

# Mémoire d'Habilitation à Diriger des Recherches

Jean-Michel LOUBES

Loubes Jean-Michel  
Laboratoire de Mathématiques de l'université Montpellier 2  
Equipe de Statistique et Probabilités  
34000 Montpellier  
[http ://www.math.univ-montp2.fr/ ~loubes](http://www.math.univ-montp2.fr/~loubes)  
Jean-Michel.Loubes@math.univ-montp2.fr

# Table des matières

<b>1</b>	<b>Régularisation par complexité</b>	<b>11</b>
1.1	M-estimation pénalisée . . . . .	11
1.2	Utilisation de la norme $l^1$ . . . . .	15
1.3	Sélection de variables . . . . .	18
1.4	Sélection de modèles . . . . .	21
<b>2</b>	<b>Statistique des Processus</b>	<b>29</b>
2.1	Modèle de fonctions multifractales . . . . .	29
2.2	Estimation Bayésienne d'une fonction multifractale . . . . .	30
2.3	Estimation des paramètres du modèle multifractal . . . . .	32
2.4	Application aux processus de Lévy . . . . .	35
<b>3</b>	<b>Problèmes inverses</b>	<b>39</b>
3.1	Présentation Générale . . . . .	39
3.2	Méthodes de régularisation de problèmes inverses . . . . .	41
3.3	Exemples de problèmes inverses non linéaires . . . . .	51
3.3.1	Application en économétrie . . . . .	51
3.3.2	Application en tomographie sismique . . . . .	54
3.4	Espaces de régularité pour les problèmes inverses . . . . .	57
<b>4</b>	<b>Un problème inverse particulier : le décalage de courbes</b>	<b>61</b>
4.1	Estimation de décalages entre courbes . . . . .	61
4.2	Etude de décalages stochastiques . . . . .	65
4.3	Processus de déformation aléatoires . . . . .	70
4.4	Classification de variétés Riemanniennes . . . . .	73
4.5	Régression inverse spatiale . . . . .	74



# Introduction Générale

Le présent mémoire d'Habilitation à Diriger des Recherches a pour but de décrire mon activité dans le domaine de la statistique non paramétrique. En particulier, mon propos vise à mettre en lumière la démarche scientifique qui m'a conduit à travailler sur des thèmes très variés, mais qui s'articulent effectivement autour de trois axes principaux : la M-estimation, les fonctions multifractales et les problèmes inverses. J'ajoute qu'au sein des 3 thèmes, j'ai essayé de maintenir un équilibre entre théorie et pratique, en puisant mes sujets de recherche à la fois dans le nombre des grandes questions en statistique mathématique et parmi les problèmes plus concrets que j'ai pu rencontrer.

En effet, l'un des avantages majeurs de la statistique est de constituer l'un des domaines où les liens entre la théorie mathématique et la pratique sont les plus forts. Cette proximité est particulièrement stimulante. Souvent, l'étude des données réelles nous fournit un cadre cohérent dans lequel évoluer et les questions à nous poser surgissent alors logiquement. Les divers travaux à accomplir et objectifs à atteindre se manifestent eux aussi de manière naturelle. Ce cadre ne limite pas les possibilités de recherche et les grands thèmes usuels en statistique sont toujours présents. Ainsi, j'ai pu prolonger mes réflexions vers des problèmes théoriques d'estimation adaptative, des propriétés probabilistes de modèles, et d'optimalité de procédures statistiques, tout en essayant d'apporter une réponse satisfaisante aux sujets rencontrés dans la pratique. J'ai pu développer de nouvelles méthodes que je me suis appliqué à mettre en œuvre par le biais d'algorithmes pour résoudre des problèmes aussi bien pratiques que théoriques.

La synthèse de mes travaux correspond à 27 articles, 12 publiés, 11 soumis et 4 en voie d'achèvement.

L'objectif principal de la statistique non paramétrique étant de reconstruire une fonction à partir d'observations, j'ai étudié le comportement asymptotique d'une classe d'estimateurs dont l'importance est fondamentale en statistique, les M-estimateurs et je me suis attaché à en améliorer les performances. Ces estimateurs sont définis comme réalisant le minimum d'une fonction de perte sous certaines contraintes. Ils ont été étudiés par de nombreux auteurs dont nous ne donnons ici qu'une liste très partielle ([120], [15], [115], ...). Ils trouvent de nombreuses applications dans les domaines les plus variés tels que la physique, la biologie, l'économie ou le traitement du signal ... En outre, la plupart des estimateurs usuels en statistique peuvent

être interprétés comme des M-estimateurs. Ainsi, la théorie des M-estimateurs fournit un cadre générique pour étudier la plupart des estimateurs employés aussi bien en théorie qu'en pratique.

Pour estimer une fonction inconnue à la vitesse optimale de convergence, on suppose d'une manière générale que l'on connaît sa régularité. C'est bien de cette hypothèse supplémentaire, peu compatible avec la réalité, que j'ai souhaité m'affranchir. Un estimateur capable d'approcher une fonction totalement inconnue à la meilleure vitesse possible sera dit *adaptatif*. Ainsi, la première partie de mon travail en statistique non paramétrique a consisté à modifier les méthodes à base de M-estimateurs afin d'obtenir des versions *adaptatives*.

A cet effet, j'ai introduit une pénalité portant sur la norme  $l^1$  des coefficients de la fonction à estimer, décomposée dans une bonne base. Les propriétés de la norme  $l^1$  permettent de choisir des signaux dont l'information est résumée par peu de gros coefficients, comme l'ont remarqué de nombreux auteurs, tout particulièrement Donoho et Johnstone [40] ou [42]. Par conséquent, minimiser une fonction de perte, sous la contrainte donnée par la norme  $l^1$ , permet d'obtenir des comportements similaires à ceux obtenus par les techniques de type seuillage décrits dans les travaux [43]. J'ai donc développé cette nouvelle méthode d'estimation dans 4 directions importantes de la statistique non paramétrique, présentée dans la partie 1.2 : pour l'estimation d'une fonction de régression, l'estimation d'une densité, l'estimation dans un modèle de problème inverse dans la partie 3.2 et finalement en sélection de variables, dans le chapitre 1.3, recherche en cours d'achèvement. L'aboutissement de ce dernier travail contribuerait à fournir une interprétation mathématique de l'algorithme LARS, un enjeu majeur de la recherche statistique actuelle.

Afin de progresser dans la compréhension des M-estimateurs pénalisés, j'ai établi des liens entre ces estimateurs et les méthodes de sélection de modèles mises au point par Birgé et Massart [15]. La question naturelle en estimation non paramétrique "Quelle méthode d'estimation choisir et pour quelle situation ?" m'a confronté aux différents critères d'évaluation des procédures statistiques d'estimation. C'est la raison pour laquelle je me suis intéressé à la nouvelle théorie des ensembles maxisets récemment développée par Cohen, De Vore, Kerkyacharyan et Picard dans [25]. Il s'agit d'une théorie visant à déterminer les différents ensembles de régularité sur lesquels une méthode d'estimation donnée sera efficace, ce qui signifie qu'elle permet d'estimer une fonction de cet espace à une vitesse de convergence préalablement choisie. Plusieurs auteurs ont calculé les maxisets des procédures d'estimation usuelles et c'est donc tout naturellement que j'ai construit les maxisets relatifs aux M-estimateurs pénalisés, résultats présentés dans la partie 1.4.

L'approfondissement de ma connaissance des méthodes d'estimation par sélection de modèles m'a conduit à utiliser ces techniques pour résoudre un problème bien réel : l'estimation de temps de parcours sur le réseau routier. Ce projet ambitieux nécessite que l'on soit capable de prévoir l'évolution journalière des vitesses de circulation sur tout le réseau routier. Ce projet de recherche a été effectué lors du co-encadrement de thèses CIFRE, et également au sein du groupe de recherche financé par l'ANR, MIST-R (Modélisation Informatique et Statistique

du Trafic Routier), dont j'assume la direction. Ce projet a permis d'unir les compétences de plusieurs laboratoires comprenant des statisticiens spécialistes de la sélection de modèles (équipe de recherche SELECT INRIA et Université Paris Sud), des probabilistes (laboratoire de statistique et probabilités de l'Université Toulouse 3), des informaticiens (laboratoire LRI de l'université Paris Sud) et enfin des spécialistes du trafic routier (laboratoire GRECIA de l'INRETS). Ce sujet de recherche m'a d'une part servi de pierre de touche pour mesurer la validité des méthodes d'estimation que j'ai développées, et d'autre part, a été à la source de certaines problématiques nouvelles que j'ai été contraint de résoudre. Pour cela, j'ai dégagé un nombre certes restreint de comportements caractéristiques pouvant néanmoins décrire avec précision les événements observés. Ce travail, présenté à la partie 1.4 met ainsi en jeu des méthodes de choix de modèles appliquées à la classification de données fonctionnelles. J'ai poursuivi l'amélioration de cette procédure de classification en utilisant dans un second temps un estimateur de maximum de vraisemblance non paramétrique appliqué aux mélanges de lois de probabilités. Bien entendu, de nombreuses applications sont encore en cours et font l'objet d'un dépôt de brevet, qui permettra une commercialisation de l'outil de prévision que j'aurais ainsi contribué à développer.

Un inconvénient majeur des hypothèses, propres aux techniques d'estimation décrites précédemment, provient du fait que l'on impose aux fonctions que l'on souhaite estimer, d'appartenir à un ensemble de régularité, portant sur sa régularité (espaces de Hölder, ...) ou celle de ses dérivées (espaces de Sobolev, de Besov, ...) par exemple. Or, dans la vie réelle, la régularité des fonctions n'a pas de sens de façon globale mais uniquement de façon locale. C'est ce formalisme, créé par S. Jaffard et appelé formalisme multifractal, que j'ai adapté dans un cadre statistique et que je présente au chapitre 2. Tout d'abord, la partie 2.1 vise à reconstruire une fonction multifractale observée dans un modèle de régression en présence d'un bruit d'observation. Comme les méthodes usuelles d'estimation ne sont pas applicables, je propose un estimateur Bayésien faisant appel à une loi a priori dépendant du modèle choisi pour représenter les fonctions multifractales, en particulier dépendant de deux paramètres, un paramètre d'intensité et un autre de lacunarité. Ensuite, dans un second temps, j'ai construit, dans la partie 2.2, un estimateur de ces paramètres et étudié les propriétés LAN de ce modèle dans la partie 2.3. Dès lors, j'ai pu conclure que la méthode d'estimation proposée était optimale. J'ai en outre construit un test permettant de vérifier si des données observées en présence d'un bruit d'observation proviennent d'un modèle multifractal. Il est possible d'utiliser le résultat de cette recherche pour estimer un grand nombre de fonctions très irrégulières rencontrées en physique (étude de turbulences ou d'écoulements de fluides) ou en économie. Le modèle gaussien reste néanmoins à la base de cette modélisation, mais cette hypothèse est irréaliste dans de nombreux cas, en particulier en finance. C'est la raison qui m'a amené à généraliser ma méthode d'estimation aux processus de Lévy fractionnaires (RHFLM). Dans la partie 2.4, je présente un estimateur de l'exposant de Hurst d'un tel processus, lorsque le signal est observé dans une situation concrète, en présence d'un bruit d'observation.

Jusqu'à ce stade de ma recherche, mon travail a porté sur l'étude de signaux observés de manière directe, or dans la plupart des problèmes concrets, le paramètre que l'on cherche à

estimer est rarement observé directement, mais au travers de son image par un opérateur. L'estimation s'avère alors plus difficile compte tenu de l'impossibilité d'inverser cet opérateur pour accéder aux observations de la fonction à reconstruire. Tout le travail du statisticien consiste alors à remplacer cette inversion par une méthode de régularisation qui puisse construire, à partir des observations indirectes, une approximation de la vraie fonction.

J'ai donc mis à profit mes connaissances en M-estimation pénalisée pour appliquer ces méthodes au cadre proposé par les problèmes inverses. J'ai obtenu une méthodologie d'estimation pour des problèmes inverses linéaires, mais aussi dans le cadre novateur des problèmes inverses non linéaires, présentée dans la partie 3.2. Les résultats sont exprimés sous la forme d'inégalités oracles, concept développé par Donoho et Johnstone dans [40]. Le principe n'est plus d'obtenir le meilleur estimateur pour une classe de régularité, mais de montrer que l'estimateur choisi est aussi bon que tous les estimateurs parmi une famille d'estimateurs donnés. En particulier, pour un bon choix de modèles, nous sommes en mesure de prouver l'*adaptivité* de l'estimateur considéré mais l'aspect le plus intéressant des inégalités oracles vient de leur caractère non asymptotique. Ainsi ce type de résultat est naturellement plus facile à appliquer en pratique que les propriétés d'adaptivité.

M'inscrivant dans cette problématique, je présente dans la partie 3.2 deux méthodes générales d'estimation pour les problèmes inverses linéaires et non linéaires. La première méthode est une version améliorée de la régularisation de Tikhonov tandis que la deuxième met à profit directement les techniques de sélection de modèles. Dans les 2 cas nous obtenons des inégalités oracles ainsi qu'une méthodologie pratique d'estimation. Je tiens à souligner que la non linéarité de l'opérateur n'entraîne pas de dégradation de la vitesse de convergence, à condition de supposer une certaine stabilité de la différentielle de l'opérateur dans un voisinage de la vraie solution.

J'ai néanmoins été amené à faire l'hypothèse, comme divers auteurs, que l'opérateur (ou à défaut sa décomposition en valeurs singulières) était connu. Or, dans les faits, l'opérateur est inconnu. J'ai donc, pour ma démarche, pris deux exemples concrets de problèmes inverses non linéaires que je me suis attaché à résoudre : l'étude de la solution d'une équation différentielle en présence de variables endogènes, problème issu de l'économétrie, et l'estimation de la nature d'un sol à partir du temps de parcours d'ondes dans ce milieu, situation rencontrée en tomographie sismique. Les résultats précédents ne s'appliquant plus, je me suis orienté vers une approche différente.

Dans le premier cas, dans la partie 3.3.1, je me suis tourné vers une méthode de résolution dans laquelle je tire profit de la forme particulière de l'opérateur pour l'estimer au moyen d'un opérateur à noyaux. Puis un estimateur de type Tikhonov m'a permis de généraliser les résultats de Florens et al. in [31] à ce cas particulier.

Le second exemple a été traité dans la partie 3.3.2. La généralité du problème m'a amené, ici, à faire appel aux techniques bayésiennes, développées par Csiszar dans [27] et popularisées par Gamboa et Gassiat [55]. J'ai utilisé ainsi l'estimateur du maximum d'entropie pour résoudre avec succès le problème posé.

Au cours de mes recherches, tant théoriques que pratiques, aussi bien théoriques que pratiques, sont apparues différentes manières d'exprimer la régularité des solutions des problèmes inverses. Ces différentes hypothèses sont propres aux domaines auxquels appartiennent les divers problèmes étudiés : statistique mathématique, économétrie et enfin analyse numérique. C'est la raison pour laquelle, j'ai cherché, dans la partie 3.4, à établir des relations entre ces divers jeux d'hypothèses pour pouvoir mieux comparer les méthodes d'estimation propres à chaque cas particulier.

Le dernier chapitre, chapitre 4, constitue un exemple de la démarche scientifique qui a animé la majeure partie de mon activité de recherche. L'étude de données fonctionnelles, en particulier les courbes d'évolution journalière des vitesses de véhicules, a fait émerger un problème concret fondamental : comment parvenir à extraire d'un ensemble de courbes une fonction représentative du phénomène étudié ? Notre problème s'étend à un grand nombre de cas réels, dans les domaines les plus variés comme le traitement du signal ou la biologie . . .

J'ai d'abord entrepris de construire un nouveau cadre dans lequel puissent s'inscrire les questions soulevées par cette problématique concrète. Dès lors, notre problème s'apparente à un problème inverse d'un nouveau type dans lequel la courbe inconnue n'est observée qu'au travers de son image par un groupe de déformations. A l'inverse des méthodes usuelles d'alignement de courbes, j'ai tiré partie de ce nouveau formalisme pour imaginer une nouvelle méthodologie semi-paramétrique et reconstruire plus efficacement la fonction inconnue. La partie 4.1 présente cette méthode d'estimation lorsque les déformations sont des translations déterministes. Elle utilise la transformée de Fourier et les propriétés des M-estimateurs pour faire coïncider les courbes sans estimer au préalable le motif à la base des déformations. Je suis parvenu à démontrer l'optimalité de la méthode proposée.

Dans la partie 4.2, j'étends ces résultats lorsque les paramètres des translations sont aléatoires. Je suis alors à même d'estimer leur loi. Ce modèle s'apparente à un modèle à effets mixtes aléatoires utilisé en pharmacocinétique pour modéliser la vitesse d'assimilation d'un médicament par un organisme vivant. Le point de vue adopté me permet de proposer un estimateur non paramétrique, présenté également dans la partie 4.2, et ainsi de s'affranchir des restrictions des modèles paramétriques, majoritairement employés en pharmacocinétique. Je tiens à souligner que cette recherche a été menée en liens étroits avec des chercheurs en biologie de l'Ecole Vétérinaire de Toulouse et des chercheurs en médecine au sein de l'équipe de recherche MONOLIX. Un prolongement naturel de ce travail est donné par l'étude des déformations quelconques. Dans la partie 4.3, je propose une méthode de reconstruction de la courbe originelle soumise à un processus quelconque de déformations aléatoires. Cette technique définit une nouvelle façon d'appréhender les données fonctionnelles grâce à un nouvel outil statistique nommé *l'espérance structurelle*. Ce travail, qui s'applique à des courbes, peut être étendu en dimensions supérieures, notamment pour l'étude d'images et trouve une continuation naturelle dans l'encadrement de 2 thèses sur cette problématique.

J'ai donc été amené à aborder les problèmes de classification de formes. Afin de nous affranchir des déformations d'un contour qui peuvent être observées dans une image, j'ai quotienté les images par le groupe des déformations. Classifier des images revient dès lors à classer

des classes d'équivalence qui présentent une structure de variété Riemannienne. Je résume ce travail dans la partie 4.4. Finalement, le lien entre les problèmes de classification et les modèles à effets mixtes est fourni par les modèles de réduction de dimension pour les modèles additifs généralisés, décrits dans [65] ou [116]. Je présente dans la partie 4.5 une méthodologie de régression inverse à noyaux appliquée à cette problématique.

“Ce qui se conçoit bien, s'énonce clairement et les mots pour le dire viennent aisément...” affirmait Boileau, mais il faut bien l'avouer, c'est là une rude gageure quand on entre dans le domaine de la recherche mathématique. J'ai néanmoins visé dans ce mémoire à présenter mon travail de recherche au cours de ces dernières années en mettant en lumière la démarche scientifique qui m'a animé, et en dégagant l'orientation générale qui m'a guidé dans les divers travaux que j'ai réalisés, tant sur le plan théorique que pratique. En cheminant au travers de nombreux problèmes, j'ai bien entendu utilisé des méthodes en statistique non paramétrique, déjà employées par de nombreux mathématiciens, mais j'ai également développé des outils qui me sont propres et qui sont décrits dans ce mémoire. Comme la recherche ne relève pas de la *génération spontanée*, je manquerais à mon devoir le plus élémentaire si je n'exprimais pas ma très profonde gratitude à tous ceux qui m'ont mis sur la voie de la recherche en statistique, qui m'ont aidé et accompagné dans la réalisation de mes travaux scientifiques. Il va de soi que toute recherche reste inachevée et ouvre la voie à bien des pistes et des interrogations. La route vers une meilleure compréhension du monde est encore longue, elle n'est pas tracée d'avance et ne peut se faire qu'en avançant. C'est le sens que je donnerai à la phrase célèbre du poète Antonio Machado “Caminante, no hay camino, el camino se hace al andar ”.

# Chapitre 1

## Régularisation par complexité

### 1.1 M-estimation pénalisée

Les M-estimateurs pénalisés jouent un rôle clef en statistique non paramétrique. Définis comme réalisant le minimum d'une fonction de perte sous certaines contraintes, ils ont été étudiés d'un point de vue théorique par de nombreux auteurs et trouvent généralement des applications dans les domaines les plus divers tels que la physique ou encore l'économie. Ils occupent une place importante dans nos travaux de recherche.

Nous nous sommes tout d'abord intéressés aux liens entre les M-estimateurs pénalisés et les méthodes d'estimation par seuillage d'une part, et par sélection de modèles d'autre part. Nous avons étudié et construit des variantes de ces estimateurs possédant des propriétés d'adaptation, c'est-à-dire permettant d'estimer une fonction à une vitesse optimale sans connaître a priori sa régularité. Ainsi, nous avons montré qu'une pénalité  $l^1$  permet d'obtenir une telle propriété, en permettant de ne sélectionner que les coefficients représentatifs de la fonction à estimer.

Nous avons poursuivi notre étude en construisant des estimateurs de ce type afin d'apporter des solutions à de nombreux problèmes rencontrés dans la pratique. En effet, ces estimateurs présentent plusieurs avantages. Tout d'abord, décrits comme la solution d'un problème variationnel, leur interprétation est aisée; ensuite, il est facile de les approcher par des méthodes numériques de minimisation. Enfin, ils mettent clairement en évidence le rôle fondamental que joue l'espace sur lequel a lieu la minimisation, notamment la complexité de cet espace. Cette complexité peut s'exprimer en terme de dimension de l'espace mais aussi en prenant en compte la régularité des fonctions le composant. Dans ce cas, l'entropie métrique est une mesure de sa complexité, cf par exemple [86].

Plus précisément, soient  $X_1, \dots, X_n$  des réalisations de variables aléatoires indépendantes à valeur dans un espace mesurable  $\mathcal{X}$ . La loi de  $X_i$  est notée  $P_i$ ,  $i = 1, \dots, n$ . Définissons  $\bar{P} = \frac{1}{n} \sum_{i=1}^n P_i$  et soit  $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  la distribution empirique associée. On suppose que la loi des observations dépend d'un paramètre d'intérêt (un paramètre dans des cas particuliers ou plus généralement d'une fonction),  $f_0$ , appartenant à un espace métrique  $\mathcal{F}$  muni de la distance  $d$ .

Notre objectif vise à estimer la fonction  $f_0$  à partir des observations  $X_i$ ,  $i = 1, \dots, n$ . Pour cela, considérons une fonction de perte  $\gamma_f : \mathcal{X} \rightarrow \mathbb{R}$  pour toute fonction  $f \in \mathcal{F}$ . La fonction de perte est choisie en fonction de la nature du problème et mesure l'adéquation entre l'estimateur que l'on cherche et les données. Ainsi, dans un modèle de régression, la fonction de perte naturelle est la norme  $l^p$  par rapport à la mesure empirique, avec  $p$  bien choisi. Dans un cadre d'estimation de densités, la distance naturelle est certes la distance de Küllback, mais d'autres choix sont envisageables. Nous nous référons sur ce point à [11] pour une étude générale des M-estimateurs. Ainsi, nous définissons l'estimateur de minimum de contraste comme la solution du problème de minimisation suivant

$$\hat{f}_{0,n} = \arg \min_{f \in \mathcal{F}} \left\{ \int \gamma_f dP_n \right\} \quad (1.1)$$

La vraie fonction, quant à elle, est obtenue à partir de la formule suivante

$$f_0 = \arg \min_{f \in \mathcal{F}} \left\{ \int \gamma_f d\bar{P} \right\}.$$

Chercher un estimateur proche des données donne souvent naissance à un estimateur trop irrégulier, qui peut même ne pas être convergent. Ainsi il est nécessaire d'ajouter des contraintes à cette minimisation. Pour cela considérons une pénalité  $I : \mathcal{F} \rightarrow \mathbb{R}^+$  qui privilégie, à l'intérieur de l'espace  $\mathcal{F}$ , les fonctions les plus régulières. Pour une suite de réels  $\lambda_n^2$  tendant vers zéro, définissons alors l'estimateur régularisé comme

$$\hat{f}_{1,n} = \arg \min_{f \in \mathcal{F}} \left\{ \int \gamma_f dP_n + \lambda_n^2 I(f) \right\}. \quad (1.2)$$

Dans certains cas, supposer que l'on connaît l'espace  $\mathcal{F}$  n'est pas une hypothèse raisonnable. Cet espace étant inconnu, il doit donc être approché. On se donne alors, une collection d'ensembles  $\mathcal{F}_m$ ,  $m \in \mathcal{M}_n$ , dont on connaît la dimension notée  $d_m$ . En pratique, on fait souvent appel à une famille orthonormale  $\psi_k$ ,  $k \geq 1$ , et on définit les espaces au moyen de cette base de la manière suivante  $\mathcal{F}_m = \{\psi_k, k \in \mathcal{K}_m, |\mathcal{K}_m| = d_m\}$ . D'autres choix sont néanmoins possibles. L'estimateur est alors obtenu en choisissant au moyen d'une pénalité notée  $\text{pen}(\mathcal{F}_m)$  et portant sur la dimension de l'espace, le meilleur espace d'approximation  $\mathcal{F}_{\hat{m}}$ , puis le meilleur estimateur  $\hat{f}_{\hat{m}}$ , à l'intérieur de cet espace. Plus précisément, pour toute suite de réels  $\alpha_n$ , l'estimateur est obtenu en minimisant le critère suivant

$$(\hat{f}_{\hat{m}}, \hat{m}) = \arg \min_{(f,m) \in \mathcal{F}_m \times \mathcal{M}_n} \left\{ \int \gamma_f dP_n + \lambda_n^2 I_{\mathcal{F}_m}(f) + \alpha_n^2 \text{pen}(\mathcal{F}_m) \right\}. \quad (1.3)$$

Cette forme d'écriture permet de replacer la plupart des estimateurs utilisés en statistique dans le cadre des M-estimateurs pénalisés. En particulier, lorsque la pénalité ne porte que sur la dimension de l'espace, nous retrouvons, pour une fonction de perte quadratique, les estimateurs de sélection de modèles étudiés par Birgé et Massart et décrits dans [14, 15]. Pour un espace de régularité choisi, nous retrouvons les estimateurs régularisés étudiés par Silverman in [115] ou van de Geer in [120].

Le comportement asymptotique de l'estimateur défini par la minimisation (1.3) est déterminé en grande partie par le choix des paramètres de régularisation  $\alpha_n$  et  $\lambda_n$ . Plus  $\lambda_n$  sera grand et plus la solution sera régulière au sens donné par la pénalité  $I$ , tandis que le choix d'un paramètre petit permettra à l'estimateur d'être plus proche des données. De façon analogue, plus  $\alpha_n$  sera grand et moins les grands espaces d'approximation seront choisis, ce qui privilégie l'erreur de variance au détriment de l'erreur de biais. Pour un choix d'un faible  $\alpha_n$ , l'espace choisi sera celui le plus important.

Le choix optimal des paramètres dépend de la régularité de la fonction  $f_0$ , or cette régularité nous est inconnue. Dans ce contexte, notre démarche utilise la complexité des espaces  $\mathcal{F}$  afin de sélectionner automatiquement les meilleurs paramètres  $\alpha_n$  et  $\lambda_n$  de telle manière que l'estimateur sélectionné ait l'erreur la plus faible parmi tous les estimateurs obtenus à partir de tous les choix possibles de paramètres. C'est le sens des inégalités de type oracle que nous cherchons à prouver dans nos travaux. En outre, si le vrai modèle appartient à la collection d'espaces d'approximation choisie, alors l'estimateur convergera à la vitesse optimale. Un tel estimateur sera dit adaptatif puisqu'il converge à la vitesse minimax, sans connaître a priori sa régularité. L'adaptativité d'une procédure est une manière d'évaluer son optimalité. Dans la littérature en statistique non paramétrique, les auteurs montrent l'adaptativité d'un grand nombre d'estimateurs, par exemple dans [40], [62] ou [14]. Au cours de nos recherches, nous nous sommes intéressés à la construction de M-estimateurs adaptatifs pour le modèle de régression, d'estimation de densités ou pour la résolution de problèmes inverses. Nous donnerons plus de précisions sur ses résultats dans la partie 1.2.

Il est vrai que le critère minimax est une approche imparfaite. D'une part, elle nécessite de choisir un espace de régularité pour les fonctions à estimer, ce choix induisant un biais dans la procédure de test. D'autre part, elle privilégie l'optimalité de l'estimateur dans la pire situation possible. Par conséquent, ce point de vue est très pessimiste et peut amener à privilégier des estimateurs dont la performance est moyenne mais constante, au détriment d'estimateurs beaucoup plus efficaces pour la plupart des cas, mais moins efficaces en un point très particulier de l'ensemble test. Ce constat a été établi dans [75] ou [25]. Ces auteurs proposent l'alternative suivante : pour un estimateur donné  $\hat{f}_n$  et une vitesse de convergence particulière  $r_n(s)$ , on cherche le plus grand espace de fonctions  $\mathcal{M}(\hat{f}_n, r_n(s))$  qui peuvent être estimées par l'estimateur choisi à la vitesse de convergence considérée. Une telle étude générale permet de comparer les différentes procédures d'estimation en termes de maxisets. Plus le maxiset est important et plus la procédure sera optimale puisqu'elle permettra de faire face à un plus grand nombre de situations. En outre cette approche permet de répondre aux questions que se posent des utilisateurs pratiques cherchant à savoir quelle méthode choisir dans une situation particulière.

Voilà pourquoi, nous avons essayé de comparer les méthodes de sélection de modèles utilisées pour estimer des fonctions dans un espace de régularité aux méthodes de pénalisation  $l^1$  développées dans ma thèse d'une part et, d'autre part aux méthodes de seuillage par ondelettes. Dans l'article [4], nous nous intéressons aux estimateurs de sélection de modèles sous leur forme générale, obtenus par Birgé et Massart, et parvenons au résultat suivant

**Théorème 1.1.1** (Maxisets pour estimateur de sélection de modèles).

Soit  $\alpha > 0$  et

$$\text{pen}(m) = KD_m/n,$$

pour  $K > (1 + \sqrt{2\kappa})^2$ . Si  $\hat{s}_{\hat{m}}$  est l'estimateur classique de sélection de modèles, défini par

$$\hat{s}_{\hat{m}} = \arg \min_{m \in \mathcal{M}_n, t \in S_m} (\gamma_n(t) + \text{pen}(m, t)), \quad (1.4)$$

pour une pénalité  $\text{pen}(t, m) = \text{pen}(m) = (1 + c) \frac{D_m L_m}{n} \sigma^2$ , alors le maxiset associé à la vitesse de convergence  $n^{-2a/(2a+1)}$  est donné par

$$\mathcal{M}(\hat{s}_{\hat{m}}, n^{-2a/(2a+1)}) = B_{2, \infty}^\alpha.$$

Ce théorème permet de répondre à la question : quel estimateur choisir ? Nous pouvons dès lors affirmer que les méthodes d'estimation par ondelettes et les méthodes de sélection de modèles ont des maxisets similaires pour les mêmes vitesses. Ce résultat est une avancée pour une meilleure compréhension des liens entre les diverses procédures d'estimation en statistique non paramétrique.

La démonstration de ce résultat s'inscrit dans le cadre des preuves classiques sur les maxisets, puisqu'elle repose sur une double inclusion.

1. Montrer que les fonctions de l'espace candidat peuvent être estimées à la bonne vitesse. Cette inclusion découle des résultats des estimateurs de sélection de modèles, démontrés par Baraud en régression [5, 6] et par Birgé et Massart en estimation de densités [8, 14] par exemple.
2. Montrer que les fonctions estimées à la bonne vitesse possèdent au moins la régularité donnée par l'espace candidat. En ce cas, cette inclusion directe demande plus de travail, et repose sur le lemme suivant :

**Lemma 1.1.2.** *Si les modèles sont en somme directe de modèles emboîtés, alors*

$$\|s - \hat{s}_{\hat{m}}\|^2 \geq \|s - s_{m_0}\|^2, \quad (1.5)$$

où  $m_0(n)$  est défini comme solution de

$$\arg \min_{m \in \mathcal{M}_K} \left\{ \|y - \hat{s}_m\|^2 + \frac{1}{4} \text{pen}_n(m) \right\}.$$

Supposons que  $\mathcal{M}_n \supset \mathcal{M}_{n/2} \supset \{0\}$ , et considérons un choix de pénalité  $\text{pen}_n(m) = \frac{K_n}{n} D_m$  avec  $0 < c \leq K_n$  et  $1 \leq K_n/K_{n/2^k} \leq ck$ . Si  $\forall n, \|s - s_{m_0}\|^2 \leq V(n) \leq C \left(\frac{K_n}{n}\right)^\beta$  avec  $0 < \beta < 1$ , alors nous aurons l'inégalité suivante

$$\|s - s_{m_0}\|^2 + \frac{1}{4} \text{pen}_n(m_0) \leq (1 + C_\beta) C \left(\frac{K_n}{n}\right)^\beta + \|s\|^2 \frac{K_n}{cn} \quad (1.6)$$

L'idée à la base de la démonstration repose sur l'inclusion des sous espaces modèles et sur les liens entre les techniques de seuillage et celles par pénalités. Ainsi l'inégalité (1.5) garantit que si l'estimateur converge à la vitesse  $V(n)$  alors l'oracle (donné par le choix du modèle  $m_0$ ) converge aussi à la même vitesse. L'inégalité (1.6) nous apporte ensuite le contrôle de la pénalité, permettant de conclure la démonstration.

## 1.2 Utilisation de la norme $l^1$

Dans nos travaux, nous nous sommes attachés très précisément à une forme bien spécifique de pénalité : la norme  $l^1$  des coefficients d'une fonction décomposée dans une bonne base. De nombreux auteurs ont étudié les différentes propriétés de cette norme, et en particulier, le fait qu'elle permet de sélectionner les représentations les plus creuses (sparse) parmi les différentes représentations d'une fonction dans des bases différentes. L'avantage de telles représentations en estimation statistique est détaillée dans [43]. Ajouter une pénalité  $l^1$  permet de restreindre le signal aux représentations formées d'un petit nombre de coefficients représentatifs et un grand nombre de coefficients nuls. Nous obtenons ainsi une procédure générale d'estimation adaptative, applicable dans un grand nombre de situations.

Dans un premier temps, dans le cadre de la régression, nous avons étudié l'adaptivité d'un estimateur construit en minimisant une fonction de perte  $\gamma(\cdot) = \|\cdot\|_p$  pour  $p \in \{1, 2\}$ , et pour pénalité  $I(\cdot)$ , la norme  $l^1$ . Pour le modèle de régression suivant

$$Y_i = \theta_0(z_i) + W_i, \quad i = 1, \dots, n$$

où  $\theta_0 \in \Theta \mathbb{R} \rightarrow \mathbb{R}$  est une fonction inconnue à reconstruire, notre étude a porté sur l'estimateur défini par

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \left[ \frac{1}{n} \sum_{i=1}^n \gamma(Y_i - \theta(z_i)) + \lambda_n^2 I_n(\theta) \right].$$

Ce type de procédure est une sélection de modèles privilégiant les modèles les plus réguliers, au sens donné par la régularité issue de la pénalité  $I(\cdot)$ . L'équilibre entre les deux termes est donné par le paramètre de lissage  $\lambda_n$ . Plus ce paramètre est faible, plus l'estimateur sera proche des données, proche au sens donné par le contraste  $\gamma$ . Au contraire de grandes valeurs de  $\lambda$  entraîneront un estimateur très régulier. Pour le choix particulier d'une pénalité  $l^1$ , nous définissons l'estimateur comme

$$\begin{aligned} \hat{\theta}_n &= \arg \min_{\theta = \alpha_1 \psi_1 + \dots + \alpha_n \psi_n} \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \theta(z_i))^2 + 2\lambda_n^2 \sum_{j=1}^n |\alpha_j| \right] \\ &= \sum_{j=1}^n \hat{\alpha}_{j,n} \psi_j. \end{aligned}$$

Nous démontrons que cet estimateur, apparenté aux estimateurs à seuillage, est adaptatif. En effet, sans connaître a priori la régularité, l'estimateur converge à la vitesse optimale qui dépend de la régularité inconnue de la fonction  $\theta_0$ . Très concrètement, nous avons prouvé le théorème suivant.

**Théorème 1.2.1** (Adaptivité de l'estimateur pénalisé  $l^1$ ).

S'il existe une constante  $K < \infty$ , telle que les erreurs  $W_1, \dots, W_n$  vérifient

$$\max_{i=1, \dots, n} \mathbf{E} \exp[W_i^2/K^2] \leq K,$$

alors

$$\mathbf{P} \left( \max_{j=1, \dots, n} |V_j| > c \sqrt{\frac{\log n}{n}} \right) \leq c \exp\left[-\frac{\log n}{c^2}\right].$$

Par conséquent pour un choix  $\lambda_n^2 = c\sqrt{\log n/n}$ , et pour  $\alpha_*$  l'estimateur oracle (le meilleur connaissant la bonne régularité), nous avons

$$\mathbf{P}(\|\hat{\alpha}_n - \alpha_0\|_n^2 > 4(\|\alpha_* - \alpha_0\|_n^2 + 4\lambda_n^4 N_n)) \leq c \exp\left[-\frac{\log n}{c^2}\right].$$

Ainsi, pour de grandes valeurs de  $n$ , nous aurons avec grande probabilité

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \leq 4(\|\alpha_* - \alpha_0\|_n^2 + 4\lambda_n^4 N_n).$$

L'estimateur adaptatif que nous proposons possède une erreur comparable à celle de l'oracle, à un terme d'erreur près. Nous pouvons remarquer que la constante  $c$  dépend dans ce cas de la variance des erreurs d'observation.

La généralité de ce théorème nous permet de l'étendre à d'autres estimateurs, notamment une version robuste de cet estimateur définie par

$$\hat{\theta}_n = \min_{\theta \in \Theta, \theta = \sum_{j=1}^n \alpha_j \psi_j} \left[ \frac{1}{n} \sum_{i=1}^n |Y_i - \theta(z_i)| + \lambda_n^2 \sum_{j=1}^n |\alpha_j| \right].$$

L'avantage d'un tel estimateur réside de ce qu'il ne dépend pas de la variance des erreurs. La constante  $c$  est alors une constante universelle. Cet estimateur peut donc être appliqué dans la pratique sous sa forme optimale. Dans ce cas, nous avons le même type de résultat.

En particulier, nous obtenons une version robuste et adaptive d'un estimateur pénalisé de régression, qui améliore les résultats proposés dans la littérature [62] par exemple, puisque les paramètres sont choisis par validation croisée.

**Théorème 1.2.2** (Adaptivité de l'estimateur robuste pénalisé).

Pour un choix  $\lambda_n = c\sqrt{\log n/n}$ , nous avons

$$\mathbf{P} \left( \|\hat{\theta}_n - \theta_0\|_{Q_n} \geq c \left( \frac{\log n}{n} \right)^{\frac{2-\rho}{4}} \right) \leq c \exp\left[-\frac{\log n}{c^2}\right].$$

Ce théorème nous permet de retrouver la vitesse de convergence en  $n^{-\frac{2s}{2s+1}}$  optimale. En effet si  $\theta_0$  appartient à un espace d'irrégularité donné par un paramètre  $\rho$  et tel que  $\sum_{j=1}^n \theta_{j,0}^\rho \leq 1$ . Alors pour un choix  $\rho = \frac{2}{2s+1}$ , nous obtenons la bonne vitesse de convergence. Ces résultats ont pu être démontrés dans [99].

La démonstration fait appel à des inégalités de concentration pour les processus empiriques et à l'écriture de l'erreur d'estimation sous la forme suivante. Pour un ensemble d'indices  $\mathcal{J}_n$  de cardinal  $N_n$ , il convient de remarquer que la pénalité  $l^1$  peut se décomposer en 2 termes  $I_N(\cdot)$  et  $I_M(\cdot)$  définis comme

$$\|f\|_1 = \sum_{j=1}^n |\alpha_j| = \sum_{j \in \mathcal{J}_n} |\alpha_j| + \sum_{j \notin \mathcal{J}_n} |\alpha_j| := I_N(f) + I_M(f).$$

Soit  $f^\star = \sum_j \alpha_{j,\star} \psi_j$  un oracle (le meilleur estimateur choisi en connaissant la régularité par exemple). C'est ainsi qu'en utilisant l'inégalité suivante

$$\left| \sum_{j \in \mathcal{J}_n} |\alpha_{j,\star}| - \sum_{j \in \mathcal{J}_n} |\hat{\alpha}_j| \right| \leq I_N(\hat{f}_n - f_\star),$$

nous obtenons :

### Inégalité fondamentale

$$cd^2(\hat{f}_n, f_0) \leq d^2(f_\star, f_0) + (1 + C)\lambda_n^2 I_N(f_\star - \hat{f}_n) + 2\lambda_n^2 I_M(f_\star). \quad (1.7)$$

Cette décomposition permet de mieux comprendre le phénomène d'adaptation lié à la norme  $l^1$ , pour un bon choix d'indices  $\mathcal{J}_n$ . Nous pouvons nous rendre compte que la distance entre l'estimateur et la fonction à reconstruire est bornée par un terme d'approximation  $d^2(f_\star, f_0)$ , le biais de l'oracle, et deux termes d'approximation  $I_N(f_\star - \hat{f}_n)$  et  $I_M(f_\star)$ . Le premier terme représente l'erreur d'approximation pour la partie de la fonction dont les coefficients appartiennent à l'ensemble  $\mathcal{J}_n$ , tandis que le second terme décrit l'erreur de la partie de l'oracle dont les coefficients appartiennent à l'ensemble  $\mathcal{J}_n^c$ . Ces deux termes sont en équilibre et dépendent de la taille de  $\mathcal{J}_n$ . Un bon choix de cet ensemble permet de préciser les vitesses de convergence en fonction de la décroissance des coefficients de la vraie fonction dans une bonne base. Ces estimateurs sont étudiés dans [114]. Nous nous sommes attachés à construire une procédure adaptative à partir d'un tel estimateur.

En estimation de densités, de telles méthodes permettent de trouver un moyen de définir un maximum de vraisemblance pénalisé, présentant aussi des propriétés d'adaptation. Si on observe  $n$  variables aléatoires i.i.d de densité  $f_0$ , considérons l'estimateur suivant pour une pénalité  $I(f)$

$$\hat{f}_n = \arg \max_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^n \log f(X_i) - \lambda_n^2 I(f) \right), \quad (1.8)$$

A cet effet, reparamétrisons le problème en utilisant une base de logsplines (cf par exemple [7]), en posant

$$\gamma_0 = \log(f_0) + b(\gamma_0),$$

avec  $b(\gamma_0) = - \int \log(f_0) d\mathbf{P}$ . Ainsi, à toute fonction  $f \in \mathcal{F}$ , nous associons une variable  $\gamma \in \Gamma$  telle que  $\gamma = \log(f) + b(\gamma)$ . Soit  $(\psi_{jk})_{j \geq 0, k = 0, \dots, 2^j - 1}$  une base d'ondelettes à support compact de  $\Gamma \subset B_{p_\infty}^s$ , de régularité  $r > s$ . Ainsi, il est possible de décomposer la log-densité

sur la base d'ondelettes. Notons  $\gamma = \sum_{j,k} \beta_{jk} \psi_{jk}$  et soit  $j_1(n)$  un niveau de résolution. Nous choisirons comme pénalité la norme  $l^1$  des coefficients d'ondelettes,

$$J(\gamma) = \sum_{j,k} |\beta_{jk}|.$$

L'estimateur que nous examinons est ainsi défini par

$$\hat{\gamma}_n = \arg \max_{\gamma = \sum_{j < j_1} \sum_{k=0}^{2^j-1} \beta_{jk} \psi_{jk} \in \Gamma} \left( \frac{1}{n} \sum_{i=1}^n \gamma(X_i) - b(\gamma) - \lambda_n^2 \sum_{j < j_1} \sum_k |\beta_{jk}| \right).$$

Cet estimateur est une version adaptative des estimateurs régularisés étudiés dans [60], [113] ou [114] pour lesquels la fonction de pénalité est une pénalité quadratique, définissant un estimateur non adaptatif.

**Théorème 1.2.3** (Adaptivité de l'estimateur pénalisé de densités).

*Supposons que  $\exists 0 < C < \infty$ ,  $\sup_{\gamma \in \Gamma} |\gamma| \leq C$ . Pour  $j_1$  tel que  $2^{j_1} = o\left(\frac{n}{\log n}\right)$ , et un choix de la suite régularisante indépendante des données  $\lambda_n^2 \geq c\sqrt{\frac{\log n}{n}}$ , l'estimateur pénalisé de la densité  $\gamma_0 \in B_{p^\infty}^s([0, 1])$  vérifie alors*

$$\|\hat{\gamma}_n - \gamma_0\|^2 = O_P \left( \frac{n}{\log n} \right)^{-\frac{2s}{2s+1}}.$$

La démonstration de ces résultats participe des mêmes techniques que celles que nous avons utilisées dans le cas de l'estimateur par régression. Elles sont décrites dans [99] ou [88].

Pénaliser une fonction de perte en utilisant une norme  $l^1$  permet, d'une part, d'obtenir des estimateurs adaptatifs qui reproduisent les techniques de seuillage doux (soft-thresholding), bien connus pour les estimateurs obtenus par projection sur des bases d'ondelettes décrits par exemple dans [64]. D'autre part, nous avons pu construire des estimateurs possédant des propriétés de robustesse, dépendant moins des erreurs d'observation, en particulier ne dépendant pas de la variance des erreurs

Nous pouvons donc étendre ces résultats au cas des problèmes inverses linéaires. Le cas où la fonction est observée au travers de son image par un opérateur linéaire mais également non linéaire est traité dans les articles [49, 89] et sera détaillé dans la partie 3.

Les propriétés de la norme  $l^1$  ont été mises à profit en sélection de variables, notamment pour expliquer un phénomène, à partir de variables explicatives. Ces techniques ont été développées par Hastie, Tibshirani dans l'article [45]. C'est donc naturellement que nous nous sommes intéressés aux techniques de sélection de modèles au moyen d'une norme  $l^1$ .

### 1.3 Sélection de variables

La pénalité avec une norme  $l^1$  a connu un succès croissant pour estimer des fonctions, mais aussi en sélection de modèles, notamment dans le cadre du choix de variables. Dans un modèle

Gaussien, l'algorithme LARS met à profit cette propriété pour sélectionner itérativement des variables explicatives. Nous essayons de prouver théoriquement les résultats algorithmiques et les conjectures de cet article en utilisant de nouveaux outils de concentration, notamment en prouvant une nouvelle inégalité de déviation pour des sommes de carrés de variables réordonnées. Ce travail a été publié dans [90] et fait partie intégrante de mes recherches actuelles. En particulier, prenons l'exemple du modèle gaussien suivant

$$y_i = \mu_i + \sigma \epsilon_i, \quad i = 1, \dots, n \quad (1.9)$$

avec  $\forall i = 1, \dots, n \quad \epsilon_i \sim \mathcal{N}(0, 1)$  i.i.d. Pour la base canonique  $(\phi_1, \dots, \phi_n)$  et l'échantillon réordonné

$$|y|_{(1)} \geq \dots \geq |y|_{(n)}.$$

Pour  $k$  donné, considérons l'estimateur par seuillage dur  $\tilde{\mu}(k) = \sum_{j=1}^k y_{(j)} \phi_{(j)}$  et par seuillage doux  $\hat{\mu}(k) = \sum_{j=1}^k \text{sgn}(y_{(j)}) (|y|_j - |y|_{(k+1)})_+ \phi_{(j)}$ .

Nous pouvons interpréter l'algorithme LARS comme un algorithme qui pour un niveau  $k$  minimise le critère suivant

$$\hat{k} = \arg \min_k C(k),$$

$$C(k) = \|y - \hat{\mu}(k)\|^2 + \text{pen}(k) = \|y - \tilde{\mu}(k)\|^2 + k|y|_{(k+1)}^2 + \text{pen}(k). \quad (1.10)$$

Alors, la conjecture LARS s'écrit

$$[\|y - \hat{\mu}(k)\|^2 - \|y\|^2 + 2k\sigma^2] \approx [\|\mu - \hat{\mu}(k)\|^2] - \|\mu\|^2.$$

Or nous pouvons prouver que cette égalité est vraie en moyenne, c'est-à-dire

$$\begin{aligned} \mathbf{E} [\|y - \hat{\mu}(k)\|^2 - \|y\|^2 + 2k\sigma^2] &= \mathbf{E} [\|\mu - \hat{\mu}(k)\|^2] - \|\mu\|^2 \\ \mathbf{E}[L] &= \mathbf{E}[R]. \end{aligned}$$

Par conséquent, ce jeu de réécriture permet de voir que la conjecture LARS pourrait être démontrée si l'on obtenait la concentration des termes précédents autour de leur espérance, à des vitesses de convergence similaires. C'est pourquoi, nous essayons par des théorèmes de concentration de prouver un résultat semblable sans l'espérance sur des ensembles de grandes probabilités. En ce cas, il nous faut parvenir à montrer des inégalités de déviation pour les quantités  $\sum_{i \leq k} |y|_{(i)}^2$  et  $|y|_{(k+1)}$ . Nous avons, à ce jour, démontré le théorème suivant

**Théorème 1.3.1** (Nouvelles inégalités de concentration).

$$\begin{aligned} \mathbf{P} \left( \sum_{i \leq k} |y|_{(i)}^2 - \mathbf{E} \left[ \sum_{i \leq k} |y|_{(i)}^2 \right] \geq cx + \sqrt{8x \mathbf{E} \left[ \sum_{i \leq k} |y|_{(i)}^2 \right]} \right) &\leq \exp(-x). \\ \mathbf{P} \left( |y|_{(i)} \geq Q^{(-1)} \left( \frac{i}{2n} \right) \left[ 1 + \frac{2\theta}{i} + \sqrt{\frac{2\theta}{i}} \right] \right) &\leq \exp(-\theta). \end{aligned}$$

Ce théorème se prouve en utilisant des techniques de concentration de type inégalités Log-Sobolev, décrites dans [83] et utilisées en statistique dans [102]. Plus précisément, nous avons prouvé que nous pouvons obtenir un nouveau contrôle précis sur la concentration de sommes de  $\chi^2$ , alors que, dans les publications actuelles, ces résultats ne sont obtenus que pour des fonctions lipschitziennes de variables aléatoires indépendantes. Les techniques de concentration de type inégalités sobolev logarithmique, développées dans [82] reposent sur l'obtention d'inégalités pour la transformée de Laplace de la distribution.

Pour cela, soit  $\xi \sim \mathcal{N}(0, Id_n)$  et  $A \in \mathcal{A}$  une matrice symétrique de projection appartenant à l'ensemble

$$\mathcal{A} = \{A \in \mathcal{M}_{n \times n}(\mathbb{R}), A^t A = A\}.$$

Considérons la fonction suivante

$$f(\xi) = \sup_{A \in \mathcal{A}} \xi^t A \xi.$$

Supposons qu'il existe  $A^* \in \mathcal{A}$ , réalisant le maximum de la fonction  $f$  sur  $\mathcal{A}$ , soit

$$\sup_A \xi^t A \xi = \xi^t A^* \xi.$$

Alors, nous aurons

$$\begin{aligned} \nabla f(\xi) &= 2A^* \xi \\ \|\nabla f(\xi)\|^2 &= 4\xi^t (A^*)^t A^* \xi = 4f(\xi). \end{aligned} \tag{1.11}$$

Si la variable  $\xi$  possède des propriétés de concentration de type Log-Sob, notons  $H$  l'entropie, nous aurons

$$H(e^{\lambda f}) \leq 2\lambda^2 \mathbf{E}(f e^{\lambda f}).$$

Soit  $F(\lambda) = \mathbf{E}(e^{\lambda f})$ . La concentration Log Sobolev entraîne que

$$H(e^{\lambda f}) \leq \frac{\lambda^2}{2} \mathbf{E}(\|\nabla\|_2^2 e^{\lambda f}).$$

En utilisant (1.11),

$$H(e^{\lambda f}) \leq 2\lambda^2 \mathbf{E}(f e^{\lambda f}), \tag{1.12}$$

nous aurons l'équation différentielle suivante :

$$\lambda F' - F \log(F) \leq 2\lambda^2 F'.$$

Alors nous obtenons

$$\lambda \tilde{F}' - \tilde{F} \log(\tilde{F}) \leq 2\lambda^2 \tilde{F}' + 2\lambda^2 \tilde{F} \mathbf{E}(f).$$

Finalement, cette équation s'écrit

$$\frac{\partial}{\partial \lambda} \left[ \frac{1}{\lambda} \log(F(\lambda)) \right] \leq 2\mathbf{E}(f).$$

En résolvant l'équation différentielle nous obtenons pour  $\tilde{F}$  :

$$\log \tilde{F}(\lambda) \leq \frac{2\lambda^2}{1-2\lambda} \mathbf{E}(f).$$

Finalement nous obtenons la majoration suivante

$$\log \tilde{F}(\lambda) \leq \frac{2\lambda^2}{1-2\lambda} \mathbf{E}(f), \quad (1.13)$$

à la base des inégalités de concentration. Par exemple, nous l'appliquons à  $n$  variables aléatoires indépendantes  $y_j \sim \mathcal{N}(0, 1)$ ,  $j = 1, \dots, n$  en définissant la fonction  $f(y) = \sum_{j \leq k} |y|_{(j)}^2$  pour obtenir le résultat (1.3.1).

Ce travail dont l'intérêt est fondamental pour une bonne compréhension des algorithmes fondés sur la perte  $l^1$  est encore en cours compte tenu des difficultés et des écueils théoriques qu'il comporte. Néanmoins, démontrer ce résultat est un des enjeux de la recherche actuelle en statistique dans ce domaine.

## 1.4 Sélection de modèles

Les méthodes par pénalités et leur étude théorique nous ont permis de découvrir les méthodes de choix de modèles. Ces méthodes sont tout particulièrement adaptées aux problèmes de clustering et de classification. En conséquence, nous avons utilisé ces techniques pour fournir des réponses aux problèmes rencontrés dans ces cas pratiques.

En effet, les techniques de classification permettent de scinder un échantillon en groupes représentatifs. Pour une distance donnée, nous construisons des groupes d'éléments tels qu'à l'intérieur d'un même groupe les données soient proches tandis que, pour 2 groupes différents, les données doivent être assez éloignées. Le problème principal réside alors dans le choix du nombre de groupes nécessaires pour obtenir une bonne représentation. Bonne représentation signifie réalisant un compromis entre précision de la description (i.e distance entre un élément et sa représentation) et généralité (taille des modèles choisis). A cet effet, nous retrouvons la problématique récurrente en sélection de modèles : construire des espaces d'approximation assez gros pour que l'erreur de biais soit faible, mais pas trop pour que l'erreur du terme de variance puisse être contrôlée. Ainsi, la théorie de la sélection de modèles fournit un cadre théorique aux méthodes de classification.

Nous avons appliqué ces méthodes dans un but pratique : prédire des temps de parcours à court et moyen terme sur un réseau routier. Plus précisément, l'enjeu est le suivant : nous observons en des points de comptage, répartis tous les 500 mètres et toutes les 3 minutes, la vitesse moyenne des véhicules parcourant le tronçon de route. Nous disposons de 2 années d'historique pour les 750 stations de comptage réparties sur le réseau routier d'Île de France. Nous noterons  $t \rightarrow V_j^s(t)$  la courbe d'évolution des vitesses de la journée  $j = 1, \dots, J$  pour la station de comptage  $s = 1, \dots, \mathcal{S}$ . La courbe est observée de 5 heures du matin à 23 heures, aux temps  $t_k$ ,  $k = 0, \dots, 180$ . Nous souhaitons construire une procédure permettant de prédire,

rapidement et en temps réel, le temps de parcours d'un usager souhaitant prendre la route dans une ou deux heures. Cette problématique procède de problèmes posés par les sociétés *Traffic First* et *Mediamobile* pour lesquelles nous avons engagé des études, et problèmes traités dans le cadre de l'ACI-NIM MIST-R dont nous assumons la direction.

Le trafic routier a été étudié en analyse numérique dans [24] et repose sur la théorie de l'écoulement des flux. Il s'agit d'une modélisation qui met en jeu la résolution locale d'équations différentielles, et fait intervenir la densité d'occupation de la chaussée. Elle permet d'obtenir de bons résultats de prévision, mais seulement en des points précis du réseau routier et à condition d'observer toutes les variables, avec néanmoins une mise en œuvre très lourde. Ainsi elles s'adaptent mal à une utilisation sur tout un réseau routier où seule la vitesse ponctuelle est observée. Par ailleurs, dès que le phénomène d'écoulement étudié change de régime (début ou fin d'un bouchon par exemple), les équations à résoudre doivent être modifiées, rendant difficile toute prévision à moyen terme de l'évolution du trafic routier. C'est la raison pour laquelle nous nous sommes tournés vers des méthodes statistiques de traitement de données fonctionnelles.

A cet effet, nous faisons l'hypothèse qu'il existe des profils types de comportement du trafic routier, c'est-à-dire des courbes caractéristiques de l'évolution journalière des vitesses, que l'on retrouve tout au long des différentes journées d'observation. Cette perspective nous amène donc à déterminer cette collection de modèles  $t \rightarrow f_i(t)$ ,  $i = 1, \dots, m$  ainsi que leur nombre  $m$ . Prédire les temps de parcours revient alors, tout d'abord, à trouver la courbe la plus proche des observations de la journée, puis à prolonger la courbe pour prédire les vitesses à venir, ou à agréger les informations des différentes courbes.

La construction de la collection de modèles d'évolution de vitesse se fait par différentes techniques de classification.

Dans une première approche nous supposons que les données  $Y_n$  sont issues d'un modèle de mélange dont il faut estimer les paramètres. Les courbes sont discrétisées et assimilées à des vecteurs de  $\mathcal{F} := \mathbb{R}^{180}$ . Les profils types sont donc des vecteurs  $f_j \in \mathcal{F}$ ,  $j = 1, \dots, m$ . Pour tout jour d'observation  $n = 1, \dots, N$ , il existe une variable aléatoire discrète  $X_j$ , à valeurs dans  $\{1, \dots, m\}$ , caractérisée par ses probabilités  $\pi_1, \dots, \pi_m$  et un bruit d'observation gaussien  $\epsilon$  de réalisations i.i.d  $\epsilon_n$ , tels que

$$Y_n = \sum_{j=1}^m \mathbf{1}_j(X_n) f_j + \epsilon_n, \quad n = 1, \dots, N. \quad (1.14)$$

A  $m$  fixé, le modèle est paramétrisé par  $\theta := \{(f_j)_{j=1, \dots, m}, (\pi_j)_{j=1, \dots, m}\}$ . L'estimation des paramètres de ce modèle est bien étudiée d'un point de vue à la fois théorique et pratique, nous pouvons nous référer par exemple à [22], [103] ou [34]. A  $m$  fixé, nous avons utilisé une version pénalisée et stochastique de l'algorithme EM pour estimer les profils caractéristiques, en maximisant en  $\theta$  la vraisemblance

$$L(y; \theta) = \sum_{n=1}^N \log \left( \sum_{j=1}^m \pi_j \phi(y_n; f_j, \sigma) \right),$$

où nous avons noté  $\phi$  la densité d'une variable aléatoire gaussienne. Nous obtenons un estimateur des paramètres  $\hat{\theta}^{(m)}$ .

Le choix du  $m$  optimal fait intervenir des méthodes de sélection de modèles. Intuitivement, le meilleur choix  $m^*$  est la valeur à partir de laquelle la vraisemblance  $m \mapsto L(\hat{\theta}^{(m)}; y, m)$  n'augmente plus de façon significative. Ce raisonnement revient à chercher le  $m$  réalisant l'équilibre entre, d'une part, l'adéquation aux données, obtenue en prenant ici, pour distance, une distance de Kullback sur les densités, et, d'autre part, la généralité de la description, donnée par le nombre de modèles sélectionné. En pratique, Lavielle in [80], construit un critère d'arrêt pour estimer le nombre de modèles optimal. Ce critère privilégie efficacement la stabilité de la procédure de sélection des groupes, et nous l'avons utilisé dans nos travaux.

Dans un second temps, nous avons fait appel à un algorithme de type classification hiérarchique [70], avec une distance bien adaptée aux phénomènes étudiés. En effet, très souvent, nous observons le même type de bouchon mais qui débute à des moments différents de la journée. Notre démarche vise à pénaliser les translations entre les courbes afin d'extraire les motifs correspondant aux types de bouchon. Pour cela, nous avons imaginé la distance suivante  $\Delta : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$  avec

$$\Delta(x, y) = \sqrt{(x - y)^t W (x - y)},$$

où  $W$  est une matrice de taille  $n \times n$  définie par  $W_{ij} = \frac{n - |i - j|}{n}$ ,  $\forall i = 1, \dots, n$ ,  $\forall j = 1, \dots, n$ .

Une fois la classification réalisée, nous obtenons des groupes de données homogènes. Nous nous trouvons confrontés à deux problèmes : extraire de chacun de ces groupes la courbe type et choisir un nombre approprié de modèles.

Obtenir le profil moyen pour chaque sous-groupe a été un de nos objectifs les plus difficiles à atteindre. En effet, si les courbes sont semblables à l'intérieur d'un groupe, elles sont assez différentes pour rendre impossible l'usage d'une simple médiane ou moyenne qui détériorait trop la forme des phénomènes étudiés. Ce problème pratique est à l'origine de toute la partie de ma recherche concernant l'alignement de courbes, que nous détaillerons dans la Section Problèmes Inverses 3.

Trouver le nombre optimal de groupes destinés à représenter au mieux les données relève de la sélection de modèles, puisque nous cherchons des groupes assez généraux mais offrant tout de même une bonne description des événements particuliers. La méthode de choix de modèles que nous proposons, prend en compte l'objectif pratique que nous nous sommes fixé (la prévision de temps de parcours) et utilise naturellement les techniques de la théorie de l'apprentissage statistique. En effet, une fois les  $m$  modèles choisis dans la classification automatique, pour un trajet donné, nous obtenons l'erreur de prévision qui dépend de ce nombre  $m$ . Choisir un  $m$  optimal revient à minimiser en  $m$  la somme des erreurs de prévision obtenue sur un échantillon d'apprentissage. La figure 1.1 montre clairement un phénomène d'overfitting et donc l'existence d'une valeur optimale pour le nombre de modèles. Cette valeur, dans cet exemple, vaut  $m^* = 11$ .

Les résultats très concrets que nous avons obtenus ont permis d'améliorer les performances des prévisions de l'évolution à court et moyen terme du trafic routier sur le réseau d'Ile de

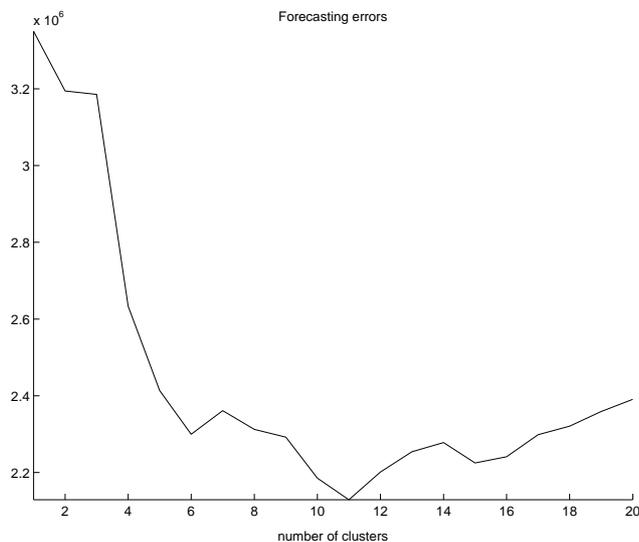


FIG. 1.1 – Erreur de prévision en fonction de  $m = 1, \dots, 20$ .

France. Nous présentons ces résultats dans la figure 2.1. Il apparait clairement que notre outil de prévision a de bien meilleurs résultats que le modèle stationnaire qui était employé jusqu'ici. La partie publique de ces résultats a été publiée dans [91]. Nous poursuivons notre étude dans cette direction afin de créer un outil de prévision qui sera commercialisé dans un futur proche. Cette recherche est développée dans le cadre de l'encadrement de la thèse CIFRE de G. Allain.

La classification hiérarchique, à partir d'une bonne distance, et des méthodes d'apprentissage statistique pour sélectionner le nombre optimal de classes, semble donner les meilleurs résultats en pratique. Cependant la classification hiérarchique crée des classes formées de profils très atypiques. En effet, choisir  $m^*$  revient à choisir une hauteur de coupure optimale dans l'arbre de classification correspondant. Mais, si l'arbre n'est pas symétrique et s'il existe de nombreuses feuilles à distance élevées d'un noeud, couper uniformément dans l'arbre entraîne la prise en compte de nombreux outliers qui ne sont pas représentatifs des tendances générales. Toutefois, éliminer ces outliers au préalable par des traitements particuliers n'est pas souhaitable, puisqu'un certain nombre de ces comportements est cependant représentatif d'événements rares mais bien connus, tels que les départs en vacances ou bien les retours de week-ends qui sont nécessaires à la prévision. Ne pas les prendre en considération reviendrait à ne plus être capable d'anticiper les événements rares de grande amplitude qui intéressent tout particulièrement les utilisateurs.

C'est pourquoi, nous avons essayé de mettre en œuvre une phase de classification plus élaborée, prenant en compte ces exigences. Pour ce but, nous avons dans un article [100] imaginé un algorithme itératif de classification. Partant d'un arbre de classification, nous cherchons, parmi toutes les feuilles appelées modèles candidats et notées  $\mathcal{M}$ , à construire un ensemble de représentants à sélectionner parmi toutes les feuilles et tous les noeuds de

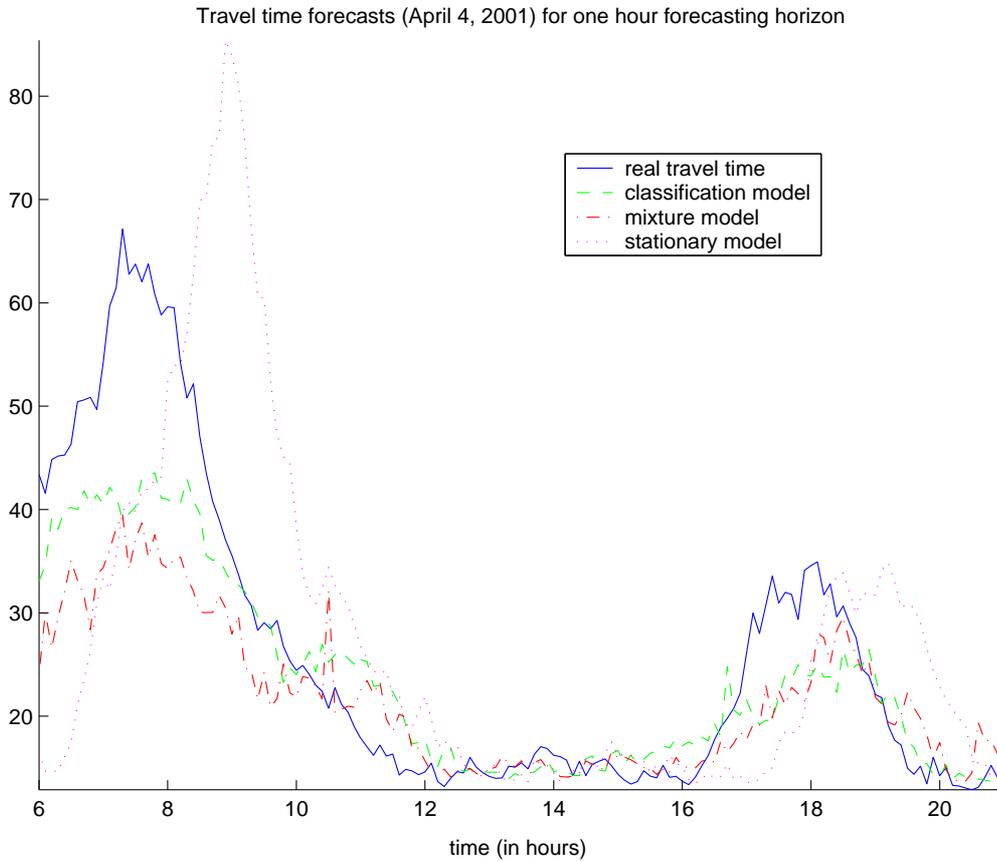


FIG. 1.2 – Erreurs de prévision

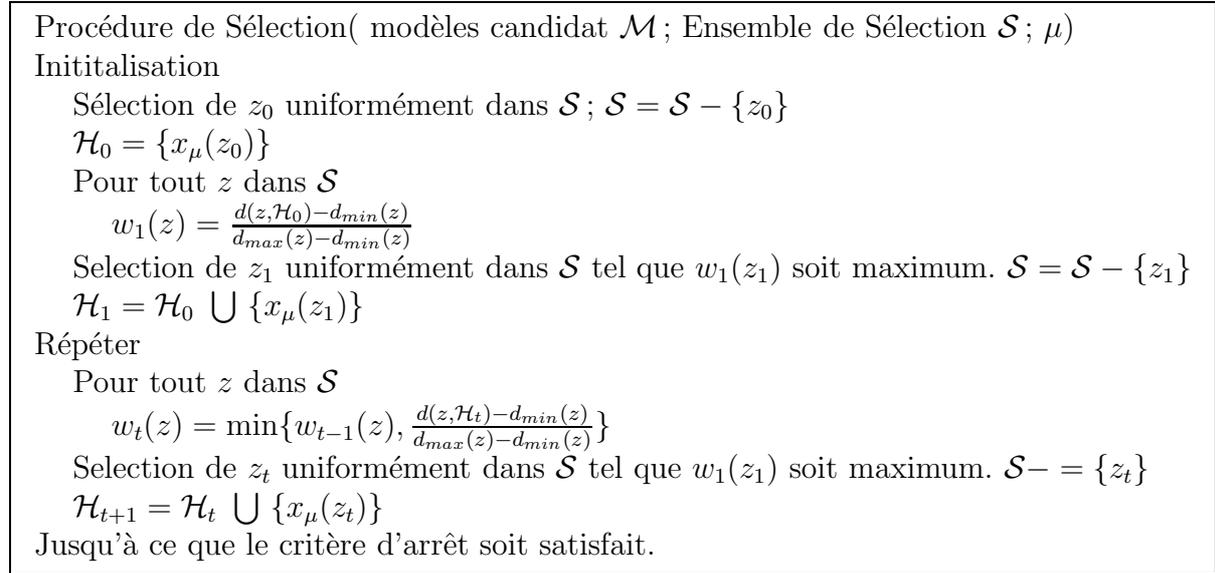


FIG. 1.3 – Selection algorithm

l'arbre. A la différence de la méthode précédente, nous nous autorisons à choisir des noeuds à différentes hauteurs dans l'arbre de classification. L'algorithme est itératif et fonctionne de la manière suivante : à chaque itération, nous choisissons une feuille  $z$  puis, dans l'ensemble de ses noeuds ascendants, nous sélectionnons un ensemble proche de la feuille choisie (au sens donné par la distance de classification  $d$ ), mais aussi recouvrant un grand nombre de feuilles, c'est-à-dire de taille  $c(\cdot)$  suffisante. Ainsi, pour un paramètre  $\mu$  à optimiser, le représentant  $x_\mu(z)$  est choisi en minimisant le critère suivant :

$$x_\mu(z) = \arg \min\{(1 - \mu)d(x, z) - \mu c(x), x \in \mathcal{M}(z)\}$$

Le noeud choisi est ajouté à l'ensemble des représentants choisis  $\mathcal{H}$ , tandis que les distances sont recalculées. Les itérations se succèdent jusqu'à l'obtention d'un ensemble représentatif de toutes les données. Tout le problème se situe donc dans la sélection de l'individu dont on souhaite améliorer la représentativité. Pour cela, notre algorithme stochastique multiobjectif privilégie à chaque fois les profils les moins bien représentés tout en pénalisant les individus dont on ne parvient à améliorer la description. Le choix se fait en fonction de poids attribués à chaque individu  $w_i$ , qui représentent l'éloignement de chaque individu par rapport à la fois à la représentation choisie mais aussi à sa meilleure représentation possible.

Dans ce système, les événements rares sont pris en compte mais pas ceux qui ne sont pas représentatifs du phénomène étudié. Cet algorithme est décrit dans 1.4. Utilisé lors de la phase de création de modèles, il assure une nette amélioration de l'outil de prévision : pour un nombre bien inférieur de modèles que ceux choisis par un algorithme de classification usuel, nous obtenons des erreurs de prévision comparables.

Nous avons ainsi construit une collection de modèles permettant d'une part de décrire le comportement du trafic routier, et d'autre part, d'utiliser ces modèles pour prédire des

temps de parcours. Nous créons alors un outil de prévision de l'évolution du trafic routier qui a l'avantage de fournir des prévisions globales sur tout le réseau. En outre, les calculs longs et coûteux en temps de calcul (construction des profils types et optimisation du nombre de classes) peuvent être réalisés dans une phase préliminaire. Les calculs propres à la prévision se font eux en-ligne et sont très rapides. De nombreuses améliorations sont, bien entendu, toujours en chantier. Par exemple, nous pouvons améliorer l'extraction des profils mais aussi utiliser des techniques d'agrégation pour construire un estimateur optimal à partir de tous les modèles construits. Un tel travail est l'objet d'un travail en cours [35].

La difficulté principale du travail de prévision du trafic routier provient des variations brusques qui apparaissent lors de l'étude de la circulation de voitures sur des axes routiers. La création d'un bouchon entraîne en effet des chutes brutales et rapides de la vitesse, ce qui se traduit par une grande irrégularité des courbes de vitesse. Ces courbes présentent des analogies avec l'étude des turbulences en écoulement des fluides [123]. De telles courbes peuvent se modéliser en utilisant des fonctions aux propriétés multifractales. Leur étude constitue le second axe de ma recherche.



# Chapitre 2

## Statistique des Processus

L'essentiel des résultats en statistique non paramétrique porte sur l'estimation de fonctions appartenant à des espaces de régularité bien connus en statistique (espaces de Sobolev et de Besov). Il existe en effet des bases adaptées permettant d'obtenir des représentations, de fonctions de ces espaces, bien adaptées à la théorie de l'estimation statistique. Nous nous référons à [43] pour de plus amples précisions.

### 2.1 Modèle de fonctions multifractales

Depuis quelques années, plusieurs tentatives ont été engagées afin de fournir des modélisations acceptables de signaux plus complexes. En particulier, une nouvelle classe de modèles a été développée afin de modéliser des fonctions très irrégulières, les fonctions multifractales. Ces fonctions sont caractérisées par la propriété que leur exposant de régularité de Hölder n'est pas constant. Leur régularité varie très rapidement et n'a de sens que de façon locale. Ces fonctions sont utilisées pour représenter l'évolution de fluides en turbulence [3], de cours boursiers [50] ou encore le trafic de données dans un réseau [84] ou [110].

C'est dans ce contexte que nous avons été amenés à considérer les modèles de séries lacunaires d'ondelettes, proposés par Jaffard dans ses travaux [69]. Soit  $\psi_{jk}$  une ondelette, considérons une fonction aléatoire  $f$  obtenue de la manière suivante. Pour deux paramètres  $(\eta, \alpha) \in [0, 1]^2$ , les coefficients aléatoires  $w_{jk}$  de la fonction  $f$  dans la base d'ondelettes suivent, pour chaque niveau de résolution  $j$  la loi suivante

$$\forall k = 0, \dots, 2^j - 1, \quad w_{jk} \sim 2^{(\eta-1)j} \delta_{2^{-\alpha}j} + (1 - 2^{(\eta-1)j}) \delta_0. \quad (2.1)$$

Il se détache qu'à chaque niveau de résolution  $j$ ,  $2^{\eta j}$  coefficients  $w_{jk}$  prennent la valeur  $2^{-\alpha j}$  tandis que les autres sont mis arbitrairement à zéro. Alors que, pour une fonction régulière, seuls comptent les gros coefficients, dans cette situation, seule la répartition des petits coefficients va déterminer le comportement de la fonction. Par exemple, les méthodes de seuillage ne permettent pas d'estimer de telles fonctions observées dans un modèle de bruit blanc. Jaffard montre dans ses travaux que de telles fonctions présentent des propriétés de multifractalité qui sont déterminées par les deux quantités qui gouvernent la loi de probabilité des coefficients aléatoires :  $\eta$ , le coefficient de lacunarité et  $\alpha$  le coefficient d'intensité. La figure 2.1 consitue

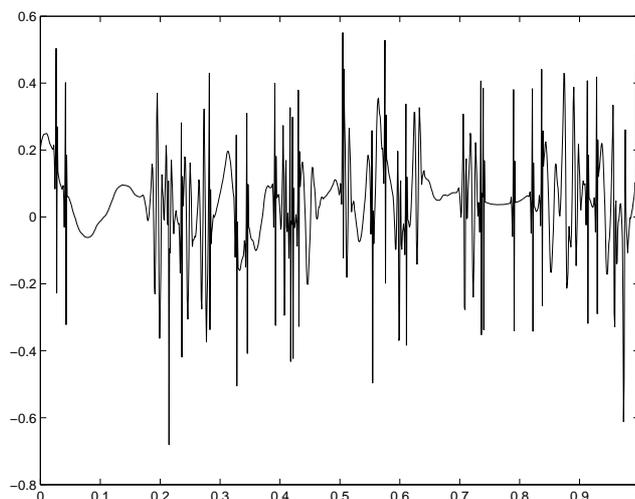


FIG. 2.1 – Fonction Multifractale

un exemple d'une réalisation d'une fonction multifractale.

Ainsi les méthodes d'estimation usuelles sont mises en défaut, mais un cadre bayésien, qui prend en compte les propriétés de l'histogramme des coefficients d'ondelettes, permet d'estimer ces fonctions dans un modèle de bruit blanc. C'est ce que nous avons pu démontrer dans notre travail.

## 2.2 Estimation Bayésienne d'une fonction multifractale

Notre objectif est de reconstruire une fonction multifractale observée dans un modèle de régression. De façon plus précise, nous observons la réalisation d'une fonction multifractale, décrite par sa décomposition en ondelettes de la forme

$$f^* = \sum_j \sum_{k=0}^{2^j-1} w_{jk}^* \psi_{jk}$$

dont les coefficients sont tirés suivant la loi décrite précédemment, loi déterminée par les paramètres  $\eta$  et  $\alpha$  vérifiant  $1 - 2\alpha > 0$ . Nous avons observé cette fonction dans un modèle de bruit blanc

$$d_{jk} = w_{jk}^* + \epsilon_{jk}, \quad j = 0, \dots, j_1 = \log_2(n), \quad k = 0, \dots, 2^j - 1. \quad (2.2)$$

où les  $\epsilon_{jk}$  sont des variables i.i.d centrées Gaussiennes de variance  $\sigma^2$ , indépendantes des  $w_{jk}$ . Donc, on observe le tableau triangulaire  $\mathbf{d}_n = (d_{jk})_{1 \leq j \leq j_1, 0 \leq k \leq 2^j - 1}$  de variables indépendantes, de loi, pour tout couple  $j, k$ , donnée par la formule suivante

$$d_{jk} \sim 2^{(\eta-1)j} \mathcal{N}(2^{-\alpha j}, \frac{\sigma^2}{n}) + (1 - 2^{(\eta-1)j}) \mathcal{N}(0, \frac{\sigma^2}{n}), \quad (2.3)$$

Concrètement, nous recherchons à estimer les coefficients  $w_{jk}^*$  à partir des observations  $d_n$ . Or les coefficients  $w_{jk}^*$  sont soit nuls, soit égaux à  $2^{-\alpha j}$  pour tout niveau de résolution  $j$ . Donc,

dans les deux cas, les coefficients sont très faibles, ce qui rend délicat leur reconstruction. Dans ce cas, nous avons utilisé une méthode Bayésienne utilisant la forme particulière des coefficients.

Quand les coefficients  $\alpha$  et  $\eta$  sont connus, considérons comme loi a priori, une loi uniforme sur l'ensemble

$$\Omega_j = \left\{ \omega = (\omega_k)_{k=0, \dots, 2^j-1} \in \{0, 2^{-\alpha_0 j}\}, \sum_{k=0}^{2^j-1} \omega_k = 2^{(\eta_0 - \alpha_0)j} \right\} \quad (j \in \mathbb{N}).$$

Soit  $j_1(n)$  le nombre de niveaux qui interviendront dans la reconstruction de l'estimateur. L'estimateur maximisant la loi a posteriori est donné par

$$\hat{w}_j = \arg \max_{w_j \in \Omega_j} p(w_j | d_j).$$

Nous pouvons remarquer que l'estimateur bayésien de la réalisation de la fonctions  $f^*$  s'écrit plus simplement sous la forme suivante

$$\hat{f}_n = \sum_{j=0}^{j_1} \sum_{k=0}^{2^j} 2^{-\alpha_0 j} 1_{|d_{jk}| \geq d_{(j, \lfloor 2^{\eta_0 j} \rfloor)}} \psi_{jk},$$

où les coefficients observés ont été réordonnés de manière décroissante

$$d_{j,(0)} \geq \dots \geq d_{j,(\lfloor 2^{\eta_0 j} \rfloor)} \geq \dots d_{j,(2^j-1)}.$$

Soit  $\Pi_1$  la projection sur l'espace  $V_{j_1}$  de l'analyse multirésolution de l'ondelette où  $2^{j_1} = n$  le nombre d'observations. Le théorème suivant donne la vitesse de convergence de l'estimateur bayésien.

**Théorème 2.2.1** (Estimation Bayésienne d'une fonction multifractale).

Supposons que  $\alpha_0 < \frac{1}{2}$ . Choisissons  $(j_1(n))$  tel que  $2^{j_1} = O\left(\left[\frac{n}{\log n^\beta}\right]^{\frac{1}{1+2\alpha_0-\eta_0}}\right)$ , avec  $\beta > 8$ .

Alors, il existe une constante positive  $c_1$  tel que

$$\mathbf{E} \left[ \|f^* - \hat{f}_n\|_2^2 \right] \leq c_1 \frac{\log n}{n}. \quad (2.4)$$

Nous retrouvons une vitesse de convergence paramétrique, dès que  $j_1(n)$  est bien choisi.

Ce résultat peut être démontré en utilisant, à la fois des techniques propres aux ondelettes, et des méthodes liées aux problèmes d'estimation dans des modèles de mélanges. En effet, les coefficients observés des fonctions multifractales suivent la loi (2.3), qui mélange deux distributions gaussiennes, avec un décalage qui tend à se combler et des variances tendant vers 0. Ainsi, le comportement asymptotique dépend du rapport de ces vitesses. En outre nous pouvons remarquer que la condition  $\alpha_0 < \frac{1}{2}$  est nécessaire pour que le signal observé possède assez d'information pour pouvoir être estimé.

### 2.3 Estimation des paramètres du modèle multifractal

D'une façon générale, l'estimation Bayésienne fait appel à une loi a priori dépendant d'hyperparamètres, un modélisant la lacunarité de la série et un autre modélisant l'intensité des coefficients. Or ces coefficients sont inconnus a priori.

Cela explique pourquoi, dans un second travail, nous avons développé une méthode d'estimation de ces coefficients, puis étudié l'influence de cette estimation préliminaire dans la procédure de reconstruction bayésienne.

Une méthode naturelle d'estimation des paramètres d'une loi est de chercher les valeurs qui maximisent la log vraisemblance. Pour cela, nous avons, tout d'abord, appliqué un algorithme EM afin de résoudre ce problème de maximisation et de déterminer les paramètres de la loi a priori. En effet, la structure de la distribution a priori des coefficients peut s'exprimer plus facilement en faisant apparaître une variable cachée du modèle, qui détermine l'appartenance des coefficients à l'un des deux types de population. Plus précisément, en normalisant les variances à chaque niveau de résolution nous obtenons la distribution suivante

$$X_{jk} := \sqrt{n}d_{jk} \sim 2^{(\eta^*-1)j} \mathcal{N}(m_j, \sigma^2) + (1 - 2^{(\eta^*-1)j}) \mathcal{N}(0, \sigma^2),$$

avec  $m_j = 2^{j_1/2 - \alpha^*j}$ ,  $j = 1, \dots, j_1$ . Dès que les coefficients sont tels que  $\eta^* - \alpha^*/2 > 0$ , les coefficients normalisés sont obtenus à partir d'un mélange de gaussiennes à 2 éléments dont les 2 composants s'éloignent asymptotiquement.

Nous sommes alors dans le cadre d'application de l'algorithme EM, cf. par exemple [103]. Ces estimateurs sont utilisés afin d'obtenir un estimateur pratique de la réalisation de la fonction  $f$ . Les performances pratiques de cet estimateur sont étudiées dans [53]. Dans la figure 2.2, nous présentons un exemple de reconstruction qui nous permet de constater l'efficacité de notre procédure d'estimation.

La deuxième étape naturelle est l'étude de l'estimation paramétrique de la loi a priori. Dans un travail plus théorique [54], nous avons construit des estimateurs empiriques de ces paramètres de la manière suivante

$$\hat{\alpha}_n = \frac{1}{j_1 \log 2} \left( \log \left[ \frac{\sum_{j=1}^{j_1} \sum_{k=0}^{2^j-1} d_{jk}}{\sum_{j=1}^{j_1} \sum_{k=0}^{2^j-1} d_{jk}^2 - \sigma^2} \right] \right) \quad (2.5)$$

$$\hat{\eta}_n = 1 + \frac{1}{\log_2 n} \log_2 \left( \frac{1}{n} \sum_{j=1}^{j_1} \sum_{k=0}^{2^j-1} 1_{\sqrt{n}d_{jk} \geq \log_2(n)} \right). \quad (2.6)$$

Le premier estimateur est un estimateur des moments du mélange. Le second estimateur tire profit de la propriété de séparation des deux composants du mélange, lorsque  $n \rightarrow +\infty$ , pour estimer la proportion de coefficients non nuls.

Leur comportement asymptotiques est décrit par le théorème suivant

**Théorème 2.3.1** (Convergence des estimateurs des hyperparamètres).

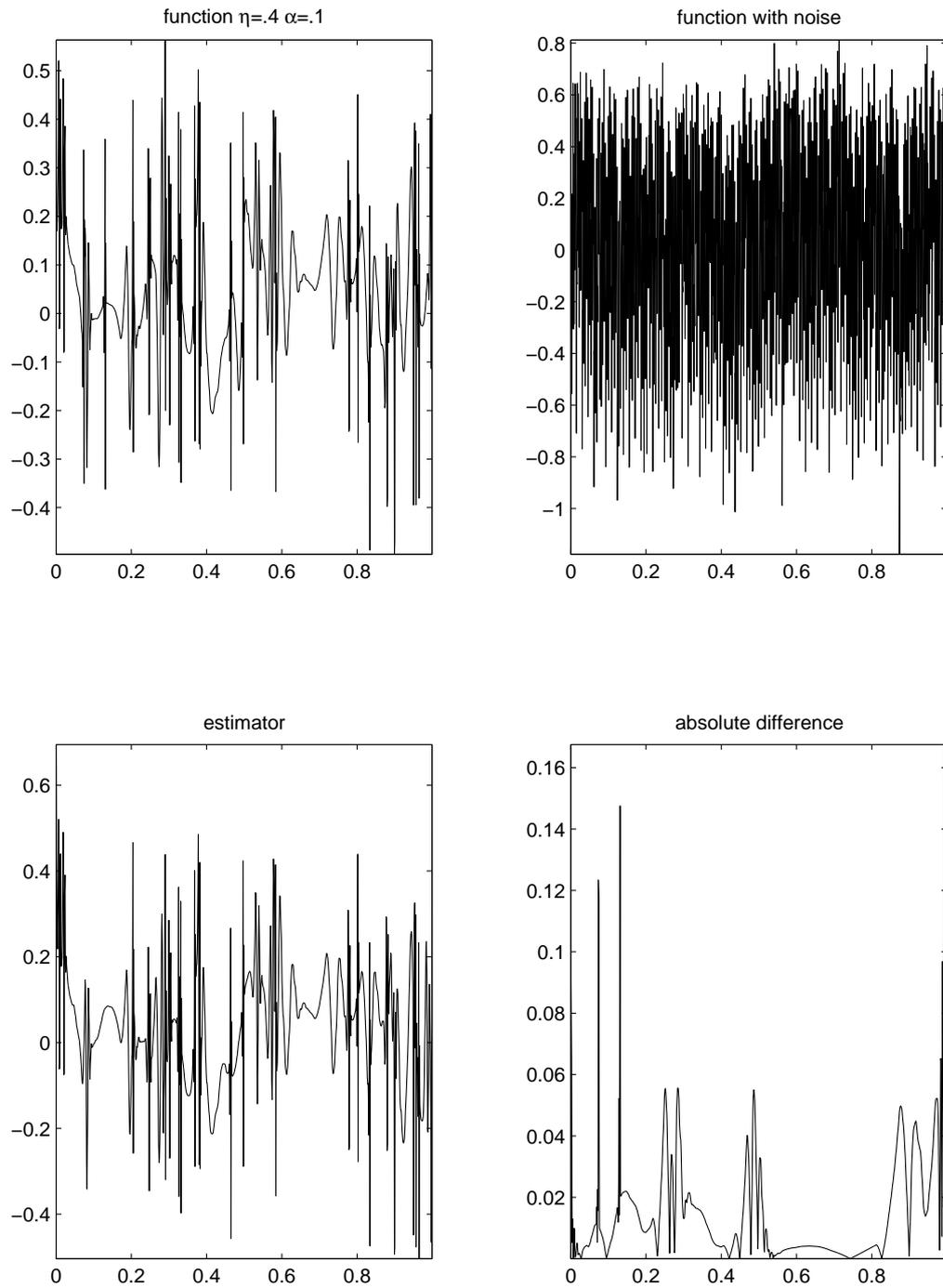


FIG. 2.2 – Reconstruction bayésienne avec paramètres estimés

Supposons que  $\alpha_0 < \frac{1}{2}$  et  $\eta_0 - 2\alpha_0 > 0$  alors les estimateurs  $\hat{\eta}_n$  et  $\hat{\alpha}_n$  sont tels que

$$\begin{aligned} n^{\frac{\eta^*}{2}} \log(n) (\tilde{\eta}_n - \eta^*) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2) \\ n^{\frac{\eta^*}{2}} \log(n) (\hat{\alpha}_n - \alpha^*) &\xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2). \end{aligned}$$

Nous constatons que la vitesse de convergence n'est pas la vitesse paramétrique ordinaire, mais dépend de la valeur du vrai paramètre  $\eta_0$ . Ce résultat nous permet d'estimer à partir des données les paramètres d'intensité et de lacunarité correspondants et de modéliser par des séries lacunaires, les réalisations de fonctions multifractales. De plus, le théorème 2.3.1, nous a permis de construire un test pour vérifier si un signal observé est multifractal, contre une hypothèse alternative de type bruit blanc.

La vitesse d'estimation des paramètres  $(\eta^*, \alpha^*)$  s'apparente à une vitesse non paramétrique. Pour comprendre une telle vitesse, nous devons adopter un point de vue semi-paramétrique. Les paramètres  $(\eta_0, \alpha_0)$  sont observés dans un modèle dépendant d'une fonction inconnue, donnée sous la forme d'une suite infinie de coefficients. Cette fonction est un paramètre de nuisance modifiant la vitesse d'estimation des paramètres.

C'est la raison pour laquelle nous nous sommes intéressés à l'optimalité de la procédure d'estimation. Afin de donner un sens à cette notion d'optimalité, nous nous sommes intéressés aux propriétés LAN (normalité asymptotique locale) du modèle multifractal paramétré par les deux paramètres d'intensité  $\alpha$  et de lacunarité  $\eta$ .

Pour cela, décomposons l'espace des paramètres  $\theta = (\alpha, \theta)' \in \Theta$  en  $\Theta = \Theta^- \cup \Theta^\pm \cup \Theta^+$ , avec  $\Theta^- = (0, 1) \times (0, 1/2)$ ,  $\Theta^\pm = (0, 1) \times \{1/2\}$ , et  $\Theta^+ = (0, 1) \times (1/2, 1)$ . Nous montrons que la famille de probabilités  $\mathcal{P}_n^- := \{P_\theta^n \mid \theta \in \Theta^-\}$  possède des propriétés LAN.

Plus précisément, pour toute suite bornée  $\tau_n = (s_n, t_n)$ , considérons la perturbation locale  $\theta_n = \theta + \nu_n(\theta)\tau_n = (\eta_n, \alpha_n)$  du paramètre  $\theta = (\eta, \alpha)$ , avec  $\nu_n(\theta) = \text{diag}(c_n(\theta), d_n(\theta)) = \text{diag}(n^{-\frac{\eta}{2}}(\log n)^{-1}, n^{-\frac{\eta}{2} + \alpha - \frac{1}{2}}(\log n)^{-1})$ . Définissons la log-vraisemblance locale par

$$L_{\theta_n/\theta}^n = \frac{dP_{\theta_n}^n}{dP_\theta^n} = \sum_{i=1}^n [\log f_{\theta_n}(X_{ni}) - \log f_\theta(X_{ni})].$$

Définissons aussi  $\phi_\alpha$  (resp.  $\phi$ ), la densité d'une gaussienne de moyenne  $n^{\frac{1}{2}-\alpha}$  (resp. 0).

**Théorème 2.3.2** (Propriété LAN du modèle probabiliste de fonctions multifractales).

La famille de probabilités  $\mathcal{P}_n^-$  est LAN, de suite centrale  $\Delta_\theta^n := (\Delta_{\theta,I}^n, \Delta_{\theta,II}^n)$ , avec

$$\Delta_{\theta,I}^n := \sum_{i=1}^n D_{\theta,I}^{ni} := n^{\frac{\eta}{2}-1} \sum_{i=1}^n \frac{\phi_\alpha - \phi}{n^{\eta-1}[\phi_\alpha - \phi] + \phi}(X_{ni}),$$

$$\Delta_{\theta,II}^n := \sum_{i=1}^n D_{\theta,II}^{ni} := \sigma^{-2} n^{\frac{\eta}{2}-1} \sum_{i=1}^n (n^{\frac{1}{2}-\alpha} - X_{ni}) \frac{\phi_\alpha}{n^{\eta-1}[\phi_\alpha - \phi] + \phi}(X_{ni}),$$

et de matrice d'information  $\Gamma := \text{diag}(1, \sigma^{-2})$ . Plus précisément, nous avons

$$L_{\theta + \nu_n(\theta)\tau_n/\theta}^n = \tau_n \Delta_{\theta}^n - \frac{1}{2}\tau_n \Gamma \tau_n + o_{\mathbb{P}}(1) \quad \text{et} \quad \Delta_{\theta}^n \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Gamma) \quad (2.7)$$

sous la loi  $\mathbb{P}_{\theta}^n$ .

Ces résultats sont décrits dans [93] et leur démonstration repose sur la méthodologie employée dans [117]. Nous remarquons que les estimateurs définis dans [54] vérifient sur l'espace de paramètre  $\Theta^{-}$

$$(\nu_n(\theta))^{-1}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Gamma^{-1}).$$

Par conséquent, il apparaît que l'estimateur du paramètre de lacunarité proposé précédemment est optimal au sens de Le Cam lorsque  $(\alpha_0, \eta_0) \in \Theta^{-}$ . Ceci nous permet de justifier l'optimalité des tests de multifractalité basés sur la lacunarité de son développement en séries d'ondelettes. C'est le sens du théorème suivant.

Considérons le problème de test suivant de niveau asymptotique  $\beta \in (0, 1)$

$$\begin{cases} \mathcal{H}_0^{(n)} : \alpha \leq \alpha_0 \\ \mathcal{H}_1^{(n)} : \alpha > \alpha_0 \end{cases},$$

pour un  $\alpha_0 < 1/2$  fixé. Alors les propriétés LAN du théorème 2.3.2 et la convergence (à la bonne vitesse) de l'estimateur  $\hat{\eta}$  prouvent le résultat suivant.

**Théorème 2.3.3** (Test optimal de multifractalité). *Soit  $\phi^{(n)}$  le test qui rejette l'hypothèse  $\mathcal{H}_0^{(n)}$  et accepte l'hypothèse  $\mathcal{H}_1^{(n)}$  si et seulement si*

$$\sigma \Delta_{\hat{\eta}_n, \alpha_0, \mathbf{II}}^n = \sigma^{-1} n^{\frac{\hat{\eta}_n}{2}-1} \sum_{i=1}^n (n^{\frac{1}{2}-\alpha_0} - X_{ni}) \frac{\phi_{\alpha_0}}{n^{\hat{\eta}_n-1} [\phi_{\alpha_0} - \phi] + \phi} (X_{ni}) > \Phi^{-1}(1 - \beta),$$

où  $\Phi$  est la fonction de répartition d'une loi normale centrée réduite. Alors, la suite de tests  $\phi^{(n)}$  est localement et asymptotiquement maxi-min au niveau asymptotique  $\beta$ .

Par conséquent, nous avons développé une procédure complète de traitement de données de type turbulence

- permettant de décider si les observations sont réellement de type multifractal,
- puis, le cas échéant, d'estimer les paramètres gouvernant les propriétés multifractales,
- et finalement de reconstruire la fonction observée.

## 2.4 Application aux processus de Lévy

Les techniques développées pour des processus multifractionnaires peuvent être adaptées à d'autres types de processus, notamment les processus de Lévy fractionnaires. Ces processus ont été introduits par Benassi et al. in [9] pour modéliser des processus en finance ou en biologie. En effet, les processus multifractal précédemment étudiés ou les processus multifractionnaires comme par exemple le mouvement brownien multifractionnaire (FBM) introduit par Mandelbrot et al. dans [101] permettent de représenter des signaux, certes très irréguliers,

mais présentant des propriétés d'une loi gaussienne. Or dans un grand nombre d'observations réelles cette hypothèse de gaussiannité des observations n'est pas vérifiée.

Dans notre étude [78], nous avons proposé un estimateur de l'exposant de Hurst  $H$  d'un processus RHFLM (Real Harmonizable Fractional Lévy Motion) défini par

$$X_H(x) = \int_{\mathbb{R}^d} \frac{e^{-ix \cdot \xi} - 1}{\|\xi\|^{H+d/2}} L(d\xi), \quad x \in \mathbb{R}^d \quad (2.8)$$

où  $\|\cdot\|$  est la norme euclidienne et  $L(d\xi)$  est une mesure de Lévy. De tels processus sont non gaussiens, localement asymptotiquement autosimilaires d'exposants de Hurst, le paramètre  $H$ . Vouloir utiliser en pratique de tels modèles implique que l'on soit capable de reconnaître un RHFLM à partir d'observations, et en particulier de construire un estimateur de  $H$  à partir de données réelles. Un premier exemple d'estimateur de l'exposant de Hurst nous est donné par l'estimateur de variation quadratique proposé par Benassi et al. dans [9]. Néanmoins cet estimateur ne converge pas lorsque les données sont observées en présence d'un bruit d'observation. C'est pourquoi nous avons étudié un estimateur de l'exposant de Hurst d'un processus RHFLM étudié dans un cadre de modélisation réaliste : le processus est observé en des temps d'observation discrets en présence d'un bruit d'observation gaussien, c'est-à-dire dans le modèle de régression suivant.

$$Y_H\left(\frac{k}{n}\right) = X_H\left(\frac{k}{n}\right) + \sigma_n \varepsilon_k, \quad k = 0, \dots, n. \quad (2.9)$$

Alors même que les estimateurs traditionnels ne sont pas convergents lorsque le processus est observé avec du bruit, nous proposons un estimateur convergent de même nature que ceux utilisés dans [54] et utilisant les coefficients d'ondelettes du processus de Lévy.

Pour une ondelette  $\psi$ , les coefficients d'ondelettes du processus de Lévy et respectivement les coefficients empiriques du processus discrétisés sont définis par la relation suivante

$$w_{jk} = 2^{j/2} \int X_H(t) \psi(2^j t - k) dt,$$

$$w_{jk}^n = \frac{1}{2^{j/2} n} \sum_{p=1}^n Y\left(\frac{p+nk}{2^j n}\right) \psi\left(\frac{p}{n}\right).$$

**Théorème 2.4.1** (Convergence de l'estimateur de Hurst d'un RHFLM).

*Définitions les coefficients obtenus par la transformée en ondelette discrète des données :  $w_{jk}$  et construisons comme estimateur de  $H$*

$$\hat{H}_{j_1(n)} = \frac{-1}{2^{j_1(n)}} \log_2 \left( \sum_{k=0}^{2^{j_1(n)}-1} w_{jk}^2 \right) \quad (2.10)$$

*Si l'ondelette possède au moins  $r = 1$  moment nul, alors*

$$\hat{H}_{j_1(n)} \xrightarrow{j_1(n) \rightarrow +\infty} H, \quad p.s.$$

Les preuves de ce résultat reposent d'une part sur la formule de calcul pour les processus RHFLM. Si  $f \in L^2(\mathbb{R})$  et  $\overline{f(\xi)} = f(-\xi)$ ,  $\int f(\xi) L(d\xi)$  sont des variables aléatoires symétriques à valeurs réelles, alors

$$\mathbf{E} \left[ e^{iu \int f(\xi) L(d\xi)} \right] = \exp \left( -\frac{b^2 u^2 \|f\|_2^2}{2} \right) \mathbf{E} \left[ e^{iau \int f(\xi) M(d\xi)} \right], \quad u \in \mathbb{R}. \quad (2.11)$$

Cette formule permet d'évaluer les moments des coefficients d'ondelettes des processus. Par ailleurs, nous utilisons des renseignements sur la loi asymptotique des vrais coefficients d'un processus RHFLM, i.e lorsque  $j \rightarrow +\infty$ ,

$$\sqrt{2^{j(1+2H)}} w_{jk} \xrightarrow{(D)} \mathcal{N} (0, \mathbf{E}(w_{00}^2)).$$

La non gaussiannité ne permet pas de préciser la vitesse de convergence, des simulations mettent en évidence l'efficacité de notre méthode.

Notre méthodologie permet néanmoins d'estimer, à partir d'observations réelles, la valeur des paramètres, nécessaires pour modéliser le signal étudié par un processus probabiliste. Elle permet également de tester la validité de ces modèles dans des cas concrets particuliers. Ce résultat pourra s'appliquer notamment en analyse de signaux économiques ou venant de phénomènes liés à l'évolution de la vitesse de véhicules en des points particuliers d'un réseau routier.

Un cadre d'application de ce travail sera donné par des signaux venant de la finance.



# Chapitre 3

## Problèmes inverses

### 3.1 Présentation Générale

Ce qui vient d'être dit jusqu'à présent ne concerne que la situation dans laquelle le signal à estimer est observé directement. Or les signaux sont souvent observés à travers leur image par un opérateur. Dès lors, estimer nécessite un inversion de cet opérateur. Cet type de modélisation permet de prendre compte bien des situations rencontrées dans divers domaines, notamment en biologie (étude de la pharmacocinétique d'un traitement, imagerie médicale), en physique (déconvolution en imagerie spatiale, tomographie, résolution d'équations différentielles), en économie (estimation en présence de variables instrumentales), ...

Dans cette approche, le problème peut être posé de la manière suivante : soit  $A$  un opérateur entre deux espaces de Hilbert  $X$  et  $Y$ , connaissant  $y$ , nous cherchons à reconstruire une fonction  $x_0 \in X$  telle que

$$Ax_0 = y.$$

$\langle \cdot, \cdot \rangle$  désigne le produit scalaire sur l'espace  $Y$ , et  $\|\cdot\|$  désigne la norme associée. Ce problème est dit bien posé s'il existe une solution unique à ce problème et si la solution dépend continûment de la fonction  $y$ . Dans le cas contraire, le problème est dit mal posé, ce qui signifie que l'inverse  $A^{-1}$  n'est pas défini en tant qu'opérateur continu sur  $Y$ . Ainsi, la solution ne dépend pas continûment de la fonction  $y$ . Dès lors, une petite perturbation sur la fonction  $y$  entraîne une forte erreur finale. Plus précisément, si l'on observe  $y^\epsilon$  proche de  $y$  au sens où  $\|y - y^\epsilon\| \leq \epsilon$ , rien ne garantit que l'erreur associée  $A^{-1}(y - y^\epsilon)$  sera négligeable. Par conséquent, notre travail doit s'attacher à définir des méthodes d'inversion approchées qui garantissent que l'inverse approché, dont la signification doit être donnée, soit proche de la vraie valeur et présente en outre des propriétés de stabilité.

La manière de définir l'erreur d'observation est donc fondamentale dans la description des problèmes inverses et requiert une modélisation spécifique. Sans terme d'erreur, il est nécessaire d'imposer un critère de stabilité. Ce type d'hypothèses apparaît lorsque l'opérateur est compliqué et que l'étude mathématique du problème est difficile, comme nous le verrons en tomographie sismique dans [48].

Une seconde approche, utilisée très souvent en analyse et en analyse numérique [46], consiste à supposer que l'erreur est une fonction appartenant à une boule de l'espace de Hilbert  $Y$ . L'erreur d'observation  $e$  est telle que  $\|e\| \leq \delta$ . Dès lors on observe  $y^\delta = y + e$  tel que  $\|y - y^\delta\| \leq \delta$ . Cette situation a été étudiée par Tikhonov et Arsenin dans [118] et leurs différentes stratégies d'estimation sont décrites dans [119].

L'approche statistique consiste à supposer que le bruit est donné par les réalisations d'une variable aléatoire à valeur dans l'espace  $Y$ . Ainsi, le modèle statistique considéré est le suivant

$$Y_i = A(x_0)(t_i) + \epsilon_i, \quad : i = 1, \dots, n \quad (3.1)$$

où  $Y_i, i = 1, \dots, n$  sont les observations réelles, les  $t_i$  sont les temps d'observations,  $x_0 : \mathbb{R} \rightarrow \mathbb{R}$  est la fonction inconnue à reconstruire dont la régularité est inconnue et  $A$  est un opérateur connu ou non entre deux espaces de Hilbert  $A : \mathbf{X} \rightarrow \mathbf{Y}$ . Les variables  $\epsilon_i$ , quant à elles, constituent le bruit d'observation. La version continue théorique de ce modèle s'écrit

$$y = Ax_0 + \epsilon.$$

La variable aléatoire  $\epsilon$  est à valeur dans l'espace de Hilbert  $Y$ . Plus précisément, cette variable est caractérisée par les relations : pour tout  $u \in Y$ ,  $\langle u, \epsilon \rangle$  est une variable aléatoire ayant des moments finis. En particulier,  $\|\epsilon\| = +\infty$ . Ainsi le bruit statistique est plus *fort* que le bruit déterministe. Cette différence fondamentale entre les deux types de modèle est particulièrement détaillée dans [96]. Le passage du modèle continu au modèle discret n'est pas aisé et devient même délicat lorsque l'opérateur n'est pas linéaire.

La difficulté des problèmes inverses réside dans le fait qu'il est nécessaire de trouver un moyen d'inverser l'opérateur  $A$  ou d'approcher son inverse alors qu'il n'est, en général, pas inversible. Cette propriété est décrite par un indice (qui sera noté  $t$ ) qui mesure la manière dont le problème est mal posé. Ce paramètre, plus connu dans la littérature anglosaxonne comme l'*index of ill-posedness*, peut s'interpréter comme le paramètre tel que  $A$  envoie un espace de Hilbert de régularité  $s$  dans un espace de régularité  $s + t$ . Plus  $t$  est élevé, plus le problème inverse sera difficile à résoudre.

Une autre interprétation est donnée en utilisant la décomposition en valeurs singulières de l'opérateur ou SVD. Si  $A$  est un opérateur compact injectif, alors  $A^*A$  est un opérateur compact sur  $X$ , possédant comme valeurs propres  $\lambda_k^2, k = 1, 2, \dots$ . Soit  $\psi_k, k \geq 1$  une base orthonormée de fonctions propres correspondantes. Remarquons que  $\|A\psi_k\| = \lambda_k$  et définissons

$$\frac{A\psi_k}{\|A\psi_k\|} = \varphi_k.$$

En ce cas, il existe  $(\lambda_k, \varphi_k, \psi_k), k \geq 1$  tels que, d'une part,  $A\psi_k = \lambda_k\varphi_k$ , et d'autre part, nous avons les décompositions suivantes

$$\forall x \in \mathbf{X}, x = \sum_{k \geq 1} x_k \psi_k$$

$$\forall y \in \mathbf{Y}, y = \sum_{k \geq 1} y_k \varphi_k.$$

Donc le modèle (3.1) peut être réécrit de la manière suivante, en projetant  $y$  sur la base des  $\varphi_k$ ,  $k \geq 1$

$$\begin{aligned} \langle y, \psi_k \rangle &= \langle Ax_0, \psi_k \rangle + \langle \epsilon, \psi_k \rangle \\ y_k &= \lambda_k x_k + \epsilon_k, \quad k \geq 1 \end{aligned} \quad (3.2)$$

où les variables aléatoires  $\epsilon_k$  sont des variables i.i.d dont la loi est déterminée par la loi de  $\epsilon$ . L'objectif porte alors sur l'estimation des coefficients  $x_k$ . Les valeurs propres  $\lambda_k$  tendent vers zéro, rendant ainsi plus compliquée cette estimation. L'indice  $t$  décrit la vitesse à laquelle les valeurs propres de l'opérateur tendent vers zéro, par exemple  $\lambda_k \sim k^{-t}$ , dans le cas de vitesses polynomiales. Si on compare ce modèle à celui où les signaux sont observés directement, les valeurs propres  $\lambda_k$  apparaissent clairement comme des paramètres qui détériorent le signal. Plus ces valeurs sont faibles et plus les coefficients seront difficiles à estimer. Le paramètre  $t$  est, par conséquent, un indice de la difficulté du problème inverse.

Si on réécrit le modèle en utilisant l'inverse généralisé  $A^{-1}$ , nous obtenons une nouvelle observation

$$z = A^{-1}y = x_0 + A^{-1}\epsilon.$$

Cette réécriture met également en exergue le fait que le problème réside dans le fait que l'inverse n'est pas continu. Notamment le bruit d'observation  $A^{-1}\epsilon$  explose pour un opérateur compact.

Nous notons que le modèle (3.2) est une version simplifiée du modèle (3.1). Tout d'abord, il ne prend pas en compte la discrétisation, cruciale en modélisation puisque l'on suppose que les fonctions propres sont orthogonales par rapport au produit scalaire de  $Y$ . Or, en toute rigueur, la projection n'a de sens que par rapport au produit scalaire donné par la mesure empirique,

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{t_i}.$$

Dans un second temps, nous pouvons noter qu'il est nécessaire de connaître la décomposition SVD de l'opérateur pour pouvoir écrire les observations sous la forme (3.2). Or, en pratique, cette décomposition n'est connue que dans très peu de cas et, d'autre part, l'opérateur est souvent soit inconnu, soit observé avec des erreurs. Enfin, lorsque l'opérateur n'est pas linéaire, il n'est plus possible d'écrire une telle décomposition.

### 3.2 Méthodes de régularisation de problèmes inverses

Nous avons vu que pour résoudre un problème inverse, il importe de régulariser l'opérateur afin d'approcher son inverse tout en contrôlant l'erreur ainsi commise.

Pour cela, nous avons, dans un premier temps, utilisé la méthodologie développée dans notre travail de thèse en proposant un estimateur fondé sur la norme  $l^1$  des coefficients de la fonction à reconstruire. Dans [88] ou [87], nous construisons la procédure suivante. Soit

$(\lambda_j; \varphi_j, \psi_j)_{j \geq 1}$  la SVD de l'opérateur linéaire  $A$  telle que les fonctions  $\psi_k$  forment un système orthonormé par rapport à la mesure empirique  $\mu_n$ . Définissons l'estimateur

$$\begin{aligned} \hat{x}_n &= \arg \min_{x = \sum_{j=1}^n x_j \psi_j \in X} \left[ \sum_{j=1}^n \left| \langle y - A(x), \frac{\varphi_j}{\lambda_j} \rangle_n \right|^2 + 2 \|\mu x\|_1 \right] \\ &= \sum_{j=1}^n \hat{x}_{j,n} \psi_j \end{aligned} \quad (3.3)$$

où  $\mu = (\mu_j, j = 1, \dots, n)$  est une suite de nombres réels permettant de contrôler, à chaque niveau de la reconstruction du signal, l'apport de chaque contribution. Ce type d'estimateur est fréquemment utilisé en traitement du signal ou en reconstruction d'images, sans qu'il n'existe de preuve de la vitesse de convergence. C'est précisément la lacune que vient combler le théorème suivant.

**Théorème 3.2.1** (Vitesse de convergence de l'estimateur pénalisé  $l^1$  pour problème inverse).  
Supposons qu'il existe  $s$  et  $0 < p < 2$  tels que

$$x_0 \in X_{s,p} = \left\{ x = \sum_j x_j \psi_j, \sum_{j=1}^n j^{p(s+\frac{1}{2}-\frac{1}{p})} x_j^p \leq 1 \right\}.$$

Pour  $0 \leq t \leq 1$ , nous obtenons

$$\|\hat{x}_n - x_0\|_n = O_{\mathbf{P}} \left( \left( \frac{n}{\log n} \right)^{-\frac{2s}{2s+2t+1}} \right). \quad (3.4)$$

Pour  $t > 1$ , la vitesse de convergence est donnée par

$$\|\hat{x}_n - x_0\|_n = O_{\mathbf{P}} \left( \left( \frac{n}{\log n} \right)^{-\frac{2s}{2s+2t+1}} n^{p(t-1)} \right). \quad (3.5)$$

La technique de démonstration repose sur les idées énoncées dans [99]. Soit l'estimateur oracle défini par  $x_* = \sum_{j=1}^n x_{j,*} \psi_j$ , obtenu par seuillage dur des coefficients de la vraie fonction, i.e

$$x_{j,*} = \begin{cases} x_{j,0}, & \text{if } |x_{j,0}| > \mu_j, \\ 0, & \text{if } |x_{j,0}| \leq \mu_j, \end{cases} \quad j = 1, \dots, n.$$

La première étape vise à assurer le contrôle de l'erreur d'estimation  $\|\hat{x}_n - x_0\|_n$  par l'erreur de l'oracle  $\|x_* - x_0\|_n$ . Pour cela, soit

$$V_j = \frac{1}{n} \sum_{i=1}^n \varphi_j(t_i) \epsilon_i, \quad j = 1, \dots, n.$$

Définissons l'événement  $B_n$  comme l'ensemble pour lequel

$$B_n = \left\{ \max_{j=1, \dots, n} |V_j| \leq c \sqrt{\frac{\log n}{n}} \right\}.$$

Définissons l'ensemble d'indices comme  $\mathcal{J}_n = \text{Card}\{j, |x_{j,0}| > \mu_j\}$ . Alors, nous pouvons montrer que la fonction de perte a été choisie de telle manière à assurer que sur  $B_n$ , nous avons

$$\|\hat{x}_n - x_0\|_n^2 \leq \|x_* - x_0\|_n^2 + 4c\sqrt{\frac{\log n}{n}} \sum_{j \in \mathcal{J}_n} \frac{|\hat{x}_j - x_{j,0}|}{\lambda_j}. \quad (3.6)$$

Sous l'hypothèse

$$\max_{i=1,\dots,n} \mathbf{E} \exp[\epsilon_i^2/K^2] \leq K,$$

nous pouvons montrer (VAN DE GEER (2000, Lemma 8.2)) que

$$\mathbf{P} \left( \max_{j=1,\dots,n} |V_j| > c\sqrt{\frac{\log n}{n}} \right) \leq c \exp\left[-\frac{\log n}{c^2}\right]. \quad (3.7)$$

Ainsi

$$\mathbf{P} \left( \|\hat{x}_n - x_0\|_n^2 > c_1 \|x_* - x_0\|_n^2 + 4c_2 \frac{\log n}{n} \sum_{j \in \mathcal{J}_n} \frac{1}{\lambda_j^2} \right) \leq c \exp\left[-\frac{\log n}{c^2}\right].$$

L'hypothèse  $x_0 \in X_{s,p}$  permet de contrôler le terme de reste  $\sum_{j \in \mathcal{J}_n} \frac{1}{\lambda_j^2}$ .

Le présent théorème prouve l'optimalité de cette méthode lorsque l'indice  $t$  est petit, donc pour des problèmes faiblement mal posés et explique pourquoi l'estimation pénalisée par une norme  $l^1$  atteint vite ses limites en pratique comme l'ont remarqué de nombreux utilisateurs de cette méthode. Nous retrouvons les résultats obtenus dans [26] et [32], connus sous le nom de méthode d'estimation de Galerkin.

Trouver une bonne méthode de régularisation est par conséquent crucial. C'est pourquoi notre choix s'est ensuite porté sur les méthodes de sélection de modèles, décrites précédemment dans la partie 1.

Considérons l'estimateur obtenu par sélection de modèles de la manière suivante

$$\hat{x}_n = \arg \min_{(x,\alpha) \in \mathcal{X} \times \Theta} (\gamma_n(y - F(x)) + \text{pen}_\alpha(x, \mathcal{X}) + \text{pen}(\alpha)), \quad (3.8)$$

où  $\mathcal{X}$  est un espace de Hilbert,  $\gamma_n(\cdot)$  est une fonction de perte empirique,  $\alpha = (\alpha_k)_{k \geq 1} \in \Theta$  est une suite décroissante de paramètres et  $\text{pen}_\alpha(\cdot, \cdot)$  une pénalité portant à la fois sur l'espace  $\mathcal{X}$  et la fonction  $x$  de  $\mathcal{X}$ . L'objectif est de trouver un estimateur proche des données, au sens défini par  $\gamma$ , possédant des propriétés de régularité définies par les termes de pénalité  $\text{pen}_\alpha(\cdot, \cdot)$  et  $\text{pen}(\cdot)$ . La suite de paramètres  $\alpha$  équilibre la contribution de tous ces termes. Plus  $\alpha_k$  est grand, plus l'estimateur sera régulier, au sens donné précédemment tandis que plus  $\alpha_k$  est petit et plus l'estimateur sera proche des données et, sans doute sera trop irrégulier.

On peut considérer deux types de pénalités  $\text{pen}_\alpha(\cdot, \cdot)$  : des pénalités sur la dimension, l'estimateur obtenu est par conséquent un estimateur de type estimateur de projection, ou bien, des pénalités sur la régularité, permettant de définir des estimateurs régularisés.

- Dans le premier cas, l'idée développée dans [8], consiste à regarder une suite de sous-espaces de complexité croissante (la complexité étant définie en terme de dimension de l'espace). Ainsi, sur chaque espace, nous pouvons étudier l'estimateur obtenu par projection.
- Dans le second cas, la fonction de régularisation  $\text{pen}_\alpha(x, \mathcal{X}) = J_\alpha(x)$  (souvent une fonction quadratique) est utilisée comme fonction de pénalité, pondérée par des poids. Cette méthode est connue sous le nom de régularisation de Tikhonov.

Dans les deux cas, le choix de la suite de paramètres  $\alpha$  détermine le comportement asymptotique de l'estimateur. Le meilleur choix (celui permettant d'estimer la fonction à une vitesse optimale) dépend d'hypothèses inconnues sur la fonction  $x_0$  (notamment sa régularité ou le lien entre sa régularité et la difficulté du problème inverse). Ces estimateurs sont bien étudiés dans la littérature des problèmes inverses, notamment dans [39] ou [47], pour des opérateurs linéaires. Dans le cadre d'opérateurs non linéaires, il existe peu de résultats sur la convergence des méthodes de régularisation. Dans [16], une adaptation de la méthode de Tikhonov est étudiée.

Notre démarche vise donc, d'une part, à adapter de telles méthodes à l'estimation dans le cadre des problèmes inverses. Mais, en même temps, nous cherchons à choisir une manière de sélectionner automatiquement la suite de paramètres, sans a priori sur la fonction à estimer, de telle manière que l'estimateur obtenu possède les mêmes propriétés que si l'on connaissait le meilleur choix à faire. L'intérêt de ce type de procédure d'estimation est, d'une part de donner une procédure adaptative d'estimation (donc facilement utilisable en pratique), et d'autre part, de permettre de généraliser ces résultats au cas d'opérateurs non linéaires.

Nous pouvons alors nous placer dans le cadre des problèmes inverses en écrivant le modèle sous la forme

$$y_i = A(x_0(t_i)) + \epsilon_i, \quad i = 1, \dots, n.$$

Pour un ensemble d'indices  $\mathcal{M}_n$ , considérons une collection de modèles  $(Y_m)_{m \in \mathcal{M}_n} \subset Y$ , de dimension  $d_m$  et projetons les données sur ces sous espaces. Pour  $d_{m_0}$  choisi indépendamment de la fonction inconnue  $x_0$ , nous noterons  $\Pi_{Y_{m_0}}$  la projection sur l'espace  $Y_{m_0}$  par rapport à la métrique empirique  $\mathbf{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{t_i}$ .

Le premier type d'estimateur que nous considérons est l'estimateur de sélection de modèles, défini comme la solution de la minimisation suivante :

$$\hat{x}_{\hat{m}} = \arg \min_{m \in \mathcal{M}_n} \left( \arg \min_{x \in X_m} \|\Pi_{X_{m_0}}^n (y - A(x))\|_n^2 + \text{pen}(m) \right) \quad (3.9)$$

avec  $\text{pen}(m) > r(1 + L)\sigma^2 \sum_j \lambda_j + \sup_j \lambda_j$ ,  $r = 2 + \theta$ , pour  $\theta > 0$ ,  $L > 0$  et  $\lambda_j$  les valeurs propres de  $(A\Pi_{Y_m})^+$ . Pour un opérateur  $F$ , nous avons noté  $F^+$  l'inverse généralisé de Moore-Penrose de  $F$ , utilisant la notation donnée dans [46]. Supposons que la fonction  $x_0$  possède la régularité donnée par la condition :

### SC condition source

Il existe un paramètre  $0 < \nu \leq 1/2$  tel que  $x_0 \in \text{Range}((T^*T)^\nu) := \mathcal{R}((T^*T)^\nu)$ .

Cette condition est couramment utilisée en analyse, cf par exemple [46] ou en économétrie [18]. Elle relie la régularité de la fonction à la décroissance des valeurs propres de l'opérateur. Sous des conditions techniques nous sommes parvenus au théorème suivant

**Théorème 3.2.2** (Inégalité oracle pour sélection de modèles dans les problèmes inverses). .

*Il existe une constante  $C$  telle que*

$$\|\hat{x}_{\hat{m}} - x_0\|^2 \leq 2\|(I - \Pi_{X_{m_0}})x_0\|^2 + C \inf_{m \in \mathcal{M}_n} [d_m^{-2t\nu} + 2(d_m^{1+2t}n)^{-1}] + \frac{\Sigma(d)}{n},$$

avec  $L_k$  une collection de poids et

$$\Sigma(d) = \sum_{k \in \mathcal{K}} 2 \left[ \sqrt{\frac{d \text{Tr}(R_{A_k}^t R_{A_k})}{\rho^2(R_{A_k})}} + 1 \right] \frac{d}{\rho^2(nR_{A_k})} e^{-\sqrt{d L_k [\text{Tr}(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})] / \rho^2(R_{A_k})}}.$$

Ce théorème établit une inégalité oracle pour l'estimateur par sélection de modèles, en d'autres termes, que l'estimateur (3.9) est optimal, au sens où il fait aussi bien que le meilleur estimateur contenu dans tout l'ensemble des estimateurs considérés.

En outre, dès que la collection d'espaces  $M_n$  est telle que  $m \in M_n$  avec  $d_m = n^{1/(4\nu p + 2p + 1)}$ , nous retrouvons la vitesse optimale  $n^{-\frac{2\nu p}{4\nu p + 2p + 1}}$ , optimalité prouvée dans [20]. Ainsi, il existe bien un bon choix de modèles  $Y_m$ , :  $m \in M_n$  pour lesquels l'estimateur de sélection de modèles est optimal au sens donné par la théorie minimax.

Nous pouvons remarquer que nous avons donné l'interprétation suivante de la procédure d'estimation pénalisée. L'équation (3.9) revient finalement à minimiser

$$\hat{x}_{\hat{m}} = \arg \min_{m \subset m_0} \arg \min_{x_m \in X_m} \{-2 \langle A_{m_0} y, A_{m_0} T x_m \rangle + \|A_{m_0} T x_m\|^2\} + \text{pen}(m) \quad (3.10)$$

$$= \arg \min_{m \subset m_0} \arg \min_{x_m \in X_m} \{-2 \langle \Pi_{X_m} A_{m_0} y, x_m \rangle + \|x_m\|^2\} + \text{pen}(m). \quad (3.11)$$

Soit  $\{e_j\}_{j \in m}$  la base canonique sur l'ensemble  $X_m = (\Pi_{Y_m}^n)^* X$ . Définissons pour tout  $m$ ,

$$x_{m,j} = \langle A_{m_0} y, e_j \rangle = \langle \mathbf{y}, A_{m_0}^t e_j \rangle, \quad j = 1, \dots, m.$$

Alors, la procédure d'estimation revient à projeter les données sur  $X_{\hat{m}}$  où  $\hat{m}$  est choisi en fonction des données, c'est-à-dire, en minimisant le contraste suivant en  $m$

$$-\sum_{j \in m} x_{m,j}^2 + r\sigma^2(1 + L_m) \left[ \sum_{j \in m} \lambda_j + \sup_{j \in m} \lambda_j \right].$$

Nous reconnaissons le contraste lié aux méthodes de seuillage fort, voir par exemple [8].

Par conséquent l'estimateur obtenu sans connaissances a priori converge à une vitesse inférieure ou égale à la vitesse que l'on obtiendrait avec un estimateur choisi en connaissant a priori la régularité de la fonction à estimer.

Dans un second temps, nous nous sommes intéressés aux estimateurs construits en minimisant une perte quadratique et une pénalité, c'est-à-dire aux estimateurs régularisés de

type Tikhonov. Soit  $\mathcal{K}_n$  une suite d'indices, pour tout  $k \in \mathcal{K}_n$  et pour toute suite de matrices diagonales  $A_k(n) \in \mathbb{M}_{m_0 \times m_0}(\mathbb{R})$ , considérons l'estimateur

$$\hat{x}_{A_k(n)} = \arg \min_{x \in X_m} [\|\Pi_{Y_{m_0}}(y - A(x))\|_n^2 + \|A_k(n)x\|^2]. \quad (3.12)$$

Définissons aussi l'opérateur de régularisation suivant

$$R_{A_k(n)} = (T^t \Pi_{Y_{m_0}} T + A_k(n)^t A_k(n) I_{m_0})^{-1} T^t \Pi_{Y_{m_0}}.$$

Posons aussi  $\gamma(x, \alpha_k) = \|R_{A_k}(y - F(x))\|^2$  et construisons la pénalité

$$\text{pen}(A_k) = r\sigma^2(1 + L_k)[\text{Tr}(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})],$$

with  $r > 2$ .

L'estimateur  $\hat{x}_{A_k}$  peut aussi s'écrire

$$\hat{x}_{A_k} = \arg \min_{k \in \mathcal{K}, x \in X_{m_0}} (\gamma(x, A_k(n)) + \text{pen}(A_k(n))).$$

Posons  $x_{A_k} = R_{A_k} T x_0$  et

$$\Sigma(d) = \sum_k 2 \left[ \sqrt{\frac{d \text{Tr}(R_{A_k}^t R_{A_k})}{\rho^2(R_{A_k})} + 1} \right] \frac{d}{\rho^2(n R_{A_k})} e^{-\sqrt{d L_k [\text{Tr}(R_{A_k}^t R_{A_k}) + \rho^2(R_{A_k})] / \rho^2(R_{A_k})}},$$

pour  $d$  choisi selon des hypothèses techniques. Le théorème suivant décrit la convergence de cet estimateur

**Théorème 3.2.3** (Inégalité oracle pour estimateur régularisé dans le cadre des problèmes inverses).

*Il existe une constante  $C$  telle que*

$$\mathbf{E} \|\hat{x}_{A_k} - x_0\|^2 \leq 2 \|(I - \Pi_{X_{m_0}})x_0\|^2 + C \inf_{k \in \mathcal{K}} [\|x_{A_k} - x_0\|^2 + 2\text{pen}(A_k)] + \frac{\Sigma(d)}{n}, \quad (3.13)$$

Nous constatons alors que l'estimateur est optimal au sens de la sélection de modèles puisque l'erreur de l'estimateur est inférieure au minimum des risques des estimateurs régularisés, à un terme d'erreur près. Plus précisément, nous pouvons supposer que  $\mathcal{K}$  est tel que  $d_{m_0} \geq \sup_{k \in \mathcal{K}} \alpha_k^{-1/(2p)}$ . Alors, sous certaines hypothèses techniques, nous avons, en calculant la trace et le rayon spectral de l'opérateur de régularisation  $R_{A_k}$ , pour  $\nu \leq 1$

$$\frac{\text{Tr}(R_{A_k}^t R_{A_k})}{\rho^2(R_{A_k})} = O(\alpha_k^{-1/(2p)}),$$

et l'inégalité devient à présent

$$\mathbf{E} \|\hat{x}_{\alpha_k} - x_0\|^2 \leq 2 \|(I - \Pi_{X_{m_0}})x_0\|^2 + C \inf_{k \in \mathcal{K}} [\alpha_k^{2\nu} + 2(\alpha_k^{1+1/2p} n)^{-1}] + \frac{\Sigma(d)}{n}. \quad (3.14)$$

L'inégalité suivante montre l'optimalité au sens minimax de l'estimateur adaptatif pour  $\nu \leq 1$ , c'est-à-dire

$$\mathbf{E}\|\hat{x}_{A_{\hat{k}}} - x_0\|^2 \leq Cn^{-\frac{4p\nu}{1+4p\nu+2p}}.$$

Nous avons démontré ces théorèmes dans les articles [49] et [89]. Sur la base de la méthode de Tikhonov, nous avons donc construit une procédure de choix de paramètre adaptative qui garantit que, sans connaître la vraie régularité de la fonction à estimer, nous pouvons malgré tout choisir le meilleur paramètre parmi tous les choix à notre disposition. Il apparaît bien que ces résultats permettent de construire des estimateurs adaptatifs (donc utilisables en pratique) à partir d'estimateurs plus simples et bien connus dans la littérature des problèmes inverses.

En particulier, ces théorèmes précisent la pénalité qui doit être choisie pour choisir les paramètres optimaux. Ce choix provient du résultat de concentration suivant que nous avons démontré pour des opérateurs linéaires, en procédant à une extension du résultat de Y. Baraud in [5]

**Théorème 3.2.4** (Inégalité de concentration pour des opérateurs linéaires).

Pour  $\eta(A) = \sup_{\|u\|=1} \sum_{i=1}^n \varepsilon_i(A^t u)_i$ , alors

$$\begin{aligned} P(\eta^2(A) \geq \sigma^2[Tr(A^t A) + \rho(A^t A)]r/2(1+L) + \sigma^2 u) \\ \leq \exp\{-\sqrt{d(1/\rho(A^t A)u + r/2L[Tr(A^t A)/\rho(A^t A) + 1])}\}. \end{aligned}$$

Ainsi le processus empirique se concentre autour de sa moyenne, obtenue à partir de la trace et du rayon spectral de l'opérateur  $A$ . Le calcul de la trace et du rayon spectral nous a permis de prouver les vitesses de convergence des estimateurs précédents. Ce résultat peut être aisément étendu à d'autres schémas de régularisation, comme la régularisation de Landweber par exemple, ou autres méthodes itératives.

Dans un second temps, nous avons généralisé ces estimateurs au cas, peu traité dans la littérature où les opérateurs ne sont pas linéaires. Il est pour cela nécessaire de linéariser le problème autour d'un point bien choisi de l'espace  $X$ , qui sera noté  $x^*$ . Plus particulièrement supposons, comme dans [73].

**AF contrôle sur la partie non linéaire de la différentielle de l'opérateur**

Il existe une constante  $c_T$ , un opérateur linéaire  $T$  (souvent  $T = F'(x^*)$ ) et un opérateur linéaire dépendant de  $x$  et de  $x'$ , qui sera nommé  $R(x, x')$  tel que pour tout  $x, x' \in B_\rho(x^*)$

$$F(x) - F(x') = TR(x, x')(x - x'),$$

avec  $\|I - R\| \leq c_T$ .

Cette condition permet d'obtenir

$$\mathcal{R}(F'(x^*))^* \subset \mathcal{R}(F'(x_0))^*$$

et nous conduit à légitimer l'hypothèse de régularité portant sur la condition source. Or, cette hypothèse est souvent vérifiée dans de nombreux cas pratiques, voir par exemple [63, 72].

Nous considérons la version non linéaire de l'estimateur étudié précédemment, obtenue en remplaçant  $F'$  par  $T_{m_0} = \Pi_{Y_{m_0}}^n T$

$$\hat{x}_{\alpha_k} = x^* + \arg \min_{x \in X_m} \|(T_{m_0}^t T_{m_0} + \alpha_k I_{m_0})^{-1} T_{m_0}^t (y - T(x))\|^2. \quad (3.15)$$

Le théorème 3.2.3 se généralise alors en le théorème suivant

**Théorème 3.2.5.** *Supposons que toutes les hypothèses précédentes sont satisfaites. Pour tout  $x \in x^* + X_m$  et tout réel  $k$  tel que  $d_{m_0} \geq \alpha_k^{-1/(2p)}$ , l'inégalité suivante est vérifiée*

$$\mathbf{E} \|\Pi_{X_{m_0}}(\hat{x}_{\alpha_k} - x_0)\|^2 \leq \frac{1}{1 - c_T} \inf_{k \in \mathcal{K}} [C(1 + c_T) \|\Pi_{X_{m_0}}(x_{\alpha_k} - x_0)\|^2 + 2\text{pen}(\alpha_k)] + \frac{\Sigma(d)}{(1 - c_T)n}, \quad (3.16)$$

où la quantité  $\Sigma(d)$  est définie précédemment.

Néanmoins, dans le cas non linéaire, le terme de biais  $\|\Pi_{X_{m_0}}(x_{\alpha_k} - x_0)\|^2$  peut être élevé. Ainsi nous nous sommes intéressés à d'autres types d'estimateur, notamment les estimateurs de sélection de modèles.

Dans ce cadre nous nous plaçons dans le cas où le design est orthonormal. Il s'agit ici d'une hypothèse qui n'est pas fondamentale mais qui permet d'énoncer les résultats plus facilement.

**ON** Supposons que les observations  $\{t_i\}_{i=1}^n$  sont telles que la matrice de Gram correspondante vérifie  $G_{m_0}^t G_{m_0} = I_{m_0}$ .

Posons  $A_{m_0} = T_{m_0}^+ \Pi_{Y_{m_0}}^n$ . Pour toute fonction fixée  $x$ , considérons l'opérateur linéaire  $D_m(x) := \Pi_{X_{m_0}} R(x) \Pi_{X_m}$ , défini sur les espaces  $D_m(x) : X_m \rightarrow X_{m_0}$ . Soit  $S_m$  la matrice dont les éléments sont donnés par

$$S_m(i, j) = \sup_{x \in B_\rho(x^*)} |A_{m_0}^t D_m(x)(i, j)|.$$

Définissons aussi  $\rho_m = \rho(S_m^t S_m)$  et  $t_m = \text{Tr}(S_m^t S_m)$ . Écrivons également  $R_m = t_m / \rho_m$ . Nous remarquons que, dans ce cas, ni  $n\rho_m$  ni  $nt_m$  ne dépendent de  $n$ . Introduisons alors  $L_m$  un poids et utilisons la notation  $\Sigma_i = \Sigma_i(m_0)$ ,  $i = 1, 2$  où nous avons posé

$$\Sigma_1 = \sum_{m \subset m_0} e^{-\sqrt{d/2r L_m (R_m + 1)}} \quad (3.17)$$

où pour tout  $q \geq 2$ ,  $C_q$  est une constante qui ne dépend que de  $q$ .

$$\Sigma_2(q) = \sum_{m \subset m_0} C_q (n\rho_m \sigma^2 / d)^q [(d/2r L_m (R_m + 1))^{q-1/2} + (d/2r L_m (R_m + 1))^{q-1}] e^{-\sqrt{d/2r L_m (R_m + 1)}} \quad (3.18)$$

Considérons enfin l'estimateur pénalisé de sélection de modèle avec la pénalité donnée par

$$\text{pen}(m) = r\sigma^2(1 + L_m)[t_m + \rho_m],$$

with  $r > 2$ . L'estimateur prend donc la forme suivante

$$\hat{x}_{\hat{m}} = x^* + \arg \min_{(m, x_m) \in m_0 \times X_m} \|A_{m_0}(y - F(x_m))\|^2 + \text{pen}(m). \quad (3.19)$$

Le théorème qui suit donne une inégalité oracle et prouve ainsi l'optimalité de cet estimateur.

**Théorème 3.2.6.** *Supposons que l'hypothèse (3.17) est satisfaite. Alors pour  $0 < \kappa < 1$ , avec probabilité supérieure à  $1 - \Sigma_1 e^{-\sqrt{d/(2n\rho_{m_0})}u}$ , nous avons*

$$\|\hat{x}_{\hat{m}} - x_0\|^2 \leq \|(I - \Pi_{X_{m_0}})(x_0 - x^*)\|^2 \quad (3.20)$$

$$+ \frac{2}{(1 - c_T - \kappa)} \inf_{m \subset m_0} \{ \arg \min_{x_m \in X_m} (1 + c_T) \|x_m - x_0\|^2 + \text{pen}(m) \} + \sigma^2 u/n \quad (3.21)$$

et pour  $q \geq 1$

$$\mathbf{E}[\|\hat{x}_{\hat{m}} - x_0\|^{2q}] \leq [\|(I - \Pi_{X_{m_0}})(x_0 - x^*)\|^2] \quad (3.22)$$

$$+ \frac{2}{1 - c_T - \kappa} \inf_{m \subset m_0} \{ \arg \min_{x_m \in X_m} (1 + c_T) \|x_0 - x_m\|^2 + \text{pen}(m) \}^q + \frac{\Sigma_2(q)}{n^q} \quad (3.23)$$

**Remarque 3.2.7.** Les résultats précédents dépendent de deux facteurs : la trace de la partie linéaire de l'opérateur de régularisation  $t_m$  et le rayon spectral  $R_m$ . Toutefois pour une fonction  $x$ , nous pouvons écrire que

$$A_{m_0} F'(x)w = A_{m_0} T R(x)w = \Pi_{X_{m_0}} R(x)w.$$

C'est ainsi que si nous connaissons des propriétés de la différentielle  $F'(\cdot)$ , il nous est possible d'obtenir un contrôle de cette quantité.

Ces résultats sont difficiles à énoncer de prime abord, mais, pour mieux comprendre le terme de pénalité, examinons le cas où  $F$  est un opérateur linéaire, c'est-à-dire, dans nos notations supposons que,  $F = T$ . Considérons  $b_j, \phi_j, \psi_j$  sa décomposition en valeurs singulières (SVD). Nous rappelons alors que nous pouvons écrire les formules suivantes,  $T\phi_j = b_j\psi_j$  et  $T^*\psi_j = b_j\phi_j$ . Soient  $Y_m$  la collection d'espaces linéaires générés par les fonctions de base  $(\phi_j)_{j \in m}$ , pour un ensemble d'indices  $m$ . Alors sous l'hypothèse **ON**, la pénalité  $\text{pen}(m)$  est proportionnelle au terme

$$\text{pen}(m) \sim 1/n \sum_{j \in m} 1/b_j^2$$

tandis que  $R_m$  est proportionnel à

$$R_m \sim \sup_{j \in m} 1/b_j^2 \times \sum_{j \in m} 1/b_j^2.$$

L'estimation par pénalité est donc équivalente à une procédure d'estimation par seuillage dur définie de la manière suivante : on garde les coefficients tels que

$$(\hat{x}_{\hat{m}})_j \geq \sqrt{\frac{Cn}{b_j^2}}, \forall j \geq 1$$

où la constante dépend de la dimension  $m_0$ .

Dans le cas non linéaire, l'estimateur s'écrit d'une façon plus complexe. Pour toute collection d'ensembles d'approximation,  $Y_m$ , la pénalité  $\text{pen}(m)$  dépendra :

- de la discrétisation du problème sous la forme de la matrice de Gram  $G_{m_0}$ ,
- de la façon dont la partie linéaire de l'opérateur  $F$ ,  $T$  agit sur l'espace maximal de projection  $Y_{m_0}$ , c'est-à-dire de la quantité,  $\inf_{v \in Y_{m_0}} \|T^*v\|/\|v\|$ ,
- et finalement, de la non linéarité de l'opérateur  $F$ , déterminée par l'écart entre la trace de la partie linéaire de l'opérateur et la trace effective,  $|t_m - \text{Tr}(A_{m_0}^t \Pi_{X_m})|$ .

Par conséquent, nous avons construit une procédure permettant d'estimer des fonctions observées à travers leur image par un opérateur non linéaire connu, en présence d'un bruit d'observation. Sans connaître a priori leur régularité, nous obtenons des estimateurs optimaux au sens où nous avons le même type d'inégalités oracle que dans le cadre linéaire.

Nous pouvons maintenant nous interroger sur le prix à payer pour résoudre ces problèmes inverses non linéaires.

Afin de mettre en œuvre notre méthode d'estimation, il est nécessaire de linéariser l'opérateur, mais les propriétés des différentielles dépendent du point en lequel, elles sont calculées. Ainsi nous avons besoin de connaître un point  $x^*$  autour duquel nous linéarisons l'opérateur. Ce point doit être bien choisi puisque nous allons chercher les solutions dans un voisinage de ce point  $B_\rho(x^*) = \{x, \|x - x^*\| \leq \rho\}$ , pour un certain  $\rho > 0$ . *Bien choisi* signifie tout d'abord que nous devons contrôler l'approximation linéaire de l'opérateur dans ce voisinage. C'est le sens de l'hypothèse **AF**. Si  $x^*$  est choisi tel que la vraie fonction vérifie  $x_0 \in B_\rho(x^*)$ , les propriétés locales de l'opérateur non linéaire  $F$  seront données par celles de l'opérateur linéaire  $T$ . Plus précisément, la difficulté de l'estimation relative au problème inverse mal posé dépendra du caractère mal posé de l'opérateur linéaire  $F'(x^*)$ .

En outre, le problème doit posséder une solution unique et être identifiable. C'est pourquoi nous supposons que nous pouvons choisir ce point initial  $x^*$  de telle manière qu'il vérifie la condition suivante

### IG Condition d'identifiabilité

$$x_0 - x^* \in \text{Ker}(T)^\perp,$$

où  $\text{Ker}(T)^\perp$  est le complément orthogonal du noyau de l'opérateur  $T$ .

Dès lors, le résultat suivant ([73]) assure l'unicité de la solution au problème d'estimation, si le point de départ est suffisamment proche

**Lemma 3.2.8.** *Sous l'hypothèse **AF** avec  $c_T < 1/2$ , considérons deux points  $x, x' \in B_\rho(x^*)$ , vérifiant  $F(x) = F(x')$  et  $x - x' \in \text{Ker}(T)^\perp$ . Alors  $x = x'$ .*

Dans la plupart des méthodes d'estimation en analyse, il est indispensable de connaître un tel point de départ, voir par exemple les travaux de [105] et une telle restriction n'est pas rédibitoire. En pratique, ce point de départ idéal peut être approché par plusieurs itérations successives de l'algorithme d'estimation. Ce type d'approche est celui utilisé dans tous les travaux portant sur les méthodes itératives d'estimation, [36] ou [71].

Finalement, sous ces deux hypothèses, nos travaux prouvent que les problèmes inverses non

linéaires peuvent être résolus avec le même type d'inégalités oracle (ou de manière équivalente avec les mêmes vitesses de convergence), que pour les problèmes inverses linéaires, en utilisant l'estimateur que nous proposons.

### 3.3 Exemples de problèmes inverses non linéaires

Nous avons étudié les problèmes inverses non linéaires d'un point de vue théorique en obtenant des vitesses optimales de convergence. Ces méthodes sont utilisables en pratique dès que l'on connaît l'opérateur  $F$ , à travers lequel la solution est déformée. Cependant, dans la plupart des problèmes rencontrés en pratique, l'opérateur est soit totalement inconnu, soit seules des approximations de l'opérateur sont connues. C'est tout particulièrement le cas en économétrie et en sismographie. Nous détaillons deux applications relatives à ces domaines.

#### 3.3.1 Application en économétrie

Nous nous sommes intéressés en économétrie à l'estimation de la variation du surplus du consommateur. Ce problème, détaillé dans [112] ou [66], peut être modélisé sous la forme du problème inverse suivant : estimer la solution  $\lambda$  d'une équation différentielle

$$\begin{cases} \lambda'(x) &= m(x, \lambda(x)) \\ \lambda(x_0) &= \lambda_0, \quad (x_0, \lambda_0) \in \mathbb{R}^2 \end{cases} \quad (3.24)$$

alors que la fonction  $m$  n'est pas observée directement. En effet, elle n'est connue qu'à travers les observations  $Y_1, \dots, Y_n$  qui proviennent du modèle

$$Y = m(Z) + U \quad (3.25)$$

où  $U$  désigne un bruit d'observation et  $Z$  les points aléatoires où est mesurée la fonction  $m$ . La particularité de ce modèle vient du fait que, contrairement à un modèle de régression, nous avons

$$bfE(U|Z) \neq 0.$$

Les variables sont dites endogènes. Toutefois, en économétrie, nous supposons qu'il existe une variable  $W$ , telle que  $E(U|W) = 0$ . Une telle variable est connue en économie sous le nom de variable instrumentale. Elles jouent un rôle capital en économie, décrit dans [18] par exemple.

Ce problème est difficile puisque, pour le résoudre, nous devons procéder en 2 étapes successives.

1. Estimer la fonction  $m$  par son estimateur  $\hat{m}_n$ . Cette étape constitue le premier problème inverse. En effet, les observations ne sont accessibles qu'à travers leur image par l'espérance conditionnelle par rapport à la variable instrumentale  $W$ .

Plus précisément, soit  $S = (Y, Z, W)$  la variable aléatoire formée de  $Y \in \mathbb{R}$ ,  $Z \in \mathbb{R}^2$  et  $W \in \mathbb{R}^2$ . Considérons alors l'espace de Hilbert  $L^2$ , des fonctions intégrables de  $S$

et notons par  $L^2(Z)$  et  $L^2(W)$  les sous espaces de  $L^2$  des fonctions à valeurs réelles de fonctions respectivement de  $Z$  et de  $W$ . Définissons les opérateurs suivants

$$\begin{aligned} T : L^2(Z) &\longrightarrow L^2(W) \\ g &\longrightarrow T(g(Z)) = \mathbf{E}[g(Z)|W] \end{aligned}$$

$$\begin{aligned} T^* : L^2(W) &\longrightarrow L^2(Z) \\ h &\longrightarrow T^*(h(W)) = \mathbf{E}[h(W)|Z] \\ r &= \mathbf{E}(Y|W) \end{aligned}$$

Ces opérateurs sont les espérances conditionnelles par rapport aux variables  $W$  et  $Z$ . Ces opérateurs linéaires vérifient :

$$\langle T(\phi(Z)), \psi(W) \rangle = \langle \phi(Z), T^*(\psi(W)) \rangle = \mathbf{E}(\phi(Z)\psi(W)).$$

Nous avons supposé que les densités de  $Z$  et de  $W$  ne s'annulent pas sur leur domaine de définition. Il s'avère par conséquent que les opérateurs  $T$  et  $T^*$  sont bien définis. Dès lors le modèle de régression à variable instrumentale peut être réécrit de la manière équivalente :

$$Tm - r = 0. \quad (3.26)$$

L'opérateur  $T$  est linéaire, mais dépend de la loi des données et doit donc être estimé. Sous certaines hypothèses de régularité, dans [31], nous montrons qu'il existe une unique solution sous la forme  $m = \hat{\Psi}(r)$ . Nous utilisons à cet effet un estimateur à noyaux de la densité pour estimer l'opérateur  $T$ . Considérons une réalisation i.i.d. du vecteur  $S$  :  $S_i := (x_i, y_i, z_i)$  pour  $i = 1, \dots, n$  et un noyau  $K$  de paramètre de lissage  $h_n$ . Construisons alors un estimateur à noyaux,  $\hat{f}_n$  estimateur de la densité  $f$  de

$$\hat{f}_n(y, z, w) = \frac{1}{n} \sum_{i=1}^n K_{y, h_n^y}(y - Y_i) K_{z, h_n^z}(z - Z_i) K_{w, h_n^w}(w - W_i)$$

Soit  $\hat{T}$  l'estimateur de l'opérateur associé.

Dès lors, nous pouvons utiliser les méthodes usuelles de résolution des problèmes inverses, par exemple la méthode de Tikhonov. Ainsi, pour une suite  $\alpha_n \rightarrow 0$ , définissons  $\hat{m}_{n, \alpha_n}$  comme solution de

$$(\alpha_n I + \hat{T}^* \hat{T}) \hat{m}_{n, \alpha_n} = \hat{r}^*. \quad (3.27)$$

$\hat{m}_{n, \alpha_n}$  est un estimateur de la fonction  $m$ . Cet estimateur a été étudié dans [31].

2. La résolution de l'équation différentielle constitue la deuxième étape. La fonction  $\lambda$  est définie par une relation implicite non linéaire. Appelons  $A$  l'opérateur défini par :

$$A(m, \lambda) = \lambda' - m(\cdot, \lambda).$$

Résoudre l'équation différentielle (3.24) équivaut à inverser l'opérateur  $A$  sous la condition initiale  $\lambda(0) = 0$ . Si  $m$  est connu, il existe une unique solution :  $\lambda(x) = \Phi[m](x)$ .

Mais comme  $m$  est inconnue, nous la remplaçons par son estimateur  $\hat{m}_{n,\alpha_n}$ , et aboutissons au système différentiel suivant

$$\begin{cases} \lambda'(x) &= \hat{m}_{n,\alpha_n}(x, \lambda(x)) \\ \lambda(0) &= 0 \end{cases} \quad (3.28)$$

dont l'estimateur  $\hat{\lambda}_{n,\alpha_n}$  est solution.

Dans [97], nous montrons que l'estimateur de ces doubles problèmes inverses peut s'écrire sous la forme suivante

$$\hat{\lambda}_{n,\alpha_n} = \Phi[\hat{m}_{n,\alpha_n}](x).$$

En fait, l'idée forte, à la base de l'étude de cet estimateur, consiste à utiliser un théorème proche du théorème de Cauchy-Lipschitz d'une part, et le théorème des fonctions implicites d'autre part, pour écrire la décomposition suivante

$$\forall x \in I, \hat{\lambda}_{n,\alpha_n}(x) - \lambda(x) = \int_0^x \left( (\hat{m}_{n,\alpha_n} - m)(t, \lambda(t)) \cdot e^{\int_t^x \frac{\partial}{\partial y} m(u, \lambda(u)) du} \right) dt + R_{2,n}(x)$$

avec  $R_{2,n}(x) = O_P(\|\hat{m}_{n,\alpha_n} - m\|^2)$  and  $\|\hat{m}_{n,\alpha_n} - m\|^2 = \int \int \mathbf{1}_D (\hat{m}_{n,\alpha_n} - m)^2(x, y) dx dy$  avec  $D = \{(x, y); |x| \leq a, |y| \leq b\}$  un voisinage compact de  $(0, 0)$ . Ainsi la convergence de l'estimateur  $\hat{\lambda}_{n,\alpha_n}$  est reliée à la convergence du premier estimateur  $\hat{m}_{n,\alpha_n}$ . Nous pouvons alors prouver le théorème suivant qui donne la vitesse de convergence de la partie linéaire de l'estimateur  $H[\hat{m}_{n,\alpha_n} - m](x)$ , où nous avons noté

$$H[\hat{m}_{n,\alpha_n} - m](x) = \int_0^x (\hat{m}_{n,\alpha_n} - m)(t, \lambda(t)) v(x, t) dt$$

avec  $v(x, t) = \exp\left(\int_t^x \frac{\partial}{\partial y} m(u, \lambda(u)) du\right)$ .

**Théorème 3.3.1** (Convergence de l'estimateur solution du système différentiel avec effets endogènes). *Supposons qu'il existe un paramètre  $0 \leq \beta \leq 2$ , tel que*

$$m \in \Phi_\beta := \{m, \|m - m_\alpha\|^2 = O(\alpha^\beta)\}.$$

*Alors si le paramètre de lissage du noyau est choisi tel que  $\frac{nh_n^{\frac{3}{2}+2}}{\log^2 n} \rightarrow \infty$ , alors quand le nombre d'observations  $n \rightarrow +\infty$ , nous obtenons la vitesse de convergence pour la norme quadratique*

$$\|H[\hat{m}_{n,\alpha_n} - m]\|^2 = \begin{cases} O_{\mathbf{P}}\left(n^{-\frac{\beta}{2+\beta}}\right), & \text{if } 0 \leq \beta \leq 1 \\ O_{\mathbf{P}}\left(n^{-\frac{3\beta-2}{2(\beta+2)}}\right), & \text{if } \beta > 1. \end{cases} \quad (3.29)$$

Ainsi, nous sommes arrivés, dans ce cas particulier de problèmes inverses non linéaires, à obtenir une vitesse de convergence de l'estimateur. Néanmoins, nous avons utilisé la forme bien particulière de l'opérateur  $T$ , défini comme espérance conditionnelle, qui peut être approché par un estimateur de type estimateur à noyau.

### 3.3.2 Application en tomographie sismique

Lorsque l'opérateur  $A$  est entièrement inconnu et sans forme particulière connue, les méthodes classiques de résolution de problèmes inverses ne s'appliquent plus. C'est également le cas lorsque la loi du bruit d'observation est inconnue. Le problème est dès lors très difficile à résoudre. Il en est ainsi pour la majorité des applications, notamment en tomographie sismique. C'est ce cas particulier que nous avons étudié dans un article [48] en utilisant des méthodes de statistique bayésienne.

Dans ce cas particulier, l'objectif de notre travail est la tomographie sismique, c'est-à-dire l'étude de la composition d'un sol, décrit dans [10] ou [23]. La nature d'un sol est caractérisée par le temps mis par une onde sismique à le traverser. C'est pourquoi nous envoyons des ondes dans le sol, notées  $i = 1, \dots, n$  et mesurons les temps de parcours de ces ondes  $T_i^{obs}$  à travers le milieu à étudier.

Le cheminement des ondes à travers le milieu  $U$  est présenté dans la figure 3.1.

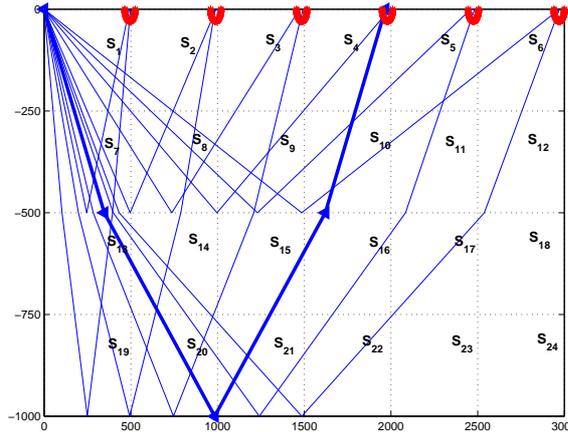


FIG. 3.1 – Tomographie sismique

A partir de ces observations, nous essayons de construire un estimateur de l'inverse de la vitesse que met une onde à traverser le milieu noté  $U$ . Considérons l'inverse du temps de propagation d'une onde dans le milieu  $S^*(x)$ ,  $x \in \mathbb{R}^2$ , fonction que l'on cherche à reconstruire. Les temps de parcours observés dépendent de  $S^*$  mais aussi des trajectoires dans le milieu qui sont prises par chaque onde  $r_i$  émise. Comme les  $r_i$  sont inconnus a priori, les temps d'observation sont observés à travers une image par un opérateur non linéaire et inconnu  $\Phi$  défini de la manière suivante

$$T = \Phi(S, r) = \int_r S(x) dl,$$

où  $r$  est une trajectoire d'onde arbitraire,  $x \in \mathbb{R}^2$  est un vecteur de position, tandis que  $dl$  est un déplacement infinitésimal le long du rayon d'onde  $r$ . Le modèle statistique s'écrit sous la

forme

$$\begin{aligned} T_i^{obs} &= \Phi(S^*, r_i) + \epsilon_i, \\ &= \int_{r_i} S^*(x) dl + \epsilon_i, \quad i = 1, \dots, n \end{aligned} \quad (3.30)$$

Ainsi donc, nous obtenons un problème inverse à opérateur non linéaire inconnu et dont la loi des erreurs d'observation n'est pas spécifiée. Nous nous trouvons alors dans le champ d'application des méthodes bayésiennes de minimum d'entropie, pour construire un estimateur du champ  $S$ . Ces méthodes ont été développées par Gamboa et Gassiat dans [55, 56].

Le Maximum d'Entropie sur la Moyenne (MEM) est une construction qui remplace le problème inverse précédent en l'approchant par une suite de problèmes en dimension finie, obtenus en discrétisant le milieu  $U$  sur lequel est définie la fonction  $S$ . L'estimateur MEM est alors obtenu en tant que limite des estimateurs discrétisés, solutions des problèmes de dimension finie. La construction du MEM est une manière naturelle et pratique d'introduire des contraintes non linéaires mais convexes sous la forme d'a priori. Pour cela, considérons une probabilité  $P$  sur  $U$  qui donne plus de poids aux zones où  $S$  est concentrée.

La fonction  $S(x, y)$  est définie sur le compact  $U \subset \mathbb{R}^2$ . Divisons  $U$  en cellules  $c_i^n, i = 1, \dots, n$  sur lesquelles nous cherchons à approcher la valeur de l'inverse du champ de vitesse  $S(x)$ , supposée constante sur chaque cellule et prenant la valeur  $s_i$ . Notre objectif est donc d'estimer les valeurs  $S = (s_1, \dots, s_n)$  à partir des temps de parcours observés  $T_1^{obs}, \dots, T_k^{obs}$  à partir d'un schéma bayésien. A chaque site  $c_j^n, j = 1, \dots, n$  du milieu  $U$ , est associée une variable aléatoire  $s_j$ . Notons  $F_n$  la distribution a priori du vecteur aléatoire correspondant  $S_n = (s_1, \dots, s_n)$ . Dans ce travail, nous avons supposé que la variable aléatoire correspondant à la distribution de l'inverse des vitesses  $S$  est positive, possédant une densité  $f$  (par rapport à la probabilité  $P$ ). Nous supposons naturellement que cette densité est bornée  $a(x) < f(x) < b(x)$ . Ces valeurs correspondent à des valeurs fournies par les expérimentations pratiques. Pour tout  $j$ , le support de la loi de  $s_j$ , notée  $F_j$  est inclus dans  $[a(c_j^n), b(c_j^n)]$ . Tous les sites sont supposés indépendants les uns des autres, et, en conséquence, la loi de  $S$  est donnée par  $F_n = \bigotimes_{j=1}^n F_j$ .

Pour une discrétisation en  $n$  cellules, nous définissons la mesure aléatoire discrète

$$\nu_n := \frac{1}{n} \sum_{j=1}^n s_j \delta_{c_j^n}. \quad (3.31)$$

obtenue à partir des valeurs de  $S_n = (s_1, \dots, s_n)$ . Les méthodes bayésiennes permettent de prouver la convergence de l'estimateur de la mesure  $\hat{\nu}_n^{bay}$ , ce qui impliquera la convergence de l'estimateur du champ de l'inverse des vitesses.

Par conséquent, nous proposons la procédure d'estimation suivante :

- Nous discrétisons le milieu et nous nous proposons d'estimer le paramètre de dimension finie  $S_n = (s_1, \dots, s_n)$ .
- Pour un champ d'inverse de vitesses bien choisi (i.e choisi selon une loi a priori),  $S_n^{(i)} = (s_1^{(i)}, \dots, s_n^{(i)})$ ,  $i = 1, \dots, N$ , nous calculons les temps de parcours des différents rayons sismiques  $T(S^{(i)})$  en utilisant un algorithme de traceurs de rayons.

- Pour un niveau donné  $\delta > 0$ , nous ne gardons que les temps de parcours admissibles, c'est-à-dire qui appartiennent à l'ensemble

$$\mathcal{I}_n = \{i = 1, \dots, N, \|T^{obs} - T^{(i)}\| \leq \delta\}.$$

- L'estimateur Bayésien est défini par

$$\hat{S}_n^{bay} = \frac{1}{\text{Card}(\mathcal{I}_n)} \sum_{i \in \mathcal{I}_n} S_n^{(i)}. \quad (3.32)$$

A l'estimateur de  $S^*$ , nous pouvons associer l'estimateur de la mesure

$$\hat{\nu}_n^{bay} = E_{\mathcal{F}_n}(\nu_n \mid \|T^{obs} - T(S_n)\|^2 < \delta^2). \quad (3.33)$$

où  $\mathcal{F}_n$  désigne la mesure empirique a priori de  $\nu_n$ . Ainsi au niveau  $n$ , nous ne gardons que les mesures positives portées par les  $c_i^n$ ,  $i = 1, \dots, n$ .

Ainsi,  $\hat{\nu}_n^{bay}$  est un estimateur Bayésien d'a priori la loi  $\mathcal{F}_n$ . Nous obtenons une suite de problèmes de dimension  $n$ . Il est important de remarquer que l'asymptotique ne porte pas ici sur le nombre d'observations, ni sur le niveau du bruit, mais sur la discrétisation du problème,  $n$ .

Pour comprendre le comportement asymptotique de l'estimateur de Bayes, nous avons besoin d'introduire l'estimateur : au niveau  $n$ , nous choisissons la distribution  $P_n^{MEM}$  de  $S_n$  en utilisant le principe de maximum d'entropie. Nous nous référons à [55] ou à [29] pour des précisions sur cette méthode.

Posons ainsi

$$\hat{\nu}_n^{MEM} := \mathbf{E}_{P_n^{MEM}}(\nu_n).$$

Alors pour  $n$  fixé,  $\hat{\nu}_n^{MEM}$  est la reconstruction par maximum d'entropie du champ  $S$  avec pour mesure a priori  $F_n$ . Lorsque  $n$  tend vers l'infini, sous certaines conditions additionnelles, nous montrons la convergence de l'estimateur  $\hat{\nu}_n^{bay}$ .

Le théorème suivant prouve la convergence de cet estimateur.

**Théorème 3.3.2.** (Gamboa, Gassiat, [55] Theorem 2.1 and 2.3) *Si  $U$  est un espace métrique compact,  $P$  une mesure de probabilité sur  $U$  and  $c_i^n$ ,  $i = 1, \dots, n$  tels que  $\frac{1}{n} \sum_{i=1}^n \delta_{c_i^n} \rightarrow P$ . Alors, sous quelques hypothèses techniques supplémentaires, l'estimateur bayésien  $\hat{\nu}_n^{bay}$ , pour  $\delta > 0$  et  $n$  grand, est bien défini. Il converge au sens de la topologie faible vers  $\hat{\nu}_\infty^{MEM}$ , le seul point d'accumulation de la suite  $\hat{\nu}_n^{MEM}$ . En outre si  $J$  est l'entropie associée,  $\hat{\nu}_\infty^{MEM}$  est un des points d'accumulation de l'ensemble des solutions du problème de minimisation*

$$\min_{\mu, \|\int \Phi d\mu - Y\| \leq \delta} J(\mu).$$

Nous avons développé un algorithme permettant de construire pratiquement un tel estimateur. Les erreurs d'estimation sont présentées dans la figure 3.2. Nous voyons que cet estimateur permet de reconstruire avec de faibles erreurs  $S^*$ . Il apparaît clairement que de

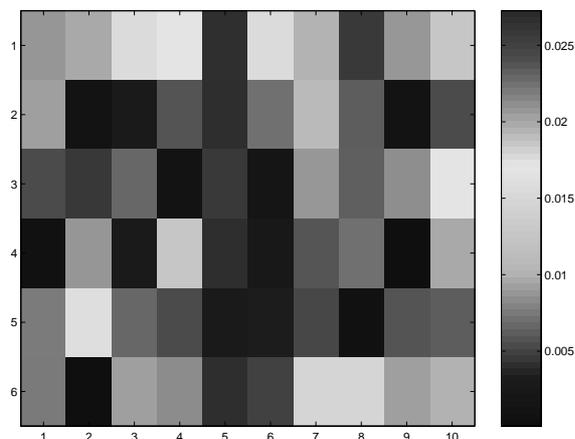


FIG. 3.2 – Erreurs d'estimation

telles méthodes bayésiennes sont efficaces lorsque l'opérateur est entièrement inconnu, donné par exemple sous la forme d'une boîte noire ou d'un code informatique.

Nous appliquons cette démarche au cadre de la tomographie océanique qui consiste à déterminer la composition de l'océan à partir de mesures faites sur des rayons lumineux se réfléchissant à travers lui. Ce travail [95] est actuellement en voie d'achèvement.

### 3.4 Espaces de régularité pour les problèmes inverses

Nous rappelons tout d'abord le problème inverse que nous étudions. Soit  $A$  un opérateur connu, admettant une décomposition en valeurs singulières (SVD) de la forme,  $(\sigma_k^2 = \lambda_k, \phi_k, \psi_k), k \geq 1$ . Nous rappelons les formules suivantes

$$\forall x \in X, A^*Ax = \sum_{k=1}^{\infty} \lambda_k \langle x, \phi_k \rangle \phi_k.$$

$$Ax = \sum_{k=1}^{\infty} \sigma_k \langle x, \phi_k \rangle \psi_k.$$

Nous cherchons à reconstruire une fonction  $x_0$  observée dans le modèle suivant

$$y = Ax_0 + \epsilon$$

où  $\epsilon$  est un bruit d'observation.

Dans les travaux théoriques sur les problèmes inverses, sont apparues différentes manières d'exprimer la régularité de la fonction à estimer aboutissant à des vitesses de convergence différentes. En effet, il existe trois types d'hypothèses différentes portant sur la régularité de la solution du problème inverse.

1. La première est la condition usuelle de régularité obtenue en supposant que  $x_0$  appartient à un espace de régularité classique Hölder, Sobolev ou Besov. Par exemple la fonction

appartient à l'espace défini par

$$W^\beta := \left\{ x = \sum_{k=1}^{\infty} \langle x, \phi_k \rangle \phi_k \in L^2([0, 1]), \sum_{k=1}^{\infty} k^{2\beta} |\langle x, \phi_k \rangle|^2 < +\infty \right\}.$$

2. La seconde porte sur la régularité de la fonction par rapport à la régularité de l'opérateur. Plus précisément, on suppose que  $x_0$  appartient à des espaces appelés sources définis par

$$X_\beta = \{x \in X, \exists \omega \in L^2, x = (A^*A)^\beta \omega\} = \mathcal{Im} [(A^*A)^\beta].$$

$$X_\beta = \left\{ x, \sum_{k=1}^{\infty} \frac{|\langle x, \phi_k \rangle|^2}{\sigma_k^{4\beta}} < +\infty \right\}. \quad (3.34)$$

Ce type d'hypothèse est courant en analyse numérique des problèmes inverses.

3. Un troisième jeu d'hypothèses s'intéresse aux fonctions convergeant à une vitesse donnée pour une méthode de régularisation préalablement choisie. Ainsi pour  $R_\alpha$  un opérateur de régularisation et  $\alpha$  une suite régularisante, définissons  $x_\alpha = R_\alpha A x_0$ , tandis que l'estimateur est donné par  $\hat{x}_{n,\alpha} = R_\alpha y$ . Supposons que l'opérateur de régularisation vérifie

**Hypothèses A** Le schéma de régularisation  $R_\alpha$  défini sur  $[0, 1]$  est tel qu'il existent des constantes  $\alpha_0$  et  $c, C, K$  telles que pour toute suite  $\alpha = (\alpha_k)_{k \geq 1}$ ,  $0 < \alpha_k < \alpha_0$ ,  $\forall k \geq 1$

$$\begin{aligned} c &\leq |1 - \lambda R_\alpha(\lambda)| \leq C, \quad \forall 0 \leq \lambda < \alpha, \\ |1 - \lambda R_\alpha(\lambda)| \frac{\lambda}{\alpha} &\leq K, \quad \forall \alpha \leq \lambda \leq 1. \end{aligned}$$

Le terme de biais mesure l'efficacité de la procédure de régularisation. C'est pourquoi, en économétrie les espaces de régularité sont définis par la relations suivante :  $\exists 0 \leq \beta \leq 1$ ,

$$x_0 \in \Phi_\beta := \{x, \|x - x_\alpha\| = O(\alpha^\beta)\}. \quad (3.35)$$

Cet espace est appelé espace de saturation. Il s'apparente à un espace maxiset déterministe associé à la vitesse  $\alpha^\beta$  et la méthode d'estimation au moyen de l'opérateur de régularisation  $R_\alpha$ .

Nous nous sommes intéressés aux liens entre ces espaces et aux différentes vitesses rencontrées dans 2 articles [89] et [98]. Le théorème suivant établit des correspondances entre les 3 types d'hypothèses.

**Théorème 3.4.1** (Comparaison entre les hypothèses de régularité pour les problèmes inverses). *Sous le jeu d'hypothèses A, nous obtenons les inclusions suivantes.*

1. Pour  $0 < \beta < 1$ ,

$$X_\beta \subset \Phi_\beta, \quad X_\beta \not\subset \Phi_t, \quad \forall t > \beta.$$

2.

$$X_1 = \Phi_1.$$

3.

$$X_s = W^{\frac{as}{2}}.$$

Ce théorème permet de mieux comprendre les vitesses de convergence obtenues précédemment. Notamment nous voyons que la correspondance  $s = 2\beta a$  permet de démontrer l'optimalité des vitesses données en analyse numérique dans un cadre statistique, et inversement.



# Chapitre 4

## Un problème inverse particulier : le décalage de courbes

Les problèmes rencontrés en classification nous ont également conduits à nous intéresser à un nouveau type de problèmes inverses : le recalage de courbes et l'étude de ces décalages.

En effet, l'étude de données fonctionnelles met souvent à jour un phénomène particulier : les méthodes de classification ou de choix de modèles permettent de rassembler les courbes en des groupes formés de fonctions similaires. Toutefois, à l'intérieur d'un même groupe subsistent de légères différences entre ces courbes, qui rendent difficile l'extraction d'une courbe type  $f$ , décrivant le phénomène lié au groupe. Ainsi, il nous faut aligner ces courbes sans connaître l'archétype  $f$  mais aussi être capable d'expliquer les raisons de ces décalages pour mieux comprendre le phénomène étudié.

Ce problème peut être formalisé dans un cadre mathématique de la manière suivante. On observe  $J$  courbes, chacune formée de  $n$  points. On constate alors qu'il existe des paramètres de déformation  $\theta_j, j = 1, \dots, J_n$  (aléatoires ou non) à valeurs dans un espace  $\Theta$  et un opérateur de déformations  $\Phi$  tels que les observations, bruitées par une erreur d'observation  $\epsilon$ , s'écrivent

$$Y_j = \Phi(f(t_{ij}), \theta_j) + \sigma\epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, J_n.$$

Les courbes à l'intérieur du groupe sont donc les  $f_j = \Phi(f, \theta_j)$  qui sont obtenues comme déformation d'un motif général  $f$ , à l'aide d'un opérateur de déformation  $\Phi$  dépendant des paramètres  $\theta_j$ .

### 4.1 Estimation de décalages entre courbes

Ce modèle combine plusieurs problèmes inverses de nature différentes.

1. en premier lieu celui de l'estimation de la fonction  $f$  à travers le recalage des courbes les unes par rapport aux autres.
2. L'estimation des paramètres de déformation  $\theta_j, j = 1, \dots, J_n$
3. enfin, dans le cadre de déformations aléatoires, l'estimation de la loi de ces déformations.

Ce problème est bien connu en data mining et de nombreux auteurs se sont intéressés aux différentes manières d'extraire un représentant général à partir de données fonctionnelles. Dans les travaux suivants [57], [124] [108] [125], les auteurs utilisent des méthodes non paramétriques pour estimer un profil moyen puis aligner les courbes sur ce profil estimé. Alternativement, ils estiment des points particuliers pour chaque courbe (par exemple les points d'inflexion) qui servent de point de repères sur chaque courbe, pouvant être alignés. Dans tous les cas, ces méthodes requièrent au préalable une étape d'estimation non paramétrique. Or, le modèle est un modèle semi-paramétrique puisque les paramètres d'intérêt  $\theta_j$  sont à estimer dans un cadre où le bruit est donné par une fonction inconnu  $f$ . C'est la raison pour laquelle nous nous sommes intéressés à des méthodes semi-paramétriques pour résoudre ce problème.

Dans un premier article *M-estimation of shifts between curves* [52], nous nous consacrons au cas particulier où les déformations sont des translations entre les courbes. Cette restriction nous permet de résoudre théoriquement le problème et de proposer une méthode pratique et efficace d'alignement des courbes. Plus précisément, nous avons développé une méthode d'estimation basée sur un contraste obtenu après avoir appliqué la transformée de Fourier aux fonctions observées. Ainsi le modèle initial

$$Y_{ij} = f(t_i - \theta_j^*) + \varepsilon_{ij}, \quad j = 1, \dots, J, \quad i = 1, \dots, n, \quad (4.1)$$

se transforme, en utilisant la transformée de Fourier discrète, en

$$d_{jl} = e^{-il\alpha_j^*} c_l(f) + w_{jl}, \quad j = 1, \dots, J, \quad l = -(n-1)/2, \dots, (n-1)/2. \quad (4.2)$$

Les coefficients  $d_{jl}$  sont observés tandis que les coefficients de Fourier de la fonction  $f$  sont notés  $c_l(f)$  et constituent un bruit fonctionnel qui empêche l'estimation directe des coefficients liés au décalage entre les courbes  $\alpha_j^*$ . L'écriture du modèle sous la forme (4.2) permet de comprendre la problématique de l'estimation des  $\alpha_j^*$  en tant que problème inverse. Les coefficients inconnus  $c_l(f)$  jouent ici le rôle des valeurs propres de l'opérateur à travers lequel sont observées les données, puisque  $c_l(f) \rightarrow 0$  lorsque  $|l| \rightarrow +\infty$ . Nous sommes bien dans le cadre des problèmes inverses à opérateurs inconnus pour lesquels il n'existe pas de méthode générale de résolution. Nous allons néanmoins tirer parti de la double asymptotique en le nombre de courbes d'une part et en le nombre d'observations par courbe d'autre part, afin d'*inverser* le problème inverse sans estimer l'opérateur.

Les propriétés de la transformée de Fourier permettent de transformer le problème initial de translation en un problème multiplicatif dans lequel on cherche à rephaser les observations. Pour cela, nous proposons une méthode semi-paramétrique d'estimation des coefficients  $\alpha_j^*$  obtenus en minimisant le contraste

$$\hat{\alpha} = \arg \min_{\alpha_j, j=1, \dots, J} \left( \frac{1}{J} \sum_{j=1}^J \sum_{l=-(n-1)/2}^{l=(n-1)/2} \delta_l^2 \left| e^{il\alpha_j} d_{jl} - \frac{1}{J} \sum_{j=1}^J e^{il\alpha_j} d_{jl} \right|^2 \right), \quad (4.3)$$

pour une suite régularisante  $\delta_l$  bien choisie. L'idée à l'origine de ce contraste, est la volonté de chercher à aligner chaque courbe sur la moyenne recalée de toutes les autres, pondérée par

une suite de réels, convenablement choisis, qui doit lisser les irrégularités de la courbe.

Nous démontrons la convergence des estimateurs, et obtenons en outre une vitesse de convergence paramétrique des estimateurs des paramètres de translation du modèle (4.2) et prouvons la normalité asymptotique suivante sous des hypothèses techniques. En particulier le paramètre  $\delta_l$  est un lissage qui ne doit pas excéder une certaine valeur.

**Théorème 4.1.1** (Comportement asymptotique des estimateurs des translations).

$$\sqrt{n}(\hat{\alpha}_n - \alpha^*) \rightarrow \mathcal{N}_{J-1}(0, \Gamma), \quad (4.4)$$

avec

$$\Gamma = \frac{\sum_{l \in \mathbb{Z}} \delta_l^4 l^2 |c_l(f)|^2}{\left(\sum_{l \in \mathbb{Z}} \delta_l^2 l^2 |c_l(f)|^2\right)^2} (I_{J-1} + U_{J-1}),$$

où,  $I_{J-1}$  est la matrice identité de dimension  $J-1$ , et  $U_{J-1}$  est la matrice carrée de dimension  $J-1$  ne comportant que des 1.

L'idée de la preuve de ce résultat, propre à l'estimation par minimisation d'un contraste, nécessite l'étude des processus empiriques suivants

i) Convergence du contraste empirique vers une fonction de contraste :

$$M_n(\alpha) \xrightarrow[n \rightarrow +\infty]{\mathbf{P}_{\alpha^*}} K(\alpha), \quad \alpha \in A, \quad (4.5)$$

où  $K(\cdot)$  possède un unique minimum au point  $\alpha^*$ .

ii) Définissons le module de continuité uniforme  $W$ , défini par

$$W(n, \eta) = \sup_{\|\alpha - \beta\| \leq \eta} |M_n(\alpha) - M_n(\beta)|.$$

On montre qu'il existe deux suites  $(\eta_k)_{k \in \mathbb{N}}$  et  $(\epsilon_k)_{k \in \mathbb{N}}$ , tendant vers zéro telles que pour  $k$  assez grand, nous avons

$$\lim_{n \rightarrow +\infty} \mathbf{P}_{\alpha^*}(W(n, \eta_k) > \epsilon_k) = 0. \quad (4.6)$$

Dès lors, nous obtenons la convergence de l'estimateur réalisant le minimum du contraste empirique  $M_n$ .

$$\hat{\alpha}_n = \arg \min_{\alpha \in \hat{A}_1} M_n(\alpha).$$

La normalité asymptotique s'obtient par un développement limité de la vraisemblance. En effet, il existe  $\bar{\alpha}_n$  dans un voisinage de  $\alpha^*$  tel que l'on puisse écrire

$$\sqrt{n}(\hat{\alpha}_n - \alpha^*) = - [\nabla^2 M_n(\bar{\alpha}_n)]^{-1} \sqrt{n} \nabla M_n(\alpha^*), \quad (4.7)$$

où  $\nabla^2$  désigne la Hessienne. La théorie des processus empiriques [120] ou [121] permet de montrer les limites suivantes

$$\begin{aligned} \sqrt{n} \nabla M_n(\alpha^*) &\xrightarrow[n \rightarrow +\infty]{\mathcal{D}} \mathcal{N}_{J-1}(0, \Gamma_0), \\ [\nabla^2 M_n(\bar{\alpha}_n)]^{-1} &\xrightarrow[n \rightarrow +\infty]{\mathbf{P}_{\alpha^*}} V, \end{aligned}$$

où  $V$  désigne une matrice symétrique positive. Il nous suffit alors de poser  $\Gamma = V' \Gamma_0 V$ , pour prouver le résultat énoncé dans le théorème.

Nous remarquons que dans ce cas, grâce au point de vue semi-paramétrique, la fonction  $f$  devient un paramètre infini de nuisance qui n'empêche pas la reconstruction paramétrique des translations. Nous mettons en oeuvre cet algorithme pour estimer la fonction  $f$ , une fois les données réalignées.

Cet algorithme est décrit et appliqué à la prévision du trafic routier dans un article publié à la revue de l'ENBIS *road traffic forecasting with semiparametric method* [92]. Le cadre de cette application est le suivant. Après classification des courbes journalières de l'évolution des vitesses en un point de mesure du trafic routier, nous obtenons des groupes qui font apparaître des types de comportement routier : nous observons par exemple des bouchons types, caractéristiques par leur durée et la vitesse minimale des véhicules qui le traversent. Mais, au sein d'un même groupe, l'heure de déclenchement du bouchon est soumise à des petites variations, sous la forme de translations. Si l'on prend comme représentant du groupe la moyenne des courbes le composant, les petits décalages entre les courbes créent un artefact qui ne ressemble pas au phénomène réellement observé. Il est par conséquent nécessaire d'aligner les courbes avant de prendre la moyenne. Nous nous trouvons ainsi dans le cadre d'application du modèle (4.2). Dans la figure 4.1, nous montrons les résultats obtenus. Le premier cadre comporte les courbes du groupe, le troisième montre les courbes réalignées tandis que le quatrième montre les deux représentants du groupe obtenus sans prendre en considération les décalages (courbe en pointillé) et par ailleurs en tenant compte (courbe en traits pleins). La courbe en pointillés est trop lissée par rapport aux phénomènes de bouchons, ce qui se traduit par des chutes très brusques des vitesses moyennes.

En utilisant une telle méthode de choix de représentant, les courbes intervenant dans les modèles de prévision sont plus proches des phénomènes physiques observés. Dès lors, les prévisions ne sous-estiment plus de façon systématique les temps de parcours et sont plus proches des temps de parcours réels.

Nous avons donc résolu le cas du recalage de courbes soumises à des effets de translation. Cette même méthodologie a été étendue dans le cadre de dilatations en utilisant les mêmes techniques. Il est alors naturel de s'intéresser au cas où les paramètres des translations ne sont plus des constantes déterministes mais des réalisations indépendantes d'une variable aléatoire  $\theta_1, \dots, \theta_J \sim \theta$  possédant une densité  $\varphi$  par rapport à la mesure de Lebesgue. Notre objectif n'est plus seulement d'estimer les décalages, mais la loi de cette variable aléatoire afin de comprendre la nature des variations entre les courbes expérimentales. Nous constatons qu'il est impossible d'utiliser la méthodologie précédente et sommes alors confrontés à un problème inverse stochastique.

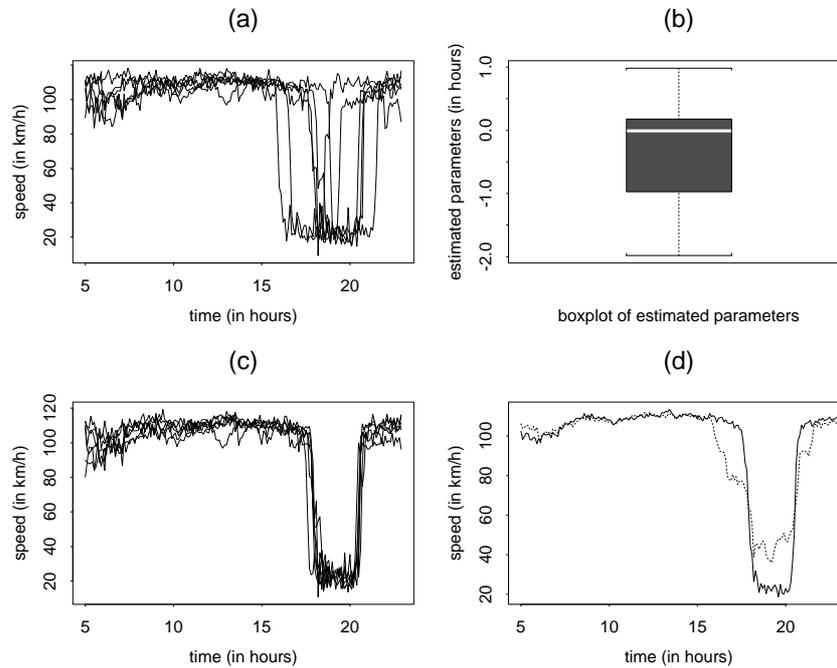


FIG. 4.1 – Estimation des décalages dans des courbes de vitesse et leur incidence dans l'extraction d'un représentant

## 4.2 Etude de décalages stochastiques

Dans le cas très général où l'opérateur  $\Phi$  est inconnu, le problème de l'estimation de la loi des paramètres de déformation est difficile. C'est pourquoi, dans un premier temps, nous nous sommes restreints à une forme particulière, pour laquelle nous précisons l'allure des déformations aléatoires. Ainsi nous avons considéré un problème d'estimation avec des paramètres de translation aléatoires d'une fonction  $f$  inconnue, mêlant les deux modèles (4.14) et (4.1). Ce nouveau modèle s'écrit alors

$$Y_{ij} = f(t_{ij} - \theta_j) + \sigma \epsilon_{ij} \quad i = 1, \dots, n, j = 1, \dots, J_n. \quad (4.8)$$

Des exemples d'application de ce type de modèle sont fournis en analyse de données lorsqu'on cherche à étudier les raisons pour lesquelles il existe des décalages entre les courbes au sein d'un même groupe, et ainsi améliorer la procédure de classification ou mieux encore interpréter les résultats obtenus.

La méthodologie que nous utilisons dans [19] fait appel aux méthodes bayésiennes et semi-paramétriques développées dans [30]. Elles se décomposent en les deux étapes suivantes

1. Pour chaque courbe observée  $j$ , nous estimons la valeur de la réalisation  $\theta_j$ , par son estimateur  $\hat{\theta}_{j,n}$ . Conditionnellement à  $\theta_j$ , nous pouvons établir des inégalités de déviation précises entre cette valeur et sa valeur estimée.

2. Nous utilisons les valeurs approchées pour toutes les courbes  $j = 1, \dots, J_n$  pour construire un estimateur à noyaux de la densité, sous la forme

$$\hat{\varphi}_n(\theta) = \frac{1}{J_n} \sum_{j=1}^{J_n} K_h(\theta - \hat{\theta}_{j,n}).$$

Cet estimateur peut être étudié grâce au contrôle précis entre l'estimateur semi-paramétrique et la réalisation du décalage  $\theta_j$  pour chaque courbe.

Ainsi, nous commençons par construire des estimateurs semi-paramétriques pour estimer les réalisations des variables  $\theta_j$  qui ne sont pas observées. Nous parvenons alors à des estimateurs  $\hat{\theta}_{j,n}$  obtenus par maximum de vraisemblance selon la démarche suivante. En premier lieu, en raisonnant courbe par courbe (et, pour alléger les notations, en oubliant l'indexation en  $j$ ), le modèle est transformé en utilisant la transformée de Fourier discrète en

$$x_k = \cos(2\pi k\theta) f_k + d_{k,n} + \frac{1}{\sqrt{n}} \xi_k, \quad (4.9)$$

$$x_k^* = \sin(2\pi k\theta) f_k + d_{k,n}^* + \frac{1}{\sqrt{n}} \xi_k^*. \quad (4.10)$$

en ayant posé

$$x_k = \frac{1}{n} \sum_{i=1}^n \cos(2\pi kt_i) Y_i, \quad \xi_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n \cos(2\pi kt_i) \epsilon_i,$$

$$x_k^* = \frac{1}{n} \sum_{i=1}^n \sin(2\pi kt_i) Y_i, \quad \xi_k^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sin(2\pi kt_i) \epsilon_i.$$

et

$$d_{k,n} = \sqrt{2} \left( \frac{1}{n} \sum_{i=1}^n \cos(2\pi kt_i) f(t_i - \theta) - \int_0^1 \cos(2\pi kt) f(t - \theta) dt \right).$$

Nous obtenons alors la version discrète de l'estimateur de maximum de vraisemblance bayésien défini dans [30]

$$\hat{\theta}_n = \arg \max_{\tau \in \Theta} \left\{ \sum_{k \geq 1} h_k \left( \frac{1}{n} \sum_{i=1}^n \cos(2\pi k(t_i - \tau)) Y_i \right)^2 \right\}, \quad (4.11)$$

où  $(h_k)$  est une suite de réels de  $[0, 1]$  qui satisfont des conditions techniques, notamment

$$(\mathbf{T}) \quad \left( \sum_{k \geq 1} (1 - h_k) (2\pi k)^2 f_k^2 \right)^2 = o \left( \sum_{k \geq 1} (1 - h_k)^2 (2\pi k)^2 f_k^2 \right).$$

Le lemme suivant fournit un contrôle optimal de l'erreur d'estimation.

**Lemma 4.2.1** (Inégalité de Déviation). *Pour  $n$  assez grand et uniformément en  $\theta_j \in \Theta$  et  $f \in F$ ,*

$$(E) \quad \mathbf{P} \left( |\hat{\theta}_{j,n} - \theta_j| \|f'\| > K \sqrt{\frac{\log n}{n}} |\theta_j| \right) \leq c_1 \exp(-c_2 K^2 \log n), \quad (4.12)$$

Soit  $R^n[h, f] = \sum_{k=1}^{\infty} (2\pi k)^2 [(1 - h_k)^2 f_k^2 + h_k^2/n]$ . Pour tout  $j = 1, \dots, J_n$  et uniformément en  $\theta_j \in \Theta$  et  $f \in F$ ,

$$\mathbf{E}[(\hat{\theta}_{j,n} - \theta_j)^2 | \theta_j] = \frac{1}{n \|f'\|^2} \left( 1 + (1 + o(1)) \frac{R^n[f, h]}{\|f'\|^2} \right). \quad (4.13)$$

Nous cherchons dès lors à estimer la densité de  $\theta$ ,  $\varphi$  en utilisant les estimateurs  $\hat{\theta}_{j,n}$ ,  $j = 1, \dots, J_n$ . Supposons que nous pouvons avoir une double asymptotique sans restriction, c'est-à-dire que le nombre de courbes et le nombre d'observations par courbes tend vers l'infini sans limitations.

**Théorème 4.2.2** (Vitesse de convergence de l'estimateur non paramétrique). *Supposons que  $\varphi$  appartient à l'espace de Sobolev  $H_s$  et supposons que le noyau  $K$  est d'ordre  $[s]$ . Alors pour un choix de fenêtre de lissage  $h_n = \alpha n^{-\frac{1}{2s+1}}$ , avec  $\alpha > 0$ , il existe une constante  $C$  telle que*

$$\mathbf{E} \|\hat{\varphi} - \varphi\|_2^2 \leq C n^{-\frac{2s}{2s+1}}.$$

La démonstration de ce résultat utilise les inégalités données par (4.12) pour comparer notre estimateur à celui que l'on aurait obtenu si l'on avait observé  $\theta_j$ .

Nous obtenons ainsi une méthode de résolution d'un problème inverse stochastique et nous obtenons une vitesse de convergence de l'estimateur. Ce travail a fait l'objet d'un article [19]. La vitesse de convergence est la même que celle obtenue en observant directement les paramètres de translation. Les méthodes semi-paramétriques employées permettent ainsi d'éliminer l'effet de la fonction  $f$  qui n'est plus qu'un paramètre de nuisance. Notre estimateur de densité est plus précis que celui calculé en utilisant comme estimateurs préliminaires, ceux obtenus par des techniques non paramétriques utilisées dans [57] par exemple.

Ces résultats nous permettent de mieux analyser le trafic routier en des points particuliers. Pour le même type de bouchon, nous avons estimé la densité des effets de translation et avons mis en évidence un effet bimodal, différenciant les jours de pluie des jours de beau temps.

Néanmoins la contrainte de la double asymptotique ne nous permet pas d'appliquer ce résultat dans un modèle issu de la biostatistique. Dans ce cas, le nombre d'observations par courbes est limité et l'asymptotique n'est obtenue que lorsque le nombre de courbes tend vers l'infini.

Plus précisément, ce cadre de recherche s'inscrit dans la théorie des modèles à effets mixtes en biostatistique. Ces modèles sont parfaitement adaptés pour décrire les effets d'un médicament en pharmacocinétique. En effet, lorsque un médicament (ou tout corps étranger) est administré à un organisme vivant, celui-ci va l'éliminer progressivement. La forme de cette élimination est commune pour chaque espèce et s'exprime comme une fonction du temps

$t \rightarrow \Phi(., t) \in \mathbb{R}$ . Cette fonction est obtenue en étudiant les échanges physiques entre les différentes parties de l'organisme, intervenant dans l'assimilation de la substance étrangère. Elle dépend de paramètres  $S_j$ ,  $j = 1, \dots, J$  (où  $J$  désigne le nombre d'individus participant à l'expérimentation). Un exemple classique est donné par  $S = (A, \alpha, B, \beta)$  et une fonction

$$\Phi(S, t) = A \exp(-\alpha t) - B \exp(-\beta t).$$

Ces paramètres, qui sont propres à chaque individu, se ressemblent néanmoins au sein d'une même espèce. Pour modéliser ce phénomène les biologistes considèrent alors que ces paramètres sont des réalisations i.i.d d'une variable aléatoire  $S$ . Comprendre l'action du traitement revient à estimer la densité de la variable aléatoire  $S$ , décrivant la réaction physique au médicament administré.

Ainsi le modèle s'écrit sous sa forme générale

$$Y_{ij} = \Phi(S_j, t_{ij}) + \sigma \epsilon_{ij}, \quad j = 1, \dots, J, \quad i = 1, \dots, n. \quad (4.14)$$

$J$  désigne le nombre d'individus tandis que  $n$  désigne le nombre d'observations pour chaque courbe. Le bruit d'observation est noté  $\epsilon_{ij}$ . Ce modèle a été étudié lorsque la densité inconnue  $\varphi$  appartient à un modèle paramétrique. Nous nous référons aux travaux de [33], [104] ou [58]. Dans le cas où la densité ne présente pas de forme particulière, peu de références existent dans la littérature. Le problème est fort complexe pour plusieurs raisons.

Tout d'abord, le modèle (4.14) est un modèle de problème inverse dans lequel on ne cherche pas seulement à estimer des paramètres, mais également la densité d'une variable sous-jacente dont les réalisations sont observées au travers de leur image par un opérateur. C'est pour cela que ce type de problème est nommé problème inverse stochastique.

Nous pouvons ensuite remarquer que ce problème s'apparente à un modèle de déconvolution dans lequel les observations ne seraient pas identiquement distribuées. En effet, en notant  $\varphi_S$  la densité de  $S$  à estimer et  $\gamma$  la densité des erreurs supposée connue, la densité d'une observation a alors la forme suivante

$$\mathcal{F}(Y_{ij}) = \int \gamma(Y - \Phi(s, t_{ij})) \varphi_S(s) ds = \mathbf{K}(\phi_S).$$

Ainsi, la densité  $\phi_S$  est obtenue comme la convolée par un noyau  $K$  qui fait intervenir la fonction  $\Phi$ . Ce point de vue est adopté dans [59] dans une approche pratique. Cependant, dans les ouvrages de recherche sur le modèle de déconvolution, le cadre du modèle (4.14) n'est pas pris en compte.

Finalement, l'expression précédente fait apparaître le modèle comme un modèle de mélange régi par la densité inconnue  $\varphi$ . C'est dans ce cadre que se sont placés les travaux [79], [17], sans toutefois établir de vrais résultats de convergence d'un estimateur non paramétrique de la densité  $\varphi_S$ .

C'est là que se situe notre objectif dans [21], où nous nous sommes placés dans le cadre d'un mélange pour étudier un estimateur non paramétrique de la log-vraisemblance pour ce modèle inverse. Pour des raisons de facilité, nous avons envisagé le cas où les temps sont aléatoires

i.i.d de loi connue et les erreurs d'observations sont gaussiennes de densité jointe  $\gamma(\sigma, n)$ . Alors, en utilisant des notations vectorielles,  $(S_i, T_i)$  est une suite de variables aléatoires i.i.d à valeurs dans  $\mathbb{R}^p \times \mathbb{R}_+^n$ , de loi  $\mu_S \otimes \mu_T$ . Le modèle se réécrit de la façon suivante

$$Y_i := f(S_i, T_i) + \sigma \epsilon_i, \quad i = 1, \dots, N \quad (4.15)$$

où  $f : \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  est une fonction connue et régulière.

La loi  $\mathcal{L}(Y_i, T_i)$  est donnée par le mélange

$$\int_{s \in \mathbb{R}^p} \gamma_{\sigma, n}(y - f(s, t)) d\mu_T(t) d\mu_S(s) dy,$$

tandis que la loi de  $\mathcal{L}(Y_i)$  de  $Y_i$  est donnée par la formule

$$\left[ \int_{s \in \mathbb{R}^p} \int_{t \in \mathbb{R}_+^n} \gamma_{\sigma, n}(y - f(s, t)) d\mu_T(t) d\mu_S(s) \right] dy,$$

Définissons alors l'opérateur

$$\mathbf{K}(\mu_S)(y, t) := \psi(t) \int_{s \in \mathbb{R}^p} \gamma_{\sigma, n}(y - f(s, t)) d\mu_S(s). \quad (4.16)$$

et la logvraisemblance empirique du modèle comme

$$\mathbf{L}_N(\mu_S) := \mathbf{P}_N \log \mathbf{K}(\mu_S), \quad (4.17)$$

en ayant noté  $\mathbf{P}_N := \frac{1}{N} \sum_{i=1}^N \delta_{(Y_i, T_i)}$ . Nous obtenons alors

$$\mathbf{L}_N(\mu_S) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n \log \int_{s \in \mathbb{R}^p} \exp \left( -\frac{1}{2\sigma^2} (Y_{i,j} - q_s(T_{i,j}))^2 \right) d\mu_S(s) + C,$$

$$C := -\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{N} \sum_{i=1}^N \log \psi(T_{i,1}, \dots, T_{i,n}).$$

Nous prouvons ensuite que cet estimateur est convergent.

**Théorème 4.2.3** (Convergence de l'estimateur du maximum de vraisemblance). *Supposons que  $\mathcal{F}_S \subset \mathcal{P}(\mathbb{R}^p)$  est un espace compact convexe, que le modèle est identifiable, que  $\mathcal{L}(T) = \psi(t) dt$ ,  $\mu_S \in \mathcal{F}_S$ , et que pour tout couple  $(y, t) \in \mathbb{R}^n \times \mathbb{R}^n$ , l'opérateur  $\mathbf{K}(\bullet)(y, t) : \mathcal{F}_S \rightarrow \mathbb{R}$  est continu. Alors l'estimateur de maximum de vraisemblance non paramétrique  $\widehat{\mu}_{S,N}$  défini par*

$$\widehat{\mu}_{S,N} := \arg \max_{\mu \in \mathcal{F}_S} \mathbf{L}_N(\mu) \quad (4.18)$$

*est bien défini, unique, et converge presque sûrement vers  $\mu_S$  lorsque  $N$  croît vers  $+\infty$ .*

D'un point de vue pratique, il n'est pas possible de regarder le maximum sur tout l'ensemble  $\mathcal{F}_S$ , c'est pourquoi nous nous sommes intéressés au maximum de vraisemblance sur des ensembles d'approximation et avons prouvé le théorème suivant.

**Théorème 4.2.4.** *Supposons que  $(\mathcal{F}_{S,m})_{m \in \mathbb{N}^*}$  est une suite d'espaces fermés et convexes de dimension finie, denses dans  $\mathcal{F}_S$ . Sous les hypothèses du théorème 4.2.3, et pour  $N$  fixé, l'estimateur approché du maximum de vraisemblance  $\widehat{\mu_{S,N,m}}$  défini par*

$$\widehat{\mu_{S,N,m}} := \arg \max_{\mu \in \mathcal{F}_{S,m}} \mathbf{L}_N(\mu). \quad (4.19)$$

*est bien défini, unique, et converge vers  $\widehat{\mu_{S,N}}$  quand  $m$  augmente vers  $+\infty$ .*

La démonstration de ces résultats repose sur les résultats préliminaires de Pfanzagl sur le maximum de vraisemblance pour les mélanges, décrits dans [106, 107].

Ainsi, lorsque  $\Phi$  est connue, nous avons obtenu un estimateur non paramétrique de la loi des effets mixtes, ainsi qu'une méthode pratique d'estimation. Obtenir la vitesse de convergence est encore un problème très délicat puisqu'il nécessiterait une analyse d'un processus empirique composé qui n'est pas étudié pour le moment. Lorsque  $\Phi$  est inconnue, l'étude revient à étudier des déformations aléatoires d'un processus.

### 4.3 Processus de déformation aléatoires

L'étude générale des déformations entre des courbes nous a amenés naturellement au cadre plus général de processus de déformations aléatoires quelconques. Notre étude a pour but de dégager le profil caractéristique d'un groupe de courbes déformées aléatoirement les unes par rapport aux autres.

Pour cela, considérons un processus de déformation défini de la manière suivante : soit  $\Omega$  un espace de probabilités et définissons

$$\begin{aligned} H : \Omega &\rightarrow \mathcal{C}([a, b]) \\ w &\mapsto H(w, \cdot), \end{aligned}$$

Précisons en outre que, pour presque tout  $w \in \Omega$ , nous supposons que

- i)  $H(w, \cdot)$  est une fonction croissante
- ii)  $H(w, a) = a$  et  $H(w, b) = b$ .

Considérons enfin des réalisations aléatoires de ces processus  $i = 1, \dots, m$   $H_i(t) = H(w_i, t)$  pour  $w_i \in \Omega$ . Soit  $f$  une fonction et des observations  $Y_{ij}$  correspondant aux courbes  $i = 1, \dots, m$  observées en des temps discrets  $t_{ij}$ ,  $j = 1, \dots, n$ . Le modèle s'écrit alors de la manière suivante

$$Y_{ij} = f \circ H_i^{-1}(t_{ij}), \quad i = 1, \dots, m, \quad j = 1, \dots, n, \quad (4.20)$$

Notre objectif est de reconstruire la fonction  $f$ .

Un tel problème n'est, en général, pas identifiable. Par conséquent, il n'est pas possible de reconstruire la fonction  $f$  de façon unique à partir des observations  $Y_{ij}$ . Il importe donc de donner un sens à la notion de profil moyen, qui puisse remplacer la fonction  $f$  et, par

conséquent, être estimée de manière unique. Cette fonction doit, à la fois, être proche des données, et respecter la forme générale des fonctions.

Pour cela, nous avons défini l'espérance structurelle  $\mathbf{ES}$ , comme la variable aléatoire la plus proche, au sens de la norme quadratique, des données déformées

$$\mathbf{ES} = \arg \min_{Z \in L^2} \mathbf{E} \|f \circ H^{-1} - Z\|^2. \quad (4.21)$$

Elle vérifie la relation suivante

$$\mathbf{ES} = \mathbf{ES} (f \circ H^{-1}) = f \circ \phi^{-1},$$

où  $\phi$  est l'espérance du processus  $H$ .

Il s'ensuit que l'espérance structurelle est un outil qui, d'une part, rend le modèle identifiable au sens où il est possible de construire un estimateur, et d'autre part, qui représente le comportement général, contenu dans les données  $Y_{ij}$ .

Nous pouvons en construire un estimateur

$$\widehat{\mathbf{ES}} = \arg \min_{Z \in L^2} \frac{1}{nm} \sum_{(i,j) \in \mathcal{I} \times \mathcal{J}} [Y_{ij} - Z(t_{ij})]^2,$$

et étudier ses propriétés asymptotiques.

Pour tout  $y \in [f(a), f(b)]$ , définissons l'estimateur de la quantité  $\phi \circ f^{-1}$  par

$$\widehat{\phi \circ f^{-1}}(y) = \frac{1}{m} \sum_{i=1}^m T_i(y), \quad (4.22)$$

avec, pour tout  $i \in \mathcal{I}$ ,

$$T_i(y) = \arg \min_{j \in \mathcal{J}} |Y_{ij} - y|. \quad (4.23)$$

Le théorème suivant démontre bien la convergence et la normalité asymptotique de l'estimateur (4.22).

**Théorème 4.3.1** (Normalité asymptotique de l'estimateur de l'espérance conditionnelle). *Sous certaines hypothèses techniques, nous prouvons que  $\widehat{\phi \circ f^{-1}}$  converge presque sûrement vers  $\phi \circ f^{-1}$*

$$\left\| \widehat{\phi \circ f^{-1}} - \phi \circ f^{-1} \right\|_{\infty} \xrightarrow[m, n \rightarrow \infty]{as} 0.$$

En outre, pour  $n = m^{\frac{1}{2} + \alpha}$  with  $\alpha > 0$ . En outre, nous obtenons l'asymptotique suivante.  $\sqrt{m} \left( \widehat{\phi \circ f^{-1}} - \phi \circ f^{-1} \right)$  converge faiblement vers un processus gaussien, centré  $G$

$$\sqrt{m} \left( \widehat{\phi \circ f^{-1}} - \phi \circ f^{-1} \right) \xrightarrow[m \rightarrow \infty]{\mathbf{D}} G,$$

dont la fonction de covariance est donnée, pour tout  $(s, t) \in [f(a), f(b)]^2$ , par

$$\text{cov}(G(s), G(t)) = r(f^{-1}(s), f^{-1}(t)),$$

où  $r$  est la fonction de covariance du processus de déformation  $H$ . En outre, un intervalle de confiance asymptotique de niveau  $(1 - \alpha)$  pour estimer  $\phi \circ f^{-1}$  est donné par

$$\left[ \widehat{\phi \circ f^{-1}}(y) - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\text{var}}(G(y))}{m}}, \widehat{\phi \circ f^{-1}}(y) + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\text{var}}(G(y))}{m}} \right],$$

où nous avons noté  $\widehat{\text{var}}(G(y)) = \widehat{\gamma \circ f^{-1}}(y) - \left\{ \widehat{\phi \circ f^{-1}}(y) \right\}^2$  et  $u_{1-\frac{\alpha}{2}}$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  pour une gaussienne standard.

Ce théorème est démontré en utilisant des résultats de la théorie des processus empiriques. Nous nous référons sur ce point aux résultats publiés dans [82]. La démonstration se trouve quant à elle dans l'article [44].

Nous obtenons ainsi un nouvel outil de recalage de courbes vers un profil stable, donné par l'espérance structurelle. De plus, pour des déformations aléatoires quelconques, notre estimateur converge à la vitesse paramétrique optimale. Nous pouvons également établir des tests asymptotiques pour vérifier l'hypothèse du décalage d'un paquet de courbes par rapport à un profil unique particulier. Finalement, notre estimateur nous offre l'avantage de pouvoir être calculé très rapidement par rapport aux méthodes existantes [85], [111] ou [109]. Nous appliquons avec succès ce principe d'estimation à l'extraction d'un profil type pour le trafic routier, aux courbes de croissance d'enfants et aussi à l'extraction de notes lors de jurys de concours. Ces résultats sont détaillés dans l'article *Non parametric estimation of the structural expectation of a stochastic increasing function*, [44].

Une extension naturelle de ces méthodes de classification de données fonctionnelles nous est fournie par l'analyse d'images. En effet, dans un grand nombre de cas, nous observons des images décalées les unes par rapport aux autres. C'est notamment le cas lorsqu'un satellite passe plusieurs fois au même endroit pour photographier un lieu, mais avec des petits décalages spatiaux. C'est également le cas en imagerie médicale (scanner, IRM) ou en reconnaissance d'images. Notre objectif dans ce cas est de fournir une procédure permettant de retrouver un profil particulier dans les différentes images et de déterminer si l'image observée est proche ou non de ce profil. Nous choisissons de considérer la cas où ce profil est un contour dans l'image.

Ainsi les objets que nous souhaitons reconnaître dans les images sont des contours soumis à des déformations. Plus précisément, en analyse de formes, les objets que nous étudions sont des courbes dans le plan qui sont observées à quelques déformations près. Ces déformations sont des combinaisons de rotations, de translations et de dilatations. Analyser la forme originale à la base des observations implique que l'on puisse regarder toutes les configurations de déformation possibles. Par conséquent, l'espace naturel dans lequel nous souhaitons nous placer est l'espace de toutes ses courbes, quotienté par le groupe des transformations du plan. Cet objet a été étudié dans la littérature, tout particulièrement dans [74] où il porte le nom d'espace des contours de Kendall. Les objets à classifier deviennent dès lors des variétés riemanniennes. Notre étude se situe alors dans le cadre de classification de variétés riemanniennes.

## 4.4 Classification de variétés Riemanniennes

Les statistiques sur les variétés ont connu un intérêt croissant au cours des dernières années. Nous citons tout particulièrement les travaux de [13], [67, 68] ou [51]. C'est dans ce contexte que nous nous sommes placés pour construire une règle consistante de classification. Tout particulièrement, nous nous plaçons sur le plan des problèmes de classification où l'on cherche à diviser en deux, un échantillon observé  $X_1, \dots, X_n$ , théorie décrite dans [38] par exemple. La classification consiste ainsi à donner une étiquette  $Y \in \{0, 1\}$  à une observation  $X \in \mathcal{X}$ . On suppose que  $X$  et  $Y$  sont des variables aléatoires.

Classifier les données revient dès lors à construire une fonction  $g : \mathcal{X} \rightarrow \{0, 1\}$ , qui associe, à chaque observation  $X$ , son étiquette. L'erreur de classification a lieu lorsque  $g(X) \neq Y$ . On peut ainsi définir la probabilité d'erreur d'un classifieur  $g$  comme la probabilité

$$L(g) = \mathbf{P}(g(X) \neq Y).$$

Le meilleur classifieur possible est connu dans la littérature [38], [122] ou [1], comme le classifieur de Bayes. Il s'exprime de la manière suivante

$$g^*(x) = \begin{cases} 0, & \text{si } \mathbf{P}(Y = 0|X = p) \geq \mathbf{P}(Y = 1|X = p) \\ 1, & \text{sinon.} \end{cases}$$

Ce classificateur est optimal dans la mesure où, quelle que soit la règle de décision choisie  $g : \mathcal{X} \rightarrow \{0, 1\}$ , on aura toujours

$$L^* = \mathbf{P}(g^*(X) \neq Y) \leq \mathbf{P}(g(X) \neq Y) = L(g).$$

La problématique en classification est de construire un classifieur, possédant une erreur la plus proche possible de cette borne optimale, et qui ne dépende pas de la loi inconnue du couple  $(X, Y)$ , contrairement au classifieur de Bayes qui en dépend.

Afin de parvenir à ce résultat, nous allons construire un classifieur empirique, fondé sur  $n$  copies indépendantes  $(X_1, Y_1), \dots, (X_n, Y_n)$  de  $(X, Y)$ . Nous ne pouvons pas espérer obtenir une erreur empirique de classification égale à l'erreur obtenue avec le classifieur de Bayes, mais nous souhaitons que l'erreur de classification empirique, définie par

$$L_n = \mathbf{P}(g_n(X) \neq Y | (X_1, Y_1), \dots, (X_n, Y_n))$$

soit aussi proche que possible de  $L^*$ , avec grande probabilité. Plus précisément, nous cherchons à construire une règle consistante de classification  $g_n$ , au sens où

$$L_n \rightarrow L^*, \text{ a.s., as } n \rightarrow +\infty.$$

Les obstacles propres à notre cas particulier résultent du fait que les objets à classifier, sont des variétés Riemanniennes compactes. Soit  $K : \mathbb{R}_+ \rightarrow \mathbb{R}$  une fonction continue et positive telle que :

$$(K1) \int_{\mathbb{R}^d} K(\|u\|) d\lambda(u) = 1,$$

$$(K2) \text{supp } K = [0; 1],$$

où  $\lambda$  est la mesure de Lebesgue sur  $\mathbb{R}^d$ .

Pour  $p$  et  $q$  deux points de  $M$ , soit  $\theta_p(q)$  la densité volumique sur  $M$ , décrite dans [12] :

$$\theta_p : q \mapsto \theta_p(q) = \frac{\mu_{exp_p^*g}(exp_p^{-1}(q))}{\mu_{g_p}},$$

i.e., le quotient de la mesure canonique de la métrique Riemannienne  $exp_p^*g$  sur  $T_p(M)$  par la mesure de Lebesgue de la structure euclidienne  $g_p$  sur  $T_p(M)$ . En terme de coordonnées géodésiques normales au point  $p$ ,  $\theta_p(q)$  est égale à la racine carrée du déterminant de la métrique  $g$  prise au point  $exp_p^{-1}(q)$ . Nous pouvons souligner que nous avons la relation suivante  $\theta_p(q) = \theta_q(p)$  [126].

Dès lors, définissons un noyau sur les variétés  $K_h(p, \cdot) : M \rightarrow \mathbb{R}_+$ , en posant

$$K_h(p, q) = \frac{1}{\theta_p(q)} \frac{1}{h^d} K\left(\frac{d_g(q, p)}{h}\right), \quad (4.24)$$

pour tout  $q \in M$ , où le paramètre de fenêtre  $h$  vérifie la condition

$$h \leq h_0 < inj_g(M),$$

où  $inj_g(M)$  désigne le rayon d'injectivité de la variété  $M$  [qui est strictement positif puisque nous nous sommes restreints aux variétés compactes].

Il est donc à présent possible de répondre à notre objectif : construire une règle de classification sur des variétés de la manière suivante :

$$g_n^0(p) = \begin{cases} 0 & \text{si } \sum_{i=1}^n \mathbf{1}(Y_i = 0) K_{h_n}(p, X_i) \geq \sum_{i=1}^n \mathbf{1}(Y_i = 1) K_{h_n}(p, X_i) \\ 1 & \text{sinon,} \end{cases} \quad (4.25)$$

Comme nous l'avons démontré dans [94], cette règle de classification est convergente.

**Théorème 4.4.1.** *Sous l'hypothèse que  $h_n \rightarrow 0$  et  $nh_n^{2d} \rightarrow \infty$  alors*

$$L_n \rightarrow L^* \text{ presque sûrement,}$$

où  $L_n$  est l'erreur de classification associée à (4.25).

Cette recherche possède de nombreuses applications, en reconnaissance de formes, actuellement en cours de développement. En particulier, il fait l'objet d'une collaboration avec Airbus pour construire un répertoire de formes élémentaires en CAO pour les utilisateurs du logiciel CATIA.

## 4.5 Régression inverse spatiale

Le problème de la réduction de dimension en classification consiste à considérer que les variables  $Y$  et  $X$  sont liées par la relation

$$Y = g(\Phi X) + \epsilon,$$

où  $\epsilon$  vérifie  $\mathbf{E}(\epsilon|X) = 0$  et  $g : \mathbb{R}^D \rightarrow \mathbb{R}$  est une fonction inconnue,  $\Phi$  est une matrice orthogonale de dimension  $d \times D$  avec  $D \leq d$ . Notre objectif est d'estimer la matrice  $\Phi$  et  $g$  au moyen d'estimateurs à noyaux. Nous cherchons à construire des estimateurs de la matrice de covariance  $\Sigma = \text{Var}(X)$  et  $\Sigma_e = \text{Var}\mathbf{E}(X|Y)$ . Dans [28], nous prouvons le théorème suivant

**Théorème 4.5.1.** *Soit  $Z_i = (X_i, Y_i)$ ,  $i \in (\mathbb{N}^*)^N$  un champ stationnaire à valeurs dans  $\mathbb{R}^d \times \mathbb{R}$ . Soit  $n \in (\mathbb{N}^*)^N$ , supposons que l'on observe  $Z_n$  et notons  $n = (n_1, \dots, n_N)$ . Définissons  $\hat{n} = n_1 \times \dots \times n_N$ . Il est possible de construire des estimateurs au moyen de noyaux  $\hat{\Sigma}_n$  et  $\hat{\Sigma}_{e,n}$  tels que*

$$\begin{aligned}\hat{\Sigma}_n - \Sigma &= O\left(\frac{\sqrt{\log \log \hat{n}}}{\hat{n}}\right) \\ \hat{\Sigma}_{e,n} - \Sigma_e &= O\left(\frac{\sqrt{\log \log \hat{n}}}{\hat{n}}\right).\end{aligned}$$

Nous montrons qu'un choix approprié d'un paramètre de lissage permet de retrouver une vitesse de type paramétrique.



# Bibliographie

- [1] C. ABRAHAM, G. BIAU, and B. CADRE. On the kernel rule classification for function classification. *Ann. Math. Statist.*, 33 :1065–1076, 2006.
- [2] A. ANTONIADIS, I. GIJBELS, and G. GRÉGOIRE. Model selection using wavelet decomposition and applications. *Biometrika*, 84(4) :751–763, 1997.
- [3] E. BACRY, A. ARNEODO, U. FRISCH, Y. GAGNE, and E. HOPFINGER. Wavelet analysis of fully developed turbulence data and measurement of scaling exponents. In *Turbulence and coherent structures (Grenoble, 1989)*, pp. 203–215. Kluwer Acad. Publ., Dordrecht, 1991.
- [4] F. AUTIN, E. LE PENNEC, J-M. LOUBES, and V. RIVOIRARD, Maxisets for model selection estimators, à paraître dans *Journal of Nonparametric Statistics*, 2007.
- [5] Y. BARAUD. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.
- [6] Y. BARAUD. Model selection for regression on a random design. *ESAIM, Probab. Stat.*, 6 :127–146, 2002.
- [7] A. BARRON and C. SHEU. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3) :1347–1369, 1991.
- [8] A. BARRON, L. BIRGÉ, and P. MASSART. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.
- [9] A. BENASSI, S. COHEN and J. ISTAS. Identification and properties of real harmonizable fractional Lévy motions. *Bernoulli* **8** (1), 97–115, 2002.
- [10] J. G. BERRYMAN, *Lectures on Nonlinear Inversion and Tomography*, University of California, 1991.
- [11] A. BERLINET, F. LIESE, and I. VAJDA. Necessary and sufficient conditions for consistency of  $M$ -estimates in regression models with general errors. *J. Statist. Plann. Inference*, 89(1-2) :243–267, 2000.
- [12] A.L. BESSE. Manifolds all of whose geodesics are closed, volume 93 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1978.
- [13] R. BHATTACHARYA and V. PATRANGENARU. Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.*, 33(3) :1225–1259, 2005.
- [14] L. BIRGÉ and P. MASSART. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

- [15] L. BIRGÉ and P. MASSART. Gaussian model selection. *Journal of the European Mathematical Society*, 2001.
- [16] N. BISSANTZ, T. HOHAGE, and A. MUNK. Nonlinear tikhonov regularization for statistical inverse problems. *preprint*, 2004.
- [17] D. BÖHNING, A review of reliable maximum likelihood algorithms for semiparametric mixture models, *J. Statist. Plann. Inference*, **47**, no. 1-2, 5–28, 1995.
- [18] M. CARRASCO, J-P. FLORENS and E. RENAULT. Linear Inverse Problems in Structural Econometrics , *Handbook of Econometrics*, 6, 2006.
- [19] I. CASTILLO and J-M. LOUBES, Estimation of the distribution of random shifts deformation, soumis à *Statistica Sinica*, 2006.
- [20] L. CAVALIER, G. K. GOLUBEV, D. PICARD, and A. B. TSYBAKOV. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3) :843–874, 2002. Dedicated to the memory of Lucien Le Cam.
- [21] D. CHAFAI and J-M. LOUBES Nonparametric likelihood estimator for a certain class of inverse problems, *Statistics and Probability Letters*, 76 :1225–1237, 2006.
- [22] J. CHEN. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1), 221–233, 1995.
- [23] J. F. CLAERBOUT, *Imaging the Earth's Interior Blakwell Scientific Publications*, 1985.
- [24] S. COHEN. Ingénierie du trafic routier. *Presses de l'Ecole Nationale des Ponts et Chaussées*. INRETS, France, 1990.
- [25] A. COHEN, R. DE VORE, G. KERKYACHARIAN and D. PICARD Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2) :167–191, 2001.
- [26] A. COHEN, M. HOFFMANN, and M. REISS. Adaptive wavelet galerkin methods for linear inverse problems. *SIAM*, 1(3) :323–354, 2003.
- [27] I. CSISZÁR, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **19** : 2032–2066, 1991.
- [28] S. DABO-NIANG, J-M. LOUBES and A-F. YAO Spatial Kernel Inverse Regression *preprint*, 2006.
- [29] D. DACUNHA-CASTELLE and F. GAMBOA. Maximum d'entropie et problème des moments. *Annales de l'IHP Proba.*, 26 :567-596, 1990.
- [30] A.S. DALALYAN, G.K. GOLUBEV, and A.B. TSYBAKOV. Bayesian maximum likelihood and semiparametric second order efficiency. *Prépublication numéro 842 du laboratoire de probabilités et modèles aléatoires*, Université Paris 6 et Paris 7, 2003.
- [31] S. DAROLLES, J-P. FLORENS, and E. RENAULT. Nonparametric instrumental regression. *preprint*, 2001.
- [32] I. DAUBECHIES, M. DEFRISE and C. DE MOL – An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math* **57**, p. 1413–1541, 2004.

- [33] M. DAVIDIAN and D. GILTINAN, Nonlinear Models for Repeated Measurement Data : An Overview and Update, *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, 387–419, 2003. <http://www4.stat.ncsu.edu/~davidian/>.
- [34] B. DELYON, M. LAVIELLE, and E. MOULINES. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1), 94–128, 1999.
- [35] A. DETAIS, J-F. DUPUY and J-M. LOUBES Time Travel Forecasting with agregation methods *preprint*, 2006.
- [36] P. DEUFLHARD, H. ENGL, and O. SCHERZER. A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions. *Inverse Probl.*, 14(5) :1081–1106, 1998.
- [37] R. DEVORE and G. G. LORENTZ, Constructive Approximation. *Springer-Verlag*, Berlin, 1993.
- [38] L. DEVROYE, L. GYÖRFI, and G. LUGOSI. A probabilistic theory of pattern recognition, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.
- [39] A. K. DEY, F. H. RUYMGAART, and B. A. MAIR. Cross-validation for parameter selection in inverse estimation problems. *Scand. J. Statist.*, 23(4) :609–620, 1996.
- [40] D. DONOHO , and I. JOHNSTONE, Minimax risk for  $l_q$  losses over  $l_p$ -balls. *Probability Theory and Related Fields*, 99 : 277–303, 1994.
- [41] D. DONOHO and I. JOHNSTONE, Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81 : 425–455, 1994.
- [42] D. DONOHO , De-noising via soft-thresholding. *IEEE Transaction in Information Theory* 41 : 613–627, 1995.
- [43] D. DONOHO, I. JOHNSTONE, G. KERKYACHARIAN, and D. PICARD. Wavelet shrinkage : asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369, 1995. With discussion and a reply by the authors.
- [44] J-F. DUPUY, J-M. LOUBES and E. MAZA. Non parametric estimation of the structural expectation of a stochastic increasing function soumis à *Scandinavian Journal of Statistics*, 2006.
- [45] B. EFRON, T. HASTIE, I. JOHNSTONE and R. TIBSHIRANI Least angle regression, *Ann. Statist.* **32**, no. 2, p. 407–499, With discussion, and a rejoinder by the authors, 2004.
- [46] H. ENGL, M. HANKE, and A. NEUBAUER. Regularization of inverse problems, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.
- [47] H. ENGL. Regularization methods for solving inverse problems. In *ICIAM 99 (Edinburgh)*, pages 47–62. Oxford Univ. Press, Oxford, 2000.
- [48] A-K. FERMIN, J-M. LOUBES, and C. LUDEÑA. Bayesian methods in seismic tomography. *International Journal of Tomography and Statistics*, 4(6) :1–18, 2006.
- [49] A-K. FERMIN, J-M. LOUBES, and C. LUDEÑA. Model selection for linear ill-posed problems. Proceedings of Oberwolfach, *Oberwolfach Workshop Statistische und Probabilistische Methoden der Modellwahl*,(47) : 59–63, 2006.

- [50] J. FILLOL. Multifractality : Theory and Evidence. An Application to the French Stock Market. *Economics Bulletin*, **3**, 31, 1–12, 2003.
- [51] N.I. FISHER, T. LEWIS, and B.J.J. EMBLETON. Statistical Analysis of Spherical Data. *Cambridge University Press*, 1993.
- [52] F. GAMBOA, J-M. LOUBES and E. MAZA M-estimation of shifts between curves in positive revision for *Annals of Statistics*, 2005.
- [53] F. GAMBOA and J-M. LOUBES Bayesian estimation of multifractal wavelet function. *Bernoulli*, (2), 2005.
- [54] F. GAMBOA, F. and J-M. LOUBES Estimation of the parameters of a multifractal wavelet function. *Test*, to appear 2006.
- [55] F. GAMBOA and E. GASSIAT, Bayesian Methods and Maximum Entropy for Ill-Posed Inverse Problems, *The Annals of Statistics*, Vol. 25, N. 1, 328-350, 1997.
- [56] F. GAMBOA New Bayesian methods for ill posed problems, *Stat. Decis.* 17, number 4, 315-337, 1999.
- [57] T. GASSER and A. KNEIP. Searching for structure in curve samples. *J. Amer. Statist. Assoc.*, 90 :1179–1188, 1995.
- [58] M. GIBALDI and D. PERRIER, *Pharmacokinetics, 2nd edition*, 1982.
- [59] C. GIUTYS, *Adaptative density estimation in deconvolution*, JASA, 1997.
- [60] I. J. GOOD and R. A. GASKINS. Nonparametric roughness penalties for probability densities. *Biometrika*, 58 :255–277, 1971.
- [61] U. GRENANDER General Pattern Theory *Oxford University Press*, 1993.
- [62] L. GYÖRFI, M. KOHLER, A. KRZYŻAK, and H. WALK *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.
- [63] M. HANKE, A. NEUBAUER, and O. SCHERZER. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.*, 72(1) :21–37, 1995.
- [64] W. HÄRDLE, G. KERKYACHARIAN, D. PICARD, and A. TSYBAKOV. *Wavelets, approximation, and statistical applications*. Springer-Verlag, New York, 1998.
- [65] T. HASTIE and R. TIBSCHIRANI *Generalized additive models* London : Chapman Hall, 1990.
- [66] J. HAUSMAN and W. K. NEWEY. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, 63 :1445–1476.
- [67] H. HENDRIKS. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *Ann. Statist.*, 18 :832–849, 1990.
- [68] H. HENDRIKS, J. JANSSEN, and F. RUYMGAART. Strong uniform convergence of density estimators on compact euclidean manifolds. *Statist. Probab. Letters*, 16 :305–311, 1993.
- [69] S. JAFFARD. On lacunary wavelet series. *The Annals of Applied Probability*, 10(1) :313–329, 2000.

- [70] M. JAMBU Classification automatique pour l'analyse des données. I, *Méthodes et algorithmes*, Dunod, 1978.
- [71] Q. JIN and U. AMATO. A discrete scheme of Landweber iteration for solving nonlinear ill-posed problems. *J. Math. Anal. Appl.*, 253(1) :187–203, 2001.
- [72] B. KALTENBACHER. On Broyden's method for the regularization of nonlinear ill-posed problems. *Numer. Funct. Anal. Optimization*, 19(7-8) :807–833, 1998.
- [73] B. KALTENBACHER. Regularization by projection with a posteriori discretization level choice for linear and nonlinear ill-posed problems. *Inverse Probl.*, 16(5) :1523–1539, 2000.
- [74] D.G. KENDALL, D. BARDEN, T.K. CARNE, and H. LE. Shape and Shape Theory. *Wiley Series in Probability and Statistics*. Wiley, 1999.
- [75] G. KERKYACHARIAN and D. PICARD Minimax or maxisets? *Bernoulli*, 8(2) :219–253, 2002.
- [76] A. KNEIP and T. GASSER. Statistical tools to analyze data representing a sample of curves. *Ann. Statist.*, 20(3) :1266–1305, 1992.
- [77] A. KNEIP, X. LI, K.B. MACGIBBON, and J.O. RAMSAY. Curve registration by local regression. *Canad. J. Statist.*, 28(1) :19–29, 2000.
- [78] C. LACAUX and J-M. LOUBES Estimation of the Hurst exponent of a RHFLM. *submitted to ALEA*, 2006.
- [79] T. LAI and M-C. SHIH, Nonparametric estimation in nonlinear mixed effects models, *Biometrika*, **90**, no. 1, 1–13, 2003.
- [80] M. LAVIELLE. On the use of penalized contrasts for solving inverse problems. *Preprint de l'université Paris Sud*, 2002.
- [81] W.H. LAWTON, E.A. SYLVESTRE and M.S. MAGGIO. Self-modeling regression. *Technometrics.*, 14 :513–532, 1972.
- [82] M. LEDOUX and M. TALAGRAND Probability in Banach spaces, Volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Berlin : Springer-Verlag.s :invM Isoperimetry and processes, 1991.
- [83] M. LEDOUX Concentration of measures and Logarithmic Sobolev inequalities, *Notes de cours de Berlin*, 1997.
- [84] W. LELAND, M. TAQQU, W. WILLINGER and D. WILSON. On the self-similar nature of Ethernet traffic. *IIE/ACM Transactions on Networking*, 2 :1–15, 1994.
- [85] X. LIU and H.-G. MÜLLER Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* 99(467), 687–699, 2004.
- [86] J-M. LOUBES, M-estimation pénalisée, *Thèse de l'université Toulouse 3*, 2001.
- [87] J-M. LOUBES  $l^1$  penalty for ill-posed inverse problems, soumis à *Communication in Statistics*, 2006.
- [88] J-M. LOUBES,  $l^1$  sparsity and application in estimation Soumis à *Notes aux Comptes Rendus de l'Académie de Sciences*, 2006.

- [89] J-M. LOUBES, and C. LUDEÑA. Penalized estimators for non linear ill-posed problems. soumis à *ESAIM PS*, 2006.
- [90] J-M. LOUBES and P. MASSART, Comments on Least Angle Regression Algorithm, *Annals of Statistics*, **32** (2) : 476–482, 2004.
- [91] J-M. LOUBES, and E. MAZA, M. LAVIELLE and L. RODRIGUEZ Road trafficking description and short term travel time forecasting, with a classification method *Canadian Journal of Statistics*, 2006
- [92] J-M. LOUBES and E. MAZA, road traffic forecasting with semiparametric method to appear in *Scientific computing world*, 2006.
- [93] J-M. LOUBES and D. PAINDAVEYNE Parametric Estimation of a multifractal function. soumis à *Journal of Statistical Planning and Inference*, 2006.
- [94] J-M. LOUBES and B. PELLETIER. Consistency of a kernel rule for classification on a Riemannian manifold soumis à *Statistica Sinica*, 2006.
- [95] J-M. LOUBES and B. PELLETIER. Bayesian estimator for ocean remote sensing. *preprint*, 2006.
- [96] J-M. LOUBES and V. RIVOIRARD Review of rates of convergence and regularity conditions for inverse problems. soumis à *Canadian journal of statistics*, 2005.
- [97] J-M. LOUBES and A. VANHEMS, Estimation of the solution of a differential equation with endogenous effects, soumis à *Journal of the Australian Econometric Society*, 2006.
- [98] J-M. LOUBES and A. VANHEMS. Saturation space for inverse problems in econometry. *Proceedings of the american econometric society*, 2004.
- [99] J.-M. LOUBES and S. VAN DE GEER,  $l^1$  thresholding penalty in adaptive M-estimation. *Statist. Neerlandica* **56**, no. 4, 454–479, 2002.
- [100] J-M. LOUBES, O. WINTENBERGER and M. SEBAG. Non parametric EM-based Clustering and Application to road traffic prediction. *preprint*, 2006.
- [101] B. MANDELBROT and J. VAN NESS. Fractional Brownian motion, fractionnal noises and applications. *Siam Review* **10**, 422–437, 1968.
- [102] P. MASSART Concentration Inequalities and Model Selection. *Cours de Saint-Flour*, Springer, 2005.
- [103] D. L.MCLEISH and C. G. SMALL. Likelihood methods for the discrimination problem. *Biometrika*, **73**(2) :397–403, 1986.
- [104] F. MENTRÉ, A. MALLET, and D.BACCAR, Optimal design in random-effects regression models, *Biometrika* **84**, no. 2, 429–442, 1997.
- [105] F. O’SULLIVAN. Convergence characteristics of methods of regularization estimators for nonlinear operator equations. *SIAM J. Numer. Anal.*, **27**(6) :1635–1649, 1990.
- [106] J. PFANZAGL, Consistency of maximum likelihood estimators for certain nonparametric families, in particular : mixtures, *J. Statist. Plann. Inference*, **19**, no. 2, 137–158, 1988.
- [107] J. PFANZAGL, Large deviation inequality for maximum likelihood estimators for certain nonparametric families, in particular : mixtures, *Ann. of Stats.*, **19**, no. 2, 137–158, 1988.

- [108] J.O. RAMSAY and X. LI. Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(2) :351–363, 1998.
- [109] J. RAMSAY and B. SILVERMAN. Applied functional data analysis. *Springer Series in Statistics*. New York : Springer-Verlag. Methods and case studies. 2002.
- [110] R. RIEDI, M. CROUSE, V. RIBEIRO, and R. BARANIUK. A multifractal wavelet model with application to network traffic. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 45(3) :992–1018, 1999.
- [111] B. RØNN. Nonparametric maximum likelihood estimation for shifted curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 63(2), 243–259, , 2001.
- [112] J.D. SARGAN. The Estimation of Economic Relationship using Instrumental Variables. *Econometrica*. 1958.
- [113] D. W. SCOTT, R. A. TAPIA, and J. R. THOMPSON. Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *Ann. Statist.*, 8(4) :820–832, 1980.
- [114] B. W. SILVERMAN. Some remarks on roughness penalty density estimators. In *Limit theorems in probability and statistics, Vol. I, II (Veszprém, 1982)*, pages 957–979. North-Holland, Amsterdam, 1984.
- [115] B. SILVERMAN. On the estimation of a probability density function by the maximum penalized likelihood method, *The Annals of Statistics* **10**, 795–810, 1982.
- [116] C. STONE. Additive regression and other nonparametric models, *The Annals of Statistics*, **13**, 689–705, 1985.
- [117] A. SWENSEN. The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *J. Multivariate Anal.* **16**, 54–70, 1985.
- [118] A. TIKHONOV, N. ARSEININ and Y. VASILYIY. Solutions of ill-posed problems. *Scripta Series in Mathematics*. New York etc. : John Wiley & Sons; Washington, D.C. : V. H. Winston & Sons. XIII, 258 p., 1977
- [119] A. N. TIKHONOV, A. S. LEONOV, and A. G. YAGOLA. Nonlinear ill-posed problems. Vol. 1, 2, volume 14 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1998. Translated from the Russian.
- [120] S. A. VAN DE GEER, Applications of empirical process theory, *Cambridge Univ. Press*, Cambridge, 2000.
- [121] A.W.VAN DER VAART. Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. *Cambridge University Press*, Cambridge, 1998.
- [122] V. VAPNIK. The nature of statistical learning theory. *Springer-Verlag*, New York, second edition, 2000.
- [123] R. VOJAK, Analyse et Modélisation Multifractales de Signaux Complexes. Applications au Trafic Routier *Thèse de l'université Paris Dauphine*, 1996.
- [124] K. WANG and T. GASSER. Alignment of curves by dynamic time warping. *Ann. Statist.*, 25(3) :1251–1276, 1997.

- [125] K. WANG and T. GASSER. Synchronizing sample curves nonparametrically. *Ann. Statist.*, 27(2) :439–460, 1999.
- [126] T.J. WILLMORE. *Riemannian geometry*. Oxford University Press, 1993.