}_N u) < \boldsymbol{\ell}(\Upsilon_0) \right\} \geq 1 - \epsilon.$$

On considère alors la différence $D_N(u) = \boldsymbol{\ell}(\Upsilon_0 + \widetilde{a}_N u) - \boldsymbol{\ell}(\Upsilon_0)$ et en utilisant que $\text{pen}(0, \cdot) = 0$, on peut écrire :

$$\begin{aligned} D_N(u) &\leq \log \mathcal{L}(\Upsilon_0 + \widetilde{a}_N u) - \log \mathcal{L}(\Upsilon_0) \\ &\quad - \sum_{m=1}^{m_N^{\boldsymbol{\beta}}} \left[\text{pen}(|\beta_{0m} + \widetilde{a}_N u_m|, \lambda_1) - \text{pen}(|\beta_{0m}|, \lambda_1) \right] \\ &\quad - \sum_{m=1}^{m_N^{\boldsymbol{\gamma}}} \left[\text{pen}(|\gamma_{0m} + \widetilde{a}_N u_m|, \lambda_2) - \text{pen}(|\gamma_{0m}|, \lambda_2) \right] \\ &= I_1 + I_2 + I_3. \end{aligned}$$

où $m_N^{\boldsymbol{\beta}}$ est telle que $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m_N^{\boldsymbol{\beta}}})$ contient les composantes non nulles du vecteur $\boldsymbol{\beta}$ et $m_N^{\boldsymbol{\gamma}}$ est telle que $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0m_N^{\boldsymbol{\gamma}}})$ contient les composantes non nulles du vecteur $\boldsymbol{\gamma}$.

- Terme I_1 :

  Par Taylor-Lagrange, on peut développer le terme I_1 :

  $$\begin{aligned} I_1 &= \widetilde{a}_N \nabla \log \mathcal{L}(\Upsilon_0) u + \frac{1}{2} \widetilde{a}_N^2 u^T \nabla^2 \log \mathcal{L}(\Upsilon_0) u + \frac{1}{6} \widetilde{a}_N^3 \nabla^T (u^T \nabla^2 \log \mathcal{L}(\Upsilon^*) u) u \\ &= I_{1.1} + I_{1.2} + I_{1.3}, \end{aligned}$$

  avec $\Upsilon^* \in]\Upsilon_0, \Upsilon_0 + u[$.
  Les trois termes ci-dessus peuvent alors être majorés séparément. D'une part, par Cauchy-Schwartz, on a :

  $$|I_{1.1}| = |\widetilde{a}_N \nabla \log \mathcal{L}(\Upsilon_0) u| \leq \widetilde{a}_N \|\nabla \log \mathcal{L}(\Upsilon_0)\| \|u\|.$$

  Or, pour tout $m = 1, \ldots, M$, on a que $\frac{\partial}{\partial \Upsilon_m} \log \mathcal{L}(\Upsilon_0) = \mathcal{O}_P(\sqrt{N})$. Donc, on en déduit que :

  $$\|\nabla \log \mathcal{L}(\Upsilon_0)\| = \mathcal{O}_P(\sqrt{MN}).$$

  Ainsi :

  $$|I_{1.1}| \leq \widetilde{a}_N \mathcal{O}_P(\sqrt{MN}) \|u\| = \mathcal{O}_P(\widetilde{a}_N^2 N) \|u\|.$$

D'autre part, on peut écrire :

$$I_{1.2} = \frac{1}{2}\widetilde{a}_N^2 u^T \nabla^2 \log \mathcal{L}(\Upsilon_0) u = \frac{1}{2} u^T \left[\frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) + \mathcal{I}(\Upsilon_0)\right] u.N\widetilde{a}_N^2 + \frac{1}{2} u^T \underbrace{\frac{1}{N}\,\mathbb{E}\big(\nabla^2 \log \mathcal{L}(\Upsilon_0)\big)}_{=-\mathcal{I}(\Upsilon_0)} u.N\widetilde{a}_N^2.$$

Or, en utilisant l'inégalité de Chebyshev et le fait que $M^4/N \to \infty$, on a que :

$$\left\| \frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) + \mathcal{I}(\Upsilon_0) \right\| = o_P(M^{-1}). \tag{B.1}$$

Et donc, comme :

$$u^T \left[\frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) + \mathcal{I}(\Upsilon_0)\right] u \leq \left\| \frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) + \mathcal{I}(\Upsilon_0) \right\| \|u\|^2 = o_P(M^{-1})\|u\|^2 = o_P(1)\|u\|^2.$$

On trouve finalement que :

$$I_{1.2} = \frac{-N\widetilde{a}_N^2}{2} u^T \mathcal{I}(\Upsilon_0) u + o_P(1) N\widetilde{a}_N^2 \|u\|^2.$$

Enfin, par Cauchy-Schwartz, on peut majorer le troisième terme de la façon suivante :

$$\begin{aligned}
|I_{1.3}| &= \left| \frac{1}{6}\widetilde{a}_N^3 \nabla^T (u^T \nabla^2 \log \mathcal{L}(\Upsilon^*) u) u \right| \\
&= \left| \frac{1}{6}\widetilde{a}_N^3 \sum_{m_1,m_2,m_3} \frac{\partial^3 \log \mathcal{L}(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2} \partial \Upsilon_{m_3}} u_{m_1} u_{m_2} u_{m_3} \right| \\
&\leq \frac{1}{6}\widetilde{a}_N^3 \left\| \frac{\partial^3 \log \mathcal{L}(\Upsilon)}{\partial \Upsilon_{m_1} \partial \Upsilon_{m_2} \partial \Upsilon_{m_3}} \right\| \|u\|^3 \\
&\leq \frac{1}{6}\widetilde{a}_N^3 \sum_{i=1}^{N} \left[\sum_{m_1,m_2,m_3} B^2_{m_1,m_2,m_3}(\mathbf{d}_i) \right]^{\frac{1}{2}} \|u\|^3.
\end{aligned}$$

Or, par l'hypothèse (H7), on a que :

$$\widetilde{a}_N \left[\sum_{m_1,m_2,m_3} B^2_{m_1,m_2,m_3}(\mathbf{d}_i) \right]^{\frac{1}{2}} = \mathcal{O}_P(\widetilde{a}_N M^{\frac{3}{2}}) = o_P(1), \quad \text{car } \widetilde{a}_N M^{\frac{3}{2}} \to 0 \text{ quand } N \to \infty.$$

D'où, finalement, pour $\|u\| = C$:

$$|I_{1.3}| \leq \frac{1}{6}\|u\|^3 o_P(1) N\widetilde{a}_N^2 \leq o_P(N\widetilde{a}_N^2)\|u\|^2.$$

- Terme I_2 :

  Concentrons-nous à présent sur le terme I_2 concernant la fonction de pénalité. De même, par un développement de Taylor, on a l'égalité suivante :

$$\begin{aligned} I_2 &= -N \sum_{m=1}^{m_N^{\boldsymbol{\beta}}} \left[\text{pen}(|\beta_{0m} + \widetilde{a}_N u_m|, \lambda_1) - \text{pen}(|\beta_{0m}|, \lambda_1) \right] \\ &= -\sum_{m=1}^{m_N^{\boldsymbol{\beta}}} \left[N\widetilde{a}_N \frac{\partial}{\partial \beta_m} \text{pen}(|\beta_{0m}|, \lambda_1) \text{sign}(\beta_{0m}) u_m + N\widetilde{a}_N^2 \frac{\partial}{\partial^2 \beta_m} \text{pen}(\beta_{0m}, \lambda_1) u_m^2 (1 + o(1)) \right] \\ &= I_{2.1} + I_{2.2}. \end{aligned}$$

  On cherche alors à majorer ces deux termes. D'une part, pour le premier terme, par Cauchy-Schwartz, on a :

$$\begin{aligned} |I_{2.1}| &\leq \sum_{m=1}^{m_N^{\boldsymbol{\beta}}} |N\widetilde{a}_N \frac{\partial}{\partial \beta_m} \text{pen}(|\beta_{0m}|, \lambda_1) \text{sign}(\beta_{0m}) u_m| \\ &\leq \sqrt{m_N^{\boldsymbol{\beta}}} N\widetilde{a}_N a_N \|u\| \\ &\leq N\widetilde{a}_N^2 \|u\|. \end{aligned}$$

  en utilisant l'inégalité aisément vérifiable que $\sqrt{m_N^{\boldsymbol{\beta}}} a_N \leq \widetilde{a}_N$.
  D'autre part, le deuxième terme peut être majoré de la façon suivante :

$$\begin{aligned} I_{2.2} &= (1 + o(1)) \sum_{m=1}^{m_N^{\boldsymbol{\beta}}} N\widetilde{a}_N^2 \frac{\partial}{\partial^2 \beta_m} \text{pen}(\beta_{0m}, \lambda_1) u_m^2 \\ &\leq (1 + o(1)) N\widetilde{a}_N^2 b_N \|u\|^2 \\ &\leq \mathcal{O}(1) N\widetilde{a}_N^2 b_N \|u\|^2. \end{aligned}$$

- Terme I_3 :

  Le même développement que pour le terme I_2 peut être réalisé pour le terme I_3 concernant les paramètres $\boldsymbol{\gamma}$.

Finalement, on obtient donc la majoration :

$$\begin{aligned} D_N(u) \leq \mathcal{O}_P(1)\widetilde{a}_N^2 N\|u\| - \frac{\widetilde{a}_N^2 N}{2} u^T \mathcal{I}(\Upsilon_0) u + o_P(1)\widetilde{a}_N^2 N\|u\|^2 + o_P(1)\widetilde{a}_N^2 N\|u\|^2 \\ + N\widetilde{a}_N^2 \|u\| + \mathcal{O}(1) N\widetilde{a}_N^2 b_N \|u\|^2. \end{aligned}$$

Et donc :

$$D_N(u) \leq \widetilde{a}_N^2 N \Big[\mathcal{O}_P(1)\|u\| - \underbrace{\frac{1}{2} u^T \mathcal{I}(\Upsilon_0) u}_{(\bigstar)} + o_P(1)\|u\|^2 + o_P(1)\|u\|^2 + \|u\| + \mathcal{O}(1) b_N \|u\|^2 \Big].$$

Or, pour $\|u\| = C$ assez grand, on a bien $\mathcal{O}_P(1)\|u\| \leq \frac{1}{2} u^T \mathcal{I}(\Upsilon_0) u$ et $b_N \to 0$ quand $N \to \infty$ (par l'hypothèse (H3')). On remarque donc que le terme $(\bigstar)$ est dominant et négatif. On a donc bien $D_N(u) < 0$ pour $\|u\| = C$ assez grand, ce qui conclut la preuve. ■

B.2.2 Preuve du théorème 8.2

La preuve de ce théorème se décompose en deux parties.

▷ On cherche d'abord à démontrer la parcimonie de l'estimateur. Pour cela, on veut démontrer le lemme suivant :

Lemme B.1. *Supposons que les hypothèses (H1) et (H5)-(H8) sont vérifiées. Supposons de plus que* $\sqrt{\frac{N}{M}}\lambda \to \infty$ *si* $\lambda \to 0$ *et* $\frac{M^5}{N} \to 0$ *quand* $N \to \infty$*. Alors, avec une probabilité tendant vers 1, pour tout* Υ_1 *tel que* $\|\Upsilon_1 - \Upsilon_0^1\|_2 = \mathcal{O}_P(\sqrt{M/N})$ *et pour toute constante* C*, on a :*

$$Q(\Upsilon_1, 0) = \max_{\|\Upsilon_2\| \leq C\sqrt{M/N}} Q(\Upsilon_1, \Upsilon_2).$$

Preuve : Soit $\epsilon_N = C\sqrt{M/N}$. Soit $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ tel que $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0^1\| = \mathcal{O}_P(\sqrt{M/N})$. On cherche à démontrer que pour tout $m = m_N^{\boldsymbol{\beta}} + 1, \ldots, M$, on a, avec une probabilité tendant vers 1 quand $N \to \infty$:

$$\begin{cases} \dfrac{\partial Q(\boldsymbol{\beta})}{\partial \beta_m} < 0 & \text{pour } \ 0 < \beta_m < \epsilon_N \\ \dfrac{\partial Q(\boldsymbol{\beta})}{\partial \beta_m} > 0 & \text{pour } \ -\epsilon_N < \beta_m < 0 \end{cases}$$

On montre ainsi que la dérivée change de signe en 0 et donc qu'il existe un extremum local en ce point.

Par la formule de Taylor, on a, pour $m = m_N^{\boldsymbol{\beta}} + 1, \ldots, M$:

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_m} &= \frac{\partial \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_m} + \sum_{r=1}^{M} \frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m}(\beta_r - \beta_{0r}) \\ &\quad + \sum_{r_1,r_2=1}^{M} \frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial \beta_{r_1} \partial \beta_{r_2} \partial \beta_m}(\beta_{r_1} - \beta_{0r_1})(\beta_{r_2} - \beta_{0r_2}) - N \frac{\partial}{\partial \beta_m} \mathrm{pen}(|\beta_m|, \lambda_1) \\ &= I_1 + I_2 + I_3 + I_4.\end{aligned}$$

avec $(\boldsymbol{\beta}^*) \in [\boldsymbol{\beta}, \boldsymbol{\beta}_0]$.
Considérons le terme I_1. Par les hypothèses (H5) et (H6), on a que :

$$I_1 = \frac{\partial \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_m} = \mathcal{O}_P(\sqrt{N}) = \mathcal{O}_P(\sqrt{MN}).$$

car $\frac{\partial \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_m}$ est une variable aléatoire d'espérance nulle et de variance en $\mathcal{O}(N)$ pour tout $m = 1, \ldots, M$.
D'autre part, on peut réécrire le terme I_2 de la manière suivante :

$$\begin{aligned}I_2 &= \sum_{r=1}^{M} \left[\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m} - \mathbb{E}\left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m} \right) \right] (\beta_r - \beta_{0r}) \\ &\quad + \sum_{r=1}^{M} \mathbb{E}\left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m} \right) (\beta_r - \beta_{0r}) \\ &= I_{2.1} + I_{2.2}.\end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwartz, on trouve :

$$\begin{aligned}|I_{2.2}| &= N \left| \sum_{r=1}^{M} [\mathcal{I}(\boldsymbol{\beta}_0)]_{m,r}(\beta_r - \beta_{0r}) \right| \\ &\leq N \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \left[\sum_{r=1}^{M} [\mathcal{I}(\boldsymbol{\beta}_0)]_{m,r}^2 \right]^{\frac{1}{2}} \\ &\leq N \times \mathcal{O}_P(\sqrt{\frac{M}{N}}) \times \mathcal{O}(1) \quad \text{par l'hypothèse (H6)} \\ &= \mathcal{O}_P(\sqrt{MN}).\end{aligned}$$

De même, pour l'autre terme, on trouve :

$$\begin{aligned}|I_{2.1}| &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \left[\sum_{r=1}^{M} \left[\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m} - \mathbb{E}\left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \beta_r \partial \beta_m} \right) \right]^2 \right]^{\frac{1}{2}} \\ &\leq \mathcal{O}_P(\sqrt{\frac{M}{N}}) \times \mathcal{O}_P(\sqrt{MN}) \quad \text{par (H6)} \\ &= \mathcal{O}_P(M) = \mathcal{O}_P(\sqrt{MN}).\end{aligned}$$

car les $\left[\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta}_0)}{\partial\beta_r \partial\beta_m}\right]_r$ sont des variables de variances en $\mathcal{O}(N)$.
On a donc finalement :

$$I_2 = \mathcal{O}_P(\sqrt{MN}).$$

Concernant le terme I_3, en utilisant la même astuce que pour I_2, on peut écrire :

$$\begin{aligned}
I_3 &= \sum_{r_1,r_2=1}^{M} \frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}(\beta_{r_1}-\beta_{0r_1})(\beta_{r_2}-\beta_{0r_2}) \\
&= \sum_{r_1,r_2=1}^{M} \left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m} - \mathbb{E}\left(\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}\right)\right](\beta_{r_1}-\beta_{0r_1})(\beta_{r_2}-\beta_{0r_2}) \\
&\quad + \sum_{r_1,r_2=1}^{M} \mathbb{E}\left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}\right](\beta_{r_1}-\beta_{0r_1})(\beta_{r_2}-\beta_{0r_2}) \\
&= I_{3.1} + I_{3.2}.
\end{aligned}$$

D'une part, par Cauchy-Schwartz :

$$\begin{aligned}
|I_{3.2}| &= \left|\sum_{r_1,r_2=1}^{M} \mathbb{E}\left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}\right](\beta_{r_1}-\beta_{0r_1})(\beta_{r_2}-\beta_{0r_2})\right| \\
&\leq N\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2 \left[\sum_{r_1,r_2=1}^{M} \mathbb{E}\left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}\right]^2\right]^{\frac{1}{2}}.
\end{aligned}$$

Or :

$$\begin{aligned}
\left[\sum_{r_1,r_2} \mathbb{E}\left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}\right]^2\right]^{\frac{1}{2}} &\leq \left[\sum_{r_1,r_2} \mathbb{E}\left[B_{m,r_1,r_2}(\mathbf{d}_i)\right]^2\right]^{\frac{1}{2}} \\
&\leq \left[\sum_{r_1,r_2} \mathbb{E}\left[B^2_{m,r_1,r_2}(\mathbf{d}_i)\right]\right]^{\frac{1}{2}} \\
&\leq \left[\sum_{r_1,r_2} C_5\right]^{\frac{1}{2}} \\
&\leq \sqrt{C_5}\mathcal{O}_P(M).
\end{aligned}$$

D'où :

$$|I_{3.2}| \leq \sqrt{C_5} \times N \times \mathcal{O}_P(M/N)\mathcal{O}_P(M) = o_P(\sqrt{MN}),$$

par la condition $M^5/N \to 0$.

De même, par Cauchy-Schwartz, on a d'autre part :

$$|I_{3.1}| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \left\{ \sum_{r_1,r_2=1}^{M} \left[\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m} - \mathbb{E}\left(\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m} \right) \right]^2 \right\}^{\frac{1}{2}}.$$

Or, on sait que, pour tout $(m, r_1, r_2) = 1, \ldots, M$, $\mathbb{E}_\Upsilon\left[B^2_{m,r_1,r_2}(\mathbf{d}_i)\right] < C_5 < \infty$, donc la variance de la variable $\frac{\partial^3 \log \mathcal{L}(\boldsymbol{\beta}^*)}{\partial\beta_{r_1}\partial\beta_{r_2}\partial\beta_m}$ est bornée.
On en déduit alors :

$$\begin{aligned} |I_{3.1}| &\leq \mathcal{O}_P(M^2) \times \mathcal{O}_P\left(\frac{M}{N}\right) \\ &\leq \mathcal{O}_P\left(\frac{M^3}{N}\right) \\ &= o_P(1) \qquad\qquad \text{car } \frac{M^5}{N} \to 0. \end{aligned}$$

Finalement, en rassemblant tous les termes, on en déduit que :

$$I_1 + I_2 + I_3 = \mathcal{O}_P(\sqrt{MN}).$$

En reprenant le critère de départ sur la dérivée du critère Q :

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_m} &= N\lambda_1 \left[\mathcal{O}_P\left(\frac{\sqrt{MN}}{N\lambda_1} \right) - \frac{1}{\lambda_1} \frac{\partial}{\partial\beta_m} \text{pen}(|\beta_m|, \lambda_1) \text{sign}(\beta_m) \right] \\ &= N\lambda_1 \left[\mathcal{O}_P\left(\frac{\sqrt{M/N}}{\lambda_1} \right) - \frac{1}{\lambda_1} \frac{\partial}{\partial\beta_m} \text{pen}(|\beta_m|, \lambda_1) \text{sign}(\beta_m) \right]. \end{aligned}$$

Or, on sait par hypothèse que $\frac{\sqrt{M/N}}{\lambda_1} \to 0$ quand $N \to \infty$ et par l'hypothèse (H1), que $\liminf_{N\to\infty} \liminf_{\beta\to 0^+} \frac{\frac{\partial}{\partial\beta}\text{pen}(\beta,\lambda_1)}{\lambda_1} > 0$. On constate alors que le signe de $\partial Q(\boldsymbol{\beta})/\partial\beta_m$ est entièrement déterminé par celui de β_m pour N assez grand, et on retrouve bien le résultat attendu.

Concernant les paramètres d'écarts-types $\boldsymbol{\gamma}$, on peut développer la même démonstration puisqu'on dispose des même hypothèses concernant ces paramètres. La seule différence réside dans la conclusion : en effet, les paramètres $\boldsymbol{\gamma}$ sont des paramètres d'écart-types et ne peuvent donc pas être négatifs. De ce fait, on n'observera pas de changement de signe au point 0. Par contre, on peut démontrer que, sous les mêmes hypothèses et pour tout $m = m_N^\gamma, \ldots, M$, on a :

$$\frac{\partial Q(\boldsymbol{\gamma})}{\partial\gamma_m} < 0 \qquad \text{pour } \; 0 < \gamma_m < \epsilon_N.$$

Ainsi, on montre que la dérivée de notre critère est strictement négative, ce qui indique que le minimum est atteint sur le bord inférieur du domaine, à savoir pour $\gamma_m = 0$.

▷ Il reste à présent à démontrer le deuxième point du Théorème (8.2), à savoir, la normalité des estimateurs.
La stratégie utilisée est la suivante, on cherche à montrer que :

$$\Big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\Big)(\widehat{\Upsilon}^1 - \Upsilon_0^1) + \mathcal{G}_N = \frac{1}{N}\nabla \log \mathcal{L}(\Upsilon_0^1) + o_P(N^{-\frac{1}{2}}),$$

Pour cela, on effectue un développement de Taylor de $\nabla Q(\widehat{\Upsilon}^1)$ autour du point Υ_0^1 :

$$\begin{aligned}\nabla Q(\widehat{\Upsilon}^1) = 0 = \nabla \log \mathcal{L}(\Upsilon_0^1) - N\nabla \text{pen}(|\Upsilon_0^1|) + \nabla^2 \log \mathcal{L}(\Upsilon_0^1)(\widehat{\Upsilon}^1 - \Upsilon_0^1) \\ - N\nabla^2 \text{pen}(\Upsilon^{1**})(\widehat{\Upsilon}^1 - \Upsilon_0^1) + \frac{1}{2}(\widehat{\Upsilon}^1 - \Upsilon_0^1)^T \nabla^2\big[\nabla \log \mathcal{L}(\Upsilon^{1*})\big](\widehat{\Upsilon}^1 - \Upsilon_0^1),\end{aligned}$$

où Υ^{1*} et Υ^{1**} sont deux points situés entre $\widehat{\Upsilon}^1$ et Υ_0^1
Et donc, on en déduit :

$$\begin{aligned}&\Big[\mathcal{H}_N(\Upsilon_0^1) + \mathcal{I}(\Upsilon_0^1)\Big](\widehat{\Upsilon}^1 - \Upsilon_0^1) + \mathcal{G}_N \\ &\qquad = \frac{1}{N}\nabla \log \mathcal{L}(\Upsilon_0^1) + \underbrace{\frac{1}{2N}(\widehat{\Upsilon}^1 - \Upsilon_0^1)^T \nabla^2\big[\nabla \log \mathcal{L}(\Upsilon^{1*})\big](\widehat{\Upsilon}^1 - \Upsilon_0^1)}_{=(I_1)} \\ &\quad + \underbrace{\left[\frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0^1) - \frac{1}{N}N\nabla^2 \text{pen}(\Upsilon^{1**}) + \mathcal{H}_N(\Upsilon_0^1) + \mathcal{I}(\Upsilon_0^1)\right](\widehat{\Upsilon}^1 - \Upsilon_0^1)}_{=(I_2)}.\end{aligned}$$

Or, d'une part, on peut montrer que :

$$\begin{aligned}|I_1| &\leq \frac{1}{N}\|\widehat{\Upsilon}^1 - \Upsilon_0^1\|_2^2 \left|\sum_{i=1}^{N}\sum_{m_1,m_2,m_3=1}^{M} B^2_{m_1,m_2,m_3}(\mathbf{D}_i)\right|^{\frac{1}{2}} \qquad \text{[par Cauchy-Schwartz]} \\ &= \frac{1}{N} \times N \times \mathcal{O}_P\left(\frac{M}{N}\right) \times \mathcal{O}_P(M^{\frac{3}{2}}) \\ &= o_P(N^{-\frac{1}{2}}).\end{aligned}$$

Et d'autre part, également par Cauchy-Schwartz :

$$\begin{aligned}|I_2| &\leq \left\|\frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) - N \times \frac{1}{N}\nabla^2 \text{pen}(\Upsilon^{**}) + \mathcal{H}_N(\Upsilon_0^1) + \mathcal{I}(\Upsilon_0^1)\right\|_2 \|(\widehat{\Upsilon}^1 - \Upsilon_0^1)\|_2 \\ &\leq \left[\left\|\frac{1}{N}\nabla^2 \log \mathcal{L}(\Upsilon_0) + \mathcal{I}(\Upsilon_0^1)\right\|_2 + \left\|\mathcal{H}_N(\Upsilon_0^1) - \mathcal{H}_N(\Upsilon^{**})\right\|_2\right] \|(\widehat{\Upsilon}^1 - \Upsilon_0^1)\|_2.\end{aligned}$$

Avec, par l'inégalité (B.1),

$$\left\| \frac{1}{N} \nabla^2 \log \mathcal{L}(\Upsilon_0^1) + \mathcal{I}(\Upsilon_0^1) \right\|_2 = o_P(M^{-\frac{1}{2}}),$$

et, par l'hypothèse (H3') :

$$\left\| \mathcal{H}_N(\Upsilon_0^1) - \mathcal{H}_N(\Upsilon^{1**}) \right\|_2 = o_P(M^{-\frac{1}{2}}).$$

car, $\mathcal{H}_N(\Upsilon_0^1) - \mathcal{H}_N(\Upsilon^{**})$ est une matrice diagonale dont le terme général est de la forme $\left[\frac{\partial}{\partial^2 \Upsilon_m} \text{pen}(\Upsilon_{0m}, \lambda) - \frac{\partial}{\partial^2 \Upsilon_m} \text{pen}(\Upsilon_m^{**}, \lambda)\right]_m$.
D'où :

$$\begin{aligned} |I_2| &\leq [o_P(M^{-\frac{1}{2}}) + o_P(M^{-\frac{1}{2}})] \times \mathcal{O}_P\left(\sqrt{\frac{M}{N}}\right) \\ &= o_P(N^{-\frac{1}{2}}). \end{aligned}$$

Et finalement, on trouve bien l'égalité recherchée, à savoir que :

$$\big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\big)(\widehat{\Upsilon}^1 - \Upsilon_0^1) + \mathcal{G}_N = \frac{1}{N} \nabla \log \mathcal{L}(\Upsilon_0^1) + o_P(N^{-\frac{1}{2}}).$$

On peut alors en déduire que :

$$\begin{aligned} \sqrt{N} \mathbf{A}_N \, \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \Big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\Big) \left[(\widehat{\Upsilon}^1 - \Upsilon_0^1) + \Big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\Big)^{-1} \mathcal{G}_N \right] \\ = \frac{1}{\sqrt{N}} \mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \nabla \log \mathcal{L}(\Upsilon_0^1) + \mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \, o_P(1). \end{aligned}$$

Ainsi, comme $\mathbf{A}_N \mathbf{A}_N^T \to \mathbf{H}$ et comme les valeurs propres de la matrice d'information de Fisher sont bornées inférieurement par l'hypothèse (H6), on a donc :

$$\mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \, o_P(1) = o_P(1)$$

D'autre part, si on note $\mathbf{D}_i = \frac{1}{\sqrt{N}} \mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \nabla \log \mathcal{L}^i(\Upsilon_0^1)$, pour tout $i = 1, \ldots, N$, alors, pour tout $\epsilon > 0$, on a :

$$\begin{aligned} \sum_{i=1}^{N} \mathbb{E}\left(\|\mathbf{D}_i\|_2^2 \mathbf{1}_{\{\|\mathbf{D}_i\|_2 > \epsilon\}} \right) &= N \, \mathbb{E}\left(\|\mathbf{D}_1\|_2^2 \mathbf{1}_{\{\|\mathbf{D}_1\|_2 > \epsilon\}} \right) \\ &\leq N \, \mathbb{E}\left(\|\mathbf{D}_1\|_2^4 \right)^{\frac{1}{2}} \mathbb{P}\left(\|\mathbf{D}_1\|_2 > \epsilon \right)^{\frac{1}{2}}, \quad \text{par Cauchy-Schwartz.} \end{aligned}$$

De là, en utilisant l'inégalité de Bienaymé-Tchebychev, on peut dire que :

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{D}_1\|_2 > \epsilon \right) &\leq \frac{\mathbb{E}\left[\|\mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \nabla \log \mathcal{L}^i(\Upsilon_0^1)\|_2^2 \right]}{N\epsilon^2} \\ &= \mathcal{O}(N^{-1}). \end{aligned}$$

car, on peut affirmer que

$$\mathbb{E}\left[\|\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\nabla\log\mathcal{L}^i(\Upsilon_0^1)\|_2^2\right]=\mathcal{O}(1).$$

En effet, par définition, on a que

$$\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\nabla\log\mathcal{L}^i(\Upsilon_0^1)\sim\mathcal{N}(0,\mathbf{I}),$$

et donc :

$$\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\nabla\log\mathcal{L}^i(\Upsilon_0^1)\sim\mathcal{N}(0,\mathbf{A}_N\mathbf{A}_N^T)\to\mathcal{N}(0,H).$$

On en déduit alors le caractère borné de l'espérance pour N assez grand. Par ailleurs, on a aussi que :

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{D}_1\|_2^4\right)&=\frac{1}{N^2}\mathbb{E}\left[\|\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\nabla\log\mathcal{L}(\Upsilon_0^1)\|_2^2\right]\\&=\frac{1}{N^2}\mathbb{E}\left[\|\mathbf{A}_N\mathbf{A}_N^T\|_2^2\|\mathcal{I}^{-1}(\Upsilon_0^1)\|_2^2\|\nabla^T\log\mathcal{L}(\Upsilon_0^1)\nabla\log\mathcal{L}(\Upsilon_0^1)\|_2^2\right]\\&\leq\frac{1}{N^2}\lambda_{vp}^{\max}(\mathbf{A}_N\mathbf{A}_N^T)\;\lambda_{vp}^{\min}\left(\mathcal{I}(\Upsilon_0^1)\right)\;\mathbb{E}\left[\|\nabla^T\log\mathcal{L}(\Upsilon_0^1)\nabla\log\mathcal{L}(\Upsilon_0^1)\|_2^2\right]\\&=\mathcal{O}\left(\frac{M^2}{N^2}\right).\end{aligned}$$

car $\|\nabla^T\log\mathcal{L}(\Upsilon_0^1)\nabla\log\mathcal{L}(\Upsilon_0^1)\|_2^2$ est le carré d'une somme de M termes bornés. Finalement, en regroupant ces deux dernières majorations, on a :

$$\sum_{i=1}^N\mathbb{E}\left(\|\mathbf{D}_i\|_2^2\mathbf{1}_{\{\|\mathbf{D}_i\|_2>\epsilon\}}\right)=N\times\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\times\mathcal{O}\left(\frac{M}{N}\right)=\mathcal{O}\left(\frac{M}{\sqrt{N}}\right)=o(1).$$

De plus :

$$\begin{aligned}\sum_{i=1}^N\mathbb{V}(\mathbf{D}_i)&=\sum_{i=1}^N\mathbb{V}\left(\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\nabla\log\mathcal{L}^i(\Upsilon_0^1)\right)\\&=\sum_{i=1}^N\mathbf{A}_N\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)\underbrace{\mathbb{V}\left(\nabla\log\mathcal{L}^i(\Upsilon_0^1)\right)}_{=\mathcal{I}(\Upsilon_0^1)}\mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1)^T\mathbf{A}_N^T\\&=\mathbf{A}_N\mathbf{A}_N^T\to\mathbf{H}.\end{aligned}$$

Ainsi, on a montré que les variables D_i vérifient bien les conditions d'application du théorème central limite de Lindeberg-Feller. Cette version du théorème est rappelée ci-dessous.

Théorème B.1. *Soit $X_1, \ldots, X_N, \ldots$ suite de variables aléatoires indépendantes, réelles centrées et de variance σ_i^2. On note $\Sigma_N^2 = \sum_{i=1}^N \sigma_i^2$ la variance de la somme et $S_N = X_1 + \ldots + X_N$. Si, pour tout $\epsilon > 0$,*

$$\frac{1}{\Sigma_N^2} \sum_{i=1}^{N} \mathbb{E}\left(X_i^2 \mathbf{1}_{\{|X_i|>\epsilon\}}\right) \xrightarrow{N\to\infty} 0,$$

alors, on a que :

$$\frac{S_N}{\Sigma_N} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Ainsi, on en déduit que

$$\frac{1}{\sqrt{N}} \mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \nabla \log \mathcal{L}(\Upsilon_0^1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{H}).$$

Et donc :

$$N \mathbf{A}_N \mathcal{I}^{-\frac{1}{2}}(\Upsilon_0^1) \big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\big) \left[\widehat{\Upsilon}^1 - \Upsilon_0^1 + \big(\mathcal{I}(\Upsilon_0^1) + \mathcal{H}_N\big)^{-1} \mathcal{G}_N\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{H}). \qquad \blacksquare$$

Annexe C

Mise à jour des paramètres pour la procédure de sélection de variables

Nous détaillons dans cette annexe les mises à jour des vecteurs de paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ dans la procédure de sélection de variables basée sur l'algorithme EM, décrite en Section 8.3.2. Ces mises à jour se ramènent à des problèmes de vraisemblance pénalisée au moyen d'une pénalité de type SCAD.

C.1 Mise à jour des paramètres d'effets fixes β_{jk}

Vis-à-vis des paramètres d'effets fixes et en reprenant la notation $\mathbf{d}_i^{\beta} = \mathbf{d}_i - \mathbf{G}_{\theta}^{1/2}\widehat{\boldsymbol{\vartheta}}_i$ pour tout $i = 1, \ldots, N$, le problème d'optimisation que l'on cherche à résoudre est donné par :

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\ \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{d}_i^{\beta} - \boldsymbol{\beta}\|^2 + \sigma_{\varepsilon}^{2\,[h]}\sum_{jk}\Big[\lambda_1|\beta_{jk}|\mathbf{1}_{\{|\beta_{jk}|\leq\lambda_1\}} - \frac{\beta_{jk}^2 - 2a\lambda_1|\beta_{jk}| + \lambda_1^2}{2(a-1)}\mathbf{1}_{\{\lambda_1<|\beta_{jk}|\leq a\lambda_1\}} + \frac{(a+1)\lambda_1^2}{2}\mathbf{1}_{|\beta_{jk}|>a\lambda_1\}}\Big].$$

Dans le cadre de la mise à jour des paramètres d'effets fixes, la principale différence par rapport au problème d'optimisation pénalisé SCAD classique réside dans le fait que nous faisons face, dans le cadre mixte, à un problème de vraisemblance pénalisée et non plus de moindres carrés pénalisés, faisant alors intervenir les effets aléatoires et le paramètre de variance σ_{ε}^2.
À ce stade, on peut alors distinguer 3 cas :

- Si $0 \leq |\widehat{\beta}_{jk}| \leq \lambda_1$, le problème d'optimisation se réduit à :

$$\underset{\boldsymbol{\beta}}{\arg\min}\quad \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{d}_i^{\beta} - \boldsymbol{\beta}\|^2 + \sigma_{\varepsilon}^{2\,[h]}\lambda_1\sum_{jk}|\beta_{jk}|,$$

dont l'expression est équivalente à une formulation du LASSO dans un cadre de maximum de vraisemblance en présence de plusieurs courbes. En dérivant le même type de calcul que pour le LASSO, basé sur la notion de sous-gradient (Tibshirani 1996), on retrouve donc une solution, proche de celle obtenue pour le LASSO, donnée par :

$$\widehat{\beta}_{jk} = \text{sign}\left(\overline{d^{\beta}_{jk}}\right)\left[\overline{d^{\beta}_{jk}} - {\sigma_{\varepsilon}^{2}}^{[h]}/N\lambda_1\right]_+,$$

et cela, si et seulement si on vérifie $|\overline{d^{\beta}_{jk}}| \leq \left(1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N\right)\lambda_1$.

- Si $\lambda_1 < |\widehat{\beta}_{jk}| \leq a\lambda_1$, il nous faut alors résoudre :

$$\underset{\boldsymbol{\beta}}{\arg\min} \quad \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{d}_i^{\beta} - \boldsymbol{\beta}\|^2 - {\sigma_{\varepsilon}^{2}}^{[h]}\,\frac{\beta_{jk}^2 - 2a\lambda_1|\beta_{jk}| + \lambda_1^2}{2(a-1)},$$

nous conduisant à la solution

$$\widehat{\beta}_{jk} = \frac{(a-1)\overline{d^{\beta}_{jk}} - \,{\sigma_{\varepsilon}^{2}}^{[h]}a\lambda_1\text{sign}\left(\overline{d^{\beta}_{jk}}\right)/N}{a - \left(1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N\right)},$$

valable pour $(1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N)\lambda_1 < |\overline{d^{\beta}_{jk}}| \leq a\lambda_1$. On a supposé pour un tel résultat que $\widehat{\beta}_{jk}$ et $\overline{d^{\beta}_{jk}}$ étaient de même signe pour tout $(j,k) \in \Lambda$. On peut aisément vérifier que ceci garantit la bonne définition de $\widehat{\beta}_{jk}$ dans le cas où $|\widehat{\beta}_{jk}| \leq a\lambda_1$.

- Enfin, si $|\widehat{\beta}_{jk}| > \lambda_1$, l'estimateur est simplement donné par

$$\widehat{\beta}_{jk} = \overline{d^{\beta}_{jk}},$$

pour $|\overline{d^{\beta}_{jk}}| > a\lambda_1$.

En regroupant tous les cas, on retrouve bien, finalement, l'expression donnée, soit :

$$\widehat{\beta}_{jk}^{[h+1]} = \begin{cases} \text{sign}\left(\overline{d^{\beta}_{jk}}\right)\left[|\overline{d^{\beta}_{jk}}| - \lambda_1{\sigma_{\varepsilon}^{2}}^{[h]}/N\right]_+ & \text{si } |\overline{d^{\beta}_{jk}}| \leq \lambda_1(1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N), \\ \frac{(a-1)\overline{d^{\beta}_{jk}} - {\sigma_{\varepsilon}^{2}}^{[h]}a\lambda_1\text{sign}\left(\overline{d^{\beta}_{jk}}\right)/N}{a - (1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N)} & \text{si } \lambda_1(1 + {\sigma_{\varepsilon}^{2}}^{[h]}/N) < \overline{d^{\beta}_{jk}} \leq a\lambda_1, \\ \overline{d^{\beta}_{jk}} & \text{si } \overline{d^{\beta}_{jk}} > a\lambda_1. \end{cases} \tag{C.1}$$

C.2 Mise à jour des variances des effets aléatoires γ_{jk}

Pour les variances associées aux effets aléatoires, le problème d'optimisation à résoudre est le suivant :

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2\sigma_\varepsilon^{2\,[h]}} \sum_{i=1}^{N} \left[\|\mathbf{d} - \boldsymbol{\beta} - \mathbf{G}_\theta^{1/2} \widehat{\boldsymbol{\vartheta}}_i\|^2 + \operatorname{tr}\left(\left(\mathbf{G}_\theta^{1/2}\right)^T \mathbb{V}(\widehat{\boldsymbol{\vartheta}}_i|\mathbf{d}_i) \mathbf{G}_\theta^{1/2} \right) \right] + \sum_{jk} \left[\lambda_2 \gamma_{jk} \mathbf{1}_{\{\gamma_{jk} \leq \lambda_1\}} - \frac{\gamma_{jk}^2 - 2a\lambda_2\gamma_{jk} + \lambda_2^2}{2(a-1)} \mathbf{1}_{\{\lambda_2 < \gamma_{jk} \leq a\lambda_2\}} + \frac{(a+1)\lambda_2^2}{2} \mathbf{1}_{\gamma_{jk} > a\lambda_2\}} \right],$$

$$\text{t.q.} \quad \gamma_{jk} > 0 \quad \forall (j,k) \in \Lambda.$$

Le problème présenté ci-dessus possède une contrainte supplémentaire portant sur la positivité des paramètres du vecteur $\boldsymbol{\gamma}$. Nous reprenons par la suite les notations définies en (8.15). L'optimisation de ce critère pénalisé nous conduit là encore à différencier trois cas :

- Si $0 \leq \widehat{\gamma}_{jk} \leq \lambda_2$, alors le problème d'optimisation se réduit à :

$$\arg\min_{\gamma_{jk}} \quad \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} \left[(d_{i,jk} - \beta_{jk} - \gamma_{jk}\sqrt{2^{-j\eta}}\widehat{\vartheta}_{i,jk})^2 + \gamma_{jk}^2 2^{-j\eta} \mathbb{V}(\widehat{\vartheta}_{i,jk}|\mathbf{d}_i) \right] + \lambda_2 \gamma_{jk},$$

$$\text{t.q.} \quad \gamma_{jk} > 0.$$

En dérivant et en annulant ce critère, on trouve alors que la mise à jour du paramètre γ_{jk} est donnée par :

$$\widehat{\gamma}_{jk} = \left\{ 2^{-j\eta} \mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2 | \mathbf{d} \right] \right\}^{-1} \left[d_{+jk}^{\gamma} - \lambda_2 \sigma_\varepsilon^{2\,[h]} \right], \tag{C.2}$$

Cette expression n'est pas toujours positive, en particulier, $\widehat{\gamma}_{jk}$ est positif si et seulement si $d_{+jk}^{\gamma} \geq \lambda_2 \sigma_\varepsilon^2$. Dans le cas contraire, l'expression (C.2) est négative et donc hors du domaine des contraintes. On considérera donc, le cas échéant, que la solution recherchée se trouve sur le bord du domaine, c'est-à-dire que $\widehat{\gamma}_{jk} = 0$.

- Si $\lambda_2 \leq \widehat{\gamma}_{jk} \leq a\lambda_2$, l'optimisation se résume à :

$$\arg\min_{\boldsymbol{\gamma}} \quad \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} \left[(d_{i,jk} - \beta_{jk} - \gamma_{jk}\sqrt{2^{-j\eta}}\widehat{\vartheta}_{i,jk})^2 + \gamma_{jk}^2 2^{-j\eta} \mathbb{V}(\widehat{\vartheta}_{i,jk}|\mathbf{d}_i) \right] - \gamma_{jk} \frac{\sigma_\varepsilon^2}{a-1} + \frac{a\sigma_\varepsilon^2 \lambda_2}{a-1},$$

conduisant à la solution :

$$\widehat{\gamma}_{jk} = \left[2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] - \frac{\sigma_\varepsilon^{2\,[h]}}{a-1}\right]^{-1}\left[d_{+jk}^{\gamma} - \frac{\sigma_\varepsilon^{2\,[h]}a\lambda_2}{a-1}\right].$$

► Enfin, si $\widehat{\gamma}_{jk} \geq a\lambda_2$, alors la solution est donnée par :

$$\left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]\right\}^{-1} d_{+jk}^{\gamma}.$$

Finalement, la mise à jour des paramètres dans l'ensemble des cas est donnée par :

$$\widehat{\gamma}_{jk}^{[h+1]} = \begin{cases} \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]\right\}^{-1}\left[d_{+jk}^{\gamma} - \lambda_2\sigma_\varepsilon^{2\,[h]}\right]_+ \\ \qquad\qquad \text{si } d_{+jk}^{\gamma} \leq \lambda_2\left[2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] + \sigma_\varepsilon^{2\,[h]}\right], \\ \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] - \frac{\sigma_\varepsilon^{2\,[h]}}{a-1}\right\}^{-1}\left[d_{+jk}^{\gamma} - \frac{\sigma_\varepsilon^{2\,[h]}a\lambda_2}{a-1}\right] \\ \qquad\qquad \text{si } \left[2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right] - \sigma_\varepsilon^{2\,[h]}\right]\lambda_2 < d_{+jk}^{\gamma} \leq a\lambda_2 2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right], \\ \left\{2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]\right\}^{-1} d_{+jk}^{\gamma} \\ \qquad\qquad \text{si } d_{+jk}^{\gamma} > a\lambda_2 2^{-j\eta}\mathbb{E}\left[(\widehat{\vartheta}_{+jk}^{[h+1]})^2|\mathbf{d}\right]. \end{cases}$$

Références

Abramovich, F., T. Sapatinas, et B. Silverman (1998). Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society Series B Stat Methodol 60*, 725–749.

Abry, P., P. Goncalves, et P. Flandrin (1995). Wavelets, spectrum analysis and 1/f processes. *Lecture Notes in Statistics "Wavelets and Statistics" 103*, 15–29.

Amato, U. et T. Sapatinas (2005). Wavelet shrinkage approaches to baseline signal estimation from repeated noisy measurements. *Advances and Applications in Statistics 51*, 21–50.

Angelini, C., D. de Canditiis, et F. Leblanc (2003). Wavelet regression estimation in nonparametric mixed effect models. *Journal of Multivariate Analysis 85*(2), 267–291.

Antoniadis, A., J. Bigot, S. Lambert-Lacroix, et F. Letue (2007). Non parametric pre-processing methods and inference tools for analyzing time-of-flight mass spectrometry data. *Current Analytical Chemistry 3*(2), 127–147.

Antoniadis, A., J. Bigot, et T. Sapatinas (2001). Wavelet estimators in nonparametric regression : A comparative simulation study. *Journal of Statistical Software 6*(6), 1–83.

Antoniadis, A., J. Bigot, et R. von Sachs (2008). A multiscale approach for statistical characterization of functional images. *Journal of Computational and Graphical Statistics 18*(1), 216–237.

Antoniadis, A. et J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association 96*(455), 939–955.

Antoniadis, A., I. Gijbels, et G. Gregoire (1997). Model selection using wavelet decomposition and applications. *Biometrika 84*(4), 751–763.

Antoniadis, A. et C. Lavergne (1995). Variance function estimation in regression by wavelet methods. *Lecture Notes in Statistics "Wavelets and Statistics" 103*, 31–42.

Antoniadis, A. et T. Sapatinas (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis 51*(10), 4793–4813.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Benjamini, Y. et Y. Hochberg (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B 57*(1), 289–300.

Biernacki, C., G. Celeux, et 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*, 561–575.

Bigot, J. (2011). Fréchet means of curves for signal averaging and application to ecg data analysis. Sumitted.

Bondell, H., A. Krishna, et S. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics 66*(4), 1069–1077.

Bouveyron, C. et C. Brunet (2013). Model-based clustering of high-dimensional data : A review. *Computational Statistics and Data Analysis*, in press.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *CART : Clasification and Regression Trees*. Chapman & Hall.

Brockmann, M., T. Gasser, et E. Herrmann (1993). Locally adaptive bandwith choice for kernel regression estimators. *Journal of American Statistical Association 88*, 1302–1309.

Buhlmann, P. et S. van de Geer (2011). *Statistics for High-Dimensional Data*. Springer.

Bunea, F., A. Tsybakov, et M. Wegkamp (2007). Aggregation for gaussian regression. *Annals of statistics 35*, 1674–1697.

Chang, W. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society. Series C 32*(3), 267–275.

Chen, S. et D. Donoho (1995). Atomic decomposition by basis pursuit. Technical report, Stanford University.

Chen, Z. et D. Dunson (2003). Random effects selection in linear mixed models. *Biometrics 59*, 762–769.

Cohen, A., I. Daubechies, B. Jawerth, et P. Vial (1993). Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris 316*(1), 417–421.

Dalhaus, R., H. Neumann, et R. von Sachs (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli 5*(5), 873–906.

Daubechies, I. (1992). *Ten lectures on Wavelets*. Society for Industrial and Applied Mathematics.

Dauxois, J., A. Pousse, et Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function : Some applications to statistical inference. *Journal of Multivariate Analysis 12*(1), 136 – 154.

Delyon, B., M. Lavielle, et E. Moulines (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics 27*(1), 94–128.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

DeVore, R. et G. Lorentz (1993). *Constructive Approximation*. Springer Verlag.

Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de Statistique Appliquée 19*(2), 19–33.

Donoho, D. et X. Huo (2002). Uncertainty principles and ideal atomics decompositions. *IEEE Transactions on Information Theory 47*, 2845–2863.

Donoho, D. et I. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

Donoho, D. et I. Johnstone (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics 26*, 879–921.

Donoho, D., I. Johnstone, G. Kerkyacharian, et D. Picard (1995). Wavelet shrinkage : asymptopia. *Journal of the Royal Statistical Society, Ser. B 57*(2), 371–394.

Duda, R. et P. Hart (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons.

Eckel-Passow, J. E., A. L. Oberg, T. M. Therneau, et H. R. Bergen (2009, Jul). An insight into high-resolution mass-spectrometry data. *Biostatistics 10*, 481–500.

Efron, B., T. Hastie, I. Johnstone, et R. Tibshirani (2004). Least angle regression. *Annals of statistics 32*(2), 407–499.

Fan, J. (1996). Test of significance based on wavelet thresholding and neyman ?s truncation. *JASA 91*, 674–688.

Fan, J. et I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman and Hall.

Fan, J. et R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. et H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*(3), 928–961.

Fisher, R. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society 22*, 700–725.

Frazier, M., B. Jawerth, et G. Weiss (1991). *Littlewood-Paley Theory and the Study of function Spaces*. Number 79. American Mathematical Society.

Fridlyand, J., A. M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. M. Ljung, A. N. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. W. Gray, F. Waldman, D. Pinkel, et D. G. Albertson (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer 6*, 96.

Fryzlewicz, P. (2008). Data-driven wavelet-fisz methodology for nonparametric function estimation. *Electronic Journal of Statistics 2*, 863–896.

Gao, H. (1997). Wavelet shrinkage estimates for heteroscedastic regression models. Technical report, MathSoft, Inc.

Gasser, T., L. Stroka, et C. Jennen-Steinmetz (1989). Residual variance and residual pattern in nonlinear regression. *Biometrika 73*, 625–633.

Giacofci, M., S. Lambert-Lacroix, G. Marot, et F. Picard (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics 69*(1), 31–40.

Green, P. et B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models : a roughness penalty approach*. Chapman & Hall.

Haar, A. (1910). Zur Theorie der orthogonalen Funktionen-Systeme. *Annals of Mathematics 69*, 331–371.

Hall, P. et M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B 68*(1), 109–126.

Härdle, W., G. Kerkyacharian, D. Picard, et A. Tsybakov (1998). *Wavelets, Approximation and Statistical Applications*. Springer.

Harville, D. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika 61*, 383–385.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association 72*, 320–340.

Hastie, T., R. Tibshirani, et J. Friedman (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics 31*(2), 423–447.

Henderson, C., O. Kempthorne, S. Searle, et C. V. Krosig (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Huang, S. et H. Lu (2000). Bayesian wavelet shrinkage for nonparametric mixed-effects models. *Statistica Sinica 10*, 1021–1040.

Istas, J. (1992). Wavelet coefficients of a gaussian process and applications. *Annales de l'institut Henri Poincare (B) Probabilites et Statistiques 28*(4), 537–556.

Jacques, J. et C. Preda (2013). Functional data clustering : a survey. Technical report, INRIA.

James, G. et C. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the Americal Statistical Association 98*, 397–408.

Johnstone, I. et B. Silverman (1997). Wavelet threshold estimators for data with

correlated noise. *Journal of the Royal Statistical Society, Ser. B 59*(2), 319–351.

Juditsky, A. et B. Delyon (1996). On minimax wavelets estimators. *Applied and computational harmonic analysis 3*, 215–228.

Kaufman, L. et P. Rousseeuw (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods 1*, 405–416.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya : The Indian Journal of Statistics, Series A 62*(1), 49–66.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society 4*(3), 502–506.

Kim, Y., H. Choi, et H.-S. Oh (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association 103*(484), 1665–1673.

Knight, K. et W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Kuhn, E. et M. Lavielle (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis 49*, 1020–1038.

Laird, N. et J. Ware (1982). Random effects models for longitudinal data. *Biometrics 38*, 963–974.

Lindstrom, M. et D. Bates (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*(404), 1014–1022.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, Volume 1, pp. 281–297. Univ. of Calif. Press.

Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition : The Sparse Way*. Academic Press.

McLachlan, G. et T. Krishnan (2008). *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Meinhausen, N. et P. Buhlmann (2004). Variable selection and high-dimensional graphs with the lasso. Technical report, ETH Zurich.

Meinhausen, N. et P. Buhlmann (2006). Consistent neighbourhood selection for high-dimensional graphs with the lasso. *Annals of Statistics 34*, 1436–1462.

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

Meng, X. (2000). Missing data : dial for ? ? ? *Journal of American Statistical Association 95*, 1325–1330.

Meng, X. et D. Rubin (1993). Maximum likelihood estimation via the ecm algorithm : A general framework. *Biometrika 80*(2), 267–278.

Meyer, Y. (1990). *Ondelettes*. Paris : Hermann.

Meynet, C. (2012). *Sélection de variables pour la classification non supervisée en grande dimension*. Ph. D. thesis, Université Paris-Sud XI.

Morris, J. S. et R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B Stat Methodol 68*, 179–199.

Pan, W. et X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*, 1145–1164.

Patterson, H. et R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Petricoin, E. F., A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, et L. A. Liotta (2002, Feb). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet 359*, 572–577.

Picard, F.and Robin, S., M. Lavielle, C. Vaisse, et J.-J. Daudin (2005). A statistical approach for array cgh data analysis. *BMC Bioinformatics 6*, 1–14.

Ramsay, J. et B. Silverman (1997). *Functional Data Analysis*. Springer, New York.

Roy, P., C. Truntzer, D. Maucourt-Boulch, T. Jouve, et N. Molinari (2011). Protein mass spectra data analysis for clinical biomarker discovery : a global review. *Briefings in Bioinformatics 12*(2), 176–186.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

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

Todd Ogden, R. (1997). On preconditioning the data for the wavelet transform when the sample size is not a power of two. *Communications in Statistics - Simulation and Computation 26*(2), 467–486.

van de Wiel, M. A., F. Picard, W. N. van Wieringen, et B. Ylstra (2011, Jan). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief. Bioinformatics 12*, 10–21.

Van Wieringen, W. N., M. A. Van De Wiel, et B. Ylstra (2008, Jul). Weighted clustering of called array CGH data. *Biostatistics 9*, 484–500.

Verbeke, G. et G. Molenberghs (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics.

von Sachs, R. et B. MacGibbon (2000). Nonparametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics 27*(3), 475–499.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Zhang, J. et G. Walter (1994). A wavelet-based kl-like expansion for wide-sense stationary random processes. *Signal Processing, IEEE Transactions on 42*(7), 1737–1745.

Zhao, P. et B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Classification non supervisée et sélection de variables dans les modèles mixtes fonctionnels

Un nombre croissant de domaines scientifiques collectent de grandes quantités de données comportant beaucoup de mesures répétées pour chaque individu. Ce type de données peut être vu comme une extension des données longitudinales en grande dimension. Le cadre naturel pour modéliser ce type de données est alors celui des modèles mixtes fonctionnels.

Nous traitons, dans une première partie, de la classification non-supervisée dans les modèles mixtes fonctionnels. Nous présentons dans ce cadre une nouvelle procédure utilisant une décomposition en ondelettes des effets fixes et des effets aléatoires. Notre approche se décompose en deux étapes : une étape de réduction de dimension basée sur les techniques de seuillage des ondelettes et une étape de classification où l'algorithme EM est utilisé pour l'estimation des paramètres par maximum de vraisemblance. Nous présentons des résultats de simulations et nous illustrons notre méthode sur des jeux de données issus de la biologie moléculaire (données omiques). Cette procédure est implémentée dans le package R "curvclust" disponible sur le site du CRAN.

Dans une deuxième partie, nous nous intéressons aux questions d'estimation et de réduction de dimension au sein des modèles mixtes fonctionnels et nous développons en ce sens deux approches. La première approche se place dans un objectif d'estimation dans un contexte non-paramétrique et nous montrons dans ce cadre, que l'estimateur de l'effet fixe fonctionnel basé sur les techniques de seuillage par ondelettes possède de bonnes propriétés de convergence. Notre deuxième approche s'intéresse à la problématique de sélection des effets fixes et aléatoires et nous proposons une procédure basée sur les techniques de sélection de variables par maximum de vraisemblance pénalisée et utilisant deux pénalités SCAD sur les effets fixes et les variances des effets aléatoires. Nous montrons dans ce cadre que le critère considéré conduit à des estimateurs possédant des propriétés oraculaires dans un cadre où le nombre d'individus et la taille des signaux divergent. Une étude de simulation visant à appréhender les comportements des deux approches développées est réalisée dans ce contexte.

Curve clustering and variable selection in mixed effects functional models. Applications to molecular biology

More and more scientific studies yield to the collection of a large amount of data that consist of sets of curves recorded on individuals. These data can be seen as an extension of longitudinal data in high dimension and are often modeled as functional data in a mixed-effects framework.

In a first part we focus on performing unsupervised clustering of these curves in the presence of inter-individual variability. To this end, we develop a new procedure based on a wavelet representation of the model, for both fixed and random effects. Our approach follows two steps : a dimension reduction step, based on wavelet thresholding techniques, is first performed. Then a clustering step is applied on the selected coefficients. An EM-algorithm is used for maximum likelihood estimation of parameters. The properties of the overall procedure are validated by an extensive simulation study. We also illustrate our method on high throughput molecular data (omics data) like microarray CGH or mass spectrometry data. Our procedure is available through the R package "curvclust", available on the CRAN website.

In a second part, we concentrate on estimation and dimension reduction issues in the mixed-effects functional framework. Two distinct approaches are developed according to these issues. The first approach deals with parameters estimation in a non parametrical setting. We demonstrate that the functional fixed effects estimator based on wavelet thresholding techniques achieves the expected rate of convergence toward the true function. The second approach is dedicated to the selection of both fixed and random effects. We propose a method based on a penalized likelihood criterion with SCAD penalties for the estimation and the selection of both fixed effects and random effects variances. In the context of variable selection we prove that the penalized estimators enjoy the oracle property when the signal size diverges with the sample size. A simulation study is carried out to assess the behaviour of the two proposed approaches.