Sélection de variables pour la classification non supervisée en grande dimension

Caroline Meynet

Synthèse

1 Sélection de variables pour la classification non supervisée en grande dimension

1.1 Modèles de mélange gaussien pour la classification non supervisée

En présence de n observations $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ décrites par p variables ($\boldsymbol{Y_i} \in \mathbb{R}^p$) et présentant des caractéristiques différentes, le but de la classification non supervisée est de partitionner ces observations en plusieurs classes de façon à regrouper entre elles les observations de caractéristiques semblables.

Pour déterminer une partition des observations, il est d'usage d'optimiser un critère pour créer des classes de telle sorte que chaque classe soit la plus homogène possible et la plus distincte possible des autres classes. En pratique, il est impossible d'explorer toutes les partitions possibles. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une "bonne" partition qui correspond en général à un optimum local. Même si le besoin de classer des objets est très ancien, seule la généralisation des outils informatiques en a permis l'automatisation dans les années 1970. Celeux et al. (1989) décrivent en détail ces algorithmes. Deux principaux types de méthodes de classification non supervisée existent : les méthodes combinatoires où le critère optimisé est une distance (K-means, classification hiérarchique), et les méthodes de modèles de mélange qui supposent que les données forment un échantillon suivant une densité de mélange (c'est-à-dire une somme pondérée de densités représentant chacune une classe), le critère optimisé étant alors un critère de maximum de vraisemblance pour ajuster le modèle aux données. Pour ces dernières méthodes, le problème de classification est reformulé en un problème d'estimation de densité.

L'objectif principal de cette thèse est de proposer une procédure de sélection des variables pertinentes pour l'obtention d'une classification des données. Comme les méthodes de modèles de mélange offrent un cadre statistique rigoureux pour déterminer le nombre de classes et les variables pertinentes pour la classification, elles s'avèrent particulièrement adaptées à notre problématique. Nous nous placerons donc dans un cadre de modèles de mélange. Nous considérerons le cas important des modèles de mélange gaussien. Nous nous restreindrons à l'étude de matrice de covariance sphérique commune à toutes les classes. Dans ce cas, les classes ne se distinguent que par la position de leur centre, qui est donnée par les vecteurs des moyennes. La densité s de l'échantillon $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ est alors

estimée par une densité de mélange de la forme

$$s_{\boldsymbol{\theta}} : \mathbb{R}^p \mapsto \mathbb{R},\ \boldsymbol{y} \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}) \tag{1}$$

où $\Phi(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I})$ désigne la densité gaussienne p-dimensionnelle définie pour tout $\boldsymbol{y} \in \mathbb{R}^p$ par

$$\begin{aligned}\Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}) &= \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu_k})^T(\boldsymbol{y} - \boldsymbol{\mu_k})\right) \\ &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_j - \mu_{kj})^2\right).\end{aligned}$$

Le vecteur des paramètres est $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Theta_K := \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+^\star$ où $\Pi_K = \{(\pi_1, \ldots, \pi_K) \in [0,1]^K;\ \sum_{k=1}^K \pi_k = 1\}$. Il rassemble les proportions π_k du mélange, les vecteurs $\boldsymbol{\mu_k}$ des moyennes représentant le centre de chaque classe et la variance σ^2 indiquant que chaque classe a une forme sphérique identique.

Supposons s estimée par $s_{\hat{\boldsymbol{\theta}}}$. Alors les observations sont classées suivant la règle suivante, appelée règle du Maximum A Posteriori (MAP). Pour tout $i \in \{1, \ldots, n\}$, pour tout $k \in \{1, \ldots, K\}$, notons

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k \, \Phi\left(\boldsymbol{Y_i} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \hat{\pi}_l \, \Phi\left(\boldsymbol{Y_i} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)} \tag{2}$$

la probabilité conditionnelle d'appartenance de l'observation $\boldsymbol{Y_i}$ à la classe k. Alors, on déclare $\boldsymbol{Y_i}$ appartenir à la classe k si $\hat{\tau}_{ik} > \hat{\tau}_{il}$ pour tout $l \neq k$.

1.2 Sélection des variables pertinentes pour la classification

On pourrait penser que plus on augmente le nombre de variables décrivant chaque observation d'un échantillon, plus on dispose d'informations concernant ces observations et plus on en facilite et on en améliore la classification. Cependant, la qualité de la classification ne dépend pas du nombre d'informations à disposition mais de la pertinence de ces informations. Parmi les variables à disposition, il s'avère souvent que seules certaines d'entre elles contiennent la structure d'intérêt des observations. Ces variables pertinentes suffisent à distinguer les différences de caractéristiques entre les observations et à les regrouper en classes. Au contraire, certaines variables peuvent ne pas avoir de lien avec la structure des observations, auquel cas la prise en compte de ces variables pour déterminer la classification risque de fausser et de détériorer la classification. Ces variables sont nuisibles pour la classification. D'autres variables, sans être nuisibles, peuvent être inutiles pour déterminer la classification si elles sont redondantes par rapport à des variables pertinentes. Supprimer ces variables inutiles permet alors d'obtenir un modèle plus simple et plus interprétable, ce qui est un point essentiel pour

les practiciens qui souhaitent comprendre le phénomène étudié au travers de la classification obtenue.

Par exemple, dans le domaine de la biologie, les chercheurs souhaitent identifier les fonctions des gènes en mesurant leur variation de niveau d'expression dans un ensemble d'expériences sur puces à ADN. Ils supposent que des gènes ayant des profils d'expression similaires ont des liens fonctionnels. Ainsi, l'objectif est de déterminer des classes de gènes co-exprimés (Eisen et al., 1998). Cependant, parmi toutes les expériences effectuées, une partie seulement d'entre elles se révèlent liées aux phénomènes biologiques étudiés. Il est alors préférable de ne considérer que ces expériences pour mettre en lumière ces phénomènes.

L'identification des variables pertinentes pour la classification est donc primordiale et l'enjeu du statisticien est de proposer des procédures de sélection de variables permettant la sélection de toutes les variables pertinentes et l'élimination de toutes les autres. La difficulté principale est de construire une méthode de sélection de variables sans savoir à quelle classe appartiennent les observations. Deux types de procédures de sélection de variables existent : les méthodes "filter" et "wrapper". Pour les premières, la sélection de variables est effectuée en amont du processus de classification (Dash et al., 2002 ; Jouve et Nicoloyannis, 2005). Pour les secondes, la sélection de variables est insérée au sein du processus de classification. Les méthodes wrapper présentent l'avantage de ne pas dissocier les problèmes de sélection de variables et de classification, ce qui permet de mieux appréhender et interpréter le rôle des variables. C'est cette seconde approche que nous envisagerons dans cette thèse. Pour les méthodes de classification basées sur des modèles de mélange gaussien, les méthodes wrapper ont principalement été introduites sous un angle bayésien. On peut par exemple citer Law et al. (2004) qui introduisent le concept de "feature saliency" pour évaluer l'importance des variables sous l'hypothèse d'indépendance entre les variables non pertinentes et pertinentes. Raftery et Dean (2006) puis Maugis et al. (2009) étendent cette procédure en s'affranchissant de l'hypothèse d'indépendance. Pan et Shen (2007) privilégient quant à eux une approche fréquentiste de sélection de variables par pénalisation ℓ_1 de la vraisemblance des modèles. C'est cette dernière idée que nous reprendrons.

1.3 Le défi de la grande dimension

La sélection de variables a pris toute son importance avec l'apparition et la multiplication des données de très grande dimension ces dernières années.

1.3.1 Données de grande dimension

Grâce aux progrès technologiques, l'acquisition de données devient de plus en plus facile techniquement et des bases de données gigantesques sont collectées quasi-quotidiennement. Par conséquent, le nombre de variables présentes dans les problèmes statistiques actuels peut maintenant atteindre des dizaines voire des centaines de milliers. Dans le même temps, pour de nombreuses applications,

le nombre d'observations se trouve réduit et peut n'être que de quelques dizaines. Dans cette thèse, nous dirons que les données considérées sont de grande dimension, et nous écrirons $p \gg n$, quand le nombre p de variables est très grand devant le nombre n d'observations.

Pour certains champs d'application tels la biologie, la climatologie, l'économétrie, la chimie quantitative, les observations peuvent même être de dimension infinie. C'est le cas lorsque les données recueillies sont de nature continue (courbes, images). En présence de telles données fonctionnelles, un objectif essentiel de la classification non supervisée de ces données est de permettre l'obtention d'une bonne estimation d'un profil type pour chaque classe. Par exemple, la demande en électricité varie selon les saisons ou les jours de la semaine, ce qui se traduit par une allure différente des courbes de consommation électrique. Ainsi, ces courbes peuvent être partitionnées en plusieurs classes suivant leur allure. Une bonne identification des classes et une bonne classification des courbes permet de fournir une bonne représentation de la courbe de la consommation électrique classe par classe. L'enjeu est d'améliorer les estimations et les prévisions de consommation électrique en tenant compte de la période de l'année ou de la semaine (Antoniadis et al., 2011).

1.3.2 Hypothèse de parcimonie

Face à ces données de grande dimension, une hypothèse souvent faite est l'hypothèse dite de parcimonie. Elle consiste à supposer que parmi les très nombreuses variables à notre disposition, peu d'entre elles (disons au maximum de l'ordre de n) sont en fait utiles pour expliquer les observations et donc pertinentes pour la classification. Cela revient à supposer que la très grande majorité des variables sont inutiles (si elles n'apportent que de l'information redondante) voire même néfastes (si elles n'ont rien à voir avec la classification) pour déterminer la classification. Cette hypothèse semble raisonnable car elle traduit le fait que la dimension impressionnante des données que nous recevons n'est qu'une illusion créée par les progrès informatiques et qu'elle ne reflète pas la réelle complexité du problème que l'on peut penser être bien inférieure.

Par exemple, en théorie du signal, de nombreux signaux a priori décrits dans un espace de dimension infinie peuvent en fait être bien approximés dans un espace de petite dimension. Une application majeure de cette propriété de parcimonie est la compression des signaux (Mallat, 1989).

1.3.3 Vers de nouvelles procédures de sélection de variables

Pour des données décrites par p variables, sélectionner un ensemble de variables pertinentes pour la classification revient à sélectionner un sous-ensemble de $\{1, \ldots, p\}$. Or, il y a 2^p tels sous-ensembles. Une recherche exhaustive du meilleur sous-ensemble de variables n'est donc pas envisageable au vu des performances informatiques actuelles. Maugis et Michel (2011a) ont été confrontés à ce problème et n'ont pas pu mettre en pratique au delà de $p \approx 10$ la théorie de sélection de variables complète (ou

au delà de $p \approx 30$ pour la sélection de variables ordonnée) qu'ils ont développée dans le cadre de la classification non supervisée par modèles de mélange gaussien.

En grande dimension, il est nécessaire d'introduire des procédures de sélection de variables alternatives à la sélection de variables complète qui soient algorithmiquement faisables. Comme la sélection de variables en grande dimension est un enjeu récent dans le cadre de la classification non supervisée, peu de méthodes existent à ce jour.

Les méthodes basées sur des modèles de mélange gaussien fournissent un cadre statistique bien adapté à la reformulation du problème de sélection de variables en un problème de sélection de modèles. En particulier, dans le cas monoclasse ($K = 1$), le mélange de densités gaussiennes (1) n'est autre qu'une densité gaussienne et le modèle correspondant peut être assimilé à un modèle de régression linéaire gaussienne avec design déterministe. Ainsi, des méthodes de sélection de variables en classification non supervisée par modèles de mélange gaussien peuvent être construites en adaptant au cas multiclasses ($K \geq 2$) des méthodes de sélection de variables testées en régression gaussienne. Par exemple, Law et al. (2004), Raftery et Dean (2006) puis Maugis et al. (2009) considèrent des méthodes analogues à la méthode stepwise utilisée pour la sélection de variables en régression, en comparant à chaque étape deux modèles emboîtés pour déterminer quelle variable doit être exclue ou inclue dans le modèle. En parallèle, Pan et Shen (2007) se sont inspirés du succès du Lasso en régression pour développer une méthode de sélection de variables par régularisation ℓ_1 de la vraisemblance observée.

2 Synthèse des travaux réalisés

Dans cette thèse, nous construisons une procédure de classification non supervisée en grande dimension reprenant l'usage de la pénalisation ℓ_1 pour sélectionner les variables pertinentes. Notre procédure se démarque de celle proposée par Pan et Shen (2007) par deux points essentiels : l'estimation des paramètres du mélange et le critère de sélection de modèles. L'amélioration apportée à l'estimation des paramètres du mélange nous permet notamment de traiter efficacement des problèmes de reconstitution de courbes dans le contexte de classification de données fonctionnelles, alors que la méthode de Pan et Shen (2007) se révèle inadaptée à ce genre de problèmes. De même que Pan et Shen (2007) se sont inspirés des propriétés de sélection de variables du Lasso constatées en régression pour établir leur procédure, c'est au vu des problèmes d'estimation du Lasso en régression que nous avons jugé nécessaire de modifier l'étape d'estimation de Pan et Shen (2007).

Le manuscrit comporte deux parties indépendantes :

1. Dans la Partie I, nous nous concentrons sur l'aspect régularisation ℓ_1 du Lasso en établissant des inégalités oracle ℓ_1 satisfaites par cet estimateur. Deux cadres sont considérés : un cadre gaussien linéaire puis un cadre gaussien non linéaire. Cette partie est purement théorique.

2. Dans la Partie II, nous exploitons les propriétés de sélection de variables du Lasso pour établir une procédure de classification non supervisée intégrant la sélection simultanée des variables pertinentes pour faire cette classification. Nous nous plaçons dans un cadre de mélange fini de densités gaussiennes multivariées en grande dimension. Cette partie mêle théorie et simulations.

Ces deux parties sont suivies d'un chapitre annexe dans lequel nous présentons deux procédures que nous avons envisagées au cours de nos recherches et qui peuvent constituer des alternatives à la procédure que nous allons présenter en Partie II. Nous comparons ces trois procédures sur des données simulées afin de motiver notre choix pour la procédure finalement retenue.

Partie I. Some ℓ_1-oracle inequalities for the Lasso in Gaussian regression models

Bien que défini comme un estimateur régularisé en norme ℓ_1, le Lasso doit principalement son succès à ses propriétés de parcimonie qui, additionnées à son caractère convexe, font de lui un substitut efficace à la régularisation en "norme" ℓ_0. Ainsi, les principaux résultats de prédiction sur cet estimateur sont des inégalités oracle le comparant à un pseudo-oracle ℓ_0. Ces résultats nécessitent des contraintes sur la matrice de Gram, qui sont en pratique difficilement vérifiées en grande dimension. Dans cette partie, nous nous focalisons sur le Lasso non pas comme procédure de sélection de variables, mais comme algorithme de régularisation ℓ_1. Dans cette optique, nous établissons des inégalités oracle ℓ_1 satisfaites par cet estimateur afin de comparer son risque de prédiction à l'oracle ℓ_1. Cela permet de fournir des résultats de prédiction complémentaires aux résultats de prédiction traditionnellement établis pour le Lasso dans le cadre de parcimonie. Nos résultats ℓ_1 s'affranchissent des contraintes sur la matrice de Gram nécessaires à l'établissement des résultats ℓ_0. De plus, ils restent valables en dehors du contexte de parcimonie.

Chapitre 2 Homogeneous Gaussian regression models

Cadre

Nous nous plaçons dans un cadre de modèles de régression gaussienne. La fonction de régression est décomposée dans un dictionnaire fini (régression linéaire gaussienne par exemple) ou infini dénombrable (ondelettes) ou infini indénombrable (réseau de neurones).

Résultats

Deux types de résultats sont établis : d'abord des inégalités oracle ℓ_1, desquelles sont ensuite déduites des vitesses de convergence. Les inégalités oracle sont obtenues en appliquant une version simplifiée d'un théorème général de sélection de modèle (Massart, 2007) où nos modèles sont des boules ℓ_1.

Nous considérons d'abord le cas de dictionnaires finis $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$. Nous démontrons le résultat suivant.

Théorème 2.3.2 Pour tout paramètre de régularisation

$$\lambda \geq 4\sigma\sqrt{\frac{1+\ln p}{n}}, \tag{3}$$

il existe une constante $C > 0$ telle que, pour tout $A > 0$, l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ défini par

$$\hat{\boldsymbol{\beta}}^*(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} \tag{4}$$

satisfait l'inégalité oracle suivante avec probabilité plus grande que $1 - 3.4\mathrm{e}^{-A}$:

$$\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*(\lambda)\|^2 + \lambda\|\hat{\boldsymbol{\beta}}^*(\lambda)\|_1 \leq C \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} + \frac{\lambda(1+A)}{\sqrt{n}}. \tag{5}$$

En intégrant par rapport à A, nous obtenons l'inégalité oracle en espérance suivante :

$$\mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*(\lambda)\|^2 + \lambda\|\hat{\boldsymbol{\beta}}^*(\lambda)\|_1 \right] \leq C \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} + \frac{\lambda}{\sqrt{n}}. \tag{6}$$

L'inégalité oracle ℓ_1 (5) est à mettre en parallèle des inégalités oracle ℓ_0 traditionellement établies. Contrairement à celles-ci, notre inégalité ne nécessite aucune hypothèse, ni sur la matrice de Gram, ni sur la parcimonie de $\boldsymbol{\beta}^*$. Dans le cas orthogonal, cette inégalité oracle permet de retrouver les vitesses de convergence optimales établies sur les espaces de Besov par Cohen et al. (2001) pour les estimateurs par seuillage doux auxquels est équivalent le Lasso.

Nous considérons ensuite le cas de dictionnaires dénombrables ordonnés $\mathcal{D} = \{\phi_1, \ldots, \phi_p, \ldots\}$ (ondelettes par exemple). Pour de tels dictionnaires, une calibration théorique du paramètre de régularisation λ comme proposée en (3) n'est plus possible car on ne dispose plus de taille finie p du dictionnaire. Nous proposons une procédure permettant la calibration de λ par choix d'un meilleur niveau de troncature du dictionnaire au sens suivant. Nous considérons la suite d'estimateurs Lasso associés à la suite de dictionnaires tronqués $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$. Nous pénalisons chacun de ces estimateurs suivant la taille du dictionnaire. Nous choisissons alors le niveau de troncature $\hat{p}$ réalisant le meilleur compromis entre qualité de l'approximation, régularisation ℓ_1 et parcimonie (taille du dictionnaire). L'estimateur ainsi obtenu correspond à l'estimateur Lasso sur le dictionnaire tronqué $\{\phi_1, \ldots, \phi_{\hat{p}}\}$. Nous appelons cet estimateur "estimateur Lasso sélectionné". Dans le cas orthogonal où les estimateurs Lasso correspondent aux estimateurs par seuillage doux, notre procédure permet de régler le problème crucial du choix du seuil.

Nous établissons une inégalité oracle pour cet estimateur Lasso sélectionné. Nous en déduisons des vitesses de convergence sur des espaces de Besov dans le cas orthogonal, puis sur des espaces

d'interpolation dans le cas non orthogonal. Dans le cas orthogonal, nous établissons des vitesses minimax prouvant que les vitesses de convergence de l'estimateur Lasso sélectionné sont optimales. En outre, cet estimateur est adaptatif aux espaces de Besov, contrairement aux estimateurs Lasso classiques.

Le cas des dictionnaires infinis utilisés pour les réseaux de neurones est finalement considéré. Une inégalité oracle ℓ_1 et des vitesses de convergence sont établies pour le Lasso.

Discussion

Contrairement aux inégalités oracle ℓ_0 usuelles considérées pour évaluer les performances du Lasso en sélection de variables, nos inégalités oracle ℓ_1 ne nécessitent aucune hypothèse. Cependant, cela n'a rien de surprenant : le Lasso est défini comme un estimateur régularisé en norme ℓ_1, on peut donc s'attendre à l'obtention d'une inégalité oracle ℓ_1 sans autre hypothèse qu'une bonne calibration de pénalisation, qui est traduite par la minoration (3) du paramètre de régularisation.

Des inégalités oracle du type de (5) et (6) ont déjà été établies (Huang et al., 2008 ; Rigollet et Tsybakov, 2011 ; Bartlett et al., 2012). L'originalité de nos résultats réside dans l'approche envisagée pour les démontrer : l'idée est de voir la procédure Lasso comme une procédure de sélection de modèle où les modèles sont des boules ℓ_1, ce qui nous permet d'exploiter la théorie sur la sélection de modèle (Massart, 2007). La linéarité de la décomposition de la fonction de régression dans le dictionnaire et le fait de considérer des erreurs gaussiennes nous permettent d'appliquer une inégalité maximale gaussienne et d'obtenir une minoration (3) du paramètre de régularisation optimale en n. Si des arguments entropiques étaient développés, on aboutirait à un résultat sous-optimal avec un terme en $\ln(n)$ en trop (cf. Chapitre 3). Pour éviter le recours aux arguments entropiques et obtenir des résultats optimaux, nous avons établi une version simplifiée (suffisante dans notre cadre) d'un théorème de sélection de modèle de Massart (2007). Nos inégalités oracle se déduisent par application directe de ce théorème simplifié (Théorème 2.A.1, annexe du Chapitre 2).

Chapitre 3 Finite mixture Gaussian regression models

Cadre

Le cadre du Chapitre 2 englobe le cas de la régression linéaire gaussienne où l'on considère n observations $(\boldsymbol{X_i}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ telles que $Y_i = \sum_{j=1}^p \beta_j^* X_{ij} + \varepsilon_i$. Pour un tel modèle, $Y_i \mid \boldsymbol{X_i} = \boldsymbol{x_i}$ suit une loi gaussienne $\mathcal{N}(\sum_{j=1}^p \beta_j^* x_{ij}, \sigma^2)$. Au Chapitre 3, nous étendons ce cadre de régression "homogène" au cadre de régression "hétérogène" en envisageant le cas où les valeurs des coefficients de régression peuvent dépendre des observations. Cette modélisation hétérogène semble plus réaliste

que la modélisation homogène surtout dans le cas de la grande dimension où les variables sont très nombreuses et où certaines d'entre elles peuvent ne pas avoir la même influence sur toutes les observations. Prendre en compte une telle situation permet alors de réduire le risque de prédiction. Cette hétérogénéité peut être modélisée par un mélange fini de K régressions gaussiennes :

$$Y_i \mid \boldsymbol{X_i} = \boldsymbol{x_i} \sim s = s_{\boldsymbol{\theta^*}} = \sum_{k=1}^{K} \pi_k^* \mathcal{N}\left(\sum_{j=1}^{p} \beta_{kj}^* x_{ij}, {\sigma_k^*}^2\right).$$

Le paramètre $\boldsymbol{\theta^*} = (\pi_k^*, \boldsymbol{\beta_k^*}, \sigma_k^*)_{1\leq k\leq K}$ englobe les proportions, les vecteurs des moyennes et les variances des K composantes du mélange. Pour $K = 1$, on retrouve le cadre de la régression linéaire gaussienne homogène.

Afin d'éviter le risque de surajustement, surtout en grande dimension, on peut considérer une régularisation ℓ_1 de la log-vraisemblance. Les paramètres de proportions et de variances sont chacun au nombre de $K \ll n$ et n'ont pas besoin d'être régularisés. Au contraire, les coefficients de régression β_{kj}^* sont au nombre de $Kp \gg n$ pour $p \gg n$ et c'est sur ces coefficients que va porter la pénalité ℓ_1. L'estimateur Lasso de la densité s associé à cette régularisation ℓ_1 est défini par

$$\hat{s}(\lambda) = \underset{s_{\boldsymbol{\theta}}}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i})\right) + \lambda|s_{\boldsymbol{\theta}}|_1 \right\}, \quad \lambda > 0, \tag{7}$$

où $|s_{\boldsymbol{\theta}}|_1 = \sum_{j=1}^{p}\sum_{k=1}^{K}|\beta_{kj}|$ pour $s_{\boldsymbol{\theta}} = \sum_{k=1}^{K}\pi_k\mathcal{N}(\sum_{j=1}^{p}\beta_{kj}x_{ij}, \sigma_k^2)$.

Résultat

Nous établissons une inégalité oracle ℓ_1 pour comparer le risque de prédiction de l'estimateur Lasso $\hat{s}(\lambda)$ défini par (7) à l'oracle ℓ_1. Dans une approche de maximisation de la vraisemblance, nous introduisons la divergence de Kullback-Leibler, notée KL, et nous considérons la fonction de perte moyenne définie pour une densité t par

$$\mathrm{KL_n}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(s(\cdot|\boldsymbol{x_i}), t(\cdot|\boldsymbol{x_i})) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \ln\left(\frac{s(y|\boldsymbol{x_i})}{t(y|\boldsymbol{x_i})}\right) s(y|\boldsymbol{x_i})\, dy.$$

Pour des raisons techniques, nous nous restreignons à un ensemble de densités $s_{\boldsymbol{\theta}}$ dans un ensemble S à paramètres $\boldsymbol{\theta}$ bornés par des constantes. Nous démontrons le résultat suivant.

Théorème 3.3.1 Si

$$\lambda \geq \kappa\, C\, K\, (\ln n)^2 \sqrt{\frac{\ln(2p+1)}{n}} \tag{8}$$

où $\kappa > 1$ est une constante absolue et $C > 0$ est une quantité dépendant des bornes imposées sur les

paramètres et des régresseurs x_{ij}, alors $\hat{s}(\lambda)$ satisfait l'inégalité oracle en espérance suivante :

$$\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}\left(s, \hat{s}(\lambda)\right)\right] \leq (1+\kappa^{-1}) \inf_{s_{\boldsymbol{\theta}} \in S} \left\{\mathrm{KL}_{\mathrm{n}}(s, s_{\boldsymbol{\theta}}) + \lambda |s_{\boldsymbol{\theta}}|_1\right\} + \lambda + C' \frac{K^{3/2} \ln n}{\sqrt{n}}, \tag{9}$$

où C' est une quantité dépendant des bornes imposées sur les paramètres.

Dans l'énoncé du Théorème 3.3.1 au Chapitre 3, les quantités C et C' sont explicitées de manière précise bien que l'optimalité de ces quantités ne soit pas garantie.

Discussion

Avant nous, Städler et al. (2010) se sont intéressés à ce cadre de régression hétérogène et à l'estimation de la densité de mélange par le Lasso. Ils ont introduit le Lasso dans le but de sélectionner les variables intervenant réellement dans un tel mélange de régressions, c'est-à-dire les variables indexées par $j \in \{1, \ldots, p\}$ tel que β^*_{kj} est non nul pour au moins une composante $k \in \{1, \ldots, K\}$ du mélange. Dans cette optique de sélection de variables, Städler et al. (2010) ont établi une inégalité oracle ℓ_0 afin de comparer les risques de prédiction du Lasso à un pseudo-oracle ℓ_0. Comme dans le cas de la régression linéaire homogène, leur résultat nécessite de fortes contraintes de non colinéarité entre les variables. De plus, afin de relier la divergence de Kullback-Leibler à la norme ℓ_2 des paramètres, ils ont introduit des hypothèses de marge faisant intervenir des quantités inconnues dont dépendent leur inégalité. Ils ont eux aussi considéré des paramètres bornés.

Pour $K = 1$, $\mathrm{KL}_{\mathrm{n}}(s,t) = \mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2\right]/2$, donc l'inégalité (6) établie au Chapitre 2 est un cas particulier de l'inégalité (9) pour $K = 1$. Cependant, nous avons établi l'inégalité (6) sans hypothèse de bornitude sur les paramètres et la borne inférieure du paramètre de régularisation (3) ne comporte pas le terme en $(\ln n)^2$ de la borne inférieure (8). Ce terme supplémentaire provient de calculs d'entropie métrique dans la démonstration.

Partie II. Variable selection for clustering based on Gaussian mixture models for high-dimensional data

Dans la partie II, nous nous plaçons dans le cadre de la classification non supervisée en grande dimension sous l'hypothèse de parcimonie, tel qu'introduit en Section 1.1. Nous envisageons une approche par modèles de mélange gaussien. Nous exploitons la parcimonie induite par la pénalisation ℓ_1 pour construire une procédure efficace de classification incluant la sélection des variables pertinentes pour déterminer cette classification. Cependant, notre procédure n'a pas recours à la pénalisation ℓ_1 de manière traditionnelle. En fait, nous n'utilisons la pénalisation ℓ_1 que pour construire de manière efficace une collection de modèles de mélange aléatoire restreinte obtenue en faisant varier le paramètre

de régularisation. Une fois cette collection de modèles obtenue, nous estimons les paramètres de chaque modèle par maximum de vraisemblance, puis nous sélectionnons un modèle grâce à un critère pénalisé ℓ_0 non asymptotique construit à partir des données suivant l'heuristique de pente de Birgé et Massart (2006). Nous nous démarquons ainsi de l'usage traditionnel de la pénalisation ℓ_1 qui consiste non seulement à construire des paquets de variables mais aussi à estimer les paramètres du mélange par sélection de l'une des solutions Lassos. Notre volonté d'éviter l'estimation par le Lasso est motivée par les performances médiocres d'estimation et de sélection de variables du Lasso usuellement constatées dans le cadre de la régression.

Dans la suite, nous conservons les notations introduites en Section 1.1.

Chapitre 4 Our Lasso-MLE procedure for variable selection in clustering

Résultats

Le point central du Chapitre 4 est la description de notre procédure de classification non supervisée en grande dimension avec sélection simultanée des variables pertinentes pour établir cette classification. Avant de décrire cette procédure, nous en motivons les étapes en analysant les points forts et les points faibles de la procédure de Pan et Shen (2007) dont nous reprenons l'idée de la pénalisation ℓ_1 pour une détection automatique des variables pertinentes pour la classification.

- **Points faibles de la procédure Lasso de Pan et Shen (2007)**

Etant données des observations $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$, Pan et Shen (2007) centrent empiriquement $\boldsymbol{Y}$ et estiment la densité $\overline{s}$[1] de l'échantillon empiriquement recentré $\overline{\boldsymbol{Y}}$ par une densité de mélange $s_{\widehat{\overline{\boldsymbol{\theta}}}}$. Afin d'obtenir un estimateur $\widehat{\overline{\boldsymbol{\theta}}} = (\hat{\pi}_k, \widehat{\overline{\mu}}_{kj}, \hat{\sigma})_{1 \leq k \leq K}$ parcimonieux en les coefficients des moyennes, ils appliquent une pénalisation ℓ_1 des coefficients des moyennes au risque empirique sur $\overline{\boldsymbol{Y}}$. Pour un nombre K de classes et un paramètre de régularisation λ fixés, l'estimateur Lasso associé est défini par

$$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = \underset{\overline{\boldsymbol{\theta}} \in \Theta_K}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}}) \right) + \lambda |\overline{\boldsymbol{\theta}}|_1 \right\} \tag{10}$$

où $|\overline{\boldsymbol{\theta}}|_1 := \sum_{j=1}^{p} \sum_{k=1}^{K} |\overline{\mu}_{kj}|$. En pratique, Pan et Shen (2007) calculent cet estimateur par un algorithme EM. En faisant varier K et λ, ils obtiennent une collection d'estimateurs $s_{\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)}}$ plus ou moins parcimonieux. Ils retiennent l'un d'entre eux par un critère de type BIC et obtiennent une partition de $\overline{\boldsymbol{Y}}$ par la règle du MAP (2).

[1] Les quantités modifiées par le recentrage empirique des données seront marquées d'une barre horizontale. C'est le cas pour l'échantillon $\boldsymbol{Y}$, la densité s des observations et les vecteurs des moyennes $\boldsymbol{\mu_k}$ pour chaque composante k du mélange. En revanche, les proportions π_k et la variance commune σ^2 ne sont pas modifiées par le recentrage empirique. Le vecteur global des paramètres $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\mu_k}, \sigma)_{1 \leq k \leq K}$, partiellement modifié, sera lui aussi marqué d'une barre horizontale.

Cette procédure nous inspire plusieurs remarques :

1. *Sélection des variables pertinentes ?*

L'idée de la sélection automatique de variables par pénalisation ℓ_1 nous semble judicieuse et prometteuse. Cependant, Pan et Shen (2007) ne justifient pas vraiment pourquoi les variables sélectionnées par minimisation de (10) sont effectivement les variables pertinentes pour la classification.

2. *Estimation de la densité de $\boldsymbol{Y}$?*

A notre connaissance, aucun résultat théorique de prédiction ou d'estimation sur le Lasso n'a été établi dans le cadre de l'estimation de densité par mélange gaussien. Cependant, les résultats théoriques traditionnellement établis en régression ne permettent de garantir de bonnes performances de prédiction et d'estimation du Lasso que sous des hypothèses difficilement vérifiées en grande dimension. Les problèmes d'estimation du Lasso ont été largement confirmés en pratique (par exemple, Connault, 2011, pour la régression ou Bertin et al., 2011, pour l'estimation de densité par décomposition dans un dictionnaire). Ils sont liés à la sous-estimation des coefficients provoquée par la régularisation ℓ_1 et il est légitime de penser que la solution Lasso de Pan et Shen (2007) souffre également de ce problème.

De plus, comme Pan et Shen (2007) recentrent préalablement les données, ils obtiennent une estimation $\widehat{\overline{s}}$ de la densité $\overline{s}$ de l'échantillon recentré et non de la densité s des données de départ. Pour obtenir une estimation de s à partir de $\widehat{\overline{s}}$, il convient d'ajouter la moyenne empirique de chacune des p variables à l'estimation des coefficients de moyenne de $\widehat{\overline{s}}$. Cela nécessite p estimations correspondant aux p moyennes empiriques. En grande dimension où $p \gg n$, cela risque de conduire à une estimation de la densité s présentant un fort risque de prédiction. En outre, en ajoutant les moyennes empiriques à chaque coefficient de moyenne, on perd la parcimonie de la solution. A cause du centrage empirique, la méthode de Pan et Shen (2007) risque de ne pas être adaptée aux problèmes de classification en grande dimension pour lesquels un objectif d'estimation se greffe à l'objectif de classification. C'est par exemple le cas pour la classification de données fonctionnelles lorsqu'une reconstruction parcimonieuse de courbes est souhaitée.

3. *Qualité de la classification ?*

D'après le point 2 ci-dessus, on peut douter de la qualité de l'estimation de la densité $\overline{s}$ par la solution Lasso $\widehat{\overline{s}}$ de Pan et Shen (2007). Or, la classification est obtenue par MAP à partir de l'estimation $\widehat{\overline{s}}$. On peut donc s'interroger sur la répercussion des problèmes d'estimation de la densité $\overline{s}$ sur la qualité de la classification.

4. *Sélection de modèle ?*

On peut douter de la pertinence d'un critère asymptotique comme BIC dans le cadre de la grande dimension où le nombre d'observations est réduit par rapport au nombre de variables.

- **Notre procédure Lasso-MLE**

Nous proposons une procédure reprenant le point fort de la procédure de Pan et Shen (2007) – à savoir la sélection de variables par la régularisation ℓ_1 – mais corrigeant un à un les quatre points faibles mentionnés ci-dessus :

1. *Sélection des variables pertinentes.*

Grâce au cadre statistique rigoureux fourni par les modèles de mélange, on peut donner une définition mathématique d'une variable pertinente pour la classification. Une variable indexée par $j \in \{1, \ldots, p\}$ telle que les coefficients de moyenne μ_{kj} sont identiques pour toutes les composantes $k \in \{1, \ldots, K\}$ du mélange ne sert à rien pour discriminer les classes. Une telle variable est non pertinente pour la classification. Au contraire, une variable qui possède au moins deux composantes de moyenne différentes est susceptible d'avoir une influence sur la classification et sera dite pertinente pour la classification. Au Chapitre 4, nous justifions que le problème de minimisation (10) proposé par Pan et Shen (2007) permet effectivement de détecter de telles variables. Nous utilisons alors la méthode de Pan et Shen (2007) pour construire des paquets de variables potentiellement pertinentes en faisant varier le nombre de classes K et le paramètre de régularisation λ du Lasso.

2. *Estimation de la densité de* $\boldsymbol{Y}$.

Pour pouvoir traiter la classification de données fonctionnelles où il est essentiel de bien estimer chaque densité gaussienne composante du mélange afin de reconstruire un profil type par classe, nous apportons deux modifications à la procédure de Pan et Shen (2007) :

(a) Nous faisons la remarque essentielle suivante. La notion de variable pertinente pour la classification n'est pas une notion induisant de la parcimonie. En effet, pour chaque variable non pertinente, un coefficient de moyenne est à estimer (le coefficient commun à toutes les classes). Même dans le cas extrême où les p variables seraient non pertinentes, cela laisse $p \gg n$ coefficients de moyenne à estimer. Pour résoudre ce problème de dégénérescence, nous introduisons une hypothèse de parcimonie. Nous supposons que parmi les variables non pertinentes, il existe un très grand nombre de variables – que nous appelons variables inactives – pour lesquelles la valeur commune de la moyenne à travers les classes est nulle[2]. Par exemple, dans le cas de la décomposition de signaux, cette hypothèse revient à supposer l'existence d'une décomposition parcimonieuse dans un dictionnaire donné (par exemple d'ondelettes) pour chaque type de signal. Pour détecter les variables inactives parmi les variables non pertinentes, nous effectuons

[2]Si l'on centre empiriquement les données comme Pan et Shen (2007), variables non pertinentes et inactives se confondent car, une fois les données recentrées, la moyenne commune des variables non pertinentes est estimée à zéro. Le centrage empirique induit de la parcimonie, mais elle est *artificielle* car on la perd en revenant à l'estimation des données de départ. Au contraire, notre hypothèse introduit une *réelle* parcimonie

une seconde pénalisation ℓ_1, mais cette fois-ci sur l'échantillon réduit aux variables non pertinentes et sans recentrage empirique préalable. A l'issue de cette seconde pénalisation ℓ_1, nous obtenons des paquets de variables potentiellement inactives parmi chaque paquet de variables potentiellement non pertinentes. Cela nous fournit une collection globale de modèles.

(b) Une fois notre collection de modèles obtenue, nous proposons une estimation des paramètres dans chaque modèle par l'estimateur du maximum de vraisemblance et non par l'estimateur Lasso. Cela revient à effectuer un seuillage dur plutôt qu'un seuillage doux des coefficients, ce qui améliore l'estimation des coefficients de moyenne non nuls. Nous améliorons ainsi l'estimation de la densité.

3. *Qualité de la classification.*

Notre choix de méthode pour la classification – à savoir une modélisation par modèles de mélange gaussien sphérique homoscédastique et une classification déduite par MAP à partir de l'estimation de la densité s du mélange – reformule le problème de classification en un problème d'estimation de la densité s. Grâce aux précautions décrites ci-dessus, notre procédure garantit une bonne estimation de la densité s de l'échantillon $\boldsymbol{Y}$. Nous pouvons donc espérer qu'il en découle une bonne classification des observations $\boldsymbol{Y_i}$, du moins si notre modélisation reflète effectivement la réelle structure des données (ce qui est par exemple le cas pour des données simulées sous les bonnes hypothèses, comme pour nos simulations au Chapitre 5).

4. *Sélection de modèle.*

Au lieu d'utiliser un critère asymptotique de type BIC, nous optons pour un critère de sélection de modèle non asymptotique par pénalisation ℓ_0, basé sur la théorie développée par Birgé et Massart (1997) et Barron et al. (1999). La recherche de la pénalité à considérer pour définir ce critère est menée au Chapitre 6.

Discussion

La procédure ci-dessus n'est pas la première à laquelle nous avons songé. Le premier défaut de la procédure de Pan et Shen (2007) que nous avons jugé indispensable de corriger est l'estimation des paramètres par le Lasso. Nous pensons que l'idée de n'utiliser le Lasso que pour construire un nombre restreint de paquets de variables en un temps qui reste raisonnable même en grande dimension, puis de réaliser l'estimation par l'estimateur du maximum de vraisemblance[3] sur les modèles engendrés par ces paquets et de sélectionner l'un d'entre eux par un critère de pénalisation ℓ_0, est à retenir,

[3]Maximum Likelihood Estimator (MLE) en anglais, d'où le nom donnée à notre procédure Lasso-MLE : "Lasso" pour indiquer que nous construisons une collection de modèles grâce à la pénalisation ℓ_1, et "MLE" pour indiquer que l'estimation et le critère de sélection de modèle sont envisagés d'un point de vue ℓ_0.

que ce soit dans notre contexte ou dans n'importe quel autre contexte. Ainsi, notre idée initiale était la suivante : centrer empiriquement les observations, utiliser la pénalisation ℓ_1 sur les observations recentrées pour créer des paquets de variables potentiellement pertinentes, en déduire une collection de modèles, estimer les paramètres sur les observations recentrées par l'estimateur du maximum de vraisemblance dans chaque modèle, choisir un modèle par un critère non asymptotique et partitionner les observations recentrées par MAP.

Par rapport à la procédure de Pan et Shen (2007), cette procédure est en particulier censée améliorer l'estimation de la densité des observations recentrées et donc la classification des observations recentrées. Cette procédure est acceptable si le seul objectif envisagé est la classification. De plus, elle est réalisable sans hypothèse de parcimonie, contrairement à notre procédure qui suppose que de nombreuses variables sont non seulement non pertinentes mais aussi inactives. Cependant, elle ne permet pas de traiter le cas où l'estimation de la densité des observations d'origine (non recentrées) fait partie du problème, comme c'est le cas pour la reconstruction de courbes multiclasses. En effet, à cause du centrage empirique, pour passer de l'estimation de la densité des observations recentrées à l'estimation des observations d'origine, il faut ajouter les p moyennes empiriques, ce qui conduit à une estimation dangereuse et non parcimonieuse.

Une autre version de notre procédure Lasso-MLE est envisageable. Supposons qu'il existe de nombreuses variables inactives. Dans la procédure que nous avons décrite ci-dessus, la construction de modèles s'opère en deux temps : nous créons des paquets de variables non pertinentes puis nous constituons des paquets de variables inactives parmi chaque paquet de variables non pertinentes. Une autre possibilité consiste à inverser l'ordre de recherche des variables en créant d'abord des paquets de variables actives puis en constituant des paquets de variables pertinentes parmi chaque paquet de variables actives. Cette alternative éjecte d'abord les variables absentes du modèle (les variables inactives) puis recherche les variables pertinentes pour la classification parmi les variables présentes dans le modèle. Elle peut paraître plus intuitive que notre procédure qui consiste à détecter les variables non pertinentes pour la classification puis à extraire les variables inactives parmi ces variables non pertinentes. Cette alternative a notamment l'avantage de rester bien définie dans la cas limite monoclasse $K = 1$: il suffit de supprimer sa deuxième étape et de ne conserver que sa première étape pour obtenir une procédure de sélection de variables dans un cadre de modèles linéaires gaussiens. Au contraire, notre procédure n'a de sens que pour $K \geq 2$ car notre première étape est focalisée sur la classification. Cependant, la mise en pratique de cette méthode alternative pose des problèmes numériques en grande dimension.

Au Chapitre A, les deux procédures alternatives mentionnées ci-dessus sont présentées et comparées à notre procédure Lasso-MLE sur des données simulées. Une analyse des performances de chacune des trois méthodes permet de comprendre notre choix pour la procédure finalement retenue.

Chapitre 5 Simulations

Au Chapitre 5, nous testons notre procédure Lasso-MLE sur des jeux de données simulées.

Résultats

D'abord, nous comparons notre procédure à deux procédures de sélection de variables en classification non supervisée par modèles de mélange gaussien qui partagent des points communs avec notre procédure. La première est la procédure Lasso de Pan et Shen (2007) dont nous avons repris le recours à la pénalisation ℓ_1 pour construire efficacement une collection aléatoire de modèles incluse dans la collection de tous les modèles possibles. La seconde est la procédure de sélection de variables complète ou ordonnée de Maugis et Michel (2011a), dont notre procédure reprend l'estimation des paramètres du mélange par maximum de vraisemblance[4] et l'utilisation de la méthode de la pente introduite par Birgé et Massart (2006) pour déterminer un critère pénalisé non asymptotique de sélection de modèle. Nos simulations nous permettent d'aboutir aux conclusions suivantes :

- L'inconvénient majeur de la procédure de Maugis et Michel (2011a) est combinatoire : à cause de la trop grande richesse de la collection de modèles en sélection de variables complète ou même ordonnée, leur procédure n'est réalisable qu'en très faible dimension. Grâce à l'écrémage de la collection de modèles complète réalisé par le Lasso, notre procédure Lasso-MLE peut être envisagée comme une solution attractive à l'extension et à l'adaptation de leur procédure de basse dimension à la grande dimension. A noter que le Lasso génère une collection de modèles suffisamment riche et pertinente pour contenir le(s) modèle(s) d'intérêt et ne pas altérer la qualité du modèle choisi par rapport au modèle choisi dans la collection plus riche de modèles considérée par Maugis et Michel (2011a), à critère de sélection de modèle identique.

- Comme attendu, le principal défaut de la procédure Lasso de Pan et Shen (2007) concerne l'estimation de la densité et le choix du modèle retenu. Ces deux problèmes ont pour cause commune la sous-estimation des moyennes des variables pertinentes par les estimateurs Lasso. En effet, d'une part, cette sous-estimation entraîne une estimation médiocre de la densité. D'autre part, pour compenser cette sous-estimation, la procédure de Pan et Shen (2007) sélectionne des modèles contenant de nombreuses variables non pertinentes pour la classification. Au contraire, comme nous prenons soin d'estimer les paramètres par l'estimateur du maximum de vraisemblance dans chaque modèle, l'estimation de la densité est bonne et notre procédure sélectionne des modèles sans (ou avec très peu de) variables non pertinentes.

Ensuite, nous testons notre procédure sur des problèmes de classification non supervisée de courbes. Nous considérons K types de courbes $f_1, \ldots, f_K$. En pratique, les courbes f_k sont décrites de manière

[4] A noter cependant que l'estimation des paramètres est réalisée sur le jeu de données brut pour notre procédure, alors qu'elle est effectuée sur le jeu de données empiriquement recentré pour la procédure de Maugis et Michel (2011a).

discrète par leurs valeurs prises sur une grille très fine comportant p points : $\boldsymbol{f_k} = (f_k(t_1), \ldots, f_k(t_p))$. Nous bruitons ces courbes par un bruit blanc gaussien et nous génèrons n observations $\boldsymbol{y_i}$ avec $n \ll p$. Nous obtenons ainsi un échantillon $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ de courbes bruitées réparties en K classes. Nous n'exécutons pas notre procédure directement sur $\boldsymbol{y}$. Au préalable, nous décomposons les courbes bruitées $\boldsymbol{y_i}$ dans une base d'ondelettes $\mathcal{B} = \{\phi_1, \ldots, \phi_p\}$, ce qui fournit un nouveau jeu de données $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ où chaque donnée $\boldsymbol{Y_i}$ est la décomposition en coefficients dans la base $\mathcal{B}$ de $\boldsymbol{y_i}$. Si la courbe $\boldsymbol{y_i}$ est obtenue par bruitage de la fonction f_k, alors sa décomposition en coefficients dans la base $\mathcal{B}$ s'écrit $\boldsymbol{Y_i} = \boldsymbol{\mu_k} + \boldsymbol{\varepsilon_i}$ où $\boldsymbol{\mu_k}$ est la décomposition en coefficients dans la base $\mathcal{B}$ de la fonction discrétisée $\boldsymbol{f_k}$ et où $\boldsymbol{\varepsilon_i}$ est de loi $\mathcal{N}(0, \sigma^2 \boldsymbol{I})$. Les variables sont les fonctions ϕ_j de la base $\mathcal{B}$. Une variable ϕ_j est non pertinente pour la classification si $\mu_{kj} = \mu_{k'j}$ pour tout $(k, k') \in \{1, \ldots, K\}^2$. Notre procédure est particulièrement adaptée pour traiter ce problème :

- Grâce au Lasso qui est capable de parcourir un panel de modèles dans la collection complète de modèles, et d'annuler des coefficients μ_{kj} jusqu'à $j = p$ même pour de très grandes valeurs de p, nous ne sommes pas obligés de choisir un niveau de troncature pour la décomposition des fonctions $\boldsymbol{f_k}$, comme c'est généralement le cas pour les méthodes de projection sur une base (Misiti et al., 2007a ; Auder et Fischer, 2011). Notre procédure visite tous les niveaux et se charge d'annuler des coefficients μ_{kj} aux meilleurs endroits $(k, j) \in \{1, \ldots, K\} \times \{1, \ldots, p\}$.

- La décomposition en ondelettes d'une fonction est généralement parcimonieuse. Ainsi, il existe de nombreux $j \in \{1, \ldots, p\}$ tels que $\mu_{kj} = 0$ pour tout $k \in \{1, \ldots, K\}$, c'est-à-dire tels que la variable ϕ_j est inactive. La détection des variables inactives permet de réduire la dimension des modèles et d'estimer les paramètres sur le jeu de données non empiriquement recentré. Cette précaution mêlée à l'estimation par maximum de vraisemblance (plutôt que par le Lasso) nous garantissent une bonne qualité d'estimation des vecteurs des moyennes $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}$. En effectuant une transformation d'ondelettes inverse, nous obtenons alors une bonne estimation des fonctions $f_1, \ldots, f_K$.

Discussion

En régression, l'algorithme LARS, calculant le chemin de régularisation entier du Lasso, se trouve implémenté dans la plupart des logiciels utilisés en statistique. Dans notre contexte de mélange gaussien en classification non supervisée, un tel algorithme n'existe pas. Nous avons repris l'idée de Pan et Shen (2007) de calculer la solution Lasso par un algorithme de type EM. Nous avons implémenté cet algorithme en MATLAB. Pan et Shen (2007) considèrent une grille régulière sur laquelle ils font varier le paramètre de régularisation du Lasso pour calculer un ensemble de solutions Lasso. Nous avons constaté qu'un réglage déterministe du pas n'est pas évident. Nous avons cherché à améliorer ce point en proposant une grille construite à partir des données (cf. Section 4.B.1). D'autre part, en grande dimension, des précautions sont à prendre pour éviter la divergence de certaines quantités,

telles les probabilités conditionnelles d'appartenance ou la log-vraisemblance. De plus, des problèmes d'estimation sont fréquents dans les modèles de très grande dimension. Pour contourner ce problème, une solution (qui est celle envisagée par Pan et Shen, 2007) est d'effectuer un centrage empirique des données. Mais cette solution empêche de traiter le problème de reconstruction parcimonieuse de courbes, qui nous semble pourtant un enjeu important. C'est pour éviter le recours au centrage empirique dans la phase d'estimation que nous avons introduit la notion parcimonieuse de variable active.

En ce qui concerne notre critère de sélection de modèle, nous devons définir une forme de pénalité pour appliquer la méthode de la pente déduite de l'heuristique de pente de Birgé et Massart (2006). Depuis les travaux fondateurs de Birgé et Massart (2006), on distingue principalement deux formes de pénalité : l'une – proportionnelle à la dimension des modèles – est valide lorsqu'on travaille avec des collections de modèles $\{S_m\}_{m\in\mathcal{M}}$ ne contenant pas ou contenant très peu de modèles de même dimension D_m (sélection de variables ordonnée en régression par exemple), l'autre – impliquant un terme logarithmique – est à considérer dans le cas de collections de modèles contenant de nombreux modèles de même dimension (sélection de variables complète en régression par exemple). Dans notre contexte, ces deux pénalités s'écrivent respectivement

$$\mathrm{pen}(m) = \kappa \frac{D_m}{n}$$

et

$$\mathrm{pen}_{\ln}(m) = \kappa_1 \frac{D_m}{n} \left(1 + \kappa_2 \ln\left(\frac{p}{D_m}\right)\right)$$

où κ, κ_1 et κ_2 sont des constantes à calibrer. Pour notre procédure, les collections de modèles sont construites à partir des données et sont donc aléatoires. Il est alors difficile de déterminer théoriquement le nombre de modèles de même dimension dans nos collections, et donc de trancher sur la présence d'un terme logarithmique dans la pénalité. Ainsi, lors de nos simulations, les deux pénalités ci-dessus sont systématiquement testées. Nous constatons que la forme de la pénalité évolue en fonction du nombre p de variables du jeu de données : pour $p \ll n$, une pénalité proportionnelle à la dimension permet la sélection d'un modèle proche de l'oracle, tandis qu'un terme logarithmique est à ajouter pour ne pas sous-pénaliser lorque $p \gg n$. Une étude théorique et pratique de la forme de la pénalité est conduite au Chapitre 6.

Chapitre 6 A non-asymptotic data-based penalized criterion

Comme nous abordons le problème de classification non supervisée par l'intermédiaire de modèles de mélange gaussien, le choix du nombre de classes ainsi que la sélection des variables pertinentes pour la classification sont reformulés en un problème global de sélection de modèle. Nous avons opté pour un critère de sélection de modèle non asymptotique dans la lignée de la théorie de sélection de

modèle développée par Birgé et Massart (1997) et Barron et al. (1999). L'enjeu du Chapitre 6 est de fournir des pistes de réflexion pour déterminer une pénalité minimale à considérer pour définir un critère pénalisé sélectionnant un modèle proche de l'oracle. L'étude de la forme de la pénalité est rendue délicate du fait du caractère aléatoire de notre collection de modèles (construite à partir des données par un algorithme Lasso). Deux points de vue sont considérés. D'un point de vue théorique, nous établissons une forme de pénalité minimale suffisante. D'un point de vue pratique, nous vérifions que cette forme de pénalité s'avère nécessaire en grande dimension. Les résultats de ce chapitre ne permettent pas de trancher définitivement sur la forme de pénalité optimale : ils sont plutôt à voir comme des éléments de réponse à la recherche de la pénalité minimale pour notre problème.

Résultats théoriques

Dans notre procédure Lasso-MLE introduite au Chapitre 4, nous estimons la densité du jeu de données par l'estimateur du maximum de vraisemblance dans chaque modèle préalablement généré par le Lasso. Nous devons donc considérer un critère de sélection de modèle dans le cadre d'estimation de densité par maximum de vraisemblance. Barron et al. (1999) et Massart (2007) ont établi un théorème général de sélection de modèle dans un tel cadre. Cependant, leur théorème est formulé pour une collection déterministe de modèles tandis que notre collection de modèles générée par le Lasso est aléatoire. Nous ne pouvons donc pas appliquer directement leur théorème. Nous adaptons leur preuve au cas d'une collection aléatoire de modèles pour obtenir un théorème général de sélection de modèle dans le cadre d'estimation de densité par une collection aléatoire d'estimateurs du maximum de vraisemblance. Ensuite, nous appliquons ce théorème général à notre collection particulière de modèles de mélange gaussien sphérique homoscédastique. Pour des raisons techniques, nous nous restreignons à des modèles à paramètres bornés. Nous obtenons un théorème dont nous donnons ici un énoncé simplifié.

Théorème 6.6.2 Soit s une densité inconnue à estimer. Soient $\{S_m\}_{m\in\widehat{\mathcal{M}}}$ une collection aléatoire de modèles de mélange gaussien à paramètres bornés. Soit $\tau > 0$ tel que $s_m \geq \mathrm{e}^{-\tau}\, s$ pour tout $m \in \widehat{\mathcal{M}}$ et pour tout $s_m \in S_m$ tels que $\mathrm{KL}(s, s_m) \leq 2\inf_{s_{\boldsymbol{\theta}}\in S_m} \mathrm{KL}(s, s_{\boldsymbol{\theta}})$. Notons $\hat{s}_m = \arg\min_{s_{\boldsymbol{\theta}}\in S_m} \gamma_n(s_{\boldsymbol{\theta}})$ l'estimateur du maximum de vraisemblance dans le modèle S_m et D_m la dimension de S_m.
Alors, il existe deux quantités $\kappa_1 > 0$ et $\kappa_2 > 0$ dépendant des bornes imposées sur les paramètres des modèles et une constante absolue $C > 1$ telles que si pour tout $m \in \widehat{\mathcal{M}}$ et pour tout $D_m \leq p \wedge n$,

$$\mathrm{pen}\,(m) \geq \kappa_1 \frac{D_m}{n}\left[1 + \kappa_2(1 \vee \tau)\ln\left(\frac{p}{D_m}\right)\right], \tag{11}$$

alors l'estimateur $\hat{s}_{\hat{m}}$ défini par $\hat{m} = \arg\min_{m\in\widehat{\mathcal{M}}}\{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\}$ satisfait l'inégalité oracle

suivante :

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{H}}^2\left(s,\hat{s}_{\hat{m}}\right)\right] \leq C\left(\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}}\left\{\inf_{s_{\boldsymbol{\theta}}\in S_m}\mathrm{KL}(s,s_{\boldsymbol{\theta}})+\mathrm{pen}(m)\right\}\right]+\frac{1\vee\tau}{n}\right),$$

où d_{H} désigne la distance de Hellinger et KL la divergence de Kullback-Leibler.

Au Chapitre 6, l'énoncé du Théorème 6.6.2 est précisé : la forme des modèles est détaillée et les quantités κ_1 et κ_2 sont explicitées en fonction des données du problème. Pour établir le Théorème 6.6.2, il est nécessaire de contrôler l'entropie à crochets de nos modèles de mélange gaussien sphérique homoscédastique. Pour cela, nous adaptons à cette forme spécifique les arguments développés par Maugis et Michel (2011b).

Discussion

Le Théorème 6.6.2 fournit une forme de pénalité minimale (11) garantissant que l'estimateur du maximum de vraisemblance pénalisé est proche de l'oracle ℓ_0. En appliquant une telle pénalité lors de notre procédure Lasso-MLE, nous sommes garantis d'obtenir un estimateur présentant un faible risque de prédiction, sans autre hypothèse que des hypothèses de bornitude des paramètres du mélange. Dans un contexte de maximisation de la vraisemblance, de telles hypothèses sont courantes (Maugis et Michel, 2011b ; Baudry, 2009 ; Städler et al., 2010). Au contraire, Pan et Shen (2007) n'ont établi aucune inégalité oracle ℓ_0 pour leur estimateur Lasso. Or, au vu des résultats connus en régression, on peut légitimement penser que le Lasso ne peut satisfaire une telle inégalité oracle que sous des hypothèses restrictives difficilement vérifiées en grande dimension.

Deux bémols concernant le Théorème 6.6.2 sont à souligner :

- La pénalité minimale (11) ne dépend pas du caractère aléatoire $\widehat{\mathcal{M}}$ de la collection de modèles générée par le Lasso. En fait, notre méthode de démonstration repose fortement sur l'inclusion de notre collection de modèles aléatoire dans une collection déterministe plus grande. Or, n'ayant aucune connaissance a priori sur les collections de modèles générées par le Lasso, nous sommes contraints de prendre la collection de modèles complète comme collection déterministe. Pour cette raison, la pénalité (11) n'est autre que la pénalité obtenue pour le problème de sélection de variables complète (Maugis et Michel, 2011b). D'après la théorie développée par Birgé et Massart (2006), le terme $\ln(p/D_m)$ présent dans la pénalité (11) est nécessaire pour compenser la grande richesse de la collection de modèles complète. Mais ce terme n'est plus nécessaire pour définir une pénalité optimale sur une collection de modèles moins riche, par exemple pour le cas de la sélection de variables ordonnée. Dans notre cas, les collections de modèles aléatoires générées par le Lasso s'avèrent en pratique bien moins riches que la collection de modèles complète, mais plus riche que la collection de modèles ordonnée. Nous pouvons

donc nous interroger sur la nécessité d'un terme en $\ln(p/D_m)$ pour définir une pénalité optimale sur la collection de modèles générée par le Lasso. Autrement dit, nous pouvons hésiter entre deux formes de pénalité, avec ou sans terme en $\ln(p/D_m)$:

$$\mathrm{pen}_{\ln}(m) = \kappa_1 \frac{D_m}{n}\left(1 + \kappa_2 \ln\left(\frac{p}{D_m}\right)\right) \quad \text{ou} \quad \mathrm{pen}(m) = \kappa \frac{D_m}{n}. \tag{12}$$

- De plus, même si la forme de la pénalité minimale (11) s'avère optimale, le Théorème 6.6.2 ne fournit pas un critère pratique de sélection de modèle car il dépend de quantités inconnues κ_1, κ_2 et τ.

Résultats pratiques

Afin de pallier les deux écueils ci-dessus, nous appliquons une méthode dérivée de la méthode heuristique dite de la "pente" introduite par Birgé et Massart (2006). La méthode de la pente est un moyen pratique pour calibrer la pénalité idéale quand la forme de celle-ci est connue à une constante multiplicative près. Elle est basée sur un mélange de théorie et d'heuristiques. Bien qu'elle n'ait été prouvée rigoureusement que dans des cadres restreints (Birgé et Massart, 2006 ; Arlot et Massart, 2008), elle a fait ses preuves d'un point de vue pratique dans de nombreux contextes (Lebarbier, 2005 ; Verzelen, 2008 ; Denis et Molinari, 2009 ; Caillerie et Michel, 2009 ; Baudry, 2009 ; Maugis et Michel, 2011a). L'idée clé de cette heuristique est de supposer que la pénalité optimale est environ le double d'une pénalité minimale qui peut être déduite graphiquement des données.

Dans cette thèse, nous apportons deux contributions à la méthode de la pente :

- Les utilisateurs de la méthode de la pente ont tendance à calibrer la pénalité minimale à partir des données puis à choisir comme pénalité optimale deux fois la pénalité minimale sans vérifier la validité de l'heuristique qui justifie ce procédé, même s'ils se trouvent dans un cadre de travail pour lequel cette heuristique n'a pas été prouvée théoriquement. Ici, nous proposons une méthode graphique facilement applicable dans n'importe quel contexte et permettant de vérifier d'un point de vue pratique la validité de l'heuristique de pente en simulant un "modèle nul". Cette méthode consiste à simuler la cible d'intérêt (dans notre cas, la densité inconnue du jeu de données) dans le plus petit modèle (au sens de l'inclusion) de la collection de modèles de façon à annuler le biais des estimateurs et à ne visualiser que la contribution de la complexité des modèles dans la forme de la pénalité. Cela permet de simplifier l'écriture de l'heuristique de pente, qui se réduit alors à des quantités calculables d'après les données. On peut alors graphiquement vérifier cette heuristique. Après la présentation de cette méthode dans un contexte général, nous l'appliquons pour vérifier la validité de l'heuristique de pente dans notre contexte.
- Pour appliquer la méthode de la pente traditionnelle introduite par Birgé et Massart (2006), on doit connaître la forme de la pénalité idéale à une constante multiplicative près. C'est le cas

pour la pénalité pen définie par (12) car seule la constante multiplicative κ est à calibrer. Par contre, ce n'est pas le cas pour la pénalité $\mathrm{pen}_{\mathrm{ln}}$ définie par (12) car deux constantes κ_1 et κ_2 sont à calibrer. Nous étendons la méthode de la pente introduite par Birgé et Massart (2006) au cas de la calibration de deux constantes. Nous fournissons une méthode de double régression robuste et une visualisation graphique semblable à celle implémentée par Baudry et al. (2011) pour le cas de la calibration d'une constante.

Outre ces deux contributions, nous appliquons la méthode de la pente traditionnelle ainsi que la méthode de la pente que nous avons développée pour tester respectivement les formes de pénalité pen et $\mathrm{pen}_{\mathrm{ln}}$ définies par (12). Ces études pratiques permettent d'aboutir aux deux conclusions suivantes :

- La forme de la pénalité idéale évolue en fonction de la dimension du problème : pour des problèmes de petite dimension ($p \ll n$), une pénalité de la forme pen est observée, alors que pour des problèmes de grande dimension ($p \gg n$), une pénalité de la forme $\mathrm{pen}_{\mathrm{ln}}$ est observée. Ainsi, comme attendu par les résultats théoriques de Birgé et Massart (2006), la richesse de la collection de modèles semble influer sur la forme de la pénalité.
- Il existe un lien étroit entre la richesse de la collection de modèles générée par le Lasso et la calibration par la méthode de la pente de la pénalité pen définie par (12). Ainsi, la méthode de la pente permet d'obtenir une pénalité qui s'adapte à la richesse de la collection de modèles aléatoire générée par le Lasso. Cet avantage incite à l'utilisation d'une pénalité calibrée à partir des données plutôt que d'une pénalité déterministe comme c'est le cas pour BIC.

Discussion

Notre étude pratique de la forme de la pénalité ne fournit qu'une conclusion partielle. Nous constatons une évolution de la forme de la pénalité en fonction de la dimension du problème, mais nous ne fournissons pas de règle générale pour déterminer laquelle des deux pénalités pen ou $\mathrm{pen}_{\mathrm{ln}}$ est à considérer. Nous pensons qu'il n'est pas souhaitable (possible ?) de chercher à établir une telle règle. Tout l'intérêt de la méthode de la pente est de fournir une pénalité adaptative au jeu de données étudié. Dans notre cadre de travail où la collection des modèles varie suivant le jeu de données, fixer de manière déterministe une forme de pénalité reviendrait à trahir l'esprit de la méthode de la pente. De la même manière qu'il est judicieux de calibrer la (les) constante(s) au lieu de fixer des constantes déterministes tels que pour les critères AIC ou BIC, nous pensons préférable de tester les deux formes de pénalité et de laisser parler les graphiques pour le choix de la forme optimale.

Conclusion et perspectives

Conclusion

A notre connaissance, l'idée d'exploiter la régularisation ℓ_1 pour la sélection de variables dans le cadre de la classification non supervisée en grande dimension n'a été envisagée avant nous que par Pan et Shen (2007) puis par Xie et al. (2008) et Zhou et al. (2009). Ainsi, le Lasso n'a été que peu exploité dans ce contexte. Cette méthode prometteuse mérite d'être approfondie et travaillée afin de l'optimiser. C'est ce que nous avons cherché à faire en modifiant la procédure Lasso de Pan et Shen (2007) pour obtenir notre procédure Lasso-MLE.

Pour évaluer les qualités et les défauts de la procédure Lasso de Pan et Shen (2007), nous avons étudié les résultats théoriques et pratiques obtenus pour le Lasso dans le cadre plus largement étudié de la régression linéaire. L'avantage algorithmique du Lasso par rapport à d'autres méthodes de sélection de variables est indéniable. Par contre, le fossé entre les fortes hypothèses nécessaires pour obtenir les résultats théoriques ℓ_0 et l'absence totale d'hypothèse pour obtenir nos résultats théoriques ℓ_1 présentés en Partie I souligne que l'on ne peut pas espérer de l'estimateur régularisé en norme ℓ_1 qu'il rivalise avec l'oracle ℓ_0. La solution intermédiaire que nous envisageons dans cette thèse – à savoir la sélection de variables par l'estimateur régularisé en norme ℓ_1 puis l'estimation par l'estimateur régularisé en "norme" ℓ_0 – nous semble une voie à retenir, que ce soit dans notre contexte ou dans tout autre contexte. Parallèlement à nos travaux, cette idée a d'ailleurs émergé chez d'autres auteurs : par exemple, Connault (2011) dans le cadre de la régression ou Bertin et al. (2011) dans le cadre de l'estimation de densité décomposée dans un dictionnaire.

Nous préconisons un critère de sélection de modèle non asymptotique et construit à partir des données, par la méthode traditionnelle de la pente pour la calibration d'une constante, ou par la méthode dérivée que nous introduisons pour la calibration de deux constantes. Outre la calibration de la pénalité optimale, la forme même de cette pénalité peut être décidée à partir des données. Cela permet d'obtenir un critère de sélection optimal dépendant du jeu de données, ce que l'on ne peut pas attendre d'un critère déterministe et asymptotique tel AIC ou BIC. Cet avantage est d'autant plus appréciable que nous travaillons avec des collections de modèles aléatoires d'une part et en grande dimension d'autre part.

Perspectives

D'un point de vue algorithmique, la grande dimension pose des problèmes numériques qui engendrent des problèmes d'estimation. Afin d'appréhender et de résoudre ces problèmes, nous avons préféré nous concentrer sur l'analyse de quelques jeux de données simulés plutôt que de traiter des jeux réels. Cependant, ce point est maintenant à envisager.

Des travaux supplémentaires seraient souhaitables afin de trancher entre les deux formes de pénalité de manière plus précise que dans cette thèse. On pourrait par exemple chercher à établir des minorations et/ou des majorations assez fines du nombre de modèles de même dimension dans nos collections de modèles aléatoires pour les insérer dans des collections déterministes approchantes. Cela faciliterait l'étude de la forme de la pénalité optimale.

La définition que nous avons donnée d'une variable pertinente pour la classification est liée à l'homogénéité des variances sur chaque classe. Dans le cas plus complexe où les matrices de covariance sont de la forme $\Sigma_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kp}^2)$, une variable est non pertinente si non seulement $\mu_{kj} = \mu_{k'j}$ mais aussi $\sigma_{kj} = \sigma_{k'j}$ pour tout $(k, k') \in \{1, \ldots, K\}^2$. On peut alors montrer qu'une pénalité de la forme $\sum_{j=1}^{p} \sum_{k=1}^{K} (|\mu_{kj}| + |\ln(\sigma_{kj}^2)|)$ est supposée détecter de telles variables. Zhou et al. (2009) ont étendu la procédure Lasso de Pan et Shen (2007) pour de tels modèles. De même, notre procédure Lasso-MLE doit pouvoir s'étendre à de telles situations. Au problème d'estimation des moyennes viendra se greffer le problème d'estimation des variances. On peut penser à adapter la notion de variable active en tenant compte non seulement des moyennes mais aussi des variances. Cela permettrait de traiter efficacement le problème de reconstruction de courbes bruitées de manière hétéroscédastique.

Références

Abraham, C., Cornillon, P. A., Matzner-Lober, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics. Theory and Applications*, 30(3) :581–595.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J. (2011). Clustering functional data using wavelets. *Arxiv preprint :1101.4744*.

Arlot, S. and Massart, P. (2008). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*.

Auder, B. and Fischer, A. (2011). Projection-based curve clustering.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113 :301–413.

Bartlett, P., Mendelson, S., and Neeman, J. (2012). ℓ_1-regularized linear regression : persistence and oracle inequalities. *Probability Theory and Related Fields*.

Baudry, J. (2009). *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Université Paris-Sud 11.

Baudry, J., Maugis, C., and Michel, B. (2011). Slope heuristics : overview and implementation. *Computing and Statistics*. INRIA RR-7728.

Bertin, K., Le Pennec, E., and Rivoirard, V. (2011). Adaptive Dantzig density estimation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 47, pages 43–74. Institut Henri Poincaré.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2) :33–73.

Caillerie, C. and Michel, B. (2009). Model selection for simplicial approximation. INRIA RR-6981.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., and Ralambondrainy, H. (1989). *Classification Automatique des Données, Environnement statistique et informatique*. Dunod.

Cohen, A., DeVore, R., Kerkyacharian, G., and Picard, D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Applied and Computational Harmonic Analysis*, 11(2) :167–191.

Connault, P. (2011). *Calibration d'algorithmes de type Lasso et analyse statistique de données métallurgiques en aéronautique*. PhD thesis, Université Paris-Sud 11.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering - a filter solution. *Proceedings of the Second IEEE International Conference on Data Mining*, pages 115–122.

Denis, M. and Molinari, N. (2009). Choix du nombre de noeuds en régression spline par l'heuristique des pentes. 41èmes Journées de Statistique SFDS, Bordeaux, France.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25) :14863–14868.

Huang, C., Cheang, G., and Barron, A. (2008). Risk of penalized least squares, greedy selection and ℓ_1-penalization for flexible function libraries. Submitted to The Annals of Statistics.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, pages 583–593.

Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9) :1154–1166.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4) :717–736.

Mallat, S. (1989). A theory for multiresolution signal decomposition : The wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7) :674–693.

Massart, P. (2007). *Concentration inequalities and model selection*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Maugis, C., Celeux, G., and Martin-Magniette, M. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics*, 65(3) :701–709.

Maugis, C. and Michel, B. (2011a). Data-driven penalty calibration : a case study for Gaussian mixture model selection. *ESAIM Probability and Statistics*, 15 :320–339.

Maugis, C. and Michel, B. (2011b). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probability and Statistics*, 15 :41–68.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2007). Clustering signals using wavelets. *Computational and Ambient Intelligence*, pages 514–521.

Raftery, A. E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473) :168–178.

Rigollet, P. and Tsybakov, A. (2011). Exponential Screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2) :731–771.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test*, 19(2) :209–256.

Verzelen, N. (2008). Data-driven neighborhood selection of a Gaussian field. INRIA RR-6798.

Xie, B., Pan, W., and Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic journal of statistics*, 2 :168–212.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3 :1473–1496.