HAL
open science

Variable selection in model-based clustering for high-dimensional data

Caroline Meynet

To cite this version:

Caroline Meynet. Variable selection in model-based clustering for high-dimensional data. General Mathematics [math.GM]. Université Paris Sud - Paris XI, 2012. English. NNT : 2012PA112234 . tel-00752613

HAL Id: tel-00752613

https://theses.hal.science/tel-00752613

Submitted on 16 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-SUD XI
École Doctorale Mathématiques de la région Paris-Sud
Laboratoire de Mathématiques de la Faculté des Sciences d'Orsay

THÈSE DE DOCTORAT SUR TRAVAUX

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS-SUD XI

Spécialité : Mathématiques

par

Caroline MEYNET

Sélection de variables pour la classification non supervisée en grande dimension

Soutenue le 09 novembre 2012 devant le jury composé de :

M.	Francis BACH	Rapporteur
M.	Gérard BIAU	Examinateur
M.	Christophe BIERNACKI	Rapporteur
M.	Gilles CELEUX	Examinateur
M.	Pascal MASSART	Directeur de thèse
Mme	Marie-Anne POURSAT	Examinatrice

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91405 Orsay CEDEX

Variable selection in model-based clustering for high-dimensional data

ABSTRACT

This thesis deals with variable selection for clustering. This problem has become all the more challenging since the recent increase in high-dimensional data where the number of variables can largely exceeds the number of observations (DNA analysis, functional data clustering...).

We propose a variable selection procedure for clustering suited to high-dimensional contexts. We consider clustering based on finite Gaussian mixture models in order to recast both the variable selection and the choice of the number of clusters into a global model selection problem. Following Pan and Shen (2007), we use the variable selection property of ℓ_1-regularization and the Lasso estimator to build a data-driven model collection in a efficient way. Our procedure differs from Pan and Shen's procedure as regards the estimation of the mixture parameters and the model selection criterion: we favor an estimation by the maximum likelihood estimators rather than by the Lasso estimators and we advocate for a non-asymptotic penalized criterion (Massart, 2007). From a theoretical viewpoint, we establish a model selection theorem for maximum likelihood estimators in a density estimation framework with a random model collection. We apply it in our context to determine a convenient penalty shape for our criterion. From a practical viewpoint, we carry out simulations to validate our procedure, for instance in the functional data clustering framework.

The basic idea of our procedure, which consists in variable selection by ℓ_1-regularization but estimation by ℓ_0-regularization, comes from theoretical results we establish in the first part of this thesis: we provide ℓ_1-oracle inequalities for the Lasso in the regression framework, which are valid with no assumption at all contrary to the usual ℓ_0-oracle inequalities in the literature, thus suggesting a gap between ℓ_1-regularization and ℓ_0-regularization.

KEYWORDS

Variable selection, data-driven non-asymptotic model selection criterion, finite Gaussian mixture models, clustering, high dimension, Lasso, ℓ_1-regularization, oracle inequalities.

Sélection de variables pour la classification non supervisée en grande dimension

RÉSUMÉ

Il existe des situations de modélisation statistique pour lesquelles le problème classique de classification non supervisée (c'est-à-dire sans information a priori sur la nature ou le nombre de classes à constituer) se double d'un problème d'identification des variables réellement pertinentes pour déterminer la classification. Cette problématique est d'autant plus essentielle que les données dites de grande dimension, comportant bien plus de variables que d'observations, se multiplient ces dernières années : données d'expression de gènes, classification de courbes...

Nous proposons une procédure de sélection de variables pour la classification non supervisée adaptée aux problèmes de grande dimension. Nous envisageons une approche par modèles de mélange gaussien, ce qui nous permet de reformuler le problème de sélection des variables et du choix du nombre de classes en un problème global de sélection de modèle. Inspirés par Pan et Shen (2007), nous exploitons les propriétés de sélection de variables de la régularisation ℓ_1 et de l'estimateur Lasso pour construire efficacement à partir des données une collection de modèles qui reste de taille raisonnable même en grande dimension. Nous nous démarquons de Pan et Shen (2007) par une estimation des paramètres du mélange par l'estimateur du maximum de vraisemblance plutôt que par le Lasso, et par une sélection de modèle par un critère pénalisé non asymptotique basé sur l'heuristique de pente introduite par Birgé and Massart (2006). D'un point de vue théorique, nous établissons un théorème de sélection de modèle pour l'estimation d'une densité par maximum de vraisemblance pour une collection aléatoire de modèles. Nous l'appliquons dans notre contexte pour trouver une forme de pénalité minimale pour notre critère pénalisé. D'un point de vue pratique, des simulations sont effectuées pour valider notre procédure, en particulier dans le cadre de la classification non supervisée de courbes.

L'idée de base de notre procédure, qui mêle sélection de variables par la régularisation ℓ_1 mais estimation par la régularisation ℓ_0, nous est inspirée par une étude théorique menée dans une première partie : nous établissons des inégalités oracle ℓ_1 pour le Lasso dans les cadres de régression gaussienne et de mélange de régressions gaussiennes, qui se démarquent des inégalités oracle ℓ_0 traditionellement établies par leur absence totale d'hypothèse.

MOTS CLÉS

Sélection de variables, critère de sélection de modèle non asymptotique, modèles de mélange gaussien, classification non supervisée, grande dimension, Lasso, régularisation ℓ_1, inégalités oracle.

Remerciements

Mes remerciements les plus chaleureux s'adressent à mon directeur de thèse Pascal Massart pour lequel j'ai une profonde estime et avec lequel j'ai beaucoup apprécié de travailler durant ces trois années. Pascal, merci de m'avoir accordé ta confiance en me proposant ce sujet de thèse si passionnant. Merci de m'avoir fait découvrir le monde de la recherche, et de m'avoir fait profiter de ta grande expérience. Merci pour ta disponibilité sans limite, tes conseils avisés et tes intuitions éclairées qui m'ont fait progresser et franchir les obstacles un à un au fil de nos réunions. Enfin, merci d'avoir guidé mon insertion dans la communauté statistique en me proposant des rencontres enrichissantes, scientifiquement et humainement.

Toute ma gratitude à Francis Bach et Christophe Biernacki pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter ma thèse, ainsi qu'à Gérard Biau, Gilles Celeux et Marie-Anne Poursat qui me font l'honneur de participer à mon jury.

Je souhaite témoigner ma reconnaissance à plusieurs professeurs de mathématiques qui ont guidé mon parcours. Je pense en particulier à mes professeurs de classe préparatoire Stéphane Aulagnier et Michel Quercia qui m'ont véritablement transmis le goût des mathématiques, et à Frédéric Pascal qui m'a prodigué de précieux conseils dès mon entrée à l'ENS de Cachan. C'est notamment grâce à lui que je me suis tournée vers la Faculté des Sciences d'Orsay, où j'ai eu l'opportunité de suivre des cours de statistiques en Master 2 de très grande qualité, qui m'ont motivée à poursuivre en thèse dans cette branche captivante des mathématiques. En tant que monitrice, j'ai eu la chance d'être encadrée par Marie-Anne Poursat avec qui j'ai pris un grand plaisir à enseigner à mon tour les statistiques. Cette expérience a confirmé mon goût prononcé pour l'enseignement.

Je tiens à adresser mes remerciements les plus sincères aux personnes qui m'ont accordé de leur temps et de leur expérience pour m'aider dans mes travaux de recherche. Je pense tout particulièrement à Gilles Celeux et Cathy Maugis avec lesquels j'ai beaucoup travaillé et échangé. Gilles, merci pour ton pragmatisme et tes suggestions clairvoyantes qui m'ont permis de résoudre des problèmes algorithmiques et de lutter contre le fléau de la grande dimension. Cathy, merci de m'avoir toujours réservé un créneau dans ton planning surchargé à chacun de tes passages à Paris. J'espère que notre collaboration permettra de donner suite à ce sujet de thèse si riche. Un merci spécial à Erwan Le

Pennec pour avoir accepté d'être mon tuteur et pour avoir discuté inégalités oracles ℓ_0 avec moi, et à Vincent Rivoirard pour son cours sur le Lasso et ses références bibliographiques qui m'ont été très profitables. Merci à tous les deux pour vos précieux conseils. Merci à Christophe Giraud et Nicolas Verzelen pour nos échanges autour de l'entropie à crochets à Polytechnique, à Gilles Stoltz et à Bertrand Thirion pour m'avoir offert l'opportunité de diffuser mes travaux, à Jean-Michel Poggi, Gérard Biau et Aurélie Fischer pour m'avoir fait profiter de leur connaissance sur la classification de courbes et les ondelettes. Thanks to Sara van de Geer, to Wei Pan and to Xiaotong Shen for answering my numerous questions about the Lasso.

Ma dernière année de thèse a été largement consacrée à la découverte des simulations algorithmiques. J'ai pu les pratiquer dans des conditions optimales grâce à l'aide précieuse de certaines personnes. Merci à Yves Misiti pour son efficacité et sa disponibilité constante, à Yves Auffray pour avoir optimisé mes codes algorithmiques, et à Bertrand Michel pour en avoir fait tourner certains sur le cluster de son laboratoire.

La bonne ambiance au sein de l'équipe de Probabilités-Statistiques d'Orsay et de l'équipe Select de l'Inria a été très appréciable. Merci à Valérie Lavigne et Katia Evrat pour leur gentillesse et leur efficacité qui ont rendu les tâches administratives plus agréables. Un grand merci à mes collègues et amis Laure, Thierry, Lionel, Olivier, Vincent, Sébastien, Ramla, Aude, Guillaume et en particulier à Aurélien, pour leur bonne humeur communicative et tous les bons moments partagés. Sans oublier mes amis statisticiens parisiens Christophe et Robbi.

Bien loin des mathématiques, mais si proche dans mon coeur, je voudrais enfin remercier ma famille pour ses encouragements depuis toujours. Laurent, merci pour toutes nos années de complicité. Papa, merci pour ta générosité. Maman, merci pour ton soutien permanent, ton optimisme et ton énergie inébranlables.

FIGURE 1 – Une illustration tirée d'un livre d'enfance – prémonitoire – de Maman : *Caroline au ranch* de Pierre Probst. Le reflet même de mes trois dernières années : épaulée par mes alliés, j'ai persévéré à apprivoiser la technique du Lasso. Je laisse à chacun d'entre vous, cités plus haut, le soin de vous retrouver dans un personnage. Merci à tous !

Contents

1. Présentation générale des résultats **13**
1.1. Sélection de variables pour la classification non supervisée en grande dimension . . . 15
1.2. Présentation du Lasso 20
1.3. Vue d'ensemble de nos résultats 30

I Some ℓ_1-oracle inequalities for the Lasso in Gaussian regression models **51**

2. Homogeneous Gaussian regression models **53**
2.1. Introduction 55
2.2. Models and notations 58
2.3. Some ℓ_1-oracle inequalities for the Lasso 60
2.4. Some rates of convergence for the Lasso 66
2.A. The Lasso as an ℓ_1-ball model selection procedure 71
2.B. Proof of the ℓ_1-oracle inequalities 76
2.C. Proofs of the rates of convergence 83
2.D. Interpolation spaces 90

3. Finite mixture Gaussian regression models **93**
3.1. Introduction 95
3.2. Notations and framework 98
3.3. An ℓ_1-oracle inequality for the Lasso 100
3.A. Proof of Theorem 3.3.1 102
3.B. Proofs of the lemmas 112

II Variable selection in model-based clustering for high-dimensional data **123**

4. Our Lasso-MLE procedure for variable selection in model-based clustering **125**
4.1. Introduction 127

4.2. Variable selection for clustering 131
4.3. Presentation of two competitor variable selection procedures 135
4.4. Some discussion on empirical centering 140
4.5. Our variable selection procedure: the Lasso-MLE procedure 149
4.6. A major application: functional data clustering 160
4.A. The EM algorithms 166
4.B. Some details about our algorithms 173
4.C. Proofs 176

5. Simulations **179**
5.1. Introduction 181
5.2. Definitions 182
5.3. Validation of the Lasso-MLE procedure on simulated data 184
5.4. Functional data clustering using wavelets 198

6. A non-asymptotic data-based model selection criterion **215**
6.1. Introduction 217
6.2. An oracle inequality for the Lasso-MLE estimator 219
6.3. The slope heuristics under the null model principle 225
6.A. Proofs 239

A. Alternatives to our Lasso-MLE procedure **253**
A.1. Two alternative procedures to our Lasso-MLE procedure 255
A.2. Performance of the three procedures on simulated data 259

Bibliographie **267**

Chapitre 1

Présentation générale des résultats

Sommaire

1.1. Sélection de variables pour la classification non supervisée en grande dimension 15

1.1.1. Modèles de mélange gaussien pour la classification non supervisée 15

1.1.2. Sélection des variables pertinentes pour la classification 16

1.1.3. Le défi de la grande dimension 17

1.2. Présentation du Lasso 20

1.2.1. Cadre historique 20

1.2.2. La pénalisation ℓ_1 comme substitut à la pénalisation ℓ_0 21

1.2.3. Le calcul des solutions Lasso 23

1.2.4. Résultat de prédiction 25

1.2.5. Résultats d'estimation et de sélection de variables 27

1.3. Vue d'ensemble de nos résultats 30

RÉSUMÉ

Dans cette thèse, nous proposons une procédure de sélection de variables pour la classification non supervisée en grande dimension, dans un cadre de modèles de mélange gaussien. La régularisation ℓ_1 et l'estimateur associé Lasso sont au coeur de nos travaux. Dans ce chapitre introductif, nous présentons notre cadre de travail, puis nous effectuons quelques rappels utiles sur le Lasso, avant de résumer nos contributions.

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

1.1.1 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

$$s_{\boldsymbol{\theta}} : \mathbb{R}^p \mapsto \mathbb{R},\ \boldsymbol{y} \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k\, \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}) \tag{1.1}$$

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

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k\, \Phi\left(\boldsymbol{Y_i} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \hat{\pi}_l\, \Phi\left(\boldsymbol{Y_i} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)} \tag{1.2}$$

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 $l \neq k$.

1.1.2 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.

1.1.3 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.

1.1.3.1 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).

1.1.3.2 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).

1.1.3.3 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 combinatoires basées sur des distances se bornent souvent à faire de la réduction dimensionnelle sans réellement sélectionner des variables. Par exemple, dans la cas de la classification de données fonctionnelles, des projections sur des bases de splines, de Fourier ou d'ondelettes permettent de passer d'un problème de dimension infinie à un problème de dimension finie (Abraham et al., 2003 ; Auder et Fischer, 2011 ; Misiti et al., 2007a). Par contre, Antoniadis et al. (2011) introduisent une réelle étape de sélection de variables en mesurant la contribution des coefficients d'ondelettes par rapport à l'énergie totale de la courbe.

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 (1.1) 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.

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). Aussi, avant de présenter les travaux de cette thèse, nous introduisons le Lasso et nous rappelons quelques résultats de prédiction, d'estimation et de sélection de variables de cet estimateur dans le cadre de la régression. A notre connaissance, aucun résultat théorique sur le Lasso n'a été établi dans notre cadre de classification non supervisée. Cependant, par extrapolation, les résultats établis dans le cadre de la régression permettent de se faire une idée des performances possibles du Lasso dans notre contexte et de comprendre les choix que nous avons fait pour construire notre procédure.

1.2 Présentation du Lasso

1.2.1 Cadre historique

Le Lasso fut introduit en parallèle par Tibshirani (1996) en régression linéaire et par Chen et al. (1999) (sous le nom de "Basis Pursuit DeNoising") pour la décomposition d'un signal dans un dictionnaire d'ondelettes. Nous nous restreignons ici à un "design" déterministe et à un bruit gaussien.

La régression linéaire gaussienne

On dispose de n observations $(\boldsymbol{x_1}, Y_1), \ldots, (\boldsymbol{x_n}, Y_n)$ de $\mathbb{R}^p \times \mathbb{R}$ indépendantes identiquement distribuées (i.i.d.) et on suppose que pour tout $i \in \{1, \ldots, n\}$,

$$Y_i = \sum_{j=1}^{p} \beta_j^* x_{ij} + \varepsilon_i. \tag{1.3}$$

Les variables aléatoires de bruit ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et les éléments x_{ij} sont déterministes. Pour tout $j \in \{1, \ldots, p\}$, l'influence de la variable $\boldsymbol{X^j} := (x_{1j}, \ldots, x_{nj})$ sur les réponses Y_i est représentée par la valeur du coefficient de régression β_j^*. Les coefficients de régression sont inconnus et à estimer.

Décomposition dans un dictionnaire

Soit $\mathcal{X}$ un ensemble mesurable. On dispose de n observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ de $\mathcal{X} \times \mathbb{R}$ i.i.d. et on suppose que pour tout $i \in \{1, \ldots, n\}$,

$$Y_i = s(x_i) + \varepsilon_i.$$

Les mesures Y_i sont des observations de $s(x_i)$ bruitées par des variables aléatoires ε_i i.i.d. $\mathcal{N}(0, \sigma^2)$ et les éléments x_i sont déterministes. La fonction $s : \mathcal{X} \mapsto \mathbb{R}$ est la fonction de régression inconnue à estimer. On considère un dictionnaire $\mathcal{D} = \{\phi_1, \ldots, \phi_p\}$, c'est-à-dire un ensemble de p fonctions $\phi_j : \mathcal{X} \mapsto \mathbb{R}$, qui constituent p variables. On suppose que s peut se décomposer dans ce dictionnaire

sous la forme $s = \sum_{j=1}^{p} \beta_j^* \phi_j$. Alors pour tout $i \in \{1, \dots, n\}$,

$$Y_i = \sum_{j=1}^{p} \beta_j^* \phi_j(x_i) + \varepsilon_i. \tag{1.4}$$

Estimer s revient à estimer les p coefficients de décomposition $\beta_1^*, \dots, \beta_p^*$.

Ecriture commune

Les deux modèles (1.3) et (1.4) peuvent s'écrire sous la forme matricielle suivante :

$$\boldsymbol{Y} = X\boldsymbol{\beta^*} + \boldsymbol{\varepsilon} \tag{1.5}$$

où $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, $\boldsymbol{\beta^*} = (\beta_1^*, \dots, \beta_p^*)$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$, $X = (x_{ij})_{1 \le i \le n, 1 \le j \le p}$ pour (1.3) ou $X = (\phi_j(x_i))_{1 \le i \le n, 1 \le j \le p}$ pour (1.4). On se place dans le cadre de la grande dimension $p \gg n$ et sous l'hypothèse de parcimonie, c'est-à-dire qu'on suppose qu'il existe une solution $\boldsymbol{\beta^*}$ possédant de nombreux coefficients β_j^* nuls. Dans le cadre de la régression linéaire, cela revient à supposer qu'un grand nombre de variables $\boldsymbol{X^j}$ sont inutiles pour expliquer la réponse $\boldsymbol{Y}$. Pour la décomposition dans un dictionnaire, cela revient à supposer l'existence d'une décomposition parcimonieuse de la fonction s dans le dictionnaire $\mathcal{D}$. Le nombre de variables intervenant réellement dans la description du modèle est

$$\|\boldsymbol{\beta^*}\|_0 := \text{card}\left\{j \in \{1, \dots, p\};\ \beta_j^* \neq 0\right\}. \tag{1.6}$$

On cherche à estimer $\boldsymbol{\beta^*}$.

1.2.2 La pénalisation ℓ_1 comme substitut à la pénalisation ℓ_0

1.2.2.1 Notion de pénalisation

Soit $(e_j)_{1 \le j \le p}$ la base canonique de $\mathbb{R}^p$ et $\mathcal{M}$ l'ensemble des parties de $\{1, \dots, p\}$. Pour tout $m \in \mathcal{M}$, on note S_m l'espace vectoriel engendré par la famille de vecteurs $\{e_j, j \in m\}$. La perte théorique d'un vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$ est $\|X\boldsymbol{\beta^*} - X\boldsymbol{\beta}\|^2$ tandis que son risque empirique est $\|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2$, où $\|\boldsymbol{v}\| = (\sum_{i=1}^{n} v_i^2/n)^{1/2}$ désigne la norme ℓ_2 renormalisée de $\mathbb{R}^n$. On considère les minimiseurs de la perte théorique et du risque empirique dans S_m :

$$\boldsymbol{\beta}_m^* = \underset{\boldsymbol{\beta} \in S_m}{\arg\min} \|X\boldsymbol{\beta^*} - X\boldsymbol{\beta}\|^2, \qquad \hat{\boldsymbol{\beta}}_m^* = \underset{\boldsymbol{\beta} \in S_m}{\arg\min} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2. \tag{1.7}$$

On cherche le modèle S_m minimisant le risque $\mathbb{E}[\|X\boldsymbol{\beta^*} - X\hat{\boldsymbol{\beta}}_m^*\|^2]$. Le problème est que la perte théorique $\|X\boldsymbol{\beta^*} - X\hat{\boldsymbol{\beta}}_m^*\|$ n'est pas accessible car elle dépend de $\boldsymbol{\beta^*}$ qui est inconnu. On peut vouloir lui substituer une quantité calculable à partir des données, et le risque empirique $\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}_m^*\|^2$ est le candidat naturel. Par construction des estimateurs $\hat{\boldsymbol{\beta}}_m^*$, la fonction $m \mapsto \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}_m^*\|^2$ est décroissante

pour l'inclusion. Ainsi, minimiser le risque empirique mène à choisir le modèle le plus complexe : S_m avec $m = \{1, \ldots, p\}$. Or, on peut montrer (Massart, 2007) que le risque admet la décomposition biais-variance suivante :

$$\mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*_m\|^2\right] = \underbrace{\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}^*_m\|^2}_{\text{biais}(m)} + \underbrace{|m|\sigma^2}_{\text{variance}(m)} ,$$

où $|m|$ désigne le cardinal de m et représente la complexité du modèle S_m. Le terme de variance augmente quand $|m|$ augmente, alors que d'après (1.7) le terme de biais diminue quand $|m|$ augmente. Ainsi, le minimum du risque est atteint pour un modèle de complexité intermédiaire qui équilibre terme de biais et terme de variance, et non pour le modèle de complexité maximale obtenu par minimisation du risque empirique. La pénalisation est une méthode qui consiste à modifier le risque empirique $\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_m\|$ en lui ajoutant un terme complémentaire $\mathrm{pen}(m) > 0$ de façon à sélectionner un modèle $S_{\hat{m}}$ de complexité intermédaire en minimisant le risque empirique pénalisé par la pénalité $\mathrm{pen}(m)$:

$$\hat{m} = \underset{m \in \mathcal{M}}{\arg\min} \left\{\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_m\|^2 + \mathrm{pen}(m)\right\}. \tag{1.8}$$

La pénalité $\mathrm{pen}(m)$ représente un prix à payer fonction de la complexité du modèle. Le critère (1.8) vise un équilibre entre apprentissage des données, mesuré par $\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_m\|^2$, et complexité du modèle, mesurée par $\mathrm{pen}(m)$.

1.2.2.2 La pénalisation ℓ_0

Historiquement (Mallows, 1973), les premières pénalités $m \mapsto \mathrm{pen}(m)$ considérées dans le problème de pénalisation (1.8) étaient proportionnelles au cardinal du modèle S_m : $\mathrm{pen}(m) = \lambda|m|$, $\lambda > 0$. Alors, par définition (1.8) de $\hat{m}$,

$$\forall\, m \in \mathcal{M}, \quad \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_{\hat{m}}\|^2 + \lambda|\hat{m}| \leq \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_m\|^2 + \lambda|m|.$$

Puis, par définition (1.7) de $\hat{\boldsymbol{\beta}}^*_m$, cela implique que

$$\forall\, m \in \mathcal{M}, \forall\, \boldsymbol{\beta} \in S_m, \quad \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_{\hat{m}}\|^2 + \lambda|\hat{m}| \leq \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda|m|.$$

Comme tout $\boldsymbol{\beta}$ dans S_m a m coordonnées non nulles, et que $\cup_{m \in \mathcal{M}} S_m = \mathbb{R}^p$, l'inégalité précédente s'écrit aussi

$$\forall\, \boldsymbol{\beta} \in \mathbb{R}^p, \quad \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}^*_{\hat{m}}\|^2 + \lambda\|\hat{\boldsymbol{\beta}}^*_{\hat{m}}\|_0 \leq \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_0,$$

où $\|\boldsymbol{\beta}\|_0 = \mathrm{card}\,\{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$ est le nombre de coordonnées non nulles de $\boldsymbol{\beta}$. Ainsi, on estime $\boldsymbol{\beta}^*$ par

$$\hat{\boldsymbol{\beta}}^*_{\hat{m}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{\|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_0\right\}. \tag{1.9}$$

En pénalisant le nombre de composantes non nulles des vecteurs $\boldsymbol{\beta}$, ce critère de pénalisation ℓ_0 permet de trouver un estimateur $\hat{\boldsymbol{\beta}}^* := \hat{\boldsymbol{\beta}}^*_m$ parcimonieux et conduit à une sélection de variables : seules les variables indexées par j tel que $\hat{\beta}^*_j \neq 0$ sont considérées comme intervenant réellement dans le problème. Cependant, le problème de minimisation (1.9) est algorithmiquement incalculable dès que p est de l'ordre de quelques dizaines.

1.2.2.3 Substitut à la pénalisation ℓ_0

Pour résoudre ce problème numérique, une idée est de chercher une autre pénalisation, similaire à la pénalisation ℓ_0, mais conduisant à un problème de minimisation dont on pourrait calculer la solution même pour de grandes valeurs de p. En remarquant que

$$\|\boldsymbol{\beta}\|_0 = \text{card}\,\{j \in \{1,\dots,p\} :\ \beta_j \neq 0\} = \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} = \lim_{q\to 0} \sum_{j=1}^{p} |\beta_j|^q = \lim_{q\to 0} \|\boldsymbol{\beta}\|_q^q,$$

on peut envisager de remplacer la pénalisation ℓ_0 par une pénalisation ℓ_q avec $q > 0$ proche de 0 et dont la solution

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_q^q \right\}, \quad \lambda > 0, \tag{1.10}$$

serait plus facilement calculable que la solution (1.9). Pour $q < 1$, $\|\cdot\|_q$ n'est pas une norme alors que pour $q \geq 1$, $\|\cdot\|_q$ est une norme et vérifie la propriété de convexité. Ainsi, pour $q \geq 1$, le problème (1.10) est un problème de minimisation convexe et peut se résoudre efficacement. Comme $q = 1$ est la valeur la plus proche de 0 vérifiant cette propriété, Tibshirani (1996) a proposé de tester la pénalisation ℓ_1 comme substitut numérique à la pénalisation ℓ_0 et a introduit l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ défini par

$$\hat{\boldsymbol{\beta}}^*(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{1.11}$$

où $\lambda > 0$ est un paramètre de régularisation à calibrer.

1.2.3 Le calcul des solutions Lasso

1.2.3.1 Cas orthogonal

Dans le cas où X est une matrice orthogonale ($X^T X/n = I$), l'estimateur Lasso (1.11) peut être explicitement calculé par sous-différentiation : pour tout $j \in \{1,\dots,p\}$,

$$\hat{\beta}^*_j(\lambda) = \text{sign}\left(\hat{\beta}^*_j\right)\left(\left|\hat{\beta}^*_j\right| - \lambda/2\right)_+ = \begin{cases} \hat{\beta}^*_j - \lambda/2 & \text{si } \hat{\beta}^*_j > \lambda/2; \\ \hat{\beta}^*_j + \lambda/2 & \text{si } \hat{\beta}^*_j < -\lambda/2; \\ 0 & \text{sinon} \end{cases} \tag{1.12}$$

où $\hat{\boldsymbol{\beta}}^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2 = X^T\boldsymbol{Y}$ est l'estimateur des moindres carrés, c'est-à-dire l'estimateur obtenu sans pénalisation.

D'après (1.12), la pénalité ℓ_1 provoque un seuillage doux des estimations des coefficients de régression obtenues par la méthode des moindres carrés. En particulier, les coefficients β_j^* tels que $|\hat{\beta}_j^*| \leq \lambda/2$ sont estimés par zéro par le Lasso et les variables correspondantes sont éliminées du modèle. Le Lasso effectue donc une sélection de variables de manière automatique. Cette sélection varie en fonction du niveau de pénalisation que l'on impose et que l'on règle avec le choix du paramètre de régularisation λ. Pour $\lambda = 0$, le risque empirique n'est pas pénalisé, l'estimateur Lasso correspond à l'estimateur des moindres carrés et toutes les variables sont sélectionnées. Plus λ augmente, plus le nombre de variables sélectionnées diminue, et pour λ assez grand, aucune variable n'est sélectionnée. Cette évolution peut être précisée. D'après (1.12), le chemin de régularisation $\lambda \in [0, +\infty[\mapsto \hat{\boldsymbol{\beta}}^*(\lambda)$ est continu et linéaire par morceaux (chacune des coordonnées est une fonction réelle continue et linéaire par morceaux). Ainsi, le calcul de l'ensemble des estimateurs Lasso ne nécessite que le calcul d'un nombre fini d'entre eux. Formellement, soient $\{j_1, \ldots, j_p\}$ tels que $|\hat{\beta}_{j_l}^*| \leq |\hat{\beta}_{j_{l+1}}^*|$. Notons $\lambda_l := 2|\hat{\beta}_{j_l}^*|$ et posons $\lambda_0 = 0$. Alors, pour tout $l \in \{1, \ldots, p-1\}$, pour tout $\lambda \in [\lambda_l, \lambda_{l+1}[$, $\hat{\boldsymbol{\beta}}^*(\lambda)$ sélectionne les variables indexées par $j \in \{j_{l+1}, \ldots, j_p\}$. Pour $\lambda \in [0, \lambda_1[$, $\hat{\boldsymbol{\beta}}^*(\lambda)$ sélectionne toutes les variables, et pour $\lambda \in [\lambda_p, +\infty[$, $\hat{\boldsymbol{\beta}}^*(\lambda)$ ne sélectionne aucune variable. Ainsi, en faisant varier $\lambda \in [0, +\infty[$, la procédure Lasso ne visite qu'une collection restreinte de paquets de variables. La collection de paramètres de régularisation $\{\lambda_l\}_{0\leq l\leq p}$ suffit à elle seule à parcourir cette collection de paquets de variables. Il suffit donc de résoudre $p+1$ problèmes de minimisation pour obtenir l'ensemble des solutions Lasso.

1.2.3.2 Cas général

Dans le cas général où X n'est pas une matrice orthogonale, la solution (1.11) ne peut plus être calculée explicitement. Cependant, Efron et al. (2004) et Zou et al. (2007) ont démontré que les propriétés intéressantes du Lasso, à savoir la parcimonie des solutions et la linéarité par morceaux du chemin de régularisation $\lambda \mapsto \hat{\boldsymbol{\beta}}^*(\lambda)$, restent vérifiées. De plus, les valeurs du paramètre de régularisation λ telles que la linéarité change peuvent être trouvées en considérant les conditions d'optimalité du premier ordre de (1.11) (Efron et al., 2004). Ainsi, comme dans le cas orthogonal, il suffit de calculer les solutions Lasso en ces valeurs particulières de λ pour en déduire toutes les solutions Lasso. Bien-sûr, ces valeurs particulières dépendent des données. En pratique, l'ensemble des solutions Lasso peuvent être calculées par un algorithme d'homotopie très efficace en considérant une version modifiée de l'algorithme LARS introduit par Efron et al. (2004). Il est alors d'usage de sélectionner l'une de ces solutions Lasso par un critère de sélection de modèle.

A ce stade, notons qu'une alternative serait de n'utiliser les avantages algorithmiques de la pénalisation ℓ_1 dans le seul but de conserver l'ensemble des paquets de variables parcouru le long du chemin

de régularisation. Ensuite, on pourrait revenir à une estimation de $\boldsymbol{\beta}^*$ par l'estimateur des moindres carrés (et non par l'estimateur Lasso) sur chacun de ces paquets de variables, puis à la sélection de l'un de ces estimateurs des moindres carrés par un critère de pénalisation en norme ℓ_0. Cette idée est mentionnée dans Efron et al. (2004) et a été explorée par Connault (2011). Pour construire notre procédure de sélection de variables pour la classification non supervisée en grande dimension, c'est cette approche alternative – adaptée à notre contexte – que nous adopterons. Cela constituera l'un des points centraux de nos travaux.

Parallèlement aux nombreuses utilisations pratiques de cet algorithme, des résultats théoriques ont été établis pour évaluer les performances de prédiction, d'estimation et de sélection de variables du Lasso. Dans le cadre de la grande dimension, des résultats de prédiction ont été établis entre autres par Greenshtein et Ritov (2004), Bunea et al. (2007a), Bickel et al. (2009), van de Geer (2008), Koltchinskii (2009), alors que des études centrées sur la sélection ont été menées par Zou (2006), Zhao et Yu (2007), Meinshausen et Bühlmann (2006), Candes et Tao (2007), Wainwright (2009), Meinshausen et Yu (2009)... Il est impossible de citer tous ces résultats. Nous n'avons sélectionné que quelques uns d'entre eux : les plus éclairants au regard du travail qui va suivre dans cette thèse.

1.2.4 Résultat de prédiction

<u>But</u>

La pénalisation ℓ_1 étant utilisée comme substitut convexe à la pénalisation ℓ_0, on peut mesurer les performances de prédiction du Lasso par rapport à l'oracle ℓ_0 en établissant des inégalités oracle. Ces inégalités sont à comparer à l'inégalité oracle ℓ_0 que l'on obtiendrait en considérant l'estimateur $\hat{\boldsymbol{\beta}}^*$ régularisé en norme ℓ_0 pour une pénalité ℓ_0 correctement choisie (voir Massart, 2007, pour la détermination d'une telle pénalité) :

$$\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*\|^2 \leq (1+\delta) \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + C(\delta)\,\sigma^2 \frac{\ln p}{n}\, \|\boldsymbol{\beta}\|_0 \right\}. \tag{1.13}$$

<u>Hypothèse</u>

Pour $\boldsymbol{v} \in \mathbb{R}^p$ et $\boldsymbol{J} \subset \{1,\ldots,p\}$, on note $\boldsymbol{v_J} \in \mathbb{R}^p$ le vecteur de mêmes coordonnées que $\boldsymbol{v}$ sur $\boldsymbol{J}$ et de coordonnées nulles sur $\boldsymbol{J^c}$. On introduit la matrice de Gram $\Psi_n = X^T X/n$.

Soient $d \in \{1,\ldots,p\}$ et $c > 0$. On dit que la *condition de valeur propre restreinte ("Restricted Eigenvalue") RE(d,c)* est vérifiée si

$$\kappa(d,c) := \min_{\substack{\boldsymbol{J}\subset\{1,\ldots,p\},\\ |\boldsymbol{J}|\leq d}} \; \min_{\substack{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\},\\ \|\boldsymbol{v_{J^c}}\|_1 \leq c\|\boldsymbol{v_J}\|_1}} \frac{\|X\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{v_J}\|} > 0. \tag{1.14}$$

Résultat

Résultat 1.2.1. [Théorème 5.1 de Bickel et al. (2009)] *Soient $d \in \{1, \ldots, p\}$ et $\delta > 0$. Supposons RE($d, 3 + 4/\delta$) définie par* (1.14) *vérifiée. Alors, il existe $C(\delta) > 0$ tel que, pour tout paramètre de régularisation $\lambda = A\sigma\sqrt{\ln(p)/n}$ avec $A > 2\sqrt{2}$, l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ défini par* (1.11) *satisfait l'inégalité oracle ℓ_0 suivante avec probabilité plus grande que $1 - p^{1-A^2/8}$:*

$$\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*(\lambda)\|^2 \leq (1+\delta) \inf_{\boldsymbol{\beta}\in\mathbb{R}^p, \|\boldsymbol{\beta}\|_0 \leq d} \left\{ \|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + \frac{C(\delta)\, A^2\sigma^2}{\kappa^2(d, 3+4/\delta)} \frac{\ln p}{n} \|\boldsymbol{\beta}\|_0 \right\}. \quad (1.15)$$

Interprétation et discussion

- de la condition de valeur propre restreinte

Expliquons d'abord l'origine du nom "valeur propre restreinte". La plus petite valeur propre de la matrice de Gram ψ_n s'écrit

$$\min_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}} \frac{\left(\boldsymbol{v}^T \Psi_n \boldsymbol{v}\right)^{1/2}}{\|\boldsymbol{v}\|} = \min_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}} \frac{\|X\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{v}\|}. \quad (1.16)$$

En comparant (1.14) et (1.16), on voit que l'ensemble $\mathbb{R}^p \setminus \{0\}$ dans (1.16) est remplacé par un ensemble restreint de vecteurs dans (1.14) et que la norme ℓ_2 de $\boldsymbol{v}$ dans (1.16) est remplacée par la norme ℓ_2 sur une partie restreinte de $\boldsymbol{v}$ dans (1.14). D'où le nom "valeur propre restreinte".

En pratique, $\boldsymbol{J}$ correspond aux variables pertinentes. Géométriquement, la condition RE(d, c) revient à supposer une faible corrélation entre les variables pertinentes et les autres. Or, en grande dimension, les corrélations entre variables sont généralement nombreuses et cette hypothèse n'est pas facilement vérifiable. De plus, pour $p \geq n$, la matrice Ψ_n est dégénérée. Sa plus petite valeur propre est nulle, ce qui s'écrit

$$\min_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}} \frac{\|X\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{v}\|} = 0. \quad (1.17)$$

Ainsi, plus les ensembles "restreints" considérés dans (1.14) se rapprochent des ensembles complets de (1.17) (plus c et d sont grands), plus $\kappa(d, c)$ définie par (1.14) s'approche de 0.

- du Résultat 1.2.1

D'après (1.15), s'il existe $\boldsymbol{\beta}$ parcimonieux ($\|\boldsymbol{\beta}\|_0$ est petit) tel que $\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|$ est petit et si $\kappa^2(d, 3+4/\delta)$ n'est pas trop petit, alors le membre de droite de l'inégalité (1.15) est petit et la solution Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ présente un petit risque de prédiction si λ est choisi assez grand. Cependant, l'inégalité (1.15) n'est pas strictement de la forme (1.13) car son terme d'erreur faisant intervenir $\kappa^2(d, 3 + 4/\delta)$ qui dépend des données, il ne se réduit pas à une quantité déterministe $C(\delta)$. En particulier, si $\kappa^2(d, 3 + 4/\delta)$ est très petit, le membre de droite de (1.15) explose et l'inégalité (1.15) ne garantit pas un faible

risque de prédiction. Pour s'assurer $\kappa^2(d, 3+4/\delta)$ assez grand, il faut prendre d petit et δ grand, mais alors l'infimum dans (1.15) est pris sur un ensemble très petit et la constante $1+\delta$ s'éloigne de la valeur idéale 1, donc l'inégalité oracle (1.15) perd en significativité. La pénalisation ℓ_1 ne permet pas de se substituer pleinement à la pénalisation ℓ_0.

1.2.5 Résultats d'estimation et de sélection de variables

La pénalisation ℓ_1 étant utilisée comme procédure de sélection de variables, il convient de s'assurer de la pertinence des variables sélectionnées par le Lasso en les comparant aux variables utiles du vrai modèle. De plus, comme la régularisation ℓ_1 fait tendre tous les coefficients vers zéro, on peut se demander à quel point la sous-estimation des coefficients détériore leur estimation.

1.2.5.1 Résultats d'estimation

But

Dans un cadre non asymptotique, les résultats d'estimation sont souvent exprimés en majorant $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_1$ ou $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_2$ ou plus rarement $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_q$. Nous nous restreignons ici à un exemple obtenu en norme ℓ_1.

Hypothèse

Comme pour les résultats de prédiction, les résultats d'estimation nécessitent des hypothèses plus ou moins restrictives sur la matrice de Gram qui supposent une faible corrélation entre les variables. L'une des hypothèses la moins forte est la condition de valeur propre restreinte (1.14).

Résultat

Résultat 1.2.2. [Théorème 6.2 de Bickel et al. (2009)] *Supposons RE($\|\boldsymbol{\beta}^*\|_0, 3$) définie par* (1.14) *vérifiée. Alors, pour tout paramètre de régularisation $\lambda = A\sigma\sqrt{\ln(p)/n}$ avec $A > 2\sqrt{2}$, l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ satisfait l'inégalité suivante avec probabilité plus grande que $1 - p^{1-A^2/8}$:*

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_1 \leq \frac{16A\sigma}{\kappa(\|\boldsymbol{\beta}^*\|_0, 3)} \|\boldsymbol{\beta}^*\|_0 \sqrt{\frac{\ln p}{n}} \,. \tag{1.18}$$

Interprétation et discussion

L'inégalité (1.18) est à voir comme un résultat d'estimation au sens où $\max_{1\leq j\leq p} |\beta_j^* - \hat{\beta}_j^*(\lambda)| \leq \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_1$, si bien qu'une petite valeur de $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*(\lambda)\|_1$ implique une bonne estimation individuelle des coefficients β_j^* par $\hat{\beta}_j^*(\lambda)$. D'après (1.18), cette estimation sera d'autant meilleure que $\|\boldsymbol{\beta}^*\|_0$ est petit, c'est-à-dire que le vrai modèle est parcimonieux, et que $\kappa(\|\boldsymbol{\beta}^*\|_0, 3)$ n'est pas trop petit, c'est-à-dire que les variables ne sont pas trop corrélées.

1.2.5.2 Résultats de sélection de variables

But

On cherche à évaluer la pertinence des variables sélectionnées par le Lasso. On distingue deux types d'étude :

- *la consistance en sélection* compare l'ensemble des variables sélectionnées par l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ représenté par $\boldsymbol{J}_\lambda = \{j \in \{1, \ldots, p\} : \ \hat{\beta}_j^*(\lambda) \neq 0\}$ à l'ensemble des variables pertinentes du vrai modèle représenté par $\boldsymbol{J} = \{j \in \{1, \ldots, p\} : \ \beta_j^* \neq 0\}$;

- *la consistance en signe* compare le signe des estimations Lasso $\hat{\beta}_j^*(\lambda)$ au signe des vrais paramètres β_j^*.

Les résultats peuvent être établis d'un point de vue asymptotique ou non asymptotique. D'un point de vue non asymptotique, la procédure Lasso est dite

- *consistante en sélection* si $\mathbb{P}\left(\exists\, \lambda_n > 0 : \ \boldsymbol{J}_{\lambda_n} = \boldsymbol{J}\right) = 1 - \eta_n$ où $(\eta_n)_{n \geq 0}$ est une suite de réels positifs tendant vers 0 ;

- *consistante en signe* si $\mathbb{P}\left(\exists\, \lambda_n > 0 : \ \text{sign}(\hat{\boldsymbol{\beta}}^*(\lambda_n)) = \text{sign}\left(\boldsymbol{\beta}^*\right)\right) = 1 - \eta_n$ où $(\eta_n)_{n \geq 0}$ est une suite de réels positifs tendant vers 0, où $\text{sign}(\boldsymbol{x}) = (\text{sign}(x_1), \ldots, \text{sign}(x_p))$ pour $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ et où $\text{sign}(x)$ est le signe de x pour $x \in \mathbb{R}$ avec la convention $\text{sign}(0) = 0$.

L'intérêt de la consistance en signe par rapport à la consistance en sélection est de comparer le signe de la corrélation entre chaque variable et la réponse $\boldsymbol{Y}$ pour l'estimation Lasso et le vrai paramètre. La consistance en signe implique la consistance en sélection.

Hypothèse

On note $X_{\boldsymbol{J}}$ la restriction de X aux colonnes d'indice $j \in \boldsymbol{J}$ et $X_{\boldsymbol{J}^c}$ la restriction de X aux colonnes d'indice $j \in \boldsymbol{J}^c$. On considère les matrices de Gram restreintes $\Psi_n(\boldsymbol{J}^c, \boldsymbol{J}) = {X_{\boldsymbol{J}^c}}^T X_{\boldsymbol{J}}/n$ et $\Psi_n(\boldsymbol{J}, \boldsymbol{J}) = {X_{\boldsymbol{J}}}^T X_{\boldsymbol{J}}/n$.

On dit que la *Condition d'Irreprésentabilité*, notée CI, est vérifiée s'il existe $C > 0$ tel que

$$\left\| \ \Psi_n(\boldsymbol{J}^c, \boldsymbol{J}) \Psi_n^{-1}(\boldsymbol{J}, \boldsymbol{J}) \ \text{sign}(\boldsymbol{\beta}_{\boldsymbol{J}}^*) \right\|_\infty \leq 1 - C$$

où $\| \cdot \|_\infty$ désigne la norme ℓ_∞ de $\mathbb{R}^{\boldsymbol{J}^c}$.

Résultats

Les premiers résultats non asymptotiques avec $p \geq n$ établis dans le cadre de la régression linéaire sont dûs à Zhao et Yu (2007). Ils ont démontré que CI est suffisante pour assurer la consistance en signe de la procédure Lasso :

Résultat 1.2.3. [Théorème 4 de Zhao et Yu (2007)] *Supposons que pour tout $n \in \mathbb{N}^\star$, $\min_{j \in \boldsymbol{J}} \beta_j^* \geq n^{-\delta/2}$ avec $0 < \delta < 1$ et que p est exponentiel en n. Alors, si CI est vérifiée, la procédure Lasso est consistante en signe (et donc en sélection).*

A notre connaissance, la nécessité de CI n'a pas été prouvée dans le cas $p \geq n$. Par contre, dans le cas $p \leq n$, si p et $\boldsymbol{\beta}^*$ sont indépendants de n, Zhao et Yu (2007) ont démontré que CI est suffisante et presque nécessaire pour que la procédure Lasso soit asymptotiquement consistante en signe au sens où $\lim_{n\to\infty} \mathbb{P}\left(\exists\, \lambda_n > 0 :\ \text{sign}(\hat{\boldsymbol{\beta}}^*(\lambda_n)) = \text{sign}\,(\boldsymbol{\beta}^*)\right) = 1$. De plus, Zou (2006) a démontré qu'une condition similaire à CI est nécessaire pour que la procédure Lasso soit asymptotiquement consistante en sélection au sens où $\lim_{n\to\infty} \mathbb{P}\left(\exists\, \lambda_n > 0 :\ \boldsymbol{J}_{\lambda_n} = \boldsymbol{J}\right) = 1$.

Interprétation et discussion

- de la Condition d'Irreprésentabilité

En pratique, $\boldsymbol{J}$ correspond aux variables pertinentes. CI est une contrainte reliant les variables non pertinentes aux variables pertinentes. Quand les signes des vrais β_j^* sont inconnus, CI revient à exiger

$$\max_{j \in \boldsymbol{J}^c} \left\{ \left\| \Psi_n(j, \boldsymbol{J}) \Psi_n^{-1}(\boldsymbol{J}, \boldsymbol{J}) \right\|_1 \right\} \leq 1 - C.$$

Cela signifie que, dans le vrai modèle, la corrélation entre chaque variable non pertinente et l'ensemble des variables pertinentes ne peut pas atteindre la valeur 1, d'où le nom "irreprésentabilité". Le Lasso sélectionne les bonnes variables si et (presque) seulement si les variables non pertinentes du vrai modèle ne peuvent pas être représentées par les variables pertinentes du modèle.

- du Résultat 1.2.3

Le Résultat 1.2.3 suggère que le Lasso doit être utilisé avec précaution comme procédure de sélection de variables car sa consistance n'est valable que sous certaines hypothèses portant sur la matrice de Gram. Les résultats asymptotiques prouvent que si CI n'est pas vérifiée, alors quelque soit le paramètre de régularisation λ, $\hat{\boldsymbol{\beta}}^*(\lambda)$ ne sélectionnera pas les bonnes variables. D'ailleurs, Zou (2006) et Zhao et Yu (2007) ont fourni des exemples où CI n'est pas satisfaite et où le Lasso ne sélectionne pas les bonnes variables.

On peut se demander pourquoi le Lasso devient inconsistant en sélection quand CI n'est plus vérifiée. En fait, ce problème de sélection de variables est étroitement lié à un problème d'estimation. Pour

induire de la parcimonie, le Lasso fait tendre les coefficients β_j^* vers 0 jusqu'à annuler les plus faibles. Le problème est que cette diminution des coefficients est trop forte si bien que, lorsqu'il existe de fortes corrélations entre variables pertinentes et non pertinentes (quand CI n'est plus vérifiée), le Lasso a tendance à piocher parmi les variables non pertinentes fortement corrélées aux variables pertinentes pour ajouter l'estimation de leurs coefficients à celle des coefficients des variables pertinentes et ainsi compenser la sous-estimation des coefficients. Par conséquent, le Lasso a tendance à sélectionner trop de variables.

Vers d'autres pistes

Afin de corriger ce défaut du Lasso à sélectionner trop de variables, des versions modifiées du Lasso ont été introduites. On peut par exemple citer le Lasso adaptatif de Zou (2006) qui a recours à des poids adaptatifs dans la pénalité ℓ_1, le Lasso seuillé de van de Geer et al. (2011) qui ne conserve que les estimations Lasso supérieures à un certain seuil, ou bien le Bolasso de Bach (2008) qui est une version bootstrap du Lasso basée sur une réplication bootstrap de l'échantillon et sur la sélection par le Bolasso des seules variables sélectionnées par le Lasso à tous les rééchantillonnages. Le Bolasso est consistant en sélection sans les hypothèses traditionnelles.

Un autre point de vue, envisagé par Connault (2011), est de ne pas se servir de la pénalisation ℓ_1 pour sélectionner un estimateur Lasso de $\boldsymbol{\beta}^*$, mais seulement pour obtenir de manière efficace un ensemble de paquets de variables en parcourant le chemin de régularisation. Ensuite, Connault (2011) estime $\boldsymbol{\beta}^*$ par l'estimateur des moindres carrés sur chaque paquet de variables atteint le long du chemin de régularisation, et sélectionne l'un de ces estimateurs des moindres carrés par un critère de pénalisation ℓ_0. C'est cette dernière voie que nous adapterons à notre contexte et que nous exploiterons dans cette thèse.

1.3 Vue d'ensemble de nos résultats

Nos travaux sont centrés sur la sélection de variables pour la classification non supervisée en grande dimension. La régularisation ℓ_1 et l'estimateur Lasso sont au coeur de cette thèse. Ce 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.

Des chapitres de ce manuscrit ont fait l'objet de publication ou de soumission d'articles :

- Les inégalites oracle ℓ_1 du Chapitre 2 sont publiées : P. Massart et C. Meynet (2011), The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, Vol.5, 669–687.
- Les vitesses de convergence du Chapitre 2 sont publiées : P. Massart et C. Meynet (2012), Somes rates of convergence for the selected Lasso estimator. *Lecture Notes in Computer Science* Vol.7568/2012, 17–33.
- Le Chapitre 3 est publié : C. Meynet (2012), An ℓ_1-oracle inequality for the Lasso in finite mixture gaussian regression models. *ESAIM Probability and Statistics*, Cambridge University Press, http ://dx.doi.org/10.1051/ps/2012016.
- Les travaux de la partie II sont en cours de soumission : C. Meynet et C. Maugis-Rabusseau (2012), A sparse variable selection procedure in model-based clustering, soumis à *Journal of the American Statistical Association*.

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, telle l'inégalité (1.15). Ces résultats nécessitent des contraintes du type de (1.14) imposées 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). En particulier, dans le cas d'un dictionnaire fini, le modèle statistique considéré est celui présenté en (1.4).

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

$$\lambda \geq 4\sigma\sqrt{\frac{1+\ln p}{n}}, \tag{1.19}$$

il existe une constante $C > 0$ telle que, pour tout $A > 0$, l'estimateur Lasso $\hat{\boldsymbol{\beta}}^*(\lambda)$ défini par (1.11) satisfait l'inégalité oracle suivante avec probabilité plus grande que $1 - 3.4\mathrm{e}^{-A}$:

$$\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*(\lambda)\|^2 + \lambda\|\hat{\boldsymbol{\beta}}^*(\lambda)\|_1 \leq C \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} + \frac{\lambda(1+A)}{\sqrt{n}}. \tag{1.20}$$

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

$$\mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}^*(\lambda)\|^2 + \lambda\|\hat{\boldsymbol{\beta}}^*(\lambda)\|_1\right] \leq C \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} + \frac{\lambda}{\sqrt{n}}. \tag{1.21}$$

L'inégalité oracle ℓ_1 (1.20) est à mettre en parallèle de l'inégalité oracle ℓ_0 (1.15). Contrairement à (1.15), 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 (1.19) 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 (1.19) du paramètre de régularisation.

Des inégalités oracle du type de (1.20) et (1.21) 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 (1.19) 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

$$\hat{s}(\lambda) = \arg\min_{s_{\boldsymbol{\theta}}} \left\{ -\frac{1}{n} \sum_{i=1}^n \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i})\right) + \lambda |s_{\boldsymbol{\theta}}|_1 \right\}, \quad \lambda > 0, \tag{1.22}$$

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 (1.22) 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

$$\lambda \geq \kappa\, C\, K\, (\ln n)^2 \sqrt{\frac{\ln(2p+1)}{n}} \tag{1.23}$$

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 :

$$\mathbb{E}\left[\mathrm{KL_n}\left(s, \hat{s}(\lambda)\right)\right] \leq (1+\kappa^{-1}) \inf_{s_{\boldsymbol{\theta}} \in S} \left\{\mathrm{KL_n}(s, s_{\boldsymbol{\theta}}) + \lambda |s_{\boldsymbol{\theta}}|_1\right\} + \lambda + C'\, \frac{K^{3/2} \ln n}{\sqrt{n}}, \tag{1.24}$$

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, traduites par une hypothèse semblable à (1.14) mais sommée sur les K composantes du mélange. 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 $K = 1$, $\mathrm{KL_n}(s,t) = \mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2\right]/2$, donc l'inégalité (1.21) établie au Chapitre 2 est un cas particulier de l'inégalité (1.24) pour $K = 1$. Cependant, nous avons établi l'inégalité (1.21) sans hypothèse de bornitude sur les paramètres et la borne inférieure du paramètre de régularisation (1.19) ne comporte pas le terme en $(\ln n)^2$ de la borne inférieure (1.23). 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 1.1.1. 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 constatées dans le cadre de la régression et rappelées en Section 1.2.5.

Dans la suite, nous conservons les notations introduites en Section 1.1.1.

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 $s_{\widehat{\boldsymbol{\theta}}}$.

[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.

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

$$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = \underset{\overline{\boldsymbol{\theta}} \in \Theta_K}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}}) \right) + \lambda |\overline{\boldsymbol{\theta}}|_1 \right\} \tag{1.25}$$

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 (1.2).

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 (1.25) 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 établis en régression que nous avons rappelés en Sections 1.2.4 et 1.2.5 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 (1.25) 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 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

[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

(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, 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

[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.

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 mono-classe $K = 1$: 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

[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).

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 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 6.2.2 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 $D_m \leq p \wedge n$,

$$\mathrm{pen}\,(m) \geq \kappa_1 \frac{D_m}{n} \left[1 + \kappa_2 (1 \vee \tau) \ln\left(\frac{p}{D_m}\right)\right], \tag{1.26}$$

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 6.2.2 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 6.2.2, 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 6.2.2 fournit une forme de pénalité minimale (1.26) 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 que nous avons rappelés en Section 1.2.4, 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 6.2.2 sont à souligner :

- La pénalité minimale (1.26) 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 (1.26) 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 (1.26) 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 $\ln(p/D_m)$:

$$\mathrm{pen}_{\ln}(m) = \kappa_1 \frac{D_m}{n}\left(1 + \kappa_2 \ln\left(\frac{p}{D_m}\right)\right) \quad \text{ou} \quad \mathrm{pen}(m) = \kappa \frac{D_m}{n}. \tag{1.27}$$

- Même si la forme de la pénalité minimale (1.26) s'avère optimale, le Théorème 6.2.2 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 (1.27) car seule la constante multiplicative κ est à calibrer. Par contre, ce n'est pas le cas pour la pénalité $\mathrm{pen}_{\ln}$ définie par (1.27) 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 $\mathrm{pen}_{\ln}$ définies par (1.27). 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 $\mathrm{pen}_{\ln}$ 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 (1.27). 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 $\Sigma_k = \mathrm{diag}(\sigma_{k1}^2, \ldots, \sigma_{kp}^2)$, 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.

Part I

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

Chapter 2

Homogeneous Gaussian regression models

Contents

2.1. Introduction **55**

2.2. Models and notations **58**

2.2.1. General framework and statistical problem 58

2.2.2. The Lasso: an ℓ_1-penalized least squares estimator 59

2.3. Some ℓ_1-oracle inequalities for the Lasso **60**

2.3.1. The Lasso for finite dictionaries 60

2.3.2. A selected Lasso estimator for infinite countable dictionaries 63

2.3.3. The Lasso for particular infinite uncountable dictionaries 65

2.4. Some rates of convergence for the Lasso **66**

2.4.1. Rates of convergence for the selected Lasso estimator 67

2.4.2. Rates of convergence for the Lasso 70

2.A. The Lasso as an ℓ_1-ball model selection procedure **71**

2.B. Proof of the ℓ_1-oracle inequalities **76**

2.B.1. Proof of Theorem 2.3.2 77

2.B.2. Proof of Theorem 2.3.3 79

2.B.3. Proof of Theorem 2.3.4 81

2.C. Proofs of the rates of convergence **83**

2.C.1. for the selected Lasso estimator in the case of orthonormal dictionaries . . 83

2.C.2. for the selected Lasso estimator in the general case 89

2.C.3. for the Lasso in the case of neural networks 89

2.D. Interpolation spaces **90**

ABSTRACT

These last years, while many efforts have been made to prove that the Lasso behaves like a variable selection procedure at the price of strong assumptions on the geometric structure of these variables, much less attention has been paid to the analysis of the performance of the Lasso as a regularization algorithm. Our first purpose here is to provide a result in this direction in the framework of Gaussian models with non-random regressors, by proving that the Lasso mimics the deterministic Lasso provided that the regularization parameter is properly chosen. This result requires no assumption at all, neither on the structure of the variables nor on the regression function.

Our second purpose is to propose a new estimator particularly adapted to deal with infinite countable dictionaries. This estimator is constructed as an ℓ_0-penalized estimator among a sequence of Lasso estimators associated to a dyadic sequence of growing truncated dictionaries. The selection procedure automatically chooses the best level of truncation of the dictionary to make the best tradeoff between approximation, ℓ_1-regularization and sparsity. We provide an oracle inequality satisfied by this selected Lasso estimator.

All the oracle inequalities presented in this chapter are obtained via the application of a single general theorem for model selection among a collection of non-linear models. The key idea that enables us to apply this general theorem is to see ℓ_1-regularization as a model selection procedure among ℓ_1-balls.

Finally, rates of convergence achieved by the Lasso and the selected Lasso estimators on a wide class of functions are derived from these oracle inequalities, showing that these estimators perform at least as well as greedy algorithms.

NOTA: The results presented in this chapter have been obtained in collaboration with Pascal Massart. A shorten version of this chapter has been published: P. Massart et C. Meynet (2011), The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, Vol.5, 669–687.

2.1 Introduction

We consider the problem of estimating a regression function s belonging to a Hilbert space $\mathbb{H}$ in a fairly general Gaussian framework which includes the fixed design regression or the white noise frameworks. Given a dictionary $\mathcal{D} = \{\phi_j\}_j$ of functions in $\mathbb{H}$, we aim at constructing an estimator $\hat{s} = \hat{\alpha}.\phi := \sum_j \hat{\alpha}_j \, \phi_j$ of s which enjoys both good statistical properties and computational performance even for large or infinite dictionaries.

For high-dimensional dictionaries, direct minimization of the empirical risk can lead to overfitting and we need to add a complexity penalty to avoid it. One could use an ℓ_0-penalty, i.e. penalize the number of non-zero coefficients $\hat{\alpha}_j$ of $\hat{s}$ so as to produce interpretable sparse models but there is no efficient algorithm to solve this non-convex minimization problem when the size of the dictionary becomes too large. On the contrary, ℓ_1-penalization leads to convex optimization and is thus computationally feasible even for high-dimensional data. Moreover, due to its geometric properties, ℓ_1-penalty tends to produce some coefficients that are exactly zero and hence often behaves like an ℓ_0-penalty. These are the main motivations for introducing ℓ_1-penalization rather than other penalizations.

In the linear regression framework, the idea of ℓ_1-penalization was first introduced by Tibshirani (1996) who considered the so-called Lasso estimator (Least Absolute Shrinkage and Selection Operator). Then, lots of studies on this estimator have been carried out, not only in the linear regression framework but also in the non-parametric regression setup with quadratic or more general loss functions (Efron et al., 2004; Zou et al., 2007; Bunea et al., 2007b, 2006; Zhang and Huang, 2008b; Bickel et al., 2009; Koltchinskii, 2009; van de Geer, 2008). For the fixed design Gaussian regression setting, if one observes n i.i.d. random couples $(x_1, Y_1), \ldots, (x_n, Y_n)$ such that

$$Y_i = s(x_i) + \sigma\xi_i, \quad i = 1, \ldots, n, \tag{2.1}$$

and consider a dictionary $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ of size p, the Lasso estimator is defined as the following ℓ_1-penalized least squares estimator

$$\hat{s}_p := \hat{s}(\lambda_p) = \underset{t \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left\{ \|Y - t\|^2 + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right\}, \tag{2.2}$$

where $\|Y - t\|^2 := \sum_{i=1}^n \left(Y_i - t(x_i)\right)^2 / n$ is the empirical risk of t, $\mathcal{L}_1(\mathcal{D}_p)$ is the linear span of $\mathcal{D}_p$ equipped with the ℓ_1-norm $\|t\|_{\mathcal{L}_1(\mathcal{D}_p)} := \inf\{\|\alpha\|_1 = \sum_{j=1}^p |\alpha_j|;\ t = \alpha.\phi = \sum_{j=1}^p \alpha_j\, \phi_j\}$ and $\lambda_p > 0$ is a regularization parameter.

Since ℓ_1-penalization is used as a convex relaxation of ℓ_0-penalization, many efforts have been made to prove that the Lasso behaves like a variable selection procedure by establishing sparsity oracle inequalities showing that the ℓ_1-solution mimics the ℓ_0-oracle (Bickel et al., 2009; Koltchinskii, 2009; van de Geer, 2008). Nonetheless, all these results require strong restrictive assumptions on the geometric structure of the variables. One typical example of such assumptions was presented in Section 1.2.4. We refer to Bühlmann and van de Geer (2009) for a detailed overview of all these restrictive assumptions.

In this chapter, we explore another approach by analyzing the performance of the Lasso as a regularization algorithm rather than a variable selection procedure. This is done by providing an ℓ_1-oracle inequality satisfied by this estimator (see Theorem 2.3.2). For the fixed design Gaussian

regression setting, this result says that if $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ with $\max_{j=1,\ldots,p} \|\phi_j\| \leq 1$, then there exists an absolute constant $C > 0$ such that for all $\lambda_p \geq 4\sigma\, n^{-1/2}(\sqrt{\ln p} + 1)$, the Lasso estimator defined by (2.2) satisfies

$$\mathbb{E}\left[\|s - \hat{s}_p\|^2 + \lambda_p \|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\right] \leq C \left[\inf_{t \in \mathcal{L}_1(\mathcal{D}_p)} \left\{\|s - t\|^2 + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} + \frac{\sigma \lambda_p}{\sqrt{n}}\right]. \tag{2.3}$$

This simply means that, provided that the regularization parameter λ_p is properly chosen, the Lasso estimator works almost as well as the deterministic Lasso. Unlike the sparsity oracle inequalities, this result requires no assumption neither on the target function s nor on the structure of the variables ϕ_j of the dictionary $\mathcal{D}_p$, except simple normalization that we can always assume by considering $\phi_j / \|\phi_j\|$ instead of ϕ_j.

We derive the ℓ_1-oracle inequality (2.3) from a model selection theorem for non-linear models, by interpreting ℓ_1-regularization as an ℓ_1-ball model selection procedure (see Section 2.A). Thanks to this approach, we can envisage to go one step further than the analysis of the Lasso for finite dictionaries and deal with infinite dictionaries.

In a second part of this chapter, we thus focus on infinite countable dictionaries. The idea is to order the dictionary $\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^\star} = \{\phi_1, \phi_2, \ldots\}$ according to the a priori knowledge we can have of the variables ϕ_j, and to consider the dyadic sequence of truncated dictionaries $\mathcal{D}_1 \subset \cdots \subset \mathcal{D}_p \subset \cdots \subset \mathcal{D}$ with $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ for $p \in \{2^J, J \in \mathbb{N}\}$. Given this sequence $(\mathcal{D}_p)_p$, we introduce the associated sequence of Lasso estimators $(\hat{s}_p)_p$ defined by (2.2), and choose $\hat{s}_{\hat{p}}$ as an ℓ_0-penalized estimator among this sequence by penalizing the size of the truncated dictionaries $\mathcal{D}_p$. This selected Lasso estimator $\hat{s}_{\hat{p}}$ is thus based on an algorithm choosing automatically the level of truncation of the dictionary $\mathcal{D}$ making the best tradeoff between approximation, ℓ_1-regularization and sparsity. Of course, although introduced for infinite dictionaries, this estimator remains well-defined for finite dictionaries and it may be profitable to use it rather than the classical Lasso for such dictionaries. From a theoretical point of view, we establish an oracle inequality satisfied by this selected Lasso estimator and we provide rates of convergence of this estimator for a wide range of function classes described by interpolation spaces $\mathcal{B}_{q,r}$. In the orthonormal case, we check that these rates of convergence are optimal by establishing a lower bound of the minimax risk. Our results prove that the selected Lasso estimator $\hat{s}_{\hat{p}}$ performs as well as the greedy algorithms described by Barron et al. (2008).

In this chapter, we also provide a few theoretical results on the performance of the Lasso for particular infinite uncountable dictionaries such as those used for neural networks. Although the Lasso solutions can not be computed in practice for such dictionaries, our purpose is just to point out that the Lasso theoretically performs as well as the greedy algorithms in Barron et al. (2008), by establishing rates of convergence based on an ℓ_1-oracle inequality similar to (2.3) satisfied by the Lasso for such dictionaries.

The chapter is organized as follows. The notations and the framework are introduced in Section 2.2. In Section 2.3, we establish three ℓ_1-oracle inequalities: one for the Lasso with finite dictionaries, one for the selected Lasso estimator with infinite countable dictionaries, and one for the Lasso with particular infinite uncountable dictionaries such as those used for neural networks. In Section 2.4, we derive from these oracle inequalities rates of convergence for the selected Lasso estimator and for the Lasso for a variety of function classes. In Appendix 2.A, we explain f the key idea that enables us to derive our three oracle inequalities from a single model selection theorem. Then, we state and prove this model selection theorem, which is a particular case of a general model selection theorem established by Massart (2007). The proofs of the three oracle inequalities and of the rates of convergence are respectively postponed until Appendix 2.B and Appendix 2.C.

2.2 Models and notations

2.2.1 General framework and statistical problem

Let $\mathbb{H}$ be a separable Hilbert space equipped with a scalar product $\langle .,. \rangle$ and its associated norm $\|.\|$. The statistical problem we consider is to estimate an unknown target function s in $\mathbb{H}$ when observing a process $(Y(t))_{t\in\mathbb{H}}$ defined by

$$Y(t) = \langle s, t\rangle + \varepsilon W(t), \quad t \in \mathbb{H}, \tag{2.4}$$

where $\varepsilon > 0$ is a fixed parameter and $(W(t))_{t\in\mathbb{H}}$ is an isonormal process, that is to say a centered Gaussian process with covariance given by $\mathbb{E}[W(u)W(t)] = \langle u, t\rangle$ for all $u, t \in \mathbb{H}$.

This framework is convenient to cover both finite-dimensional models and the infinite-dimensional white noise model as described in the following examples.

Example 2.2.1. [Fixed design Gaussian regression model] Let $\mathcal{X}$ be a measurable space. One observes n i.i.d. random couples $(x_1, Y_1), \ldots, (x_n, Y_n)$ of $\mathcal{X} \times \mathbb{R}$ such that

$$Y_i = s(x_i) + \sigma\xi_i, \quad i = 1, \ldots, n, \tag{2.5}$$

where the covariates $x_1, \ldots, x_n$ are deterministic elements of $\mathcal{X}$, the errors ξ_i are i.i.d. $\mathcal{N}(0, 1)$, $\sigma > 0$ and $s : \mathcal{X} \mapsto \mathbb{R}$ is the unknown regression function to be estimated. If one considers $\mathbb{H} = \mathbb{R}^n$ equipped with the scalar product $\langle u, v\rangle = \sum_{i=1}^n u_i\, v_i/n$, defines $y = (Y_1, \ldots, Y_n)$, $\xi = (\xi_1, \ldots, \xi_n)$ and denotes $t = (t(x_1), \ldots, t(x_n))$ for every $t : \mathcal{X} \mapsto \mathbb{R}$, then $W(t) := \sqrt{n}\,\langle \xi, t\rangle$ defines an isonormal Gaussian process on $\mathbb{H}$ and $Y(t) := \langle y, t\rangle$ satisfies (2.4) with $\varepsilon = \sigma/\sqrt{n}$. In this case,

$$\|t\| = \sqrt{\frac{1}{n}\sum_{i=1}^n t^2(x_i)}\,. \tag{2.6}$$

Example 2.2.2. [The white noise framework] For $x \in [0,1]$, one observes $\zeta(x)$ given by the stochastic differential equation

$$d\zeta(x) = s(x)\,dx + \varepsilon\,dB(x) \text{ with } \zeta(0) = 0,$$

where B is a standard Brownian motion, s is a square-integrable function and $\varepsilon > 0$. Define $W(t) = \int_0^1 t(x)\,dB(x)$ for every $t \in \mathbb{L}_2([0,1])$. Then, W is an isonormal process on $\mathbb{H} = \mathbb{L}_2([0,1])$, and $Y(t) = \int_0^1 t(x)\,d\zeta(x)$ obeys to (2.4) if $\mathbb{H}$ is equipped with its usual scalar product $\langle s,t\rangle = \int_0^1 s(x)t(x)\,dx$. Typically, s is a signal and $d\zeta(x)$ represents the noisy signal received at time x. This framework easily extends to a d-dimensional setting if one considers some multivariate Brownian sheet B on $[0,1]^d$ and takes $\mathbb{H} = \mathbb{L}_2([0,1]^d)$.

2.2.2 The Lasso: an ℓ_1-penalized least squares estimator

To solve the general statistical problem (2.4), we introduce a dictionary $\mathcal{D}$, i.e. a given finite or infinite set of functions $\phi_j \in \mathbb{H}$ that arise as candidate basis functions for estimating the target function s, and consider estimators $\hat{s} = \hat{\alpha}.\phi := \sum_{j,\,\phi_j\in\mathcal{D}} \hat{\alpha}_j\,\phi_j$ in the linear span of $\mathcal{D}$. All the matter is to choose a "good" linear combination in the following meaning. It makes sense to aim at constructing an estimator as the best approximating point of s by minimizing $\|s-t\|$ or, equivalently, $-2\langle s,t\rangle + \|t\|^2$. However s is unknown, so one may instead minimize the empirical least squares criterion

$$\gamma(t) := -2Y(t) + \|t\|^2. \tag{2.7}$$

But, for high-dimensional data, direct minimization of the empirical least squares criterion can lead to overfitting. To avoid it, one can rather consider a penalized risk minimization problem and estimate s by

$$\hat{s} \in \arg\min_t \left\{\gamma(t) + \mathrm{pen}(t)\right\}, \tag{2.8}$$

where $\mathrm{pen}(t)$ is a positive penalty to be chosen. Here, we focus on ℓ_1-penalization defined by $\mathrm{pen}(t) = \lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}$ where

$$\|t\|_{\mathcal{L}_1(\mathcal{D})} := \inf\left\{\|\alpha\|_1 = \sum_{j,\,\phi_j\in\mathcal{D}} |\alpha_j| \quad \text{such that } t = \alpha.\phi\right\}$$

and $\lambda > 0$ is the regularization parameter to be tuned. We estimate s by the associated estimator – the Lasso – defined by

$$\hat{s}(\lambda) = \arg\min_{t\in\mathcal{L}_1(\mathcal{D})} \left\{\gamma(t) + \lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\},$$

where $\mathcal{L}_1(\mathcal{D})$ denotes the set of functions t in the linear span of $\mathcal{D}$ with finite ℓ_1-norm $\|t\|_{\mathcal{L}_1(\mathcal{D})}$.

2.3 Some ℓ_1-oracle inequalities for the Lasso

While many efforts have been made to prove that the Lasso behaves like a variable selection procedure at the price of strong (though unavoidable) assumptions on the geometric structure of the dictionary (Bickel et al., 2009; Bühlmann and van de Geer, 2009), much less attention has been paid to the analysis of the performance of the Lasso as a regularization algorithm. The analysis we propose below goes in this direction.

2.3.1 The Lasso for finite dictionaries

2.3.1.1 Definition of the Lasso estimator

Consider the generalized linear Gaussian model and the statistical problem (2.4). Assume that $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ is a finite dictionary of size p. Then, any function t in the linear span of $\mathcal{D}_p$ has finite ℓ_1-norm

$$\|t\|_{\mathcal{L}_1(\mathcal{D}_p)} := \inf \left\{ \|\alpha\|_1 = \sum_{j=1}^{p} |\alpha_j| \; ; \; \alpha \in \mathbb{R}^p \text{ such that } t = \alpha.\phi \right\} \tag{2.9}$$

and thus belongs to $\mathcal{L}_1(\mathcal{D}_p)$. We estimate s by the Lasso estimator $\hat{s}_p$ defined by

$$\hat{s}_p := \hat{s}(\lambda_p) = \underset{t \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left\{ \gamma(t) + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right\}, \tag{2.10}$$

where $\lambda_p > 0$ is a regularization parameter and $\gamma(t)$ is defined by (2.7).

Example 2.3.1. Let us specify Definition (2.10) for the classical fixed design Gaussian regression setting presented in Example 2.2.1. Define $y = (Y_1, \ldots, Y_n)$. Then,

$$\gamma(t) = -2Y(t) + \|t\|^2 = -2\langle y, t\rangle + \|t\|^2 = \|y - t\|^2 - \|y\|^2.$$

So, we deduce from (2.10) that the Lasso satisfies

$$\hat{s}_p = \underset{t \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left\{ \|y - t\|^2 + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right\}. \tag{2.11}$$

For all $t \in \mathcal{L}_1(\mathcal{D}_p)$, set $A_t := \{\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p : \; t = \alpha.\phi = \sum_{j=1}^{p} \alpha_j \, \phi_j\}$. We get from (2.9)

that

$$\begin{aligned}\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|y-t\|^2+\lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} &= \inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|y-t\|^2+\lambda_p\inf_{\alpha\in A_t}\|\alpha\|_1\right\}\\ &= \inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\inf_{\alpha\in A_t}\left\{\|y-t\|^2+\lambda_p\|\alpha\|_1\right\}\\ &= \inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\inf_{\alpha\in A_t}\left\{\|y-\alpha.\phi\|^2+\lambda_p\|\alpha\|_1\right\}\\ &= \inf_{\alpha\in\mathbb{R}^p}\left\{\|y-\alpha.\phi\|^2+\lambda_p\|\alpha\|_1\right\}.\end{aligned}$$

Therefore, we get from (2.11) that $\hat{s}_p=\hat{\alpha}_p.\phi$ where $\hat{\alpha}_p=\arg\min_{\alpha\in\mathbb{R}^p}\{\|y-\alpha.\phi\|^2+\lambda_p\|\alpha\|_1\}$. This does correspond to the Lasso definition found in the literature for the fixed design Gaussian regression setting with finite dictionaries of size p (Bickel et al., 2009).

2.3.1.2 An ℓ_1-oracle inequality

Here, we provide an ℓ_1-oracle inequality satisfied by the Lasso for finite dictionaries. It highlights the fact that, provided that the regularization parameter λ_p is properly chosen, the Lasso, which is the solution of the ℓ_1-penalized empirical risk minimization problem, behaves nearly as well as the deterministic Lasso, that is to say the solution of the ℓ_1-penalized true risk minimization problem.

Theorem 2.3.2. *Assume that* $\max_{j=1,\dots,p}\|\phi_j\|\leq 1$ *and that*

$$\lambda_p\geq 4\varepsilon\left(\sqrt{\ln p}+1\right). \tag{2.12}$$

Let $\hat{s}_p$ *be the Lasso estimator defined by* (2.10).
Then, there exists an absolute positive constant C *such that, for all* $z>0$*, with probability larger than* $1-3.4\,\mathrm{e}^{-z}$*,*

$$\|s-\hat{s}_p\|^2+\lambda_p\|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\leq C\left[\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2+\lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}+\lambda_p\,\varepsilon(1+z)\right]. \tag{2.13}$$

Integrating (2.13) *with respect to* z *leads to the following* ℓ_1*-oracle inequality in expectation,*

$$\mathbb{E}\left[\|s-\hat{s}_p\|^2+\lambda_p\|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\right]\leq C\left[\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2+\lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}+\lambda_p\varepsilon\right]. \tag{2.14}$$

Proof. Page 77. □

These last years, the Lasso has essentially been developed as an approach to sparse recovery based on convex optimization and thus the main focus on this estimator has been on the establishment of ℓ_0-oracle inequalities so as to study its performance as a variable selection procedure (see recall in Section 1.2.4). In contrast, Theorem 2.3.2 does not take into account sparsity and rather provides

information about the performance of the Lasso as an ℓ_1-regularization algorithm. Note that the ℓ_1-oracle inequalities of Theorem 2.3.2 are valid for regularization parameters of the same order (2.12) as the regularization parameters allowing ℓ_0-oracle inequalities (Bickel et al., 2009). Contrary to the ℓ_0-results that require some restrictive assumptions on the dictionary and that are interesting only if the target function can be well approximated by a sparse function in the linear span of the dictionary (see Section 1.2.4), the ℓ_1-oracle inequalities (2.13) and (2.14) are established with no assumption neither on the target function nor on the structure of the variables ϕ_j of the dictionary $\mathcal{D}_p$, except simple normalization that we can always assume by considering $\phi_j/\|\phi_j\|$ instead of ϕ_j. Thus, we are guaranteed that the Lasso always achieves high performance as regards ℓ_1-regularization.

Let us mention that ℓ_1-oracle inequalities similar to (2.13) or (2.14) have been provided by a few authors such as Huang et al. (2008), Rigollet and Tsybakov (2011), Bartlett et al. (2012). Yet, all these results present dissimilarities with Theorem 2.3.2. Let us have a look at these differences.

Rigollet and Tsybakov (2011) propose an oracle inequality for the Lasso similar to (2.13) which is valid under the same assumption as for Theorem 2.3.2, i.e. simple normalization of the variables ϕ_j. Yet, their bound in probability can not be integrated to get an bound in expectation as the one we propose at (2.14). Indeed, the constant measuring the level of confidence of their risk bound appears inside the infimum term as a multiplicative factor of the ℓ_1-norm whereas the constant z measuring the level of confidence of our risk bound (2.13) appears as an additive constant outside the infimum term so that the bound in probability (2.13) can easily be integrated with respect to z, which leads to the bound in expectation (2.14). Besides, the lower bound of the regularization parameter λ_p proposed by Rigollet and Tsybakov (2011) depends on the level of confidence z, so their choice of the Lasso estimator $\hat{s}_p = \hat{s}(\lambda_p)$ also depends on this level of confidence. On the contrary, our lower bound (2.12) does not depend on z, so the result (2.13) is satisfied with high probability by an estimator $\hat{s}_p = \hat{s}(\lambda_p)$ independent of the level of confidence of this probability.

Bartlett et al. (2012) also obtain an ℓ_1-oracle inequality for the Lasso in linear regression, but they consider random design $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ rather than fixed design. Therefore, their analysis requires a uniform concentration phenomenon that forces them to make strong assumptions, namely that both X and Y are bounded almost surely by a constant independent of n. Moreover, they get a lower bound on the regularization parameter with an extra $\ln(n)$ factor compared to (2.12).

Huang et al. (2008) also consider random design. Rather than assuming that Y is bounded as it is done by Bartlett et al. (2012), they suppose that the errors satisfy some Bernstein's moment condition. Nevertheless, they assume that the target function is bounded by a constant, and their risk bound is not satisfied by the Lasso itself but only by a truncated Lasso estimator.

Let us point a weakness of our result: the ℓ_1-oracle inequalities (2.13) and (2.14) are proved with

undetermined constant C whereas the ℓ_1-oracle inequalities in both Rigollet and Tsybakov (2011) and Bartlett et al. (2012) are sharp, i.e. with $C = 1$.

In Section 2.A, we describe the key observation that has enabled us to establish Theorem 2.3.2. In a nutshell, the basic idea is to view the Lasso as the solution of a penalized least squares model selection procedure over a countable model collection consisting of ℓ_1-balls. Inequalities (2.13) and (2.14) are then deduced from a model selection theorem (Theorem 2.A.1, Section 2.A). Thanks to this approach, we can go one step further than the analysis of the Lasso for finite dictionaries: as we shall see now, we can also deal with infinite dictionaries.

2.3.2 A selected Lasso estimator for infinite countable dictionaries

In many applications such as microarray data analysis or signal reconstruction, we are now faced with situations where the number of variables of the dictionary is always increasing and can even be infinite. So, it is desirable to find competitive estimators for such infinite dimensional problems. Yet, except in rare situations where the variables have a specific structure (see Remark 2.3.3 on neural networks), it is difficult to extend the results established for the Lasso for finite dictionaries to infinite dictionaries. Indeed, for infinite dictionaries, there is no longer finite size p. As a consequence, one can no longer calibrate the regularization parameter as it is done in (2.12). Here, we propose a procedure to calibrate the regularization parameter by providing an optimal size $\hat{p}$ in a sense described below.

To deal with an infinite countable dictionary $\mathcal{D}$, one may order the variables of the dictionary, write the dictionary $\mathcal{D} = \{\phi_j\}_{j\in\mathbb{N}^\star} = \{\phi_1, \phi_2, \ldots\}$ according to this order, then truncate $\mathcal{D}$ at a given level p to get a finite subdictionary $\{\phi_1, \ldots, \phi_p\}$ and finally estimate s by the Lasso estimator $\hat{s}_p$ over this subdictionary. This procedure implies two difficulties. First, one has to put an order on the variables of the dictionary. Second, all the matter is to decide at which level one should truncate the dictionary to make the best tradeoff between approximation and complexity. Here, our purpose is to resolve this last dilemma by proposing a selected Lasso estimator based on an algorithm choosing automatically the best level of truncation of the dictionary once the variables have been ordered. Of course, the algorithm, and thus the estimation of s, depend on the preliminary order put on the variables. Ordering the variables can be more or less difficult according to the problem under consideration. For some applications, such as decomposition in wavelet dictionaries, the variables may be naturally ordered.

For orthonormal dictionaries, the Lasso estimators are soft-thresholding estimators with a fixed threshold. Then, the selected Lasso estimator is a soft-thresholding estimator with an adaptive threshold automatically chosen by the algorithm constructing this estimator. So, our procedure provides a new contribution to the crucial choice of the threshold when working with soft-thresholding estimators.

2.3.2.1 Definition of the selected Lasso estimator

We still consider the generalized linear Gaussian model and the statistical problem (2.4) introduced in Section 2.2. To solve this problem, we use a dictionary $\mathcal{D} = \{\phi_j\}_j$ and seek for an estimator $\hat{s} = \hat{\alpha}.\phi = \sum_{j,\, \phi_j \in \mathcal{D}} \hat{\alpha_j}\, \phi_j$ solution of the penalized risk minimization problem,

$$\hat{s} \in \underset{t \in \mathcal{L}_1(\mathcal{D})}{\arg\min} \left\{ \gamma(t) + \mathrm{pen}(t) \right\}, \tag{2.15}$$

where $\mathrm{pen}(t)$ is a suitable positive penalty. Here, we assume that the dictionary is infinite countable and that it is ordered:

$$\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^\star} = \{\phi_1, \phi_2, \ldots\}.$$

Given this order, we consider the sequence of truncated dictionaries $(\mathcal{D}_p)_{p \in \mathbb{N}^\star}$ where

$$\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\} \tag{2.16}$$

is the subdictionary of $\mathcal{D}$ truncated at level p, and the associated sequence of Lasso estimators defined in Section 2.3.1.1,

$$\hat{s}_p = \underset{t \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left\{ \gamma(t) + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right\}, \tag{2.17}$$

where $(\lambda_p)_{p \in \mathbb{N}^\star}$ is a sequence of regularization parameters specified below. Now, we choose a final estimator as an ℓ_0-penalized estimator among a subsequence of the Lasso estimators $(\hat{s}_p)_{p \in \mathbb{N}^\star}$. Specifically, denote by Λ the set of dyadic integers,

$$\Lambda = \left\{ 2^J ; J \in \mathbb{N} \right\}, \tag{2.18}$$

and define

$$\hat{p} = \underset{p \in \Lambda}{\arg\min} \left\{ \gamma(\hat{s}_p) + \lambda_p \|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} + \mathrm{pen}(p) \right\} \tag{2.19}$$

$$= \underset{p \in \Lambda}{\arg\min} \left\{ \underset{t \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left\{ \gamma(t) + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right\} + \mathrm{pen}(p) \right\}, \tag{2.20}$$

where $\mathrm{pen}(p)$ is a penalty to be chosen to penalize the size p of the truncated dictionary $\mathcal{D}_p$ for all $p \in \Lambda$. Then, we choose $\hat{s}_{\hat{p}}$ as final estimator. From (2.20) and the fact that $\mathcal{L}_1(\mathcal{D}) = \cup_{p \in \Lambda} \mathcal{L}_1(\mathcal{D}_p)$, we see that this selected Lasso estimator $\hat{s}_{\hat{p}}$ is a penalized least squares estimator solution of (2.15) where, for any $p \in \Lambda$ and $t \in \mathcal{L}_1(\mathcal{D}_p)$, $\mathrm{pen}(t) = \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} + \mathrm{pen}(p)$ is a combination of both ℓ_1-regularization and complexity penalization. We also see from (2.19) that the algorithm automatically chooses the rank $\hat{p}$ so that $\hat{s}_{\hat{p}}$ makes the best tradeoff between approximation, ℓ_1-regularization and sparsity.

Remark 1. From a theoretical point of view, one could define $\hat{s}_{\hat{p}}$ as an ℓ_0-penalized estimator among the whole sequence of Lasso estimators $(\hat{s}_p)_{p\in\mathbb{N}^\star}$ (or more generally among any subsequence of $(\hat{s}_p)_{p\in\mathbb{N}^\star}$) instead of $(\hat{s}_p)_{p\in\Lambda}$. Nonetheless, to compute $\hat{s}_{\hat{p}}$ efficiently, it is interesting to limit the number of computations of the sequence of Lasso estimators $\hat{s}_p$, especially if we choose a complexity penalty $\mathrm{pen}(p)$ that does not grow too fast with p. In the sequel, we shall consider a penalty $\mathrm{pen}(p)$ proportional to $\ln(p)$ (see Theorem 2.3.3). So, a dyadic truncation $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\} = \{\phi_1, \ldots, \phi_{2^J}\}$ of the dictionary $\mathcal{D}$ enables to get a complexity penalty $\mathrm{pen}(p) \propto \ln p = J \ln 2$ growing linearly at each step J of the algorithm, thus leading to a more efficient algorithm.

Although our primary motivation for introducing the selected Lasso estimator is to adapt the Lasso to infinite dictionaries, note that the selected Lasso estimator remains well-defined for finite dictionaries. For a dictionary of size p_0, one can estimate s by the selected Lasso estimator $\hat{s}_{\hat{p}}$ rather than by the Lasso estimator $\hat{s}_{p_0}$. The definition of $\hat{s}_{\hat{p}}$ guarantees that $\hat{s}_{\hat{p}}$ makes a better tradeoff between approximation, ℓ_1-regularization and sparsity than $\hat{s}_{p_0}$. Besides, $\hat{s}_{\hat{p}}$ is sparser than $\hat{s}_{p_0}$ since $\hat{p} \leq p_0$. In particular, $\hat{s}_{\hat{p}}$ and $\hat{s}_{p_0}$ coincide when $\hat{p} = p_0$.

2.3.2.2 An oracle inequality for the selected Lasso estimator

By applying the same model selection theorem (Theorem 2.A.1) as for Theorem 2.3.2, we can provide a risk bound satisfied by the estimator $\hat{s}_{\hat{p}}$ with properly chosen penalties λ_p and $\mathrm{pen}(p)$ for all $p \in \Lambda$. The sequence of ℓ_1-regularization parameters $(\lambda_p)_{p\in\Lambda}$ is simply chosen from the lower bound (2.12), while a convenient choice for the ℓ_0-penalty is $\mathrm{pen}(p) \propto \ln p$.

Theorem 2.3.3. *Assume that* $\sup_{j\in\mathbb{N}^\star} \|\phi_j\| \leq 1$. *Set for all* $p \in \Lambda$,

$$\lambda_p = c_1 \varepsilon \left(\sqrt{\ln p} + 1\right), \qquad \mathrm{pen}(p) = c_2 \varepsilon^2 \ln p, \tag{2.21}$$

where $c_1 \geq 4$ *and* $c_2 > c_1/\sqrt{\ln 2}$. *Let* $\hat{s}_{\hat{p}}$ *be the selected Lasso estimator defined by* (2.20). *Then, there exists an absolute constant* $C > 0$ *such that*

$$\begin{aligned} &\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p})\right] \\ &\leq C \left[\inf_{p\in\Lambda} \left\{ \inf_{t\in\mathcal{L}_1(\mathcal{D}_p)} \left\{\|s - t\|^2 + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} + \mathrm{pen}(p) \right\} + \varepsilon^2 \right]. \end{aligned} \tag{2.22}$$

Proof. Page 79. □

2.3.3 The Lasso for particular infinite uncountable dictionaries

As explained at the beginning of Section 2.3.2, it is generally difficult to establish theoretical results on the performance of the Lasso for infinite dictionaries. Yet, it can be easier to prove such results

for some particular infinite dictionaries whose structure is nice enough. For example, it is the case for neural networks in the fixed design Gaussian regression setting. A neural network is a real-valued function defined on $\mathbb{R}^d$ belonging to the linear span of the dictionary $\mathcal{D} = \{\phi_{a,b}\,;\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$ where

$$\phi_{a,b} : \mathbb{R}^d \mapsto \mathbb{R}, \quad x \mapsto \mathbb{1}_{\{\langle a,x\rangle + b > 0\}} \,. \tag{2.23}$$

Given a training sequence $(x_1, Y_1), \ldots, (x_n, Y_n)$ such that $Y_i = s(x_i) + \sigma\xi_i$, the Lasso estimator over the set of neural network estimators is

$$\hat{s} := \hat{s}(\lambda) = \underset{t \in \mathcal{L}_1(\mathcal{D})}{\arg\min} \left\{ \|Y - t\|^2 + \lambda \|t\|_{\mathcal{L}_1(\mathcal{D})} \right\}, \tag{2.24}$$

where $\lambda > 0$ is a regularization parameter, $\|Y - t\|^2 = \sum_{i=1}^n (Y_i - t(x_i))^2 / n$ is the empirical risk of t and $\mathcal{L}_1(\mathcal{D})$ is the linear span of $\mathcal{D}$ equipped with the ℓ_1-norm

$$\|t\|_{\mathcal{L}_1(\mathcal{D})} = \inf \left\{ \|\alpha\|_1 = \sum_{a \in \mathbb{R}^d, b \in \mathbb{R}} |\alpha_{a,b}| \,;\ t = \alpha.\phi = \sum_{a \in \mathbb{R}^d, b \in \mathbb{R}} \alpha_{a,b}\, \phi_{a,b} \right\}.$$

Although the dictionary $\mathcal{D}$ is infinite uncountable, we are able to establish an ℓ_1-oracle inequality satisfied by the Lasso, which is similar to the ℓ_1-oracle inequality provided in Theorem 2.3.2 for finite dictionaries. This is possible thanks to the particular structure of the dictionary $\mathcal{D}$ which is only composed of functions derived from the Heaviside function. This property enables us to achieve theoretical results without truncating the whole dictionary into finite subdictionaries, contrary to Section 2.3.2 where arbitrary infinite countable dictionaries were considered. The following ℓ_1-oracle inequality is again a direct application of the model selection Theorem 2.A.1 (Section 2.A).

Theorem 2.3.4. *Assume that*

$$\lambda \geq \kappa\sigma\sqrt{d/n} \tag{2.25}$$

for some absolute constant $\kappa > 0$ large enough. Let $\hat{s}$ be the Lasso estimator defined by (2.24). *Then, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[\|s - \hat{s}\|^2 + \lambda\|\hat{s}\|_{\mathcal{L}_1(\mathcal{D})}\right] \leq C \left[\inf_{t \in \mathcal{L}_1(\mathcal{D})} \left\{ \|s - t\|^2 + \lambda \|t\|_{\mathcal{L}_1(\mathcal{D})} \right\} + \lambda \frac{\sigma}{\sqrt{n}} \right].$$

Proof. Page 81. □

2.4 Some rates of convergence for the Lasso

We now establish rates of convergence for the selected Lasso estimator and the Lasso estimator. They are derived from the oracle inequalities of Theorem 2.3.3 and Theorem 2.3.4.

2.4.1 Rates of convergence for the selected Lasso estimator

Here, we provide rates of convergence for the selected Lasso estimator. First, we restrict to orthonormal dictionaries for a target function s in the intersection between a weak $\mathcal{L}_q$ space and a Besov space (see definitions below). Moreover, we establish lower bounds of the minimax risk to check that the rates of convergence achieved by this estimator are optimal. Then, we extend our upper bounds to the non-orthonormal case.

We keep the same framework and notations as in Section 2.3.2. In particular, we still consider a Hilbert space $\mathbb{H}$ and an infinite countable dictionary $\mathcal{D} = \{\phi_j\}_{j\in\mathbb{N}^\star}$ which is a basis of $\mathbb{H}$.

2.4.1.1 Orthonormal case

Here, we assume that $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$.

Definition of the spaces

Assume that s belongs to $w\mathcal{L}_q(R)$ for some $1 < q < 2$ and $R > 0$, that is to say $s = \sum_{j=1}^{\infty} \alpha_j \, \phi_j$ with coefficients α_j in the weak ℓ_q-balls of radius R:

$$\sup_{\eta>0} \left(\eta^q \sum_{j=1}^{\infty} \mathbb{1}_{\{|\alpha_j|>\eta\}} \right) \leq R^q. \tag{2.26}$$

So as to control the size of the high-level components of s in the orthonormal basis $\mathcal{D}$, also assume that s belongs to the Besov space $\mathcal{B}^r_{2,\infty}(R)$ with radius R:

$$\sup_{J\in\mathbb{N}^\star} \left(J^{2r} \sum_{j=J}^{\infty} \alpha_j^2 \right) \leq R^2. \tag{2.27}$$

Upper bound of the quadratic risk

Proposition 2.4.1. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of the Hilbert space $\mathbb{H}$. Let $1 < q < 2$, $r > 0$, $R > 0$ such that $R\varepsilon^{-1} \geq \mathrm{e}$, and assume that $s \in w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. Consider the selected Lasso estimator $\hat{s}_{\hat{p}}$ defined by* (2.20) *with parameters λ_p and $\mathrm{pen}(p)$ given by* (2.21)*. Then, there exists $C_{q,r} > 0$ depending only on q and r such that the quadratic risk of $\hat{s}_{\hat{p}}$ satisfies*

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2\right] \leq C_{q,r}\, R^q \left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}. \tag{2.28}$$

Proof. Page 83. □

Remark 2. The assumption $R\varepsilon^{-1} \geq \mathrm{e}$ of Proposition 2.4.1 is not restrictive since it only means that we consider non-degenerate situations where the signal to noise ratio is large enough, which is the only

interesting case to use the selected Lasso estimator. Indeed, if $R\varepsilon^{-1}$ is too small, then the estimator equal to zero will always be better than any other non-zero estimators, in particular the selected Lasso estimator.

In the orthonormal case, the selected Lasso estimator is a soft-thresholding estimator with an adaptive threshold. So, the bound (2.28) is to be compared with the rates of convergence achieved by the soft-thresholding estimators with a fixed threshold (i.e. the classical Lasso estimators) when the target function belongs to $w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. From Rivoirard (2006, Theorem 1), the rates achieved by the soft-thresholding estimators with a fixed threshold determined by the level of truncation of the dictionary chosen by the statistician strongly depend on the parameter of smoothness r and are valid only for values of r large enough compared to the level of truncation. On the contrary, Proposition 2.4.1 shows that the rates achieved by the selected Lasso estimator are valid whatever the value of $r > 0$ and that this smoothness parameter has little effect on the rates since it only appears through the multiplicative factor $C_{q,r}$. Proposition 2.4.1 thus highlights the major advantage of the selected Lasso estimator over the classical Lasso estimators which is its adaptability to the unknown parameters of smoothness q and r of the target function. This adaptability comes from the fact that the selected Lasso estimator is constructed from an algorithm choosing an adaptive level of truncation of the dictionary.

Lower bound of the minimax risk

We now establish a lower bound of the minimax risk over the balls $w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ to prove that the rates of convergence (2.28) are optimal. We even establish a stronger result by providing the lower bound of the minimax risk over the smaller balls $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$, where we denote by $\mathcal{L}_q(R)$ the set of functions whose coefficients in the orthonormal basis $\mathcal{D} = \{\phi_j\}_{j\in\mathbb{N}^\star}$ are in the ℓ_q-ball of radius R, that is to say functions $\sum_{j=1}^\infty \alpha_j\,\phi_j$ such that $\sum_{j=1}^\infty |\alpha_j|^q \leq R^q$.

Proposition 2.4.2. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$. Let $1 < q < 2$, $0 < r < 1/q - 1/2$ and $R > 0$ such that $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, \varsigma^2)$ where*

$$\varsigma := \frac{1}{r} - q\left(1 + \frac{1}{2r}\right) > 0. \tag{2.29}$$

Then, there exists an absolute constant $\kappa > 0$ such that the minimax risk over $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ satisfies

$$\inf_{\tilde{s}} \sup_{s\in\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)} \mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq \kappa\,\varsigma^{1-\frac{q}{2}}\,R^q\left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}, \tag{2.30}$$

where the infimum is taken over all possible estimators $\tilde{s}$.

Proof. Page 87. □

Remark 3. The constraint $r < 1/q - 1/2$ of Proposition 2.4.2 is necessary to work on the intersection between an $\mathcal{L}_q$-ball and a Besov ball. Indeed, assume that $r > 1/q - 1/2$. For all $R > 0$, put $R' = (1 - 2^{r\varsigma})^{1/q} R$ where ς is defined by (2.29). Then, it is easy to check that $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R)$. Thus, $\mathcal{B}^r_{2,\infty}(R') = \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R')$. Moreover, $R' < R$, so $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{B}^r_{2,\infty}(R)$ and $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. Consequently, $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset \mathcal{B}^r_{2,\infty}(R)$: the intersection $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ is no longer a real intersection between an $\mathcal{L}_q$-ball and a Besov ball but rather a Besov ball itself.

The upper bound (2.28) and the lower bound (2.30) match up to a constant. This proves that the selected Lasso estimator is simultaneously approximately minimax over $w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ for suitable signal to noise ratio $R\varepsilon^{-1}$ in the orthonormal case.

2.4.1.2 General case

Here, we no longer assume that $\mathcal{D}$ is orthonormal. We say a few words on the rates of convergence of the selected Lasso estimator in the non-orthonormal case. We extend the upper bound (2.28) of the quadratic risk of this estimator when assuming that the target function belongs to some real interpolation spaces that are extensions of the spaces $w\mathcal{L}_q \cap \mathcal{B}^r_{2,\infty}$ considered in the orthonormal case.

Definition of the interpolation spaces

We introduce a whole range of interpolation spaces $\mathcal{B}_{q,r}$ that are intermediate spaces between subsets of $\mathcal{L}_1(\mathcal{D})$ and the Hilbert space $\mathbb{H}$.

Definition 2.4.3. [Spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$] Let $R > 0$, $r > 0$, $1 < q < 2$ and $\nu = 1/q - 1/2$.
We say that $u \in \mathbb{H}$ belongs to $\mathcal{L}_{1,r}$ if there exists $C > 0$ such that for all $p \in \mathbb{N}^\star$, there exists $u_p \in \mathcal{L}_1(\mathcal{D}_p)$ such that

$$\|u_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C$$

and

$$\|u - u_p\| \leq C p^{-r}. \tag{2.31}$$

The smallest C such that this holds defines a norm $\|u\|_{\mathcal{L}_{1,r}}$ on the space $\mathcal{L}_{1,r}$.

We say that u belongs to $\mathcal{B}_{q,r}(R)$ if, for all $\delta > 0$,

$$\inf_{t \in \mathcal{L}_{1,r}} \left\{ \|u - t\| + \delta \|t\|_{\mathcal{L}_{1,r}} \right\} \leq R\, \delta^{2\nu}. \tag{2.32}$$

We say that $u \in \mathcal{B}_{q,r}$ if there exists $R > 0$ such that $u \in \mathcal{B}_{q,r}(R)$. In this case, the smallest R such that $u \in \mathcal{B}_{q,r}(R)$ defines a norm on the space $\mathcal{B}_{q,r}$ and is denoted by $\|u\|_{\mathcal{B}_{q,r}}$.

Remark 4. The abstract interpolation spaces $\mathcal{B}_{q,r}$ are in fact natural extensions of the spaces $w\mathcal{L}_q \cap \mathcal{B}^r_{2,\infty}$ for non-orthonormal dictionaries. Indeed, if $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$, then, for all $1 < q < 2$ and $r > 0$, there exists $C_{q,r} > 0$ depending only on q and r such that, for all $R > 0$,

$$w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset \mathcal{B}_{q,r}(C_{q,r}\, R). \tag{2.33}$$

Proof. Page 90. □

Upper bound of the quadratic risk

Proposition 2.4.4. *Assume that* $\sup_{j\in\mathbb{N}^\star} \|\phi_j\| \leq 1$. *Let* $1 < q < 2$, $r > 0$, $R > 0$ *such that* $R\varepsilon^{-1} \geq \mathrm{e}$ *and assume that* $s \in \mathcal{B}_{q,r}(R)$. *Consider the selected Lasso estimator* $\hat{s}_{\hat{p}}$ *defined by* (2.20) *with parameters* λ_p *and* $\mathrm{pen}(p)$ *given by* (2.21).
Then, there exists $C_{q,r} > 0$ *depending only on* q *and* r *such that the quadratic risk of* $\hat{s}_{\hat{p}}$ *satisfies*

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2\right] \leq C_{q,r}\, R^q \left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}. \tag{2.34}$$

Proof. Page 89. □

Proposition 2.4.4 is to be compared with Proposition 2.4.1 established in the orthonormal case. Taking into account the inclusion (2.33) and noting that the upper bounds of the quadratric risk (2.28) and (2.34) are exactly of the same order and valid under the same assumption on the signal to noise ratio, we can conclude that Proposition 2.4.4 extends the result established in Proposition 2.4.1. Yet, we shall provide an independent proof of Proposition 2.4.1 in Appendix 2.C.1.1 to see how things work in the simpler orthonormal case.

Proposition 2.4.4 highlights the high performance of the selected Lasso estimator compared with other existing estimators in the theory of approximation and learning. In particular, (2.34) proves that the selected Lasso estimator performs as well as the greedy algorithms for which Barron et al. (2008) have provided similar rates of convergence. Besides, since the construction of the selected Lasso estimator is based on an adaptive truncation of the dictionary, this estimator has the great advantage of being adaptive to the unknown parameters of smoothness q and r of the target function, whereas the greedy algorithms achieve their rates of convergence only for restricted values of the parameter r depending on the level of truncation of the dictionary (Barron et al., 2008, Corollary 3.7).

2.4.2 Rates of convergence of the Lasso for neural networks

Here, we provide rates of convergence of the Lasso for the infinite dictionaries used for neural networks when the target function s belongs to some interpolation space between $\mathcal{L}_1(\mathcal{D})$ and $\mathbb{R}^n$. We keep the notations used in Section 2.3.3.

Definition of the spaces

In Section 2.4.1.1, we introduced the spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$ because they were adapted to the truncation of the dictionary. In Section 2.3.3, we saw that, for the specific case of neural networks, no truncation of the dictionary is necessary to establish results for the Lasso. Therefore, we no longer consider the spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$ here. The spaces $\mathcal{L}_{1,r}$ are replaced by the whole space $\mathcal{L}_1(\mathcal{D})$. The spaces $\mathcal{B}_{q,r}$ are replaced by bigger spaces $\mathcal{B}_q$ that are real interpolation spaces between $\mathcal{L}_1(\mathcal{D})$ and $\mathbb{R}^n$, which coincide with $w\mathcal{L}_q$ when the dictionary is orthonormal:

Definition 2.4.5. [Space $\mathcal{B}_q$] If $1 < q < 2$, $\nu = 1/q - 1/2$ and $R > 0$, we say that a function u belongs to $\mathcal{B}_q(R)$ if we have the following control of the $K_{\mathcal{D}}$-functional of u for all $\delta > 0$:

$$K_{\mathcal{D}}(u, \delta) := \inf_{t \in \mathcal{L}_1(\mathcal{D})} \left\{ \|u - t\| + \delta \|t\|_{\mathcal{L}_1(\mathcal{D})} \right\} \leq R\, \delta^{2\nu}. \tag{2.35}$$

Upper bound of the quadratic risk

Proposition 2.4.6. *Let $R \geq \sigma d^{1/4} n^{-1/2}$. Assume that $s \in \mathcal{B}_q(R)$. Consider the Lasso estimator $\hat{s}$ defined by* (2.24) *with a regularization parameter λ checking* (2.25).
Then, the quadratic risk of $\hat{s}$ satisfies

$$\mathbb{E}\left[\|s - \hat{s}\|^2\right] \leq C_q\, R^q \left(\frac{d}{n}\right)^{1-\frac{q}{2}}, \tag{2.36}$$

where $C_q > 0$ depends only on q.

Proof. Page 89. □

This result shows that the Lasso is adaptive to the unknown parameter of smoothness q of the target function and that it theoretically performs as well as the greedy algorithms introduced by Barron et al. (2008) for neural networks. Our result can be seen as the analog in the Gaussian framework of Barron et al.'s result which is stated under the assumption that the output variable Y is bounded but not necessarily Gaussian.

Appendices

2.A The Lasso as an ℓ_1-ball model selection procedure

Here, we describe the idea that has enabled us to establish the oracle inequalities of Theorem 2.3.2, Theorem 2.3.3 and Theorem 2.3.4 as an application of a single model selection theorem. Then, we state and prove this model selection theorem. We keep the notations introduced in Section 2.2.

The basic idea is to view the Lasso as the solution of an ℓ_1-penalized least squares model selection procedure over a properly defined countable model collection. The key observation that enables to make this connection is the simple fact that $\mathcal{L}_1(\mathcal{D}) = \bigcup_{R>0}\{t \in \mathcal{L}_1(\mathcal{D});\ \|t\|_{\mathcal{L}_1(\mathcal{D})} \leq R\}$, so that for any finite or infinite dictionary $\mathcal{D}$, the Lasso estimator $\hat{s}$ satisfies

$$\gamma(\hat{s}) + \lambda\|\hat{s}\|_{\mathcal{L}_1(\mathcal{D})} = \inf_{t\in\mathcal{L}_1(\mathcal{D})}\left\{\gamma(t) + \lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\} = \inf_{R>0}\left\{\inf_{\|t\|_{\mathcal{L}_1(\mathcal{D})}\leq R}\gamma(t) + \lambda R\right\}.$$

To get a countable model collection, we discretize the family of ℓ_1-balls $\{t \in \mathcal{L}_1(\mathcal{D});\ \|t\|_{\mathcal{L}_1(\mathcal{D})} \leq R\}$ by setting for all $m \in \mathbb{N}^\star$,

$$S_m = \left\{t \in \mathcal{L}_1(\mathcal{D});\ \|t\|_{\mathcal{L}_1(\mathcal{D})} \leq m\varepsilon\right\}.$$

We define $\hat{m}$ as the smallest integer such that $\hat{s}$ belongs to $S_{\hat{m}}$, i.e.

$$\hat{m} = \left\lceil \frac{\|\hat{s}\|_{\mathcal{L}_1(\mathcal{D})}}{\varepsilon} \right\rceil. \tag{2.37}$$

It is now easy to derive from the definitions of $\hat{m}$ and $\hat{s}$, from the fact that $\mathcal{L}_1(\mathcal{D}) = \bigcup_{m\in\mathbb{N}^\star} S_m$ and from the definition of S_m that

$$\begin{aligned}
\gamma(\hat{s}) + \lambda\hat{m}\varepsilon &\leq \gamma(\hat{s}) + \lambda\left(\|\hat{s}\|_{\mathcal{L}_1(\mathcal{D})} + \varepsilon\right) \\
&= \inf_{t\in\mathcal{L}_1(\mathcal{D})}\left\{\gamma(t) + \lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\} + \lambda\varepsilon \\
&= \inf_{m\in\mathbb{N}^\star}\left\{\inf_{t\in S_m}\left\{\gamma(t) + \lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\}\right\} + \lambda\varepsilon \\
&\leq \inf_{m\in\mathbb{N}^\star}\left\{\inf_{t\in S_m}\gamma(t) + \lambda m\varepsilon\right\} + \lambda\varepsilon,
\end{aligned}$$

that is to say

$$\gamma(\hat{s}) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\mathbb{N}^\star}\left\{\inf_{t\in S_m}\gamma(t) + \mathrm{pen}(m)\right\} + \rho \tag{2.38}$$

with $\mathrm{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$. This means that the Lasso $\hat{s}$ is a ρ-approximate penalized least squares estimator over the model collection of ℓ_1-balls $\{S_m;\ m \in \mathbb{N}^\star\}$. This property will enable us to derive ℓ_1-oracle inequalities for the Lasso from the theory on model selection via penalization. Specifically, we have established the following model selection theorem, which is adapted to provide oracle inequalities for estimators fulfilling (2.38):

Theorem 2.A.1. *Let $\{S_m\}_{m\in\mathcal{M}}$ be a countable collection of convex and compact subsets of a Hilbert space $\mathbb{H}$. Define, for any $m \in \mathcal{M}$,*

$$\Delta_m := \mathbb{E}\left[\sup_{s_m\in S_m} W(s_m)\right], \tag{2.39}$$

with W defined by (2.4)*, and consider weights $\{x_m\}_{m\in\mathcal{M}}$ such that*

$$\Sigma := \sum_{m\in\mathcal{M}} \mathrm{e}^{-x_m} < \infty.$$

Let $K > 1$ and assume that, for any $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq 2K\varepsilon \left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right). \tag{2.40}$$

Given non-negative ρ_m, $m \in \mathcal{M}$, define a ρ_m-approximate penalized least squares estimator as any $\hat{s} \in S_{\hat{m}}$, $\hat{m} \in \mathcal{M}$, such that

$$\gamma(\hat{s}) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\mathcal{M}} \left\{ \inf_{s_m\in S_m} \gamma(s_m) + \mathrm{pen}(m) + \rho_m \right\}.$$

Then, there is a positive constant $C(K)$ such that for all $s \in \mathbb{H}$ and $z > 0$, with probability larger than $1 - \Sigma\,\mathrm{e}^{-z}$,

$$\begin{aligned} &\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \\ &\leq C(K) \left[\inf_{m\in\mathcal{M}} \left\{ \inf_{s_m\in S_m} \|s - s_m\|^2 + \mathrm{pen}(m) + \rho_m \right\} + (1+z)\varepsilon^2 \right]. \end{aligned} \tag{2.41}$$

Integrating this inequality with respect to z leads to the following risk bound

$$\begin{aligned} &\mathbb{E}\left[\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m})\right] \\ &\leq C(K) \left[\inf_{m\in\mathcal{M}} \left\{ \inf_{s_m\in S_m} \|s - s_m\|^2 + \mathrm{pen}(m) + \rho_m \right\} + (1+\Sigma)\varepsilon^2 \right]. \end{aligned} \tag{2.42}$$

Remark 5. Theorem 2.A.1 is a particular case of Theorem 4.18 in Massart (2007): the function ϕ_m of Theorem 4.18 (Massart, 2007) bounding the expectation of the supremum of the normalized empirical process is a constant function (which is equal to Δ_m) in Theorem 2.A.1. This supposes that we do not use the localization refinement used in Theorem 4.18 (Massart, 2007) to improve the risk minimization method developed by Vapnik (1982). This non-localized version of Theorem 4.18 (Massart, 2007) is sufficient to derive the oracle inequalities satisfied by the Lasso and the selected Lasso estimator in our framework. By considering the solution $D_m = \Delta_m/\varepsilon$ of (4.72) in Theorem 4.18 (Massart, 2007), one recovers the same lower bound of the penalty function in both Theorem 2.A.1 and Theorem 4.18 (Massart, 2007).

Proof. The proof is based on the concentration inequality for the suprema of Gaussian processes established in Boucheron et al. (2012).

Fix $m \in \mathcal{M}$. Since S_m is assumed to be a convex and compact subset, we can consider $\overline{s}_m$ the projection of s onto S_m, that is the unique element of S_m such that $\|s - \overline{s}_m\| = \inf_{s_m\in S_m} \|s - s_m\|$.

By definition of $\hat{s}$, we have

$$\gamma(\hat{s}) + \mathrm{pen}(\hat{m}) \leq \gamma(\overline{s}_m) + \mathrm{pen}(m) + \rho_m.$$

Since $\|s\|^2 + \gamma(s_m) = \|s - s_m\|^2 - 2\varepsilon W(s_m)$, this implies that

$$\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \leq \|s - \overline{s}_m\|^2 + 2\varepsilon \left(W(\hat{s}) - W(\overline{s}_m)\right) + \mathrm{pen}(m) + \rho_m. \tag{2.43}$$

For all $m' \in \mathcal{M}$, let $y_{m'}$ be a positive number whose value will be specified below and define for every $s_{m'} \in S_{m'}$

$$2w_{m'}(s_{m'}) = \left(\|s - \overline{s}_m\| + \|s - s_{m'}\|\right)^2 + y_{m'}^2. \tag{2.44}$$

Finally, set

$$V_{m'} = \sup_{s_{m'} \in S_{m'}} \left(\frac{W(s_{m'}) - W(\overline{s}_m)}{w_{m'}(s_{m'})} \right).$$

Taking these definitions into account, we get from (2.43) that

$$\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \leq \|s - \overline{s}_m\|^2 + 2\varepsilon w_{\hat{m}}(\hat{s}) V_{\hat{m}} + \mathrm{pen}(m) + \rho_m. \tag{2.45}$$

The essence of the proof is the control of the random variables $V_{m'}$ for all possible values of m'. To this end, we may use the concentration inequality for the suprema of Gaussian processes (Boucheron et al., 2012) which ensures that, given $z > 0$, for all $m' \in \mathcal{M}$,

$$\mathbb{P}\left[V_{m'} \geq \mathbb{E}\left[V_{m'}\right] + \sqrt{2v_{m'}(x_{m'} + z)}\right] \leq \mathrm{e}^{-(x_{m'}+z)}, \tag{2.46}$$

where

$$v_{m'} = \sup_{s_{m'} \in S_{m'}} \mathrm{Var}\left[\frac{W(s_{m'}) - W(\overline{s}_m)}{w_{m'}(s_{m'})}\right] = \sup_{s_{m'} \in S_{m'}} \frac{\|s_{m'} - \overline{s}_m\|^2}{w_{m'}^2(s_{m'})}.$$

From (2.44), $w_{m'}(s_{m'}) \geq \left(\|s - \overline{s}_m\| + \|s - s_{m'}\|\right) y_{m'} \geq \|s_{m'} - \overline{s}_m\| y_{m'}$, so $v_{m'} \leq y_{m'}^{-2}$ and summing the inequalities (2.46) over $m' \in \mathcal{M}$, we get that for every $z > 0$, there is an event Ω_z with $\mathbb{P}(\Omega_z) > 1 - \Sigma \mathrm{e}^{-z}$ such that on Ω_z, for all $m' \in \mathcal{M}$,

$$V_{m'} \leq \mathbb{E}\left[V_{m'}\right] + y_{m'}^{-1}\sqrt{2(x_{m'} + z)}. \tag{2.47}$$

Let us now bound $\mathbb{E}\left[V_{m'}\right]$. We may write

$$\mathbb{E}\left[V_{m'}\right] \leq \mathbb{E}\left[\frac{\sup_{s_{m'} \in S_{m'}} \left(W(s_{m'}) - W(\overline{s}_{m'})\right)}{\inf_{s_{m'} \in S_{m'}} w_{m'}\left(s_{m'}\right)}\right] + \mathbb{E}\left[\frac{\left(W(\overline{s}_{m'}) - W(\overline{s}_m)\right)_+}{\inf_{s_{m'} \in S_{m'}} w_{m'}(s_{m'})}\right]. \tag{2.48}$$

But from the definition of $\overline{s}_{m'}$, we have for all $s_{m'} \in S_{m'}$

$$2w_{m'}(s_{m'}) \geq (\|s-\overline{s}_m\| + \|s-\overline{s}_{m'}\|)^2 + y_{m'}^2 \geq \|\overline{s}_{m'} - \overline{s}_m\|^2 + y_{m'}^2 \geq \left(y_{m'}^2 \vee 2y_{m'}\|\overline{s}_{m'} - \overline{s}_m\|\right).$$

Hence, on the one hand, via (2.39) and recalling that W is centered, we get

$$\mathbb{E}\left[\frac{\sup_{s_{m'}\in S_{m'}}\left(W(s_{m'}) - W(\overline{s}_{m'})\right)}{\inf_{s_{m'}\in S_{m'}} w_{m'}(s_{m'})}\right] \leq 2\, y_{m'}^{-2}\, \Delta_{m'},$$

and on the other hand, using the fact that $\left(W(\overline{s}_{m'}) - W(\overline{s}_m)\right)/\|\overline{s}_{m'} - \overline{s}_m\|$ is a standard normal variable, we get

$$\mathbb{E}\left[\frac{\left(W(\overline{s}_{m'}) - W\left(\overline{s}_m\right)\right)_+}{\inf_{s_{m'}\in S_{m'}} w_{m'}(s_{m'})}\right] \leq y_{m'}^{-1}\, \mathbb{E}\left[\frac{W(\overline{s}_{m'}) - W(\overline{s}_m)}{\|\overline{s}_m - \overline{s}_{m'}\|}\right]_+ \leq y_{m'}^{-1}\,(2\pi)^{-1/2}.$$

Collecting these inequalities, we get from (2.48) that for all $m' \in \mathcal{M}$,

$$\mathbb{E}\left[V_{m'}\right] \leq 2\Delta_{m'} y_{m'}^{-2} + (2\pi)^{-1/2}\, y_{m'}^{-1}.$$

Hence, setting $\delta = ((4\pi)^{-1/2} + \sqrt{z})^2$, (2.47) implies that on the event Ω_z, for all $m' \in \mathcal{M}$,

$$\begin{aligned} V_{m'} &\leq y_{m'}^{-1}\left[2\Delta_{m'} y_{m'}^{-1} + \sqrt{2x_{m'}} + (2\pi)^{-1/2} + \sqrt{2z}\right] \\ &= y_{m'}^{-1}\left[2\Delta_{m'} y_{m'}^{-1} + \sqrt{2x_{m'}} + \sqrt{2\delta}\right]. \end{aligned} \tag{2.49}$$

Given $K' \in (1, \sqrt{K}]$ to be chosen later, we now define

$$y_{m'}^2 = 2K'^2\varepsilon^2\left[\left(\sqrt{x_{m'}} + \sqrt{\delta}\right)^2 + K'^{-1}\varepsilon^{-1}\Delta_{m'} + \sqrt{K'^{-1}\varepsilon^{-1}\Delta_{m'}}\left(\sqrt{x_{m'}} + \sqrt{\delta}\right)\right].$$

With this choice of $y_{m'}$, it is not hard to check that (2.49) warrants that on the event Ω_z, $\varepsilon V_{m'} \leq K'^{-1}$ for all $m' \in \mathcal{M}$, which in particular implies that $\varepsilon V_{\hat{m}} \leq K'^{-1}$, and we get from (2.45) and (2.44) that

$$\begin{aligned} &\|s-\hat{s}\|^2 + \mathrm{pen}(\hat{m}) \\ &\leq \|s-\overline{s}_m\|^2 + 2K'^{-1} w_{\hat{m}}(\hat{s}) + \mathrm{pen}(m) + \rho_m \\ &= \|s-\overline{s}_m\|^2 + K'^{-1}\left[(\|s-\overline{s}_m\| + \|s-\hat{s}\|)^2 + y_{\hat{m}}^2\right] + \mathrm{pen}(m) + \rho_m. \end{aligned} \tag{2.50}$$

Moreover, using repeatedly the inequality $(a+b)^2 \leq (1+\beta)a^2 + (1+\beta^{-1})b^2$ for various values of $\beta > 0$, we derive that, on the one hand,

$$(\|s-\overline{s}_m\| + \|s-\hat{s}\|)^2 \leq \sqrt{K'}\left(\|s-\hat{s}\|^2 + \frac{\|s-\overline{s}_m\|^2}{\sqrt{K'}-1}\right),$$

and, on the other hand,

$$K'^{-1} y_{\hat{m}}^2 \leq 2K'^2 \varepsilon^2 \left[\varepsilon^{-1} \Delta_{\hat{m}} + x_{\hat{m}} + \sqrt{\varepsilon^{-1} \Delta_{\hat{m}} x_{\hat{m}}} + B(K') \left(\frac{1}{2\pi} + 2z \right) \right]$$

where $B(K') = (K'-1)^{-1} + (4K'(K'^2-1))^{-1}$.
Hence, setting $A(K') = 1 + K'^{-1/2}(\sqrt{K'}-1)^{-1}$, we deduce from (2.50) that on the event Ω_z,

$$\begin{aligned} &\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \\ &\leq A(K')\|s - \overline{s}_m\|^2 + K'^{-1/2}\|s - \hat{s}\|^2 + 2K'^2 \varepsilon \left[\Delta_{\hat{m}} + \varepsilon x_{\hat{m}} + \sqrt{\varepsilon \Delta_{\hat{m}} x_{\hat{m}}} \right] \\ &\quad + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 K'^2 B(K') \left(\frac{1}{2\pi} + 2z \right). \end{aligned}$$

Because of Condition (2.40) on the penalty function, this implies that

$$\begin{aligned} &\left(1 - K'^{-1/2}\right) \|s - \hat{s}\|^2 + \left(1 - K'^2 K^{-1}\right) \mathrm{pen}(\hat{m}) \\ &\leq A(K') \|s - \overline{s}_m\|^2 + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 K'^2 B(K') \left(\frac{1}{2\pi} + 2z \right). \end{aligned}$$

Now choosing $K' = K^{2/5}$, we get that

$$\begin{aligned} &\left(1 - K^{-1/5}\right) \left(\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \right) \\ &\leq A(K^{2/5})\|s - \overline{s}_m\|^2 + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 K^{4/5} B(K^{2/5}) \left(\frac{1}{2\pi} + 2z \right). \end{aligned}$$

So, there exists a positive constant $C(K)$ depending only on K such that for all $z > 0$, on the event Ω_z,

$$\|s - \hat{s}\|^2 + \mathrm{pen}(\hat{m}) \leq C(K) \left(\inf_{m \in \mathcal{M}} \left\{ \|s - \overline{s}_m\|^2 + \mathrm{pen}(m) + \rho_m \right\} + \varepsilon^2 (1 + z) \right),$$

which proves (2.41). Integrating this inequality with respect to z leads to the risk bound (2.42). □

2.B Proof of the ℓ_1-oracle inequalities

Theorem 2.3.2, Theorem 2.3.3 and Theorem 2.3.4 are direct applications of Theorem 2.A.1. Indeed, using the key observation that the Lasso estimator is an approximate penalized least squares estimator over a collection of ℓ_1-balls with a convenient penalty (see Section 2.A), it only remains to determine a lower bound of this penalty to guarantee Condition (2.40) and then to apply the conclusion of Theorem 2.A.1.

2.B.1 Proof of Theorem 2.3.2

Fix $p \in \mathbb{N}^\star$. Let $\mathcal{M} = \mathbb{N}^\star$ and consider the collection of ℓ_1-balls for $m \in \mathcal{M}$,

$$S_m = \left\{ s_m \in \mathcal{L}_1(\mathcal{D}_p);\, \|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon \right\}.$$

From (2.38), the Lasso estimator $\hat{s}_p$ is a ρ-approximate penalized least squares estimator over the collection $\{S_m;\ m \in \mathbb{N}^\star\}$ for $\mathrm{pen}(m) = \lambda_p m\varepsilon$ and $\rho = \lambda_p \varepsilon$. So, it only remains to determine a lower bound of λ_p such that $\mathrm{pen}(m)$ satisfies Condition (2.40).

Let $s_m \in S_m$ and consider $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p$ such that $s_m = \alpha.\phi = \sum_{j=1}^p \alpha_j\, \phi_j$ and $\|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)} = \|\alpha\|_1$. The linearity of W and the definition of S_m imply that

$$W(s_m) = \sum_{j=1}^p \alpha_j\, W(\phi_j) \leq \sum_{j=1}^p |\alpha_j|\, |W(\phi_j)| \leq m\varepsilon \max_{j=1,\ldots,p} |W(\phi_j)|\,.^1 \tag{2.51}$$

Recalling that W is isonormal (see (2.4)), we have $\mathrm{Var}\,[W(\phi_j)] = \|\phi_j\|^2 \leq 1$ for all $j \in \{1, \ldots, p\}$. So, the variables $W(\phi_j)$ and $(-W(\phi_j))$, $j \in \{1, \ldots, p\}$, are $2p$ centered normal variables with variance less than 1 and thus (Massart, 2007, Lemma 2.3),

$$\mathbb{E}\left[\max_{j=1,\ldots,p} |W(\phi_j)|\right] = \mathbb{E}\left[\left(\max_{j=1,\ldots,p} W(\phi_j)\right) \vee \left(\max_{j=1,\ldots,p} (-W(\phi_j))\right)\right] \leq \sqrt{2\ln(2p)}\,.$$

Therefore, we deduce from (2.51) that

$$\Delta_m := \mathbb{E}\left[\sup_{s_m \in S_m} W(s_m)\right] \leq m\varepsilon\sqrt{2\ln(2p)} \leq \sqrt{2} m\varepsilon \left(\sqrt{\ln p} + \sqrt{\ln 2}\right). \tag{2.52}$$

Now, choose weights of the form $x_m = \delta m$ where $\delta > 0$ is specified below. Then, $\sum_{m \geq 1} \mathrm{e}^{-x_m} = 1/\left(\mathrm{e}^\delta - 1\right) := \Sigma_\delta < +\infty$.

Define $K = 4\sqrt{2}/5 > 1$ and $\delta = (1 - \sqrt{\ln 2})/K$. By using the inequality $2\sqrt{ab} \leq a/2 + 2b$, we

[1] Note that we are turning a problem with a possible infinite set S_m into a problem with a finite set $\{\phi_1, \ldots, \phi_p\}$. This crucial point will enable us to apply a maximal inequality for a finite family of Gaussian random variables. This is why the non-localized Theorem 2.A.1 is sufficient for our framework. If we work directly with the infinite sets S_m and use metric entropy arguments to find a function upper bounding the expectation of the supremum of the empirical process (2.39), then we find an extra-$\ln(n)$ factor in the minimal penalty function allowing the ℓ_1-oracle inequality. This is the problem encountered by Bartlett et al. (2012).

deduce from (2.52) and the definition of x_m that

$$\begin{aligned}
2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right) &\leq K\varepsilon\left(\frac{5}{2}\Delta_m + 4x_m\varepsilon\right) \\
&\leq 4m\varepsilon^2\left(\sqrt{\ln p} + \sqrt{\ln 2} + K\delta\right) \\
&\leq 4m\varepsilon^2\left(\sqrt{\ln p} + 1\right) \\
&\leq \lambda_p m\varepsilon
\end{aligned}$$

as soon as

$$\lambda_p \geq 4\varepsilon\left(\sqrt{\ln p} + 1\right). \tag{2.53}$$

For such values of λ_p, Condition (2.40) on the penalty function is satisfied and we may apply Theorem 2.A.1 with $\mathrm{pen}(m) = \lambda_p m\varepsilon$ and $\rho = \lambda_p\varepsilon$. Taking into account the definition of $\hat{m}$ at (2.37) and noting that $\varepsilon^2 \leq \lambda_p\varepsilon/4$ for λ_p satisfying (2.53), we get from (2.41) that there exists $C > 0$ such that for all $z > 0$, with probability larger than $1 - \Sigma_\delta\,\mathrm{e}^{-z} \geq 1 - 3.4\,\mathrm{e}^{-z}$,

$$\begin{aligned}
&\|s - \hat{s}_p\|^2 + \lambda_p\|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \\
&\leq C\left[\inf_{m\in\mathbb{N}^\star}\left\{\inf_{\|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)}\leq m\varepsilon}\|s - s_m\|^2 + \lambda_p m\varepsilon\right\} + \lambda_p\varepsilon + (1+z)\varepsilon^2\right] \\
&\leq C\left[\inf_{m\in\mathbb{N}^\star}\left\{\inf_{\|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)}\leq m\varepsilon}\|s - s_m\|^2 + \lambda_p m\varepsilon\right\} + \lambda_p\varepsilon(1+z)\right].
\end{aligned} \tag{2.54}$$

Finally, to get the desired bound (2.13), for all $t \in \mathcal{L}_1(\mathcal{D}_p)$, consider $m_t = \lceil \|t\|_{\mathcal{L}_1(\mathcal{D}_p)}/\varepsilon\rceil \in \mathbb{N}^\star$ so that $\|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m_t\varepsilon$. Then,

$$\begin{aligned}
\inf_{m\in\mathbb{N}^\star}\left\{\inf_{\|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)}\leq m\varepsilon}\|s - s_m\|^2 + \lambda_p m\varepsilon\right\} &\leq \|s - t\|^2 + \lambda_p m_t\varepsilon \\
&\leq \|s - t\|^2 + \lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)} + \lambda_p\varepsilon,
\end{aligned} \tag{2.55}$$

and combining (2.54) with (2.55) leads to

$$\|s - \hat{s}_p\|^2 + \lambda_p\|\hat{s}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq 2C\left[\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s - t\|^2 + \lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} + \lambda_p\varepsilon(1+z)\right].$$

Similarly, we get the risk bound (2.14) from (2.42). □

2.B.2 Proof of Theorem 2.3.3

Let $\mathcal{M} = \mathbb{N}^\star \times \Lambda$ and consider the set of ℓ_1-balls for all $(m,p) \in \mathcal{M}$,

$$S_{m,p} = \left\{ s_m \in \mathcal{L}_1(\mathcal{D}_p);\, \|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon \right\}.$$

Define $\hat{m}$ as the smallest integer such that $\hat{s}_{\hat{p}}$ belongs to $S_{\hat{m},\hat{p}}$, i.e.

$$\hat{m} = \left\lceil \frac{\|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})}}{\varepsilon} \right\rceil. \tag{2.56}$$

Let $c = 1 - c_1/(c_2\sqrt{\ln 2})$. From (2.56) and (2.21), using that for all $p \in \Lambda$, $\sqrt{\ln p} \leq (\ln p)/\sqrt{\ln 2}$, the definitions of $\hat{m}$, $\lambda_{\hat{p}}$, $\mathrm{pen}(\hat{p})$, c and $\hat{s}_{\hat{p}}$ and the fact that $\mathcal{L}_1(\mathcal{D}_p) = \bigcup_{m\in\mathbb{N}^\star} S_{m,p}$, we get that

$$\begin{aligned}
&\gamma(\hat{s}_{\hat{p}}) + \lambda_{\hat{p}}\hat{m}\varepsilon + c\,\mathrm{pen}(\hat{p}) \\
&\leq \gamma(\hat{s}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \lambda_{\hat{p}}\varepsilon + c\,\mathrm{pen}(\hat{p}) \\
&\leq \gamma(\hat{s}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + c_1\varepsilon^2\left(\sqrt{\ln\hat{p}} + 1\right) + cc_2\varepsilon^2\ln\hat{p} \\
&\leq \gamma(\hat{s}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \left(\frac{c_1}{c_2\sqrt{\ln 2}} + c\right)c_2\varepsilon^2\ln\hat{p} + c_1\varepsilon^2 \\
&\leq \gamma(\hat{s}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) + c_1\varepsilon^2 \\
&\leq \inf_{p\in\Lambda}\left\{\inf_{s_m\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\gamma(s_m) + \lambda_p\|s_m\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} + \mathrm{pen}(p)\right\} + c_1\varepsilon^2 \\
&\leq \inf_{p\in\Lambda}\left\{\inf_{m\in\mathbb{N}^\star}\left\{\inf_{s_m\in S_{m,p}}\gamma(s_m) + \lambda_p m\varepsilon\right\} + \mathrm{pen}(p)\right\} + c_1\varepsilon^2 \\
&\leq \inf_{(m,p)\in\mathcal{M}}\left\{\inf_{s_m\in S_{m,p}}\gamma(s_m) + \lambda_p m\varepsilon + \mathrm{pen}(p)\right\} + c_1\varepsilon^2,
\end{aligned}$$

that is to say

$$\gamma(\hat{s}_{\hat{p}}) + \mathrm{pen}(\hat{m},\hat{p}) \leq \inf_{(m,p)\in\mathcal{M}}\left\{\inf_{s_m\in S_{m,p}}\gamma(s_m) + \mathrm{pen}(m,p) + \rho_p\right\},$$

with $\mathrm{pen}(m,p) := \lambda_p m\varepsilon + c\,\mathrm{pen}(p)$ and $\rho_p := (1-c)\,\mathrm{pen}(p) + c_1\varepsilon^2$ (thanks to the assumption $c_2 > c_1/\sqrt{\ln 2}$, we have $c \in]0,1[$, so $\mathrm{pen}(m,p) > 0$ and $\rho_p > 0$). This means that $\hat{s}_{\hat{p}}$ is a ρ_p-approximate penalized least squares estimator over the model collection $\{S_{m,p}\,;\,(m,p)\in\mathcal{M}\}$. By applying Theorem 2.A.1, this property will enable us to derive a risk bound satisfied by $\hat{s}_{\hat{p}}$ provided that $\mathrm{pen}(m,p)$ is large enough.

Let us now choose weights of the form $x_{m,p} = \delta m + \beta\ln p$ where $\delta > 0$ and $\beta > 0$ are numerical

constants specified later. Then,

$$\begin{aligned}\Sigma_{\delta,\beta} := \sum_{(m,p)\in\mathcal{M}} \mathrm{e}^{-x_{m,p}} &= \left(\sum_{m\in\mathbb{N}^\star} \mathrm{e}^{-\delta m}\right)\left(\sum_{p\in\Lambda} \mathrm{e}^{-\beta \ln p}\right)\\ &= \left(\sum_{m\in\mathbb{N}^\star} \mathrm{e}^{-\delta m}\right)\left(\sum_{J\in\mathbb{N}} \mathrm{e}^{-\beta \ln 2^J}\right)\\ &= \frac{1}{(\mathrm{e}^{\delta}-1)\,(1-2^{-\beta})} < +\infty.\end{aligned}$$

Moreover, for all $(m,p)\in\mathcal{M}$, we can prove similarly as (2.52) that

$$\Delta_{m,p} := \mathbb{E}\left[\sup_{s_m\in S_{m,p}} W(s_m)\right] \leq \sqrt{2}m\varepsilon\left(\sqrt{\ln p}+\sqrt{\ln 2}\right).$$

Now, define $K = c_1[\sqrt{2}(2+(c_1-2)^{-1})]^{-1}$ ($K>1$ thanks to the assumption $c_1\geq 4$), $\delta = (1-\sqrt{\ln 2})/K > 0$ and $\beta = (c_2 c)/(c_1 K) > 0$. By taking into account these definitions and using the inequality $2\sqrt{ab}\leq \eta a + \eta^{-1} b$ with $\eta = (c_1-2)^{-1}$, $a=\Delta_{m,p}$ and $b=\varepsilon x_{m,p}$, we get that

$$\begin{aligned}&2K\varepsilon\left(\Delta_{m,p}+\varepsilon x_{m,p}+\sqrt{\Delta_{m,p}\varepsilon x_{m,p}}\right)\\ &\leq K\varepsilon\left(\left(2+(c_1-2)^{-1}\right)\Delta_{m,p}+c_1 x_{m,p}\varepsilon\right)\\ &\leq c_1\varepsilon^2\left(m\sqrt{\ln p}+m\sqrt{\ln 2}+K\delta m+K\beta\ln p\right)\\ &\leq c_1\varepsilon^2\left(m\left(\sqrt{\ln p}+1\right)+\frac{c_2 c}{c_1}\ln p\right)\\ &\leq \lambda_p m\varepsilon + c\,\mathrm{pen}(p)\\ &= \mathrm{pen}(m,p).\end{aligned}$$

Thus, Condition (2.40) is satisfied and we can apply Theorem 2.A.1 with $\mathrm{pen}(m,p) = \lambda_p m\varepsilon + c\,\mathrm{pen}(p)$ and $\rho_p = (1-c)\,\mathrm{pen}(p) + c_1\varepsilon^2$, which leads to:

$$\begin{aligned}&\mathbb{E}\left[\|s-\hat{s}_{\hat{p}}\|^2+\lambda_{\hat{p}}\hat{m}\varepsilon + c\,\mathrm{pen}(\hat{p})\right]\\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left\{\inf_{s_m\in S_{m,p}}\|s-s_m\|^2+\lambda_p m\varepsilon+\mathrm{pen}(p)\right\}+\left(c_1+1+\Sigma_{\delta,\beta}\right)\varepsilon^2\right]\\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left\{\inf_{s_m\in S_{m,p}}\|s-s_m\|^2+\lambda_p m\varepsilon+\mathrm{pen}(p)\right\}+\varepsilon^2\right],\end{aligned} \tag{2.57}$$

where $C>0$ denotes some absolute constant. The infimum of this risk bound can easily be extended to $\inf_{p\in\Lambda}\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}$. Indeed, let $p_0\in\Lambda$ and $t\in\mathcal{L}_1(\mathcal{D}_{p_0})$, and consider $m_t = \lceil \|t\|_{\mathcal{L}_1(\mathcal{D}_{p_0})}/\varepsilon\rceil \in \mathbb{N}^\star$

so that $t \in S_{m_t,p_0}$. Then,

$$\begin{aligned}
&\inf_{(m,p)\in\mathcal{M}} \left\{ \inf_{s_m \in S_{m,p}} \|s - s_m\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p) \right\} \\
&\leq \|s-t\|^2 + \lambda_{p_0} m_t \varepsilon + \mathrm{pen}(p_0) \\
&\leq \|s-t\|^2 + \lambda_{p_0} \left(\|t\|_{\mathcal{L}_1(\mathcal{D}_{p_0})} + \varepsilon \right) + \mathrm{pen}(p_0) \\
&\leq \|s-t\|^2 + \lambda_{p_0} \|t\|_{\mathcal{L}_1(\mathcal{D}_{p_0})} + \left(\frac{c_1}{c_2\sqrt{\ln 2}} + 1 \right) \mathrm{pen}(p_0) + c_1\varepsilon^2. \qquad (2.58)
\end{aligned}$$

We deduce from (2.57) and (2.58) that there exists $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \hat{m}\varepsilon + c\,\mathrm{pen}(\hat{p}) \right] \\
&\leq C \left[\inf_{p\in\Lambda} \left\{ \inf_{t\in\mathcal{L}_1(\mathcal{D}_p)} \left(\|s-t\|^2 + \lambda_p \|t\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \mathrm{pen}(p) \right\} + \varepsilon^2 \right]. \qquad (2.59)
\end{aligned}$$

But from the fact that $c \in]0,1[$ and from (2.56), we have

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \|\hat{s}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) \right] \leq \frac{1}{c}\, \mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \hat{m}\varepsilon + c\,\mathrm{pen}(\hat{p}) \right]. \qquad (2.60)$$

Combining (2.59) with (2.60) leads to the result. □

2.B.3 Proof of Theorem 2.3.4

Let us recall that, for $\delta > 0$, the δ-packing number $\mathcal{N}(\delta, \mathcal{G}, N(.))$ of a set $\mathcal{G}$ with respect to a norm $N(.)$ is the maximal $m \in \mathbb{N}^\star$ such that there exist $g_1, \ldots, g_m \in \mathcal{G}$ with $N(g_i - g_j) \geq \delta$ for all $1 \leq i < j \leq m$. The δ-entropy number is defined by $H(\delta, \mathcal{G}, N(.)) = \ln\left(\mathcal{N}(\delta, \mathcal{G}, N(.))\right)$.

Lemma 2.B.1. *Let $\delta > 0$ and let $\mathcal{D} = \{\phi_{a,b}; a \in \mathbb{R}^d, b \in \mathbb{R}\}$ be a dictionary of neurons where $\phi_{a,b}$ is defined by* (2.23)*. Then,*

$$\int_0^1 \sqrt{H(\delta, \mathcal{D}, \|.\|)}\; d\delta \leq C\sqrt{d+1},$$

where $C > 0$ is an absolute constant ($C \geq 22$ is convenient).

Proof. The result just comes from the fact that $\mathcal{D}$ is a subset of the boolean n-cube with Vapnik-Chervonenkis dimension $d+1$. Indeed, for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, denote by $A_{a,b}$ the affine half-space of $\mathbb{R}^d$ defined by $A_{a,b} = \{x \in \mathbb{R}^d : \langle a, x\rangle + b > 0\}$ and consider the associated VC class $\mathcal{A} = \{A_{a,b}\,;\, a \in \mathbb{R}^d, b \in \mathbb{R}\}$ which is of dimension $d+1$. Also introduce

$$\mathcal{A}(x_1^n) := \left\{ \left(\mathbb{1}_{\{x_1 \in A\}}, \ldots, \mathbb{1}_{\{x_n \in A\}} \right) ; A \in \mathcal{A} \right\} \subset \{0,1\}^n$$

equipped with the ℓ_1-norm $\|.\|_{1,n}$ defined by $\|\zeta\|_{1,n} = \sum_{i=1}^n |\zeta_i|/n$ for all $\zeta = (\zeta_1, \ldots, \zeta_n) \in \mathcal{A}(x_1^n)$.

Then, for all $\phi_{a,b} \in \mathcal{D}$, (2.23) implies that $\phi_{a,b} = \mathbb{1}_{A_{a,b}}$ and $\|\phi_{a,b}\| = \sqrt{\|\zeta_{a,b}\|_{1,n}}$ where $\zeta_{a,b} = (\mathbb{1}_{\{x_1 \in A_{a,b}\}}, \ldots, \mathbb{1}_{\{x_n \in A_{a,b}\}}) \in \mathcal{A}(x_1^n)$. Thus, we get that

$$H\left(\delta, \mathcal{D}, \|.\|\right) \leq H\left(\sqrt{\delta}, \mathcal{A}(x_1^n), \|.\|_{1,n}\right).$$

Moreover, we easily get from the upper bound of the entropy for a VC class of dimension $d+1$ provided by Haussler (1995) that

$$\int_0^1 \sqrt{H\left(\sqrt{\delta}, \mathcal{A}(x_1^n), \|.\|_{1,n}\right)}\, d\delta \leq C\sqrt{d+1},$$

where $C > 0$ is an absolute constant ($C \geq 22$ is convenient), hence the result. □

Proof of Theorem 2.3.4

Define $\varepsilon = \sigma/\sqrt{n}$. Consider the collection of ℓ_1-balls for $m \in \mathbb{N}^\star$,

$$S_m = \left\{s_m \in \mathcal{L}_1(\mathcal{D});\, \|s_m\|_{\mathcal{L}_1(\mathcal{D})} \leq m\varepsilon\right\}.$$

From Section 2.A, the Lasso $\hat{s}$ is a ρ-approximate penalized least squares estimator over the collection $\{S_m;\ m \in \mathbb{N}^\star\}$ for $\mathrm{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$. So, it only remains to determine a lower bound of λ such that $\mathrm{pen}(m)$ satisfies Condition (2.40) and to apply the conclusion of Theorem 2.A.1.

Let $s_m \in S_m$. For all $\tau > 0$, there exist coefficients $\alpha_{a,b}$ such that $s_m = \sum_{a,b} \alpha_{a,b}\, \phi_{a,b}$ and $\sum_{a,b} |\alpha_{a,b}| \leq m\varepsilon + \tau$. By linearity of W, we get that

$$W(s_m) = \sum_{a,b} \alpha_{a,b}\, W(\phi_{a,b}) \leq \sup_{a,b} |W(\phi_{a,b})| \sum_{a,b} |\alpha_{a,b}| \leq (m\varepsilon + \tau) \sup_{a,b} |W(\phi_{a,b})|.$$

Then, by Dudley's Criterion (Massart, 2007, Theorem 3.18), we have

$$\begin{aligned}\Delta_m := \mathbb{E}\left[\sup_{s_m \in S_m} W(s_m)\right] &\leq (m\varepsilon + \tau)\, \mathbb{E}\left[\sup_{a,b} |W(\phi_{a,b})|\right] \\ &\leq 12(m\varepsilon + \tau) \int_0^S \sqrt{H\left(\delta, \mathcal{D}, \|.\|\right)}\, d\delta,\end{aligned}$$

where $S^2 = \sup_{a,b} \mathbb{E}[W^2(\phi_{a,b})] = \sup_{a,b} \|\phi_{a,b}\|^2 = \sup_{a,b}(\sum_{i=1}^n \phi_{a,b}^2(x_i)/n) \leq 1$ from (2.23). So, $S \leq 1$ and we get from Lemma 2.B.1 that there exists $c > 0$ ($c \geq 264$ is convenient) such that

$$\Delta_m \leq 12(m\varepsilon + \tau) \int_0^1 \sqrt{H\left(\delta, \mathcal{D}, \|.\|\right)}\, d\delta \leq c(m\varepsilon + \tau)\sqrt{d+1}. \tag{2.61}$$

Now if we choose weights $x_m = cm$, then $\sum_{m \geq 1} \mathrm{e}^{-x_m} := \Sigma_c < +\infty$, and using the inequality

$2\sqrt{ab} \leq a+b$ we get from (2.61) that for all $K > 1$,

$$\begin{aligned}
2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right) &\leq 3K\varepsilon\left(\Delta_m + \varepsilon x_m\right) \\
&\leq 3K\varepsilon\left(c(m\varepsilon+\tau)\sqrt{d+1} + cm\varepsilon\right) \\
&\leq 3cK\varepsilon(m\varepsilon+\tau)\left(\sqrt{d+1}+1\right) \\
&\leq 3c(\sqrt{2}+1)K\varepsilon(m\varepsilon+\tau)\sqrt{d}.
\end{aligned}$$

Since this inequality is true for all $\tau > 0$, we get when τ tends to 0 that there exists $\kappa > 0$ ($\kappa = 3c(\sqrt{2}+1)K$) such that

$$2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right) \leq \kappa m\varepsilon^2\sqrt{d} \leq \lambda m\varepsilon$$

as soon as $\lambda \geq \kappa\varepsilon\sqrt{d}$. For such values of λ, Condition (2.40) on the penalty function is satisfied and me may apply Theorem 2.A.1 with $\mathrm{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$ for all $m \in \mathbb{N}^\star$. We end the proof similarly as the proof of Theorem 2.3.2. □

2.C Proofs of the rates of convergence

2.C.1 Convergence rates of the selected Lasso estimator for orthonormal dictionaries

2.C.1.1 Proof of the upper bound: Proposition 2.4.1

We know from Theorem 2.3.3 that the quadratic risk of the selected Lasso estimator $\hat{s}_{\hat{p}}$ is bounded by

$$\mathbb{E}\left[\|s-\hat{s}_{\hat{p}}\|^2\right] \leq C\left[\inf_{p\in\Lambda}\left\{\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2 + \lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\} + \mathrm{pen}(p)\right\} + \varepsilon^2\right], \tag{2.62}$$

where C is an absolute positive constant. Let us bound $\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2 + \lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}$ for all $p \in \Lambda$ thanks to the following lemma.

Lemma 2.C.1. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of the Hilbert space $\mathbb{H}$ and that there exist $1 < q < 2$, $r > 0$ and $R > 0$ such that $s \in w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. For all $p \in \mathbb{N}^\star$ and $\lambda > 0$, define*

$$s_{p,\lambda} := \operatorname*{arg\,min}_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2 + \lambda\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}.$$

Then, there exist $C_q > 0$ depending only on q and $C_r > 0$ depending only on r such that for all $p \in \mathbb{N}^\star$ and $\lambda > 0$,

$$\|s_{p,\lambda}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_q R^q \lambda^{1-q}$$

and

$$\|s - s_{p,\lambda}\|^2 \leq C_r R^2 p^{-2r} + C_q R^q \lambda^{2-q}.$$

The proof of Lemma 2.C.1 uses the two following easy calculations.

Lemma 2.C.2. *For all $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$ and $\delta > 0$,*

$$\sum_{j=1}^{p} a_j^2 \, \mathbb{1}_{\{|a_j| \leq \delta\}} \leq 2 \sum_{j=1}^{p} \int_0^{\delta} t \, \mathbb{1}_{\{|a_j| > t\}} \, dt.$$

Proof.

$$\begin{aligned}
& 2 \sum_{j=1}^{p} \int_0^{\delta} t \, \mathbb{1}_{\{|a_j| > t\}} \, dt \\
&= 2 \sum_{j=1}^{p} \left[\left(\int_0^{\delta} t \, \mathbb{1}_{\{|a_j| > t\}} \, dt \right) \mathbb{1}_{\{|a_j| > \delta\}} + \left(\int_0^{\delta} t \, \mathbb{1}_{\{|a_j| > t\}} \, dt \right) \mathbb{1}_{\{|a_j| \leq \delta\}} \right] \\
&= 2 \sum_{j=1}^{p} \left[\left(\int_0^{\delta} t \, dt \right) \mathbb{1}_{\{|a_j| > \delta\}} + \left(\int_0^{|a_j|} t \, dt \right) \mathbb{1}_{\{|a_j| \leq \delta\}} \right] \\
&= \sum_{j=1}^{p} \left(\delta^2 \, \mathbb{1}_{\{|a_j| > \delta\}} + a_j^2 \, \mathbb{1}_{\{|a_j| \leq \delta\}} \right) \\
&\geq \sum_{j=1}^{p} a_j^2 \, \mathbb{1}_{\{|a_j| \leq \delta\}}.
\end{aligned}$$

□

Lemma 2.C.3. *For all $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$ and $\delta > 0$,*

$$\sum_{j=1}^{p} |a_j| \mathbb{1}_{\{|a_j| > \delta\}} = \delta \sum_{j=1}^{p} \mathbb{1}_{\{|a_j| > \delta\}} + \sum_{j=1}^{p} \int_{\delta}^{+\infty} \mathbb{1}_{\{|a_j| > t\}} \, dt.$$

Proof.

$$\sum_{j=1}^{p} \int_{\delta}^{+\infty} \mathbb{1}_{\{|a_j| > t\}} \, dt = \sum_{j=1}^{p} \left(\int_{\delta}^{|a_j|} dt \right) \mathbb{1}_{\{|a_j| > \delta\}} = \sum_{j=1}^{p} \left(|a_j| - \delta \right) \mathbb{1}_{\{|a_j| > \delta\}}.$$

□

Proof of Lemma 2.C.1.

Let us denote by $\{\alpha_j^*\}_{j \in \mathbb{N}^\star}$ the coefficients of the target function s in the basis $\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^\star}$, so that

$s = \alpha^*.\phi = \sum_{j\in\mathbb{N}^\star} \alpha_j^* \phi_j$.
For all $p \in \mathbb{N}^\star$, set $A_p := \{\alpha = (\alpha_j)_{j\in\mathbb{N}^\star};\ \alpha_j \in \mathbb{R},\ \alpha_j = 0 \text{ for } j \geq p+1\}$.
Let $\lambda > 0$. Since $s_{p,\lambda} \in \mathcal{L}_1(\mathcal{D}_p)$, there exists $\alpha^{p,\lambda} \in A_p$ such that $s_{p,\lambda} = \alpha^{p,\lambda}.\phi$. Moreover, from (2.9) and the orthonormality of the basis $\mathcal{D}$,

$$\alpha^{p,\lambda} = \underset{\alpha\in A_p}{\arg\min} \left\{\|\alpha^*.\phi - \alpha.\phi\|^2 + \lambda\|\alpha\|_1\right\} = \underset{\alpha\in A_p}{\arg\min} \left\{\|\alpha^* - \alpha\|^2 + \lambda\|\alpha\|_1\right\}. \tag{2.63}$$

By calculating the subdifferential of the function $\alpha \in \mathbb{R}^p \mapsto \|\alpha^* - \alpha\|^2 + \lambda\|\alpha\|_1$, we get that the solution of the convex minimization problem (2.63) is $\alpha^{p,\lambda} = (\alpha_1^{p,\lambda}, \dots, \alpha_p^{p,\lambda}, 0, \dots, 0, \dots)$ where for all $j \in \{1, \dots, p\}$,

$$\alpha_j^{p,\lambda} = \begin{cases} \alpha_j^* - \lambda/2 & \text{if } \alpha_j^* > \lambda/2, \\ \alpha_j^* + \lambda/2 & \text{if } \alpha_j^* < -\lambda/2, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have

$$\begin{aligned} \|s - s_{p,\lambda}\|^2 &= \|\alpha^* - \alpha^{p,\lambda}\|^2 \\ &= \sum_{j=1}^{\infty} \left(\alpha_j^* - \alpha_j^{p,\lambda}\right)^2 \\ &= \sum_{j=p+1}^{\infty} \alpha_j^{*2} + \sum_{j=1}^{p} \alpha_j^{*2} \mathbb{1}_{\{|\alpha_j^*| \leq \lambda/2\}} + \sum_{j=1}^{p} \frac{\lambda^2}{4} \mathbb{1}_{\{|\alpha_j^*| > \lambda/2\}} \\ &\leq \underbrace{\sum_{j=p+1}^{\infty} \alpha_j^{*2}}_{(i)} + \underbrace{\sum_{j=1}^{p} \alpha_j^{*2} \mathbb{1}_{\{|\alpha_j^*| \leq \lambda/2\}}}_{(ii)} + \underbrace{\frac{\lambda}{2} \sum_{j=1}^{p} |\alpha_j^*| \mathbb{1}_{\{|\alpha_j^*| > \lambda/2\}}}_{(iii)}, \end{aligned} \tag{2.64}$$

while

$$\|s_{p,\lambda}\|_{\mathcal{L}_1(\mathcal{D}_p)} = \sum_{j=1}^{\infty} |\alpha_j^{p,\lambda}| = \sum_{j=1}^{p} \left(|\alpha_j^*| - \frac{\lambda}{2}\right) \mathbb{1}_{\{|\alpha_j^*| > \lambda/2\}} \leq \sum_{j=1}^{p} |\alpha_j^*| \mathbb{1}_{\{|\alpha_j^*| > \lambda/2\}} = (iii). \tag{2.65}$$

Now, s is assumed to belong to $\mathcal{B}_{2,\infty}^r(R)$, so (2.27) implies that (i) is bounded by

$$\sum_{j=p+1}^{\infty} \alpha_j^{*2} \leq R^2(p+1)^{-2r} \leq 2^{-2r} R^2 p^{-2r}. \tag{2.66}$$

Let us now bound (ii) and (iii) thanks to the assumption $s \in w\mathcal{L}_q(R)$. By applying Lemma 2.C.2 and Lemma 2.C.3 with $a_j = \alpha_j^*$ for all $j \in \{1, \dots, p\}$ and $\delta = \lambda/2$, and by using the fact that $\sum_{j=1}^{p} \mathbb{1}_{\{|\alpha_j^*| > t\}} \leq \sum_{j=1}^{\infty} \mathbb{1}_{\{|\alpha_j^*| > t\}} \leq R^q t^{-q}$ for all $t > 0$ if $s \in w\mathcal{L}_q(R)$, we get that (ii) is bounded

by

$$\sum_{j=1}^{p} \alpha_j^{*2} \mathbb{1}_{\{|\alpha_j^*| \leq \lambda/2\}} \leq \frac{2^{q-1}}{2-q} R^q \lambda^{2-q}, \tag{2.67}$$

while (iii) is bounded by

$$\sum_{j=1}^{p} |\alpha_j^*| \mathbb{1}_{\{|\alpha_j^*| > \lambda/2\}} \leq \frac{q\, 2^{q-1}}{q-1} R^q \lambda^{1-q}. \tag{2.68}$$

Gathering (2.65) and (2.68) on the one hand and (2.64), (2.66), (2.67) and (2.68) on the other hand, we get that there exists $C_q > 0$ depending only on q and $C_r > 0$ depending only on r such that $\|s_{p,\lambda}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_q R^q \lambda^{1-q}$ and $\|s - s_{p,\lambda}\|^2 \leq C_r R^2 p^{-2r} + C_q R^q \lambda^{2-q}$. □

Proof of Proposition 2.4.1

We deduce from Theorem 2.3.3 and Lemma 2.C.1 that there exists some constant $C_{q,r} > 0$ depending only on q and r such that the quadratic risk of $\hat{s}_{\hat{p}}$ is bounded by

$$\begin{aligned} \mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2\right] &\leq C_{q,r} \left[\inf_{p \in \Lambda} \left\{ R^2 p^{-2r} + R^q \left(\varepsilon \left(\sqrt{\ln p} + 1\right)\right)^{2-q} + \varepsilon^2 \ln p \right\} + \varepsilon^2 \right] \\ &\leq C_{q,r} \inf_{p \in \Lambda \setminus \{1\}} \left\{ R^2 p^{-2r} + R^q (\varepsilon \sqrt{\ln p})^{2-q} + \varepsilon^2 \ln p \right\}, \end{aligned} \tag{2.69}$$

where we use the fact that, for all $p \geq 2$, $\sqrt{\ln p} + 1 \leq (1 + 1/\sqrt{\ln 2})\sqrt{\ln p}$ and $\varepsilon^2 \leq \varepsilon^2 (\ln p)/\ln 2$. Now, we choose p such that the terms inside the infimum are of the same order. Denote by $\lceil x \rceil$ the smallest integer greater than x. Define $J_{q,r} = \lceil (2-q)(2r)^{-1} \log_2(R\varepsilon^{-1}) \rceil$ and $p_{q,r} = 2^{J_{q,r}}$. Since we have assumed $R\varepsilon^{-1} \geq \mathrm{e}$, then $p_{q,r} \in \Lambda \setminus \{1\}$ and we deduce from (2.69) that

$$\mathbb{E}\left[\|s - \hat{s}_{\hat{p}}\|^2\right] \leq C_{q,r} \left(R^2 p_{q,r}^{-2r} + R^q (\varepsilon \sqrt{\ln p_{q,r}})^{2-q} + \varepsilon^2 \ln p_{q,r} \right). \tag{2.70}$$

Now, let us give an upper bound of each term of the right-hand side of (2.70). From the fact that $2 \leq \mathrm{e} \leq R\varepsilon^{-1}$ and by definition of $p_{q,r}$, on the one hand we have $p_{q,r} \geq (R\varepsilon^{-1})^{(2-q)/(2r)}$, while on the other hand we have

$$\ln p_{q,r} \leq \ln 2 + \frac{2-q}{2r} \ln\left(R\varepsilon^{-1}\right) \leq \left(1 + \frac{2-q}{2r}\right) \ln\left(R\varepsilon^{-1}\right) := A_{q,r} \ln\left(R\varepsilon^{-1}\right)$$

where $A_{q,r} > 0$ depends only on q and r. Thus, we get that

$$R^2 p_{q,r}^{-2r} \leq R^2 \left(R\varepsilon^{-1}\right)^{q-2} = R^q \varepsilon^{2-q} \tag{2.71}$$

while

$$R^q \left(\varepsilon \sqrt{\ln p_{q,r}}\right)^{2-q} \leq {A_{q,r}}^{1-\frac{q}{2}} R^q \left(\varepsilon \sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q} \tag{2.72}$$

and

$$\varepsilon^2 \ln p_{q,r} \leq A_{q,r}\, \varepsilon^2 \ln\left(R\varepsilon^{-1}\right). \tag{2.73}$$

Now, these three bounds are upper bounded by $C_{q,r} R^q (\varepsilon\sqrt{\ln(R\varepsilon^{-1})})^{2-q}$ where $C_{q,r} > 0$ depends only on q and r. Indeed, $R\varepsilon^{-1} \geq \mathrm{e}$ and $2-q>0$, so (2.71) is bounded by $R^q(\varepsilon\sqrt{\ln(R\varepsilon^{-1})})^{2-q}$. Moreover, the right-hand side of (2.73) can be written $A_{q,r}\left(g((R\varepsilon^{-1})^2)\right)^{q/2} R^q(\varepsilon\sqrt{\ln(R\varepsilon^{-1})})^{2-q}$ with $g :]0,+\infty[\mapsto \mathbb{R},\ x\mapsto \ln(x)/(2x)$. Using the fact that $g(x^2)\leq 1/(2x)$ for all $x>0$ and that $R\varepsilon^{-1}\geq \mathrm{e}$, we get that (2.73) is bounded by $A_{q,r}(2\mathrm{e})^{-q/2}R^q(\varepsilon\sqrt{\ln(R\varepsilon^{-1})})^{2-q}$.

Then, we deduce from (2.70) that there exists $C_{q,r}>0$ depending only on q and r such that

$$\mathbb{E}\left[\|s-\hat{s}_{\hat{p}}\|^2\right] \leq C_{q,r}\, R^q \left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.$$

□

2.C.1.2 Proof of the lower bound : Proposition 2.4.2.

Define

$$M=\varepsilon\sqrt{\varsigma\ln\left(R\varepsilon^{-1}\right)},\quad J=\left\lfloor \frac{2-q}{2r}\,\log_2\left(RM^{-1}\right)\right\rfloor,\quad K=\left\lfloor q\log_2\left(RM^{-1}\right)\right\rfloor.$$

Set $p=2^J$ and $d=2^K$. Let us first check that M is well-defined and that $d\leq p$ under the assumptions of Proposition 2.4.2. Under the assumption $r<1/q-1/2$, we have $u>0$, and since $R\varepsilon^{-1}\geq \mathrm{e}^2\geq \mathrm{e}$, M is well-defined. Moreover, since $r<1/q-1/2$, we have $(2-q)/(2r)>q$, so it only remains to check that $RM^{-1}\geq \mathrm{e}$ to prove that $d\leq p$. We shall in fact prove the following stronger result:

Claim 2.C.1. *If* $R\varepsilon^{-1}\geq \max(\mathrm{e}^2,\varsigma^2)$*, then* $R\varepsilon^{-1}/\left(\ln(R\varepsilon^{-1})\right)\geq\varsigma$.

This result implies that, under the assumption $R\varepsilon^{-1}\geq\max(\mathrm{e}^2,\varsigma^2)$,

$$RM^{-1}=\frac{R\varepsilon^{-1}}{\sqrt{\varsigma\ln\left(R\varepsilon^{-1}\right)}}=\sqrt{R\varepsilon^{-1}}\sqrt{\frac{R\varepsilon^{-1}}{\varsigma\ln\left(R\varepsilon^{-1}\right)}}\geq \mathrm{e}\times 1\geq \mathrm{e}.$$

Let us prove Claim 2.C.1. Introduce the function $g :]0,+\infty[\mapsto\mathbb{R},\ x\mapsto x/\ln(x)$. It is easy to check that $g(x^2)\geq x$ for all $x>0$ and that g is non-decreasing on $[\mathrm{e},+\infty[$. Now, assume that $R\varepsilon^{-1}\geq\max(\mathrm{e}^2,\varsigma^2)$. Using the properties of g, we deduce that, if $\varsigma\geq\mathrm{e}$ then $R\varepsilon^{-1}\geq\varsigma^2\geq\mathrm{e}^2\geq\mathrm{e}$ and $R\varepsilon^{-1}/(\ln(R\varepsilon^{-1}))=g(R\varepsilon^{-1})\geq g(\varsigma^2)\geq\varsigma$, while if $\varsigma<\mathrm{e}$ then $R\varepsilon^{-1}\geq\mathrm{e}^2\geq\mathrm{e}$ and $R\varepsilon^{-1}/(\ln(R\varepsilon^{-1}))=g(R\varepsilon^{-1})\geq g(\mathrm{e}^2)\geq\mathrm{e}>\varsigma$, hence Claim 2.C.1.

Now, consider the hypercube $A(p,d,M)$ defined by

$$\left\{\sum_{j=1}^{\infty}\alpha_j\,\phi_j;\ (\alpha_1,\dots,\alpha_p)\in[0,M]^p,\ \alpha_j=0 \text{ for } j\geq p+1,\ \sum_{j=1}^{p}\mathbb{1}_{\{\alpha_j\neq 0\}}=d\right\}$$
$$=\left\{M\sum_{j=1}^{\infty}\beta_j\,\phi_j;\ (\beta_1,\dots,\beta_p)\in[0,1]^p,\ \beta_j=0 \text{ for } j\geq p+1,\ \sum_{j=1}^{p}\mathbb{1}_{\{\beta_j\neq 0\}}=d\right\}.$$

The essence of the proof is just to check that $A(p,d,M)\subset\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$. This shall enable us to bound from below the minimax risk over $\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$ by the lower bound of the minimax risk over $A(p,d,M)$ provided by Birgé and Massart (2001).

Let $u\in A(p,d,M)$. Write $u=\sum_{j=1}^{\infty}\alpha_j\phi_j=M\sum_{j=1}^{\infty}\beta_j\phi_j$.

$$\sum_{j=1}^{\infty}|\alpha_j|^q=M^q\sum_{j=1}^{p}\beta_j^q\,\mathbb{1}_{\{\beta_j\neq 0\}}\leq M^q\sum_{j=1}^{p}\mathbb{1}_{\{\beta_j\neq 0\}}\leq M^q d\leq M^q\left(RM^{-1}\right)^q\leq R^q.$$

Thus, $u\in\mathcal{L}_q(R)$.
Let $J_0\in\mathbb{N}^\star$. If $J_0>p$, then

$$J_0^{2r}\sum_{j=J_0}^{\infty}\alpha_j^2\leq J_0^{2r}\sum_{j=p+1}^{\infty}\alpha_j^2=0\leq R^2.$$

If $J_0\leq p$, then

$$J_0^{2r}\sum_{j=J_0}^{\infty}\alpha_j^2=J_0^{2r}M^2\sum_{j=J_0}^{p}\beta_j^2\mathbb{1}_{\{\beta_j\neq 0\}}\leq J_0^{2r}M^2\sum_{j=J_0}^{p}\mathbb{1}_{\{\beta_j\neq 0\}}\leq p^{2r}M^2 d\leq R^2.$$

Thus, $u\in\mathcal{B}^r_{2,\infty}(R)$.
Therefore, $A(p,d,M)\subset\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$ and

$$\inf_{\tilde{s}}\sup_{s\in\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)}\mathbb{E}\left[\|s-\tilde{s}\|^2\right]\geq\inf_{\tilde{s}}\sup_{s\in A(p,d,M)}\mathbb{E}\left[\|s-\tilde{s}\|^2\right]. \tag{2.74}$$

Now, from Birgé and Massart (2001, Theorem 5), the minimax risk over $A\left(p,d,M\right)$ satisfies

$$\begin{aligned}\inf_{\tilde{s}}\sup_{s\in A(p,d,M)}\mathbb{E}\left[\|s-\tilde{s}\|^2\right]&\geq\kappa d\min\left\{M^2,\varepsilon^2\left(1+\ln\left(\frac{p}{d}\right)\right)\right\}\\&\geq\kappa\frac{\left(RM^{-1}\right)^q}{2}\min\left\{M^2,\varepsilon^2\left(1+\ln\left(\frac{p}{d}\right)\right)\right\}\end{aligned} \tag{2.75}$$

where $\kappa>0$ denotes some absolute constant.

Moreover, we have

$$\begin{aligned}\varepsilon^2\left(1+\ln\left(\frac{p}{d}\right)\right) &\geq \varepsilon^2\left(1+\ln\left[\frac{\left(RM^{-1}\right)^{\frac{2-q}{2r}}}{2\left(RM^{-1}\right)^q}\right]\right)\\ &\geq \varepsilon^2\ln\left[\left(RM^{-1}\right)^{\varsigma}\right]\\ &= \varepsilon^2\ln\left[\left(R\varepsilon^{-1}\right)^{\varsigma}\left(\varepsilon M^{-1}\right)^{\varsigma}\right]\\ &= M^2+\varepsilon^2\ln\left[\left(\varepsilon M^{-1}\right)^{\varsigma}\right]\\ &= M^2-\frac{\varsigma}{2}\varepsilon^2\ln\left[\varsigma\ln\left(R\varepsilon^{-1}\right)\right]. \end{aligned} \tag{2.76}$$

But the assumption $R\varepsilon^{-1}\geq\max(\mathrm{e}^2,\varsigma^2)$ implies that (2.76) is greater than $M^2/2$. Indeed, first note that

$$M^2-\frac{\varsigma}{2}\varepsilon^2\ln\left[\varsigma\ln\left(R\varepsilon^{-1}\right)\right]\geq M^2/2 \;\Leftrightarrow\; \frac{R\varepsilon^{-1}}{\ln(R\varepsilon^{-1})}\geq\varsigma, \tag{2.77}$$

and then apply Claim 2.C.1. Thus, we deduce from (2.74), (2.75), (2.76) and (2.77) that there exists some $\kappa>0$ such that

$$\inf_{\tilde{s}}\sup_{s\in\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)}\mathbb{E}\left[\|s-\tilde{s}\|^2\right]\geq\kappa R^qM^{2-q}=\kappa\varsigma^{1-\frac{q}{2}}R^q\left(\varepsilon\sqrt{\ln(R\varepsilon^{-1})}\right)^{2-q}.$$

□

2.C.2 Rates of convergence of the selected Lasso estimator in the general case

Sketch of the proof of Proposition 2.4.4

Proposition 2.4.4 is deduced from the oracle inequality (2.22) in Theorem 2.3.3. First, the proof consists in bounding $\inf_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2+\lambda_p\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}$ for all $p\in\mathbb{N}^\star$, just as it is done in Lemma 2.C.1 in the orthonormal case. This first step is very similar to Corollary 3.7 in Barron et al. (2008). Then, an additional step is needed to adapt the truncation of the dictionary according to the unknown parameters of smoothness q and r of the target function. This second step is similar to the proof of Proposition 2.4.1 in the orthonormal case. We refer the interested reader to Massart and Meynet (2010, proof of Proposition 5.7) for a detailed proof of Proposition 2.4.4. □

2.C.3 Rates of convergence of the Lasso for neural networks

Sketch of the proof of Proposition 2.4.6

The rates of convergence (2.36) can easily be deduced from Theorem 2.3.4 by first checking that the ℓ_1-penalized risk of the deterministic Lasso $\inf_{t\in\mathcal{L}_1(\mathcal{D})}\left\{\|s-t\|^2+\lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\}$ is linked to the

$K_{\mathcal{D}}$-functional of s by the following relation:

$$\inf_{t\in\mathcal{L}_1(\mathcal{D})}\left\{\|s-t\|^2+\lambda\|t\|_{\mathcal{L}_1(\mathcal{D})}\right\}\leq\inf_{\delta>0}\left\{K^2_{\mathcal{D}}(s,\delta)+\frac{\lambda^2}{4\delta^2}\right\}, \tag{2.78}$$

and then by bounding this last quantity thanks to (2.35) when s belongs to $\mathcal{B}_q(R)$, which leads to an upper bound of the right-hand side of (2.78) of order $R^q\lambda^{2-q}+\lambda\sigma/\sqrt{n}$. We finally get (2.36) by taking into account the assumption on R. □

2.D Interpolation spaces

Proof of Remark 4.
Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of the Hilbert space $\mathbb{H}$ and that there exist $1<q<2$, $r>0$ and $R>0$ such that $s\in w\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$. For all $p\in\mathbb{N}^\star$ and $\lambda>0$, define

$$s_{p,\lambda}:=\operatorname*{arg\,min}_{t\in\mathcal{L}_1(\mathcal{D}_p)}\left\{\|s-t\|^2+\lambda\|t\|_{\mathcal{L}_1(\mathcal{D}_p)}\right\}.$$

The proof is divided in two main parts. First, we choose λ such that $s_{p,\lambda}\in\mathcal{L}_{1,r}$. Secondly, we choose p such that $\|s-s_{p,\lambda}\|+\delta\|s_{p,\lambda}\|_{\mathcal{L}_{1,r}}\leq C_{q,r}R\,\delta^{2\nu}$ for all $\delta>0$, some $C_{q,r}>0$ and $\nu=1/q-1/2$, which means that $s\in\mathcal{B}_{q,r}(C_{q,r}R)$.

Let us first choose λ such that $s_{p,\lambda}\in\mathcal{L}_{1,r}$. From Lemma 2.C.1, we have

$$\|s-s_{p,\lambda}\|\leq\sqrt{C_rR^2p^{-2r}+C_qR^q\lambda^{2-q}}\leq\sqrt{C_r}\,Rp^{-r}+\sqrt{C_q}\,R^{q/2}\lambda^{1-q/2}.$$

Now choose λ such that $\sqrt{C_r}\,Rp^{-r}=\sqrt{C_q}\,R^{q/2}\lambda^{1-q/2}$, that is to say

$$\lambda_p:=R\left(\sqrt{C_qC_r^{-1}}\,p^r\right)^{-\frac{2}{2-q}}. \tag{2.79}$$

Then, we have

$$\|s-s_{p,\lambda_p}\|\leq2\sqrt{C_r}\,Rp^{-r}. \tag{2.80}$$

Let us now check that $s_{p,\lambda_p}\in\mathcal{L}_{1,r}$. Define

$$C_p:=\max\left\{4\sqrt{C_r}\,R,\max_{p'\in\mathbb{N}^\star,\,p'\leq p}\|s_{p',\lambda_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})}\right\}. \tag{2.81}$$

Let $p'\in\mathbb{N}^\star$. By definition of $s_{p',\lambda_{p'}}$, we have $s_{p',\lambda_{p'}}\in\mathcal{L}_1(\mathcal{D}_{p'})$. If $p'\leq p$, then we deduce from

(2.80) and (2.81) that

$$\|s_{p,\lambda_p} - s_{p',\lambda_{p'}}\| \leq \|s_{p,\lambda_p} - s\| + \|s - s_{p',\lambda_{p'}}\| \leq 2\sqrt{C_r}\, R\left(p^{-r} + p'^{-r}\right) \leq C_p p'^{-r},$$

and we have $\|s_{p',\lambda_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq C_p$ by definition of C_p. If $p' > p$, then $\mathcal{L}_1(\mathcal{D}_p) \subset \mathcal{L}_1(\mathcal{D}_{p'})$ and $s_{p,\lambda_p} \in \mathcal{L}_1(\mathcal{D}_{p'})$ with $\|s_{p,\lambda_p}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq \|s_{p,\lambda_p}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_p$ and $\|s_{p,\lambda_p} - s_{p,\lambda_p}\| = 0 \leq C_p p'^{-r}$. So, $s_{p,\lambda_p} \in \mathcal{L}_{1,r}$.

Now, it only remains to choose a convenient $p \in \mathbb{N}^\star$ to prove that $s \in \mathcal{B}_{q,r}(C_{q,r}R)$ for some $C_{q,r}$. Let us first give an upper bound of $\|s_{p,\lambda_p}\|_{\mathcal{L}_{1,r}}$ for all $p \in \mathbb{N}^\star$. By definition of $\|s_{p,\lambda_p}\|_{\mathcal{L}_{1,r}}$ and the above upper bounds, we have $\|s_{p,\lambda_p}\|_{\mathcal{L}_{1,r}} \leq C_p$. So, we just have to bound C_p. Let $p' \in \mathbb{N}^\star, p' \leq p$. From Lemma 2.C.1, there exists $C_q > 0$ depending only on q such that $\|s_{p',\lambda_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq C_q R^q \lambda_{p'}^{1-q}$. So, we get from (2.79) that

$$\begin{aligned}
\|s_{p',\lambda_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} &\leq C_q R \left(\sqrt{C_q C_r^{-1}}\, p'^r\right)^{\frac{2(q-1)}{2-q}} \\
&\leq C_q R \left(\sqrt{C_q C_r^{-1}} p^r\right)^{\frac{2(q-1)}{2-q}} \\
&= C_q^{\frac{1}{2-q}} C_r^{-\frac{(q-1)}{2-q}} R p^{\frac{2(q-1)r}{2-q}},
\end{aligned}$$

and we deduce from (2.81) that

$$C_p \leq \max\left\{4\sqrt{C_r}\, R, C_q^{\frac{1}{2-q}} C_r^{-\frac{(q-1)}{2-q}} R p^{\frac{2(q-1)r}{2-q}}\right\} \leq C_{q,r} R p^{\frac{2(q-1)r}{2-q}}$$

where $C_{q,r} > 0$ depends only on q and r. Thus, we have

$$\|s_{p,\lambda_p}\|_{\mathcal{L}_{1,r}} \leq C_{q,r} R p^{\frac{2(q-1)r}{2-q}}. \tag{2.82}$$

Then, we deduce from (2.80) and (2.82) that for all $p \in \mathbb{N}^\star$ and $\delta > 0$,

$$\begin{aligned}
\inf_{t \in \mathcal{L}_{1,r}} \left\{\|s - t\| + \delta \|t\|_{\mathcal{L}_{1,r}}\right\} &\leq \|s - s_{p,\lambda_p}\| + \delta \|s_{p,\lambda_p}\|_{\mathcal{L}_{1,r}} \\
&\leq 2\sqrt{C_r}\, R p^{-r} + \delta C_{q,r} R p^{\frac{2(q-1)r}{2-q}}.
\end{aligned} \tag{2.83}$$

Now, we choose p such that p^{-r} and $\delta\, p^{\frac{2(q-1)r}{2-q}}$ are of the same order. More precisely, set $p = 2^J$ where $J = \left\lceil (2-q)(qr)^{-1} \log_2(\delta^{-1})\right\rceil$. With this value of p, we get that there exists $C'_{q,r} > 0$ depending only on q and r such that (2.83) is upper bounded by $C'_{q,r} R\, \delta^{(2-q)/q} = C'_{q,r} R\, \delta^{2\nu}$. This means that $s \in \mathcal{B}_{q,r}(C'_{q,r} R)$, hence (2.33). □

Chapter 3

Finite mixture Gaussian regression models

Contents

3.1. Introduction **95**

3.2. Notations and framework **98**

3.2.1. The models 98

3.2.2. Boundedness assumption on the mixture parameters 99

3.2.3. The Lasso estimator 99

3.3. An ℓ_1-oracle inequality for the Lasso **100**

3.A. Proof of Theorem 3.3.1 . . . . . **102**

3.A.1. Statement of the main results 102

3.A.2. Proof of the main results 104

3.B. Proofs of the lemmas **112**

3.B.1. Proof of Lemma 3.A.3 112

3.B.2. Proof of Lemmas 3.B.4, 3.B.5 and 3.B.6 115

ABSTRACT

We consider a finite mixture of Gaussian regressions for high-dimensional heterogeneous data where the number of covariates may be much larger than the sample size. We estimate the unknown conditional mixture density by an ℓ_1-penalized maximum likelihood estimator. We provide an ℓ_1-oracle inequality satisfied by this Lasso estimator with the Kullback-Leibler loss. In particular, we give a condition on the regularization parameter of the Lasso to obtain such an oracle inequality.

Our aim is twofold: to extend the ℓ_1-oracle inequality established in Chapter 2 in the homogeneous Gaussian linear regression case, and to present a complementary result to Städler et al. (2010) by studying the Lasso for its ℓ_1-regularization properties rather than for its variable selection properties.

We deduce our oracle inequality from a finite mixture Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation that we establish by following Vapnik's structural risk minimization method (Vapnik, 1982) and theory on model selection for maximum likelihood estimators (Massart, 2007).

NOTA: This chapter is submitted to ESAIM Probability and Statistics.

3.1 Introduction

In applied statistics, a tremendous number of applications deal with relating a random response variable $Y \in \mathbb{R}$ to a set of explanatory variables or covariates $\boldsymbol{X} \in \mathbb{R}^p$ through a regression-type model. As a consequence, linear regression is one of the most studied fields in statistics. Due to computer progress and development of state of the art technologies such as DNA microarrays, we are faced with high-dimensional data where the number of variables can be much larger than the sample size. To solve this problem, the sparsity scenario – which consists in assuming that the coefficients of the high-dimensional vector of covariates are mostly 0 – has been widely studied. These last years, a great deal of attention has been focused on the ℓ_1-penalized least squares estimator of parameters, called the Lasso (Tibshirani, 1996). This interest has been motivated by the geometric properties of the ℓ_1-norm: ℓ_1-penalization tends to produce sparse solutions and can be thus used as a convex surrogate for the non-convex ℓ_0-penalization. Thus, the Lasso has essentially been developed for sparse recovery based on convex optimization. In this sparsity approach, many results, such as ℓ_0-oracle inequalities, have been proved to study the performance of this estimator as a variable selection procedure (Bickel et al., 2009; van de Geer, 2008; Koltchinskii, 2009). Nonetheless, all these results need strong restrictive eigenvalue assumptions on the Gram matrix that can be far from being fulfilled in practice (see

Bühlmann and van de Geer, 2009, for an overview of these assumptions). In parallel, a few results on the performance of the Lasso for its ℓ_1-regularization properties have been established (Bartlett et al., 2012; Huang et al., 2008; Massart and Meynet, 2011; Rigollet and Tsybakov, 2011). In particular, in Chapter 2, we provided an ℓ_1-oracle inequality for the Lasso in the framework of fixed design Gaussian regression (see Theorem 2.3.2). Contrary to the ℓ_0-results that require strong assumptions on the regressors, our ℓ_1-result is valid with no assumption at all.

In linear regression, the homogeneity assumption that the regression coefficients are the same for different observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ is often inadequate and restrictive. It seems all the more true for the case of high-dimensional data where the number p of covariates can be much larger than the number n of observations: at least a fraction of covariates may exhibit a different influence on the response among various observations (i.e. subpopulations) and parameters may change for different subgroups of observations. Thus, addressing the issue of heterogeneity in high-dimensional data is important in many practical applications. For instance, Städler et al. (2010) prove that substantial prediction improvements are possible by incorporating a heterogeneity structure to the model. Such heterogeneity can be modeled by a finite mixture of regressions: if $s(\cdot|\boldsymbol{X})$ is the density of Y conditionally to the covariate $\boldsymbol{X}$, we can look for an estimator of s as a finite mixture of regressions. Here, we restrict to the important case of Gaussian models. Thus, we consider models defined by a mixture of K Gaussian densities,

$$s_{\boldsymbol{\theta}}(Y_i \mid \boldsymbol{X_i} = \boldsymbol{x_i}) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(Y_i - \boldsymbol{\beta_k^T x_i}\right)^2}{2\sigma_k^2}\right)$$

with parameters $\boldsymbol{\theta} = (\pi_k, \beta_{kj}, \sigma_k)_{1 \le k \le K, 1 \le j \le p}$.

During the last years, both theoretical studies and experiments have been carried out on finite mixture regression models (Young and Hunter, 2006; Grün et al., 2007; Städler et al., 2010; Beninel et al., 2012). For instance, Städler et al. (2010) consider an ℓ_1-penalization approach by introducing a Lasso estimator to estimate the density s:

$$\hat{s}(\lambda) = \arg\min_{s_{\boldsymbol{\theta}}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_i)\right) + \lambda |s_{\boldsymbol{\theta}}|_1 \right\}, \tag{3.1}$$

where $|s_{\boldsymbol{\theta}}|_1 = \sum_{j=1}^{p} \sum_{k=1}^{K} |\beta_{kj}|$ is the ℓ_1-norm of the mean parameters only. In a sparsity viewpoint, Städler et al. (2010) provide an ℓ_0-oracle inequality satisfied by this Lasso estimator. Their oracle inequality is based on the same restricted eigenvalue conditions used in the homogeneous linear regression case recalled in Section 1.2.4. Moreover, the $\log$-likelihood function used for maximum likelihood estimation requires additional mathematical arguments in comparison to the quadratic loss used in the homogeneous linear regression case. In particular, Städler et al. (2010) have to introduce

some margin assumptions to link the Kullback-Leibler loss function to the ℓ_2-norm of the parameters and get optimal rates of convergence of order $|s_{\boldsymbol{\theta}}|_0/n$.

In this chapter, we propose another approach that does not take into account sparsity. We rather study the performance of the Lasso estimator in the framework of finite mixture Gaussian regression models for its ℓ_1-regularization properties, thus extending the results presented in Chapter 2 for homogeneous Gaussian linear regression models. As in Chapter 2, we restrict to the fixed design case, that is to say to non-random regressors. We provide an ℓ_1-oracle inequality satisfied by the Lasso with no assumption, neither on the regressors, nor on the margin. This can be achieved thanks to the fact that we are only looking for rates of convergence of order $|s_{\boldsymbol{\theta}}|_1/\sqrt{n}$ rather than $|s_{\boldsymbol{\theta}}|_0/n$. We give a lower bound of the regularization parameter λ of the Lasso in (3.1) to guarantee such an oracle inequality:

$$\lambda \geq CK\,(\ln n)^2 \sqrt{\frac{\ln(2p+1)}{n}} \tag{3.2}$$

where C is a positive quantity depending on the mixture parameters and on the regressors, whose value is specified at (3.7).

Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates can grow with respect to n and can be much larger than n. The number K of clusters in the mixture is fixed. We pay a great attention to obtain a lower bound (3.2) of λ with optimal dependence on p, that is to say $\sqrt{\ln(2p+1)}$ just as for homogeneous Gaussian linear regression in Chapter 2.

Our oracle inequality is deduced from a finite mixture Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation that we establish by following both Vapnik's structural risk minimization method (Vapnik, 1982) and the theory around model selection (Cohen and Pennec, 2011; Massart, 2007). Just as in Chapter 2, the key idea that enables us to deduce our ℓ_1-oracle inequality from such a model selection theorem is to view the Lasso as the solution of a penalized maximum likelihood model selection procedure over a countable collection of ℓ_1-ball models.

The chapter is organized as follows. The notations and the framework are introduced in Section 3.2. In Section 3.3, we state the main result of this chapter, which is an ℓ_1-oracle inequality satisfied by the Lasso in finite mixture Gaussian regression models. Section 3.A is devoted to the proof of this oracle inequality. In particular, we state and prove the general finite mixture Gaussian regression model selection theorem from which this result is derived. Finally, some lemmas are proved in Section 3.B.

3.2 Notations and framework

3.2.1 The models

Our statistical framework is a finite mixture of Gaussian regressions for high-dimensional data where the number of covariates can be much larger than the sample size. We observe n independent couples $((\boldsymbol{x_i}, Y_i))_{1\leq i\leq n}$ of variables. We are interested in estimating the law of the random variable $Y_i \in \mathbb{R}$ conditionally to the fixed one $\boldsymbol{x_i} \in \mathbb{R}^p$. We assume that, conditionally to the $\boldsymbol{x_i}$s, the Y_is are independent identically distributed with density $s(\cdot|\boldsymbol{x_i})$ which is a finite mixture of K Gaussian densities:

$$
\begin{aligned}
&Y_i \mid \boldsymbol{x_i} \text{ independent}\\
&Y_i \mid \boldsymbol{x_i} = \boldsymbol{x} \ \sim \ s(y|\boldsymbol{x})dy\\
&s(y|\boldsymbol{x}) = s_{\boldsymbol{\theta^*}}(y|\boldsymbol{x}) = \sum_{k=1}^{K} \frac{\pi_k^*}{\sqrt{2\pi}\sigma_k^*} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k^*}^T\boldsymbol{x}\right)^2}{2\sigma_k^{*2}}\right),\\
&\boldsymbol{\theta^*} = (\pi_1^*, \ldots, \pi_K^*, \boldsymbol{\beta_1^*}, \ldots, \boldsymbol{\beta_K^*}, \sigma_1^*, \ldots, \sigma_K^*) \in \left(\Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+^K\right),\\
&\Pi_K = \left\{(\pi_1, \ldots, \pi_K)\,;\ \pi_k > 0 \text{ for } k \in \{1, \ldots, K\} \text{ and } \sum_{k=1}^{K} \pi_k = 1\right\}.
\end{aligned}
$$

We want to estimate the conditional density function s from the observations. To this aim, we consider a collection of finite mixture Gaussian regression models as follows:

$$
\begin{aligned}
&s_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k}^T\boldsymbol{x}\right)^2}{2\sigma_k^2}\right),\\
&\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_K}, \sigma_1, \ldots, \sigma_K) \in \left(\Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+^K\right).
\end{aligned}
$$

For all $k \in \{1, \ldots, K\}$, $\boldsymbol{\beta_k}$ is the vector of regression coefficients and σ_k is the standard deviation in mixture component k. The π_ks are the mixture proportions.

For all $\boldsymbol{x} \in \mathbb{R}^p$, we define the parameter $\boldsymbol{\theta}(\boldsymbol{x})$ of the conditional density $s_{\boldsymbol{\theta}}(.|\boldsymbol{x})$ by

$$
\boldsymbol{\theta}(\boldsymbol{x}) = \left(\pi_1, \ldots, \pi_K, \boldsymbol{\beta_1}^T\boldsymbol{x}, \ldots, \boldsymbol{\beta_K}^T\boldsymbol{x}, \sigma_1, \ldots, \sigma_K\right) \in \mathbb{R}^{3K}.
$$

For all $k \in \{1, \ldots, K\}$, $\boldsymbol{\beta_k}^T\boldsymbol{x}$ is the mean coefficient of the mixture component k for the conditional density $s_{\boldsymbol{\theta}}(.|\boldsymbol{x})$.

Since we work conditionally to the covariates $(\boldsymbol{x_i})_{1\leq i\leq n}$, our results are expressed with quantities

depending on them. In particular, we set

$$\|\boldsymbol{x}\|_{\max,n} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\max_{j=1,\dots,p} x_{ij}^2}\,. \tag{3.3}$$

3.2.2 Boundedness assumption on the mixture parameters

For technical reasons, we restrict our study to bounded parameter vectors $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\beta_k}, \sigma_k)_{1\le k\le K}$. We assume that there exist deterministic positive quantities a_π, a_β, A_β, a_σ and A_σ such that the parameter vectors belong to the bounded space

$$\Theta = \left\{ \begin{array}{l} \boldsymbol{\theta} = (\pi_1, \dots, \pi_K, \boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_K}, \sigma_1, \dots, \sigma_K)\,;\ \forall\, k \in \{1, \dots, K\} : \\ a_\pi \le \pi_k,\ a_\beta \le \inf_{\boldsymbol{x}\in\mathbb{R}^p} \left|\boldsymbol{\beta_k}^T\boldsymbol{x}\right| \le \sup_{\boldsymbol{x}\in\mathbb{R}^p} \left|\boldsymbol{\beta_k}^T\boldsymbol{x}\right| \le A_\beta,\ a_\sigma \le \sigma_k \le A_\sigma \end{array} \right\}. \tag{3.4}$$

We denote by S the set of conditional densities $s_{\boldsymbol{\theta}}$ in this model:

$$S = \left\{ s_{\boldsymbol{\theta}}\,;\ \boldsymbol{\theta} \in \left(\Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+^K\right) \cap \Theta \right\}.$$

Besides, we assume that the true density $s = s_{\boldsymbol{\theta}^*}$ belongs to S, that is to say that the true parameter $\boldsymbol{\theta}^*$ belongs to the bounded space Θ.

3.2.3 The Lasso estimator

In a maximum likelihood approach, the loss function taken into consideration is the Kullback-Leibler information, which is defined for a density t by

$$\mathrm{KL}(s,t) = \int_{\mathbb{R}} \ln\left(\frac{s(y)}{t(y)}\right) s(y)\, dy$$

if sdy is absolutely continuous with respect to tdy and $+\infty$ otherwise. Since we work with conditional densities, we define the following adapted Kullback-Leibler information $\mathrm{KL_n}$ that takes into account the structure of conditional densities: for fixed covariates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, we consider the average loss function

$$\mathrm{KL_n}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(s(\cdot|\boldsymbol{x_i}), t(\cdot|\boldsymbol{x_i})) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \ln\left(\frac{s(y|\boldsymbol{x_i})}{t(y|\boldsymbol{x_i})}\right) s(y|\boldsymbol{x_i})\, dy. \tag{3.5}$$

The maximum likelihood approach suggests to estimate s by the conditional density $s_{\boldsymbol{\theta}}$ that maximizes the likelihood conditionally to $(\boldsymbol{x_i})_{1\le i\le n}$, or equivalently that minimizes the empirical contrast which is $-\sum_{i=1}^{n}\ln(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i}))/n$. Choosing such an estimator $s_{\boldsymbol{\theta}}$ boils down to choosing a parameter $\boldsymbol{\theta} = (\pi_k, \beta_{kj}, \sigma_k)_{1\le k\le K, 1\le j\le p}$ minimizing the empirical contrast. It requires the estimation of

$pK + 2K$ coefficients. But we are mostly interested in high-dimensional data with $p \gg n$, and thus $pK + 2K \gg n$. For such high-dimensional problems, we have to regularize the maximum likelihood estimator in order to obtain reasonably accurate estimates. Since there are only K proportion coefficients π_k and K standard deviation coefficients σ_k with $K \ll n$, no regularization on these parameters is necessary. On the opposite, there are Kp mean parameters β_{kj} with $Kp \gg n$, so these parameters must be regularized. Here, we consider ℓ_1-regularization and its associated so-called Lasso estimator which is the ℓ_1-norm penalized maximum likelihood estimator defined by

$$\hat{s}(\lambda) := \underset{s_{\boldsymbol{\theta}} \in S}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i})\right) + \lambda |s_{\boldsymbol{\theta}}|_1 \right\}, \tag{3.6}$$

where $\lambda > 0$ is a regularization parameter and $|s_{\boldsymbol{\theta}}|_1 := \sum_{j=1}^{p} \sum_{k=1}^{K} |\beta_{kj}|$ for $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\beta_k}, \sigma_k)_{1 \le k \le K}$.

3.3 An ℓ_1-oracle inequality for the Lasso

Here, we state the main result of this chapter. Theorem 3.3.1 provides an ℓ_1-oracle inequality for the Lasso in finite mixture Gaussian regression models.

Theorem 3.3.1. *Denote $a \wedge b = \min\{a, b\}$. Assume that*

$$\lambda \ge \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\beta^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)}\right) \tag{3.7}$$

for some absolute constant $\kappa \ge 360$. Then, there exists an absolute positive constant κ' such that the Lasso estimator $\hat{s}(\lambda)$ defined by (3.6) *satisfies the following ℓ_1-oracle inequality:*

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL_n}\left(s, \hat{s}(\lambda)\right)\right] \le (1+\kappa^{-1}) \inf_{s_{\boldsymbol{\theta}} \in S} \{\mathrm{KL_n}(s, s_{\boldsymbol{\theta}}) + \lambda |s_{\boldsymbol{\theta}}|_1\} + \lambda \\ + \frac{\kappa' \sqrt{K}}{\sqrt{n}} \left[K \frac{\left(1 + (A_\beta + A_\sigma)^2\right)}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\beta^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) + \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{A_\beta^2}{a_\sigma^2} \right].\end{aligned} \tag{3.8}$$

Proof. Page 112. We have not looked for optimizing the constants in Theorem 3.3.1. Thus, we do not explicit the value of κ' and the lower bound on κ is sufficient but not optimal. □

Let us make a few comments on this result.

Theorem 3.3.1 provides information about the performance of the Lasso as an ℓ_1-regularization algorithm. It highlights the fact that, provided that the regularization parameter λ is properly chosen, the Lasso estimator, which is the solution of the ℓ_1-penalized empirical risk minimization problem,

behaves as well as the deterministic Lasso, that is to say the solution of the ℓ_1-penalized true risk minimization problem, up to an error term of order $(\ln n)^2/\sqrt{n}$. This ℓ_1-result is complementary to the ℓ_0-oracle inequality established by Städler et al. (2010) whose is rather stated in a sparsity approach looking at the Lasso as a variable selection procedure.

Let us stress that we present here an ℓ_1-oracle inequality with no assumption, neither on the regressors, nor on the margin. This represents a great advantage compared to the ℓ_0-oracle inequality in Städler et al. (2010) which requires some restricted eigenvalue conditions. Besides, Städler et al.'s result needs some margin assumptions. Although one may prove that these margin assumptions are actually fulfilled for some undetermined quantities thanks to theoretical arguments such as continuity or differentiability of the functions into consideration, it seems difficult to calculate explicit values of these quantities. Thus, one has no idea of the concrete values of these quantities. Yet, the ℓ_0-oracle inequality established by Städler et al. (2010) strongly depends on these unknown quantities. Therefore, their result is hardly interpretable. On the contrary, the only assumption used to establish Theorem 3.3.1 is the boundedness of the mixture parameters, which is anyway also assumed by Städler et al. (2010) and which is quite usual when working with maximum likelihood estimation (Baudry, 2009; Maugis and Michel, 2011b), at least to tackle the problem of the unboundedness of the likelihood at the boundary of the parameter space (Redner and Walker, 1984; McLachlan and Peel, 2000) and to prevent it from divergence. In fact, Städler et al. (2010) must make their eigenvalue assumption so as to bound the ℓ_2-norm of the parameter vector on its support and they add assumptions on the margin in order to link the loss function to the ℓ_2-norm of the parameters and get optimal rates of convergence $|s_{\boldsymbol{\theta}}|_0/n$ in a sparsity viewpoint. On the opposite, since we focus on an ℓ_1-regularization approach, we are just looking for rates of convergence of order $|s_{\boldsymbol{\theta}}|_1/\sqrt{n}$ and we can avoid such restrictive vague assumptions.

Both our ℓ_1-oracle inequality and the ℓ_0-oracle inequality in Städler et al. (2010) are valid for regularization parameters of the same order as regards the sample size n and the number p of covariates, that is $(\ln n)^2\sqrt{\ln(2p+1)/n}$. This means that if one considers a Lasso estimator with such a regularization parameter, then, although one can not be sure that the Lasso indeed performs well as regards variable selection (because one has no idea of the unknown constants present in Städler et al., 2010), one is at least guaranteed that the Lasso will act as a good ℓ_1-regularizator.

Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates can grow with respect to n and can be much larger than n. A great attention has been paid to obtain a lower bound (3.7) of λ with optimal dependence on p, which is the only parameter not to be fixed. We recover the same dependence $\sqrt{\ln(2p+1)}$ as for the homogeneous linear regression in Chapter 2. Contrary to Städler et al. (2010), we give an explicit dependence not only on n and p, but also on the number K of clusters in the mixture as well as on the regressors and all the quantities bounding the mixture parameters of the model. Nonetheless, we are aware of the fact that these

dependences may not be optimal. In particular, we get a linear dependence on K in (3.7), while we might think that the true minimal dependence is only $\sqrt{K}$ (see Remark 6 for more details).

For $K = 1$, $\mathrm{KL_n}(s,t) = \mathbb{E}\left[\|X\boldsymbol{\beta}^* - X\boldsymbol{\beta}\|^2\right]/2$, so Inequality (3.8) enables to recover Inequality (2.14) for linear homogeneous regression. Nonetheless, we have established Inequality (2.14) with no boundedness assumption on the parameters. Moreover, there is an extra-$(\ln n)^2$ factor in the lower bound (3.7) of the regularization parameter λ in the heterogeneous regression case compared to the lower bound (2.12) in the homogeneous regression case. In fact, in Chapter 2, we managed to obtain an optimal bound of order $\sqrt{\ln(p)/n}$ of the regularization parameter thanks to linearity arguments adapted to the quadratic loss function. In contrast, in this chapter, we have not managed to extend these arguments to deal with the non-linear Kullback-Leibler information and we have rather envisaged entropy arguments based on Dudley's entropy bound (Dudley, 2010), which results in this extra-$(\ln n)^2$ factor[1].

Appendices

3.A Proof of Theorem 3.3.1

3.A.1 Statement of the main results

To prove Theorem 3.3.1, we first establish an ℓ_1-ball mixture regression model selection theorem for density estimation in the Gaussian framework, which is stated below as Theorem 3.A.1. Then, by looking at the Lasso as the solution of a penalized maximum likelihood model selection procedure over a countable collection of ℓ_1-ball models, Theorem 3.3.1 is an immediate consequence of Theorem 3.A.1.

Theorem 3.A.1. *Assume we observe $((\boldsymbol{x_i}, Y_i))_{1\leq i\leq n}$ with unknown conditional Gaussian mixture density s. For all $m \in \mathbb{N}^\star$, consider the ℓ_1-ball*

$$S_m = \{s_{\boldsymbol{\theta}} \in S;\ |s_{\boldsymbol{\theta}}|_1 \leq m\} \tag{3.9}$$

and let $\hat{s}_m$ be a η_m-empirical risk minimizer in S_m for some $\eta_m \geq 0$:

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(\hat{s}_m(Y_i|\boldsymbol{x_i})) \leq \inf_{s_m\in S_m}\left\{-\frac{1}{n}\sum_{i=1}^{n}\ln(s_m(Y_i|\boldsymbol{x_i}))\right\} + \eta_m. \tag{3.10}$$

[1]While finishing to write this thesis, we have learnt about a very new result found by van de Geer (2012) that enables to free from this extra-$(\ln n)^2$ factor, by developing chaining arguments based on Talagrand's approach (Talagrand, 1996, 2005) and specialized to the geometry of ℓ_1 balls. van de Geer's result relies on a multivariate version of the contraction theorem provided that some componentwise Lipschitz condition is satisfied, which is the case for the finite mixture Gaussian regression setting we consider here.

Assume that for all $m \in \mathbb{N}^\star$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ *with*

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\beta^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)}\right) \tag{3.11}$$

for some absolute constant $\kappa \geq 360$. Then, any penalized likelihood estimate $\hat{s}_{\hat{m}}$ with $\hat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_{\hat{m}}(Y_i|\boldsymbol{x_i})) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^\star} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_m(Y_i|\boldsymbol{x_i})) + \mathrm{pen}(m) \right\} + \eta \tag{3.12}$$

for some $\eta \geq 0$ satisfies

$$\begin{aligned} &\mathbb{E}\left[\mathrm{KL_n}(s, \hat{s}_{\hat{m}})\right] \\ &\leq (1+\kappa^{-1}) \inf_{m \in \mathbb{N}^\star} \left\{ \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \mathrm{pen}(m) + \eta_m \right\} \\ &\quad + \frac{\kappa' \sqrt{K}}{\sqrt{n}} \left[\frac{K\left(1 + (A_\beta + A_\sigma)^2\right)}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\beta^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) + \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{A_\beta^2}{a_\sigma^2} \right] + \eta, \end{aligned} \tag{3.13}$$

where κ' is an absolute positive constant.

Proof. Page 111. □

Theorem 3.A.1 can be deduced from the two following propositions.

Proposition 3.A.1. *Assume we observe $((\boldsymbol{x_i}, Y_i))_{1 \leq i \leq n}$ with unknown conditional density s. Let $M_n > 0$ and consider the event*

$$\mathcal{T} := \left\{ \max_{i=1,\ldots,n} |Y_i| \leq M_n \right\}.$$

For all $m \in \mathbb{N}^\star$, consider the ℓ_1-ball

$$S_m = \{ s_{\boldsymbol{\theta}} \in S;\ |s_{\boldsymbol{\theta}}|_1 \leq m \}, \tag{3.14}$$

and let $\hat{s}_m$ be a η_m-empirical risk minimizer in S_m for some $\eta_m \geq 0$:

$$-\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_m(Y_i|\boldsymbol{x_i})) \leq \inf_{s_m \in S_m} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln(s_m(Y_i|\boldsymbol{x_i})) \right\} + \eta_m.$$

Assume that for all $m \in \mathbb{N}^\star$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ *with*

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\beta)^2}{a_\sigma^2}\right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)}\right) \tag{3.15}$$

for some absolute constant $\kappa \geq 36$. Then, any penalized likelihood estimate $\hat{s}_{\hat{m}}$ with $\hat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_{\hat{m}}(Y_i|\boldsymbol{x_i})) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^\star} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln(\hat{s}_m(Y_i|\boldsymbol{x_i})) + \mathrm{pen}(m) \right\} + \eta \tag{3.16}$$

for some $\eta \geq 0$ satisfies

$$\begin{aligned} \mathbb{E}\left[\mathrm{KL_n}(s, \hat{s}_{\hat{m}}) \mathbb{1}_{\mathcal{T}}\right] \leq (1+\kappa^{-1}) \inf_{m \in \mathbb{N}^\star} \left\{ \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \mathrm{pen}(m) + \eta_m \right\} \\ + \frac{\kappa' K^{3/2}}{\sqrt{n}} \frac{\left(1 + (A_\beta + A_\sigma)^2\right)}{(a_\sigma \wedge a_\pi)} \left(1 + \frac{(M_n + A_\beta)^2}{a_\sigma^2}\right) + \eta, \end{aligned} \tag{3.17}$$

where κ' is an absolute positive constant.

Proof. Page 105. □

Proposition 3.A.2. *Assume we observe $((\boldsymbol{x_i}, Y_i))_{1 \leq i \leq n}$ with unknown conditional Gaussian mixture density s. Let $M_n > 0$ and consider the event*

$$\mathcal{T}^C := \left\{ \max_{i=1,\ldots,n} |Y_i| > M_n \right\}.$$

For all $m \in \mathbb{N}^\star$, consider $\hat{m}$ defined by (3.16).
Then,

$$\mathbb{E}\left[\mathrm{KL_n}(s, \hat{s}_{\hat{m}}) \mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{2K} \left(\ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{2A_\beta^2}{a_\sigma^2}\right) \mathrm{e}^{-\frac{M_n\left(M_n - 2A_\beta\right)}{4A_\sigma^2}} \sqrt{n}\,.$$

Proof. Page 108. □

3.A.2 Proof of the main results

The main result is Proposition 3.A.1. Its proof is based on Vapnik's structural risk minimization method (Vapnik, 1982, 1990). To obtain an upper bound of the empirical process in expectation, we use concentration inequalities combined with symmetrization arguments, in the spirit of the proof of Massart's general model selection theorem for maximum likelihood estimators (Massart, 2007, Theorem 7.11). Nonetheless, Massart's arguments are lightened because we are just looking for low rates of convergence.

3.A.2.1 Proof of Proposition 3.A.1

For any measurable function $f : \mathbb{R} \mapsto \mathbb{R}$, consider its empirical norm

$$\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} f^2(Y_i|\boldsymbol{x_i})}\,, \tag{3.18}$$

its conditional expectation $\mathbb{E}_X\left[f\right] = \int_{\mathbb{R}} f(y|\boldsymbol{x})s(y|\boldsymbol{x})\,dy$, as well as its empirical process

$$P_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(Y_i|\boldsymbol{x_i})\,, \tag{3.19}$$

and the recentred process

$$\nu_n(f) := P_n(f) - \mathbb{E}_X\left[P_n\left(f\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[f(Y_i|\boldsymbol{x_i}) - \int_{\mathbb{R}} f(y|\boldsymbol{x_i})s(y|\boldsymbol{x_i})\,dy\right]. \tag{3.20}$$

Let $\delta_{\mathrm{KL}} > 0$, $\eta > 0$.
Fix $m \in \mathbb{N}^\star$. Let $\eta_m > 0$. There exist two functions $\hat{s}_m$ and $\overline{s}_m$ in S_m such that

$$P_n(-\ln \hat{s}_m) \leq \inf_{s_m \in S_m} P_n(-\ln s_m) + \eta_m, \tag{3.21}$$

$$\mathrm{KL_n}(s, \overline{s}_m) \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \delta_{\mathrm{KL}}. \tag{3.22}$$

Define

$$\mathcal{M}(m) = \left\{m' \in \mathbb{N}^\star \middle| P_n(-\ln \hat{s}_{m'}) + \mathrm{pen}(m') \leq P_n(-\ln \hat{s}_m) + \mathrm{pen}(m) + \eta\right\}. \tag{3.23}$$

Introduce the set

$$F_m = \left\{f_m = -\ln\left(\frac{s_m}{s}\right)\;;\; s_m \in S_m\right\} \tag{3.24}$$

and put

$$\hat{f}_m := -\ln\left(\frac{\hat{s}_m}{s}\right), \quad \overline{f}_m := -\ln\left(\frac{\overline{s}_m}{s}\right). \tag{3.25}$$

For every $m' \in \mathcal{M}(m)$, we get from (3.23), (3.25) and (3.21) that

$$P_n\left(\hat{f}_{m'}\right) + \mathrm{pen}(m') \leq P_n\left(\hat{f}_m\right) + \mathrm{pen}(m) + \eta \leq P_n\left(\overline{f}_m\right) + \mathrm{pen}(m) + \eta + \eta_m,$$

which implies by (3.20) that

$$\mathbb{E}_X\left[P_n\left(\hat{f}_{m'}\right)\right] + \mathrm{pen}(m') \leq \mathbb{E}_X\left[P_n\left(\overline{f}_m\right)\right] + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\hat{f}_{m'}\right) + \eta + \eta_m.$$

By taking into account (3.5), (3.19) and (3.22), we get

$$\mathrm{KL_n}(s, \hat{s}_{m'}) + \mathrm{pen}(m') \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\hat{f}_{m'}\right) + \eta + \eta_m + \delta_{\mathrm{KL}}. \tag{3.26}$$

Thus, all the matter is to control $-\nu_n(\hat{f}_{m'}) = \nu_n(-\hat{f}_{m'})$. To cope with the randomness of $\hat{f}_{m'}$, we control the deviation of $\sup_{f_{m'} \in F_{m'}} \nu_n\left(-f_{m'}\right)$ thanks to the following lemma.

Lemma 3.A.3. *Let $M_n > 0$. Consider the event*

$$\mathcal{T} := \left\{ \max_{i=1,\ldots,n} |Y_i| \leq M_n \right\}.$$

Put

$$B_n = \frac{1}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\beta)^2}{a_\sigma^2}\right) \tag{3.27}$$

and

$$\Delta_{m'} := m' \|\boldsymbol{x}\|_{\max,n} \ln n \sqrt{K \ln(2p+1)} + 6\left(1 + K(A_\beta + A_\sigma)\right). \tag{3.28}$$

Then, on the event $\mathcal{T}$, for all $m' \in \mathbb{N}^\star$, for all $u > 0$, with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-u}$,

$$\sup_{f_{m'} \in F_{m'}} \left|\nu_n\left(-f_{m'}\right)\right| \leq \frac{4B_n}{\sqrt{n}} \left[9\sqrt{K}\,\Delta_{m'} + \sqrt{2}(1 + K(A_\beta + A_\sigma))\sqrt{u}\right]. \tag{3.29}$$

Proof. Page 112. □

We derive from (3.26) and (3.29) that on the event $\mathcal{T}$, for all $m \in \mathbb{N}^\star$, for all $m' \in \mathcal{M}(m)$, for all $u > 0$, with $\mathbb{P}_X$-probability larger than $1 - \mathrm{e}^{-u}$,

$$\begin{aligned}
\mathrm{KL_n}(s, \hat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta + \eta_m + \delta_{\mathrm{KL}} \\
&\quad + \frac{4B_n}{\sqrt{n}} \left[9\sqrt{K}\,\Delta_{m'} + \sqrt{2}(1 + K(A_\beta + A_\sigma))\sqrt{u}\right] \\
&\leq \inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta + \eta_m + \delta_{\mathrm{KL}} \\
&\quad + \frac{4B_n}{\sqrt{n}} \left(9\sqrt{K}\,\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + A_\sigma\right)\right)^2 + \sqrt{K}u\right),
\end{aligned} \tag{3.30}$$

where we get the last inequality by using $2ab \leq \alpha a^2 + \alpha^{-1} b^2$ for $\alpha = 1/\sqrt{K}$.

It remains to sum up the tail bounds (3.30) over all the possible values of $m \in \mathbb{N}^\star$ and $m' \in \mathcal{M}(m)$. To get an inequality valid on a great probability set, we need to choose adequately the value of the parameter u depending on $m \in \mathbb{N}^\star$ and $m' \in \mathcal{M}(m)$. Let $z > 0$. By applying (3.30) to $u = z + m + m'$

for all $m \in \mathbb{N}^\star$ and $m' \in \mathcal{M}(m)$, we get that, and on the event $\mathcal{T}$, with $\mathbb{P}_X$-probability larger than

$$1-\sum_{(m,m')\in\mathbb{N}^\star\times\mathcal{M}(m)} \mathrm{e}^{-(z+m+m')} \geq 1-\sum_{(m,m')\in\mathbb{N}^\star\times\mathbb{N}^\star} \mathrm{e}^{-(z+m+m')} = 1-\mathrm{e}^{-z}\left(\sum_{m\in\mathbb{N}^\star}\mathrm{e}^{-m}\right)^2 \geq 1-\mathrm{e}^{-z},$$

$$\begin{aligned}
&\mathrm{KL_n}(s,\hat{s}_{m'}) + \mathrm{pen}(m') \\
&\leq \inf_{s_m\in S_m} \mathrm{KL_n}(s,s_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) + \eta + \eta_m + \delta_{\mathrm{KL}} \\
&\quad + \frac{4B_n}{\sqrt{n}}\left(9\sqrt{K}\,\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1+K\left(A_\beta+A_\sigma\right)\right)^2 + \sqrt{K}(z+m+m')\right).
\end{aligned}$$

holds simultaneously for all $m \in \mathbb{N}^\star$ and $m' \in \mathcal{M}(m)$. It can also be written

$$\begin{aligned}
&\mathrm{KL_n}(s,\hat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) - \delta_{\mathrm{KL}} \\
&\leq \inf_{s_m\in S_m} \mathrm{KL_n}(s,s_m) + \left[\mathrm{pen}(m) + \frac{4B_n}{\sqrt{n}}\sqrt{K}m\right] + \eta_m \\
&\quad + \left[\frac{4B_n}{\sqrt{n}}\sqrt{K}\left(9\Delta_{m'}+m'\right) - \mathrm{pen}(m')\right] + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1+K\left(A_\beta+A_\sigma\right)\right)^2 + \sqrt{K}z\right) + \eta.
\end{aligned}$$

By taking into account the definition (3.28) of $\Delta_{m'}$, it gives

$$\begin{aligned}
&\mathrm{KL_n}(s,\hat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) - \delta_{\mathrm{KL}} \\
&\leq \inf_{s_m\in S_m} \mathrm{KL_n}(s,s_m) + \left[\mathrm{pen}(m) + \frac{4B_n}{\sqrt{n}}\sqrt{K}m\right] + \eta_m \\
&\quad + \left[\frac{4B_n}{\sqrt{n}}\sqrt{K}\left(9\|\boldsymbol{x}\|_{\max,n}\ln n\sqrt{K\ln(2p+1)}+1\right)m' - \mathrm{pen}(m')\right] \\
&\quad + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1+K\left(A_\beta+A_\sigma\right)\right)^2 + 54\sqrt{K}\left(1+K(A_\beta+A_\sigma)\right) + \sqrt{K}z\right) + \eta. \qquad (3.31)
\end{aligned}$$

Now, let $\kappa \geq 1$ and assume that, for all $\tilde{m} \in \mathbb{N}^\star$, $\mathrm{pen}(\tilde{m})$ satisfies $\mathrm{pen}(\tilde{m}) = \lambda\tilde{m}$ with

$$\lambda \geq 4\kappa\frac{B_n}{\sqrt{n}}\sqrt{K}\left(9\|\boldsymbol{x}\|_{\max,n}\ln n\sqrt{K\ln(2p+1)}+1\right).$$

Then, (3.31) implies

$$\begin{aligned}
&\mathrm{KL_n}(s,\hat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) - \delta_{\mathrm{KL}} \\
&\leq \inf_{s_m\in S_m} \mathrm{KL_n}(s,s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m \\
&\quad + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1+K\left(A_\beta+A_\sigma\right)\right)^2 + 54\sqrt{K}\left(1+K(A_\beta+A_\sigma)\right) + \sqrt{K}z\right) + \eta.
\end{aligned}$$

Then, by using the inequality $2ab \leq \alpha a^2 + \alpha^{-1} b^2$ for $\alpha = \sqrt{K}$, we get

$$\begin{aligned}
&\mathrm{KL_n}(s, \hat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) - \delta_{\mathrm{KL}} \\
\leq &\inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m \\
&+ \frac{4B_n}{\sqrt{n}}\left(27K^{3/2} + \frac{55}{2\sqrt{K}}\left(1 + K\left(A_\beta + A_\sigma\right)\right)^2 + \sqrt{K}z\right) + \eta. \qquad (3.32)
\end{aligned}$$

Now consider $\hat{m}$ defined by (3.16). From (3.16) and (3.23), $\hat{m}$ belongs to $\mathcal{M}(m)$ for all $m \in \mathbb{N}^\star$. So, from (3.32), on the event $\mathcal{T}$, for all $z > 0$, with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-z}$,

$$\begin{aligned}
&\mathrm{KL_n}(s, \hat{s}_{\hat{m}}) - \nu_n\left(\overline{f}_m\right) - \delta_{\mathrm{KL}} \\
\leq &\inf_{m \in \mathbb{N}^\star}\left\{\inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right\} \\
&+ \frac{4B_n}{\sqrt{n}}\left(27K^{3/2} + \frac{55}{2\sqrt{K}}\left(1 + K\left(A_\beta + A_\sigma\right)\right)^2 + \sqrt{K}z\right) + \eta. \qquad (3.33)
\end{aligned}$$

We end the proof by integrating (3.33) with respect to z. Since $\mathbb{E}\left(\nu_n\left(\overline{f}_m\right)\right) = 0$, we get when δ_{KL} tends to zero:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{KL_n}(s, \hat{s}_{\hat{m}})\,\mathbb{1}_{\mathcal{T}}\right] \\
\leq &\inf_{m \in \mathbb{N}^\star}\left\{\inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right\} \\
&+ \frac{4B_n}{\sqrt{n}}\left(27K^{3/2} + \frac{55}{2\sqrt{K}}\left(1 + K\left(A_\beta + A_\sigma\right)\right)^2 + \sqrt{K}\right) + \eta \\
\leq &\inf_{m \in \mathbb{N}^\star}\left\{\inf_{s_m \in S_m} \mathrm{KL_n}(s, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right\} \\
&+ \frac{112B_n}{\sqrt{n}} K^{3/2}\left(1 + \left(A_\beta + A_\sigma\right)^2\right) + \eta.
\end{aligned}$$

By taking into account the value (3.27) of B_n, we obtain (3.17). □

3.A.2.2 Proof of Proposition 3.A.2

By Cauchy-Schwarz Inequality,

$$\mathbb{E}\left[\mathrm{KL_n}(s, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{\mathbb{E}\left[\mathrm{KL_n^2}(s, \hat{s}_{\hat{m}})\right]}\sqrt{\mathbb{P}\left(\mathcal{T}^C\right)}\,. \qquad (3.34)$$

Let us bound the two terms of the right-hand side of (3.34).
For the first term, let us bound $\mathrm{KL}(s(.|\boldsymbol{x}), s_{\boldsymbol{\theta}}(.|\boldsymbol{x}))$ for all $s_{\boldsymbol{\theta}} \in S$ and $\boldsymbol{x} \in \mathbb{R}^p$.
Let $s_{\boldsymbol{\theta}} \in S$ and $\boldsymbol{x} \in \mathbb{R}^p$. Denote $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\beta_k}, \sigma_k)_{1 \leq k \leq K}$. Recall that the true density s is assumed

to be itself a finite mixture of Gaussian regressions with $s = s_{\boldsymbol{\theta}^*}$. Let $y \in \mathbb{R}$. Since the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ belong to the bounded space Θ defined by (3.4) and since $\sum_{k=1}^K \pi_k = \sum_{k=1}^K \pi_k^* = 1$, on the one hand, we have

$$s(y|\boldsymbol{x}) = \sum_{k=1}^K \frac{\pi_k^*}{\sqrt{2\pi}\sigma^*} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k^*}^T \boldsymbol{x}\right)^2}{2\sigma_k^{*2}}\right) \leq \frac{1}{\sqrt{2\pi}a_\sigma},$$

and, on the other hand, we have

$$\begin{aligned} s_{\boldsymbol{\theta}}(y|\boldsymbol{x})) &= \sum_{k=1}^K \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k}^T \boldsymbol{x}\right)^2}{2\sigma_k^2}\right) \\ &\geq \sum_{k=1}^K \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{y^2 + \left(\boldsymbol{\beta_k}^T \boldsymbol{x}\right)^2}{\sigma_k^2}\right) \\ &\geq \frac{1}{\sqrt{2\pi}A_\sigma} \exp\left(-\frac{y^2 + A_\beta^2}{a_\sigma^2}\right). \end{aligned}$$

Thus,

$$\ln\left(\frac{s(y|\boldsymbol{x})}{s_{\boldsymbol{\theta}}(y|\boldsymbol{x})}\right) \leq \ln\left(\frac{A_\sigma}{a_\sigma} \exp\left(\frac{y^2 + A_\beta^2}{a_\sigma^2}\right)\right) = \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{y^2 + A_\beta^2}{a_\sigma^2}$$

and

$$\mathrm{KL}(s(.|\boldsymbol{x}), s_{\boldsymbol{\theta}}(.|\boldsymbol{x})) = \int_{\mathbb{R}} \ln\left(\frac{s(y|\boldsymbol{x})}{s_{\boldsymbol{\theta}}(y|\boldsymbol{x})}\right) s(y|\boldsymbol{x})\, dy \leq \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{A_\beta^2}{a_\sigma^2} + \frac{1}{a_\sigma^2} \int_{\mathbb{R}} y^2\, s(y|\boldsymbol{x})\, dy\,.$$

But

$$\int_{\mathbb{R}} y^2\, s(y|\boldsymbol{x})\, dy = \mathbb{E}\left(Y^2|\boldsymbol{X} = \boldsymbol{x}\right) \leq [\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})]^2 \leq \left[\sum_{k=1}^K \pi_k^*\, \boldsymbol{\beta_k^*}^T \boldsymbol{x}\right]^2 \leq A_\beta^2.$$

So,

$$\mathrm{KL}(s(.|\boldsymbol{x}), s_{\boldsymbol{\theta}}(.|\boldsymbol{x})) \leq \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{2A_\beta^2}{a_\sigma^2}.$$

Then, for all $s_{\boldsymbol{\theta}} \in S$,

$$\mathrm{KL_n}(s, s_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \mathrm{KL}\left(s(.|\boldsymbol{x_i}), s_{\boldsymbol{\theta}}(.|\boldsymbol{x_i})\right) \leq \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{2A_\beta^2}{a_\sigma^2}$$

and

$$\sqrt{\mathbb{E}\left[\mathrm{KL_n^2}(s, \hat{s}_{\hat{m}})\right]} \leq \ln\left(\frac{A_\sigma}{a_\sigma}\right) + \frac{2A_\beta^2}{a_\sigma^2}. \tag{3.35}$$

Let us now provide an upper bound of $\mathbb{P}(\mathcal{T}^C)$.

$$\mathbb{P}\left(\mathcal{T}^C\right) = \mathbb{E}\left(\mathbb{1}_{\mathcal{T}^C}\right) = \mathbb{E}\left[\mathbb{E}_X\left(\mathbb{1}_{\mathcal{T}^C}\right)\right] = \mathbb{E}\left[\mathbb{P}_X\left(\mathcal{T}^C\right)\right] \tag{3.36}$$

with

$$\mathbb{P}_X\left(\mathcal{T}^C\right) = \mathbb{P}_X\left(\bigcup_{i=1}^n \{|Y_i| > M_n\}\right) \leq \sum_{i=1}^n \mathbb{P}_X(|Y_i| > M_n). \tag{3.37}$$

For all $i \in \{1, \ldots, n\}$, $Y_i \mid \boldsymbol{X_i} = \boldsymbol{x_i} \sim \sum_{k=1}^K \pi_k^* \mathcal{N}(\boldsymbol{\beta_k^*}^T \boldsymbol{x_i}, \sigma_k^{*2})$. So, from (3.37), we just need to provide an upper bound of $\mathbb{P}\left(|Y_{\boldsymbol{x}}| > M_n\right)$ with $Y_{\boldsymbol{x}} \sim \sum_{k=1}^K \pi_k^* \mathcal{N}(\boldsymbol{\beta_k^*}^T \boldsymbol{x}, \sigma_k^{*2})$ for $\boldsymbol{x} \in \mathbb{R}^p$. First using Chernoff's Inequality for a centered Gaussian variable (Massart, 2007), and then using the fact that $\boldsymbol{\theta}^*$ belongs to the bounded space Θ defined by (3.4) and that $\sum_{k=1}^K \pi_k^* = 1$, we get

$$\begin{aligned}
\mathbb{P}\left(|Y_{\boldsymbol{x}}| > M_n\right) &= \int_{\mathbb{R}} \mathbb{1}_{\{|y|>M_n\}} \sum_{k=1}^K \pi_k^* \frac{1}{\sqrt{2\pi}\sigma_k^*} \mathrm{e}^{-\frac{1}{2}\left(\frac{y-\boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right)^2} dy \\
&= \sum_{k=1}^K \pi_k^* \int_{\mathbb{R}} \mathbb{1}_{\{|y|>M_n\}} \frac{1}{\sqrt{2\pi}\sigma_k^*} \mathrm{e}^{-\frac{1}{2}\left(\frac{y-\boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right)^2} dy \\
&= \sum_{k=1}^K \pi_k^* \mathbb{P}\left(|Y_{\boldsymbol{x},k}| > M_n\right) \quad \text{with } Y_{\boldsymbol{x},k} \sim \mathcal{N}\left(\boldsymbol{\beta_k^*}^T \boldsymbol{x}, \sigma_k^{*2}\right) \\
&= \sum_{k=1}^K \pi_k^* \left[\mathbb{P}\left(U > \frac{M_n - \boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right) + \mathbb{P}\left(U > \frac{M_n + \boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right)\right] \quad \text{with } U \sim \mathcal{N}(0,1) \\
&\leq \sum_{k=1}^K \pi_k^* \left[\mathrm{e}^{-\frac{1}{2}\left(\frac{M_n - \boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right)^2} + \mathrm{e}^{-\frac{1}{2}\left(\frac{M_n + \boldsymbol{\beta_k^*}^T \boldsymbol{x}}{\sigma_k^*}\right)^2}\right] \\
&\leq 2\sum_{k=1}^K \pi_k^* \mathrm{e}^{-\frac{1}{2}\left(\frac{M_n - \left|\boldsymbol{\beta_k^*}^T \boldsymbol{x}\right|}{\sigma_k^*}\right)^2} \\
&\leq 2\sum_{k=1}^K \pi_k^* \mathrm{e}^{-\frac{M_n^2 + \left(\boldsymbol{\beta_k^*}^T \boldsymbol{x}\right)^2 - 2M_n\left|\boldsymbol{\beta_k^*}^T \boldsymbol{x}\right|}{2\sigma_k^{*2}}} \\
&\leq 2K\mathrm{e}^{-\frac{M_n^2 + a_\beta^2 - 2M_n A_\beta}{2A_\sigma^2}} \\
&\leq 2K\mathrm{e}^{-\frac{M_n\left(M_n - 2A_\beta\right)}{2A_\sigma^2}}.
\end{aligned} \tag{3.38}$$

We derive from (3.36), (3.37) and (3.38) that

$$\mathbb{P}(\mathcal{T}^C) \leq 2K\mathrm{e}^{-\frac{M_n\left(M_n - 2A_\beta\right)}{2A_\sigma^2}} n. \tag{3.39}$$

Finally, we get from (3.34), (3.35) and (3.39) that

$$\mathbb{E}\left[\mathrm{KL_n}(s,\hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{2K}\left(\ln\left(\frac{A_\sigma}{a_\sigma}\right)+\frac{2A_\beta^2}{a_\sigma^2}\right)\mathrm{e}^{-\frac{M_n\left(M_n-2A_\beta\right)}{4A_\sigma^2}}\sqrt{n}\,. \tag{3.40}$$

□

3.A.2.3 Proof of Theorem 3.A.1

Let $M_n > 0$ and $\kappa \geq 36$. Assume that, for all $m \in \mathbb{N}^\star$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi}\left(1+\frac{(M_n+A_\beta)^2}{a_\sigma^2}\right)\frac{\sqrt{K}}{\sqrt{n}}\left(1+\|x\|_{\max,n}\ln(n)\sqrt{K\ln(2p+1)}\right). \tag{3.41}$$

We derive from Proposition 3.A.1 and Proposition 3.A.2 that there exists an absolute positive constant κ' such that any penalized likelihood estimate $\hat{s}_{\hat{m}}$ defined by (3.12) satisfies

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL_n}(s,\hat{s}_{\hat{m}})\right] &= \mathbb{E}\left[\mathrm{KL_n}(s,\hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}\right]+\mathbb{E}\left[\mathrm{KL_n}(s,\hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^C}\right]\\ &\leq (1+\kappa^{-1})\inf_{m\in\mathbb{N}^\star}\left\{\inf_{s_m\in S_m}\mathrm{KL_n}(s,s_m)+\mathrm{pen}(m)+\eta_m\right\}\\ &\quad+\frac{\kappa' K^{3/2}}{\sqrt{n}}\frac{\left(1+(A_\beta+A_\sigma)^2\right)}{a_\sigma\wedge a_\pi}\left(1+\frac{(M_n+A_\beta)^2}{a_\sigma^2}\right)+\eta\\ &\quad+\sqrt{2K}\left(\ln\left(\frac{A_\sigma}{a_\sigma}\right)+\frac{2A_\beta^2}{a_\sigma^2}\right)\mathrm{e}^{-\frac{M_n\left(M_n-2A_\beta\right)}{4A_\sigma^2}}\sqrt{n}\,.\end{aligned} \tag{3.42}$$

To get Inequality (3.13), it only remains to optimize Inequality (3.42) with respect to M_n. Since the two terms depending on M_n in (3.42) have opposite monotony with respect to M_n, we are looking for a value of M_n such that these two terms are of the same order with respect to n. Consider $M_n = A_\beta + \sqrt{A_\beta^2 + 4A_\sigma^2\ln n}$ the positive solution of the equation $X(X-2A_\beta) - 4A_\sigma^2\ln n = 0$. Then, on the one hand,

$$\mathrm{e}^{-\frac{M_n\left(M_n-2A_\beta\right)}{4A_\sigma^2}}\sqrt{n} = \mathrm{e}^{-\ln n}\sqrt{n} = \frac{1}{\sqrt{n}}\,.$$

On the other hand, by using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\frac{1}{\sqrt{n}}\left(1+\frac{(M_n+A_\beta)^2}{a_\sigma^2}\right) \leq \frac{1}{\sqrt{n}}\left(1+\frac{2\left(5A_\beta^2+4A_\sigma^2\ln n\right)}{a_\sigma^2}\right),$$

hence (3.13).

The upper bound (3.11) of the tuning parameter λ is obtained from the upper bound (3.41) and the fact that $(M_n + A_\beta)^2 \leq 2(5A_\beta^2 + 4A_\sigma^2 \ln n)$ for $M_n = A_\beta + \sqrt{A_\beta^2 + 4A_\sigma^2 \ln n}$. □

3.A.2.4 Proof of Theorem 3.3.1

Let $\lambda > 0$. Define $\hat{m}$ as the smallest integer such that $\hat{s}(\lambda)$ belongs to $S_{\hat{m}}$, i.e. $\hat{m} = \lceil |\hat{s}(\lambda)|_1 \rceil$. Then, by using the definition of $\hat{m}$, Definition (3.6) of $\hat{s}(\lambda)$ and Definition (3.9) of S_m, we get

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{s}(\lambda)(Y_i|\boldsymbol{x_i})\right) + \lambda\hat{m} &\leq -\frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{s}(\lambda)(Y_i|\boldsymbol{x_i})\right) + \lambda|\hat{s}(\lambda)|_1 + \lambda \\
&= \inf_{s_{\boldsymbol{\theta}} \in S} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i})\right) + \lambda|s_{\boldsymbol{\theta}}|_1 \right\} + \lambda \\
&= \inf_{m \in \mathbb{N}^\star} \inf_{s_{\boldsymbol{\theta}} \in S, |s_{\boldsymbol{\theta}}|_1 \leq m} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x_i})\right) + \lambda|s_{\boldsymbol{\theta}}|_1 \right\} + \lambda \\
&\leq \inf_{m \in \mathbb{N}^\star} \left\{ \inf_{s_m \in S_m} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_m(Y_i|\boldsymbol{x_i})\right) \right\} + \lambda m \right\} + \lambda,
\end{aligned}$$

which implies

$$-\frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{s}(\lambda)(Y_i|\boldsymbol{x_i})\right) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^\star} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{s}_m(Y_i|\boldsymbol{x_i})\right) + \mathrm{pen}(m) \right\} + \eta,$$

with $\mathrm{pen}(m) = \lambda m$, $\eta = \lambda$ and $\hat{s}_m$ defined by (3.10) with $\eta_m = 0$. Thus, $\hat{s}(\lambda)$ satisfies (3.12) and Theorem 3.3.1 follows from Theorem 3.A.1. □

3.B Proofs of the lemmas

3.B.1 Proof of Lemma 3.A.3

Let $m \in \mathbb{N}^\star$. From (3.20), we have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left| \frac{1}{n}\sum_{i=1}^{n} \left(f_m(Y_i|\boldsymbol{x_i}) - \mathbb{E}_X\left[f_m(Y_i|\boldsymbol{x_i})\right] \right) \right|. \tag{3.43}$$

To control the deviation of such a quantity, we shall combine concentration with symmetrization arguments. We shall first use the following concentration inequality.

Lemma 3.B.1. (Boucheron et al., 2012) Let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let F be a class of real-valued functions on $\mathcal{Z}$. Assume that there exists a

non-random constant R_n such that $\sup_{f\in F}\|f\|_n \leq R_n$. Then, for all $u>0$,

$$\mathbb{P}\left(\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i)-\mathbb{E}\left(f(Z_i)\right)\right| > \mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i)-\mathbb{E}\left(f(Z_i)\right)\right|\right] + 2\sqrt{2}R_n\sqrt{\frac{u}{n}}\right) \leq \mathrm{e}^{-u}. \tag{3.44}$$

Then, $\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i)-\mathbb{E}\left(f(Z_i)\right)\right|\right]$ can be bounded thanks to the following symmetrization argument.

Lemma 3.B.2. (van der Vaart and Wellner, 1996, Lemma 2.3.6) Let $Z_1,\ldots,Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let F be a class of real-valued functions on $\mathcal{Z}$. Let $(\varepsilon_1,\ldots,\varepsilon_n)$ be a Rademacher sequence independent of $(Z_1,\ldots,Z_n)$. Then,

$$\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i)-\mathbb{E}\left(f(Z_i)\right)\right|\right] \leq 2\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right]. \tag{3.45}$$

From (3.45), the problem boils down to providing an upper bound of $\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right]$. To this aim, we shall apply the following lemma, which is adapted from (Massart, 2007, Lemma 6.1).

Lemma 3.B.3. (Massart, 2007, Lemma 6.1) Let $Z_1,\ldots,Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let F be a class of real-valued functions on $\mathcal{Z}$. Let $(\varepsilon_1,\ldots,\varepsilon_n)$ be a Rademacher sequence independent of $(Z_1,\ldots,Z_n)$. Define a non-random constant R_n such that

$$\sup_{f\in F}\|f\|_n \leq R_n. \tag{3.46}$$

Then, for all $J\in\mathbb{N}^\star$,

$$\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right] \leq R_n\left(\frac{6}{\sqrt{n}}\sum_{j=1}^{J} 2^{-j}\sqrt{\ln\left[1+\mathcal{N}\left(2^{-j}R_n,F,\|.\|_n\right)\right]}+2^{-J}\right), \tag{3.47}$$

where $\mathcal{N}(\delta,F,\|.\|_n)$ stands for the δ-packing number of the set of functions F equipped with the metric induced by the norm $\|.\|_n$.

From (3.43), we propose to apply a conditional version of Lemma 3.B.1, Lemma 3.B.2 and Lemma 3.B.3 to $F=F_m$, $(Z_1,\ldots,Z_n)=(Y_1|\boldsymbol{x_1},\ldots,Y_n|\boldsymbol{x_n})$ and $f(Z_i)=f_m(Y_i|\boldsymbol{x_i})$ so as to control $\sup_{f_m\in F_m}\left|\nu_n\left(-f_m\right)\right|$. On the one hand, we see from (3.46) that we need an upper bound of $\sup_{f_m\in F_m}\|f_m\|_n$. On the other hand, we see from (3.47) that we need to bound the entropy of the set of functions F_m equipped with the metric induced by the norm $\|.\|_n$. Such bounds are provided by the two following lemmas.

Let $M_n > 0$. Consider the event

$$\mathcal{T} := \left\{ \max_{i=1,\dots,n} |Y_i| \leq M_n \right\}$$

and put

$$B_n = \frac{1}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\beta)^2}{a_\sigma^2} \right).$$

Lemma 3.B.4. *On the event $\mathcal{T}$, for all $m \in \mathbb{N}^\star$,*

$$\sup_{f_m \in F_m} \|f_m\|_n \leq R_n := 2B_n \left(1 + K(A_\beta + A_\sigma)\right). \tag{3.48}$$

Proof. Page 116. □

Lemma 3.B.5. *Let $\delta > 0$ and $m \in \mathbb{N}^\star$. On the event $\mathcal{T}$, we have the following upper bound of the δ-packing number of the set of functions F_m equipped with the metric induced by the norm $\|.\|_n$:*

$$\mathcal{N}\left(\delta, F_m, \|.\|_n\right) \leq (2p+1)^{\frac{4B_n^2 K^2 m^2 \|\boldsymbol{x}\|_{\max,n}^2}{\delta^2}} \left(1 + \frac{8B_n K A_\sigma}{\delta}\right)^K \left(1 + \frac{8B_n}{\delta}\right)^K.$$

Proof. Page 117. □

By using the upper bounds provided by Lemma 3.B.4 and Lemma 3.B.5, we can apply Lemma 3.B.3 to get an upper bound of $\mathbb{E}_X \left[\sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_m(Y_i|\boldsymbol{x_i})\right|\right]$. It gives the following result.

Lemma 3.B.6. *Let $m \in \mathbb{N}^\star$. Consider $(\varepsilon_1, \dots, \varepsilon_n)$ a Rademacher sequence independent of $(Y_1, \dots, Y_n)$. Then, on the event $\mathcal{T}$,*

$$\mathbb{E}_X \left[\sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_m(Y_i|\boldsymbol{x_i}) \right| \right] \leq 18\sqrt{K} \frac{B_n}{\sqrt{n}} \Delta_m, \tag{3.49}$$

where

$$\Delta_m := m\|\boldsymbol{x}\|_{\max,n} \ln n \sqrt{K \ln(2p+1)} + 6\left(1 + K(A_\beta + A_\sigma)\right).$$

Proof. Page 120. □

Now, by using (3.49) and by applying Lemma 3.B.1 and Lemma 3.B.2 to $F = F_m$, $(Z_1, \dots, Z_n) = (Y_1|\boldsymbol{x_1}, \dots, Y_n|\boldsymbol{x_n})$ and $f(Z_i) = f_m(Y_i|\boldsymbol{x_i})$, we get from (3.43) that for all $m \in \mathbb{N}^\star$, for all $u > 0$,

with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-u}$,

$$\begin{aligned}
\sup_{f_m \in F_m} |\nu_n(-f_m)| &= \sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f_m(Y_i|\boldsymbol{x_i}) - \mathbb{E}_X \left[f_m(Y_i|\boldsymbol{x_i}) \right] \right) \right| \\
&\leq \mathbb{E}_X \left[\sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f_m(Y_i|\boldsymbol{x_i}) - \mathbb{E}_X \left[f_m(Y_i|\boldsymbol{x_i}) \right] \right) \right| \right] + 2\sqrt{2} R_n \sqrt{\frac{u}{n}} \\
&\leq 2\mathbb{E}_X \left[\sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_m(Y_i|\boldsymbol{x_i}) \right| \right] + 2\sqrt{2} R_n \sqrt{\frac{u}{n}} \\
&\leq \frac{4B_n}{\sqrt{n}} \left[9\sqrt{K}\Delta_m + \sqrt{2}(1 + K(A_\beta + A_\sigma))\sqrt{u} \right]
\end{aligned}$$

where we get the last inequality by taking into account Definition (3.48) of R_n.
Hence Lemma 3.A.3. □

3.B.2 Proof of Lemmas 3.B.4, 3.B.5 and 3.B.6

Proofs of both Lemma 3.B.4 and Lemma 3.B.5 require an upper bound of the uniform norm of the gradient of $\ln(s_{\boldsymbol{\theta}})$ for $s_{\boldsymbol{\theta}} \in S$. Let us thus begin by providing such an upper bound.

Lemma 3.B.7. *For finite mixture Gaussian regression models as described in Section 3.2.1,*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^p} \sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{\partial \ln\left(s_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})\right)}{\partial \boldsymbol{\theta}} \right\|_\infty \leq G(\cdot)$$

with

$$G : \mathbb{R} \mapsto \mathbb{R}, \quad y \mapsto \frac{1}{a_\sigma \wedge a_\pi} \left(1 + \frac{(|y| + A_\beta)^2}{a_\sigma^2} \right). \tag{3.50}$$

Proof. Let $s_{\boldsymbol{\theta}} \in S$ with $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\beta_k}, \sigma_k)_{1 \leq k \leq K}$. For all $\boldsymbol{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$,

$$\ln\left(s_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right) = \ln\left(\sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k}^T \boldsymbol{x}\right)^2}{2\sigma_k^2} \right) \right) = \ln\left(\sum_{k=1}^{K} f_k(\boldsymbol{x}, y) \right)$$

where we put

$$f_k(\boldsymbol{x}, y) = \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y - \boldsymbol{\beta_k}^T \boldsymbol{x}\right)^2}{2\sigma_k^2} \right) \geq 0.$$

For all $l \in \{1, \ldots, K\}$, by using the fact that $f_l(\boldsymbol{x}, y)/(\sum_{k=1}^{K} f_k(\boldsymbol{x}, y)) \leq 1$ and the fact that $\boldsymbol{\theta}$ belongs to the bounded space Θ defined by (3.4), we have

$$\left| \frac{\partial \ln\left(s_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right)}{\partial(\beta_l^T \boldsymbol{x})} \right| = \left| \frac{f_l(\boldsymbol{x}, y)}{\sum_{k=1}^{K} f_k(\boldsymbol{x}, y)} \frac{y - \boldsymbol{\beta_l}^T \boldsymbol{x}}{\sigma_l^2} \right| \leq \frac{|y| + A_\beta}{a_\sigma^2},$$

$$\left|\frac{\partial \ln\left(s_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right)}{\partial \sigma_l}\right| = \left|\frac{f_l(\boldsymbol{x},y)}{\sum_{k=1}^K f_k(\boldsymbol{x},y)}\left(-\frac{1}{\sigma_l}+\frac{\left(y-\boldsymbol{\beta_l}^T\boldsymbol{x}\right)^2}{\sigma_l^3}\right)\right| \leq \frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\beta)^2}{a_\sigma^2}\right),$$

$$\left|\frac{\partial \ln\left(s_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right)}{\partial \pi_l}\right| = \frac{f_l(\boldsymbol{x},y)}{\pi_l\sum_{k=1}^K f_k(\boldsymbol{x},y)} \leq \frac{1}{a_\pi}.$$

By using that $1+\eta^2 \geq \eta$ for $\eta \in \mathbb{R}$, we deduce from the three above bounds that, for all $y \in \mathbb{R}$,

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{R}^p}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{\partial \ln\left(s_{\boldsymbol{\theta}}(y|\boldsymbol{x})\right)}{\partial \boldsymbol{\theta}}\right| &\leq \max\left\{(\frac{|y|+A_\beta}{a_\sigma^2}, \frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\beta)^2}{a_\sigma^2}\right), \frac{1}{a_\pi}\right\} \\
&\leq \max\left\{\frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\beta)^2}{a_\sigma^2}\right), \frac{1}{a_\pi}\right\} \\
&\leq \frac{1}{a_\pi\wedge a_\sigma}\left(1+\frac{(|y|+A_\beta)^2}{a_\sigma^2}\right).
\end{aligned}$$

□

3.B.2.1 Proof of Lemma 3.B.4

Let $m \in \mathbb{N}^\star$ and $f_m \in F_m$. From (3.24), there exists $s_m \in S_m$ such that $f_m = -\ln(s_m/s)$. For all $\boldsymbol{x} \in \mathbb{R}^p$, denote by $\boldsymbol{\theta}(\boldsymbol{x}) = (\pi_k, \boldsymbol{\beta_k}^T\boldsymbol{x}, \sigma_k)_{1\leq k\leq K}$ the parameters of the density $s_m(\cdot|\boldsymbol{x})$. Recall that there exists $\boldsymbol{\theta^*}$ such that $s = s_{\boldsymbol{\theta^*}}$. First applying Taylor's Inequality and then Lemma 3.B.7 on the event $\mathcal{T} = \{\max_{i=1,\dots,n}|Y_i| \leq M_n\}$, we get for all $i \in \{1,\dots,n\}$,

$$\begin{aligned}
|f_m(Y_i|\boldsymbol{x_i})|\,\mathbb{1}_\mathcal{T} &= |\ln\left(s_m(Y_i|\boldsymbol{x_i})\right) - \ln\left(s(Y_i|\boldsymbol{x_i})\right)|\;\mathbb{1}_\mathcal{T} \\
&\leq \sup_{\boldsymbol{x}\in\mathbb{R}^p}\sup_{\varphi\in\Theta}\left|\frac{\partial \ln\left(s_\varphi(Y_i|\boldsymbol{x})\right)}{\partial\varphi}\right| \|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta^*}(\boldsymbol{x_i})\|_1\;\mathbb{1}_\mathcal{T} \\
&\leq \frac{1}{a_\sigma\wedge a_\pi}\left(1+\frac{(|Y_i|+A_\beta)^2}{a_\sigma^2}\right)\|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta^*}(\boldsymbol{x_i})\|_1\;\mathbb{1}_\mathcal{T} \\
&\leq \underbrace{\frac{1}{a_\sigma\wedge a_\pi}\left(1+\frac{(M_n+A_\beta)^2}{a_\sigma^2}\right)}_{:=B_n}\|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta^*}(\boldsymbol{x_i})\|_1 \\
&\leq B_n\sum_{k=1}^K\left(\left|\boldsymbol{\beta_k}^T\boldsymbol{x_i} - \boldsymbol{\beta_k^*}^T\boldsymbol{x_i}\right| + |\sigma_k-\sigma_k^*| + |\pi_k-\pi_k^*|\right).
\end{aligned}$$

Now, $\boldsymbol{\theta}$ and $\boldsymbol{\theta^*}$ belong to the bounded space Θ defined by (3.4) and $\sum_{k=1}^K\pi_k = \sum_{k=1}^K\pi_k^* = 1$, so

$$|f_m(Y_i|\boldsymbol{x_i})|\,\mathbb{1}_\mathcal{T} \leq B_n\left(2KA_\beta + 2KA_\sigma + 2\right) \leq 2B_n\left(1+K(A_\beta+A_\sigma)\right),$$

and thus $\|f_m\|_n\,\mathbb{1}_\mathcal{T} \leq 2B_n\left(1+K(A_\beta+A_\sigma)\right)$. □

3.B.2.2 Proof of Lemma 3.B.5

Let $m \in \mathbb{N}^\star$ and $f_m \in F_m$. From (3.24), there exists $s_m \in S_m$ such that $f_m = -\ln(s_m/s)$. Introduce s'_m in S_m and put $f'_m = -\ln(s'_m/s)$. For all $\boldsymbol{x} \in \mathbb{R}^p$, denote by $\boldsymbol{\theta}(\boldsymbol{x}) = (\pi_k, \boldsymbol{\beta_k}^T\boldsymbol{x}, \sigma_k)_{1\le k\le K}$ and $\boldsymbol{\theta}'(\boldsymbol{x}) = (\pi'_k, \boldsymbol{\beta'_k}^T\boldsymbol{x}, \sigma'_k)_{1\le k\le K}$ the parameters of the densities $s_m(\cdot|\boldsymbol{x})$ and $s'_m(\cdot|\boldsymbol{x})$ respectively. First applying Taylor's Inequality and then Lemma 3.B.7 on the event $\mathcal{T} = \{\max_{i=1,\dots,n} |Y_i| \le M_n\}$, we get for all $i \in \{1,\dots,n\}$,

$$\begin{aligned}
&\left|f_m(Y_i|\boldsymbol{x_i}) - f'_m(Y_i|\boldsymbol{x_i})\right| \mathbb{1}_{\mathcal{T}} \\
&= \left|\ln\left(s_m(Y_i|\boldsymbol{x_i})\right) - \ln\left(s'_m(Y_i|\boldsymbol{x_i})\right)\right| \mathbb{1}_{\mathcal{T}} \\
&\le \sup_{\boldsymbol{x}\in\mathbb{R}^p}\sup_{\varphi\in\Theta}\left|\frac{\partial \ln\left(s_\varphi(Y_i|\boldsymbol{x})\right)}{\partial\varphi}\right| \left\|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta}'(\boldsymbol{x_i})\right\|_1 \mathbb{1}_{\mathcal{T}} \\
&\le \frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(|Y_i| + A_\beta)^2}{a_\sigma^2}\right) \left\|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta}'(\boldsymbol{x_i})\right\|_1 \mathbb{1}_{\mathcal{T}} \\
&\le \underbrace{\frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(M_n + A_\beta)^2}{a_\sigma^2}\right)}_{:=B_n} \left\|\boldsymbol{\theta}(\boldsymbol{x_i}) - \boldsymbol{\theta}'(\boldsymbol{x_i})\right\|_1 \\
&\le B_n\left(\sum_{k=1}^K \left|\boldsymbol{\beta_k}^T\boldsymbol{x_i} - \boldsymbol{\beta'_k}^T\boldsymbol{x_i}\right| + \|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right).
\end{aligned}$$

Then, using $(a+b)^2 \le 2(a^2+b^2)$ and applying Cauchy-Schwarz Inequality leads to

$$\begin{aligned}
&\left(f_m(Y_i|\boldsymbol{x_i}) - f'_m(Y_i|\boldsymbol{x_i})\right)^2 \mathbb{1}_{\mathcal{T}} \\
&\le 2B_n^2\left[\left(\sum_{k=1}^K \left|\boldsymbol{\beta_k}^T\boldsymbol{x_i} - \boldsymbol{\beta'_k}^T\boldsymbol{x_i}\right|\right)^2 + \left(\|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right)^2\right] \qquad (3.51)\\
&\le 2B_n^2\left[K\sum_{k=1}^K \left(\boldsymbol{\beta_k}^T\boldsymbol{x_i} - \boldsymbol{\beta'_k}^T\boldsymbol{x_i}\right)^2 + \left(\|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right)^2\right] \qquad (3.52)\\
&\le 2B_n^2\left[K\sum_{k=1}^K \left(\sum_{j=1}^p \beta_{kj}x_{ij} - \sum_{j=1}^p \beta'_{kj}x_{ij}\right)^2 + \left(\|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right)^2\right]
\end{aligned}$$

and

$$\begin{aligned}
&\|f_m - f'_m\|_n^2 \mathbb{1}_{\mathcal{T}} \\
&\le 2B_n^2 \underbrace{K\sum_{k=1}^K \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p \beta_{kj}x_{ij} - \sum_{j=1}^p \beta'_{kj}x_{ij}\right)^2}_{(a)} + 2B_n^2\left(\|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right)^2.
\end{aligned}$$

So, for all $\delta > 0$, if $(a) \leq \delta^2/(4B_n^2)$, $\|\sigma - \sigma'\|_1 \leq \delta/(4B_n)$ and $\|\pi - \pi'\|_1 \leq \delta/(4B_n)$, then $\|f_m - f'_m\|_n^2 \leq \delta^2$.

To bound (a), we write

$$(a) = Km^2 \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\beta_{kj}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\beta'_{kj}}{m} x_{ij} \right)^2 . \tag{3.53}$$

Since $s_m \in S_m$ defined by (3.14), we have

$$\sum_{j=1}^{p} \left| \frac{\beta_{kj}}{m} \right| \leq 1. \tag{3.54}$$

Thus, by applying Lemma 3.B.8 stated below to $\boldsymbol{\beta_k}/m = (\beta_{kj}/m)_{1 \leq j \leq p}$ for all $k \in \{1, \ldots, K\}$, we deduce that there exists a family $\mathcal{B}$ of $(2p+1)^{4B_n^2 K^2 m^2 \|\boldsymbol{x}\|_{\max,n}^2/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $k \in \{1, \ldots, K\}$, for all $\boldsymbol{\beta_k}$, there exists $\boldsymbol{\beta'_k} \in \mathcal{B}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\beta_{kj}}{m} x_{ij} - \frac{\beta'_{kj}}{m} x_{ij} \right)^2 \leq \frac{\delta^2}{4B_n^2 K^2 m^2},$$

which implies that $(a) \leq \delta^2/(4B_n^2)$.

Now, define $B_1^K(R) = \{\boldsymbol{v} \in \mathbb{R}^K; \|\boldsymbol{v}\|_1 \leq R\}$. Since $\boldsymbol{\theta}$ belongs to the bounded space Θ defined by (3.4), $\|\sigma\|_1 = \sum_{k=1}^{K} |\sigma_k| \leq KA_\sigma$ and $\|\pi\|_1 = \sum_{k=1}^{K} \pi_k = 1$, so $\sigma \in B_1^K(KA_\sigma)$ and $\pi \in B_1^K(1)$.

Therefore, we get that, on the event $\mathcal{T}$,

$$\begin{aligned} \mathcal{N}(\delta, F_m, \|.\|_n) &\leq \text{card}(\mathcal{B})\, \mathcal{N}\left(\frac{\delta}{4B_n}, B_1^K(KA_\sigma), \|.\|_1 \right) \mathcal{N}\left(\frac{\delta}{4B_n}, B_1^K(1), \|.\|_1 \right) \\ &\leq (2p+1)^{\frac{4B_n^2 K^2 m^2 \|\boldsymbol{x}\|_{\max,n}^2}{\delta^2}} \left(1 + \frac{8B_n KA_\sigma}{\delta} \right)^K \left(1 + \frac{8B_n}{\delta} \right)^K \end{aligned} \tag{3.55}$$

where we get the last inequality by using the fact that $\mathcal{N}(\delta, B_1^K(R), \|.\|_1) \leq (1 + 2R/\delta)^K$ (see for instance Pisier, 1999). □

Remark 6. Note that S_m defined by (3.14) is such that $\sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}| \leq m$. Yet, in (3.54), we only use that $\max_{k=1,\ldots,K} \sum_{j=1}^{p} |\beta_{kj}| \leq m$. To use the whole assumption $\sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}| \leq m$, we should consider $\sum_{k=1}^{K} \sum_{j=1}^{p} (\beta_{kj} - \beta'_{kj}) x_{ij}/m$ instead of $\sum_{j=1}^{p} (\beta_{kj} - \beta'_{kj}) x_{ij}/m$ in (3.53). This could be possible if $\sum_{k=1}^{K} |\boldsymbol{\beta_k^T x_i} - \boldsymbol{\beta'_k}^T \boldsymbol{x_i}|$ was replaced by $|\sum_{k=1}^{K} \boldsymbol{\beta_k^T x_i} - \boldsymbol{\beta'_k}^T \boldsymbol{x_i}|$ in the right-hand side of (3.51). This would require to consider the single parameter $\sum_{k=1}^{K} \boldsymbol{\beta_k^T x}$ in place of the K parameters $(\boldsymbol{\beta_1^T x}, \ldots, \boldsymbol{\beta_K^T x})$ in the parameter $\boldsymbol{\theta}(\boldsymbol{x})$. But it seems difficult to differentiate

$\ln(s_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}))$ with respect to $\sum_{k=1}^{K} \boldsymbol{\beta}_{\boldsymbol{k}}^T \boldsymbol{x}$. Yet, if one managed to do that, this would avoid to use Cauchy-Schwarz Inequality and the K-factor in (3.52) would be eliminated. In this case, the term $(2p+1)^{4B_n^2K^2m^2\|\boldsymbol{x}\|_{\max,n}^2/\delta^2}$ in (3.55) would be improved by $(2p+1)^{4B_n^2Km^2\|\boldsymbol{x}\|_{\max,n}^2/\delta^2}$. Then, taking the square root of the entropy number in (3.47), the term $m\|\boldsymbol{x}\|_{\max,n} \ln n\sqrt{K\ln(2p+1)}$ in (3.28) would be replaced by $m\|\boldsymbol{x}\|_{\max,n} \ln n\sqrt{\ln(2p+1)}$, and the lower bound of the regularization parameter λ in (3.15) would be proportional to $\sqrt{K}$ instead of K.

Lemma 3.B.8. *Let $\delta > 0$ and $(x_{ij})_{1\leq i\leq n, 1\leq j\leq p} \in \mathbb{R}^{np}$. There exists a family $\mathcal{B}$ of $(2p+1)^{\|\boldsymbol{x}\|_{\max,n}^2/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $\eta \in \mathbb{R}^p$ checking $\|\eta\|_1 \leq 1$, there exists $\eta' \in \mathcal{B}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\left(\eta_j - \eta'_j\right)x_{ij}\right)^2 \leq \delta^2. \tag{3.56}$$

Proof. Consider the set of functions $\mathcal{F} = \{f_0, f_1^+, \ldots, f_p^+, f_1^-, \ldots, f_p^-\}$ defined by

$$\begin{cases} f_0 \equiv 0, \\ f_j^+ : \mathbb{R}^p \mapsto \mathbb{R},\ \boldsymbol{x} = (x_1, \ldots, x_p) \mapsto x_j, & j = 1, \ldots, p, \\ f_j^- : \mathbb{R}^p \mapsto \mathbb{R},\ \boldsymbol{x} = (x_1, \ldots, x_p) \mapsto -x_j, & j = 1, \ldots, p. \end{cases}$$

Denote by $\mathcal{C}_{\mathcal{F}}$ the convex hull of $\mathcal{F}$. Let $\delta > 0$. By applying Lemma 2.6.11 of van der Vaart and Wellner (1996) to $\mathcal{F}$, which is of cardinality $2p+1$, we deduce that there exists a packing family $\mathcal{G} \subset \mathcal{C}_{\mathcal{F}}$ of cardinality $(2p+1)^{(\text{diam}\mathcal{F}/\delta)^2}$ for $(\mathcal{C}_{\mathcal{F}}, \|.\|_n)$ where diam$\mathcal{F}$ is the diameter of $\mathcal{F}$ for $\|.\|_n$. Here, diam$\mathcal{F} = \|\boldsymbol{x}\|_{\max,n}$ defined by (3.3).
Now, let $\eta \in \mathbb{R}^p$ such that $\|\eta\|_1 \leq 1$ and introduce the function

$$f_\eta : \mathbb{R}^p \mapsto \mathbb{R},\ \boldsymbol{x} = (x_1, \ldots, x_p) \mapsto \sum_{j=1}^{p} \eta_j x_j. \tag{3.57}$$

For all $\boldsymbol{x} \in \mathbb{R}^p$,

$$f_\eta(\boldsymbol{x}) = \left(\sum_{j:\eta_j>0} |\eta_j| f_j^+(\boldsymbol{x}) + \sum_{j:\eta_j<0} |\eta_j| f_j^-(\boldsymbol{x}) + \left(1 - \sum_{j:\eta_j\neq 0} |\eta_j|\right) f_0(\boldsymbol{x})\right),$$

with $\sum_{j:\eta_j>0} |\eta_j| + \sum_{j:\eta_j<0} |\eta_j| + (1 - \sum_{j:\eta_j\neq 0} |\eta_j|) = 1$, $(1 - \sum_{j:\eta_j\neq 0} |\eta_j|) \geq 0$ and $|\eta_j| \geq 0$ for all $j \in \{1, \ldots, p\}$. So, f_η belongs to $\mathcal{C}_{\mathcal{F}}$ and there exists f'_η in $\mathcal{G}$ such that $\|f_\eta - f'_\eta\|_n \leq \delta$, that is to say

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_\eta(\boldsymbol{x_i}) - f'_\eta(\boldsymbol{x_i})\right)^2 \leq \delta^2. \tag{3.58}$$

Since f'_η belongs to $\mathcal{C}_{\mathcal{F}}$, there exist coefficients $\left(\alpha_0, \alpha_1^+, \ldots, \alpha_p^+, \alpha_1^-, \ldots, \alpha_p^-\right)$ such that $f'_\eta = \alpha_0 f_0 +$

$\sum_{j=1}^{p} \alpha_j^+ f_j^+ + \alpha_j^- f_j^-$, and for all $\boldsymbol{x} \in \mathbb{R}^p$,

$$f'_\eta(\boldsymbol{x}) = \alpha_0 f_0(\boldsymbol{x}) + \sum_{j=1}^{p} \alpha_j^+ f_j^+(\boldsymbol{x}) + \alpha_j^- f_j^-(\boldsymbol{x}) = \sum_{j=1}^{p} (\alpha_j^+ - \alpha_j^-) x_j = \sum_{j=1}^{p} \eta'_j x_j \tag{3.59}$$

if we put $\eta'_j := \alpha_j^+ - \alpha_j^-$ for all $j \in \{1, \ldots, p\}$. This way, for each function f'_η, we define $\eta' \in \mathbb{R}^p$ associated to f'_η. This leads to the construction of a family $\mathcal{B}$ of $(2p+1)^{\|\boldsymbol{x}\|^2_{\max,n}/\delta^2}$ vectors of $\mathbb{R}^p$. Inequality (3.56) is obtained from (3.58), (3.57) and (3.59). □

3.B.2.3 Proof of Lemma 3.B.6

Let $m \in \mathbb{N}^\star$. From Lemma 3.B.4, on the event $\mathcal{T}$, $\sup_{f_m \in F_m} \|f_m\|_n$ is bounded by

$$R_n := 2B_n \left(1 + K(A_\beta + A_\sigma)\right). \tag{3.60}$$

Besides, we deduce from Lemma 3.B.5 that on the event $\mathcal{T}$, for all $J \in \mathbb{N}^\star$,

$$\begin{aligned}
&\sum_{j=1}^{J} 2^{-j} \sqrt{\ln\left[1 + \mathcal{N}\left(2^{-j} R_n, F_m, \|.\|_n\right)\right]} \\
\leq &\sum_{j=1}^{J} 2^{-j} \sqrt{\ln\left[2\mathcal{N}\left(2^{-j} R_n, F_m, \|.\|_n\right)\right]} \\
\leq &\sum_{j=1}^{J} 2^{-j} \left[\sqrt{\ln 2} + \frac{2^{j+1} B_n K m \|\boldsymbol{x}\|_{\max,n}}{R_n} \sqrt{\ln(2p+1)}\right. \\
&\qquad \left. + \sqrt{K \ln\left[\left(1 + \frac{2^{j+3} B_n K A_\sigma}{R_n}\right)\left(1 + \frac{2^{j+3} B_n}{R_n}\right)\right]}\right].
\end{aligned} \tag{3.61}$$

But, from (3.60), $R_n \geq 2B_n \max(KA_\sigma, 1)$. And $1 \leq 2^{j+2}$ for all $j \in \mathbb{N}^\star$. So, (3.61) implies

$$\begin{aligned}
&\sum_{j=1}^{J} 2^{-j} \sqrt{\ln\left[1 + \mathcal{N}\left(2^{-j} R_n, F_m, \|.\|_n\right)\right]} \\
\leq &\sum_{j=1}^{J} 2^{-j} \left[\sqrt{\ln 2} + \frac{2^{j+1} B_n K m \|\boldsymbol{x}\|_{\max,n}}{R_n} \sqrt{\ln(2p+1)} + \sqrt{K \ln\left[(2^{j+3}) \times (2^{j+3})\right]}\right] \\
\leq &\sum_{j=1}^{J} 2^{-j} \left[\sqrt{\ln 2} + \frac{2^{j+1} B_n K m \|\boldsymbol{x}\|_{\max,n}}{R_n} \sqrt{\ln(2p+1)} + \sqrt{K}\sqrt{2(j+3)\ln 2}\right] \\
\leq &\frac{2B_n K m \|\boldsymbol{x}\|_{\max,n}}{R_n} \sqrt{\ln(2p+1)} J + \sqrt{K}\sqrt{2\ln 2} \sum_{j=1}^{J} 2^{-j}\sqrt{j} + \sqrt{\ln 2}(1 + \sqrt{6K}).
\end{aligned}$$

By using that $2^{-j}\sqrt{j} \leq (\sqrt{\mathrm{e}}/2)^j$ for all $j \in \mathbb{N}^\star$, we get

$$\begin{aligned}
&\sum_{j=1}^{J} 2^{-j}\sqrt{\ln\left[1+\mathcal{N}\left(2^{-j}R_n, F_m, \|.\|_n\right)\right]} \\
&\leq \frac{2B_n K m\|\boldsymbol{x}\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}J + \sqrt{\ln 2}\left(1+\sqrt{K}\left(\sqrt{6}+\frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)\right) \\
&\leq \frac{2B_n K m\|\boldsymbol{x}\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}J + \sqrt{K\ln 2}\left(1+\sqrt{6}+\frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right). \qquad (3.62)
\end{aligned}$$

Then, we derive from (3.47) and (3.62) that for all $J \in \mathbb{N}^\star$,

$$\begin{aligned}
&\mathbb{E}_X\left[\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_m(Y_i|\boldsymbol{x_i})\right|\right] \\
&\leq R_n\left[\frac{6}{\sqrt{n}}\left(\frac{2B_n K m\|\boldsymbol{x}\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}J + \sqrt{K\ln 2}\left(1+\sqrt{6}+\frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)\right)+2^{-J}\right]. \qquad (3.63)
\end{aligned}$$

Now, let us choose $J = \ln n/\ln 2$ so that the two terms depending on J in (3.63) are of the same order. For this value of J, $2^{-J} \leq 1/n$, and we deduce from (3.63) and (3.60) that

$$\begin{aligned}
&\mathbb{E}_X\left[\sup_{f_m\in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_m(Y_i|\boldsymbol{x_i})\right|\right] \\
&\leq \frac{12B_n K m\|\boldsymbol{x}\|_{\max,n}}{\sqrt{n}}\sqrt{\ln(2p+1)}\,\frac{\ln n}{\ln 2} \\
&\quad + 2B_n\left(1+K(A_\beta+A_\sigma)\right)\left(6\sqrt{\ln 2}\left(1+\sqrt{6}+\frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)\frac{\sqrt{K}}{\sqrt{n}}+\frac{1}{n}\right) \\
&\leq \frac{18B_n K m\|\boldsymbol{x}\|_{\max,n}}{\sqrt{n}}\sqrt{\ln(2p+1)}\,\ln n \\
&\quad + 2\frac{\sqrt{K}}{\sqrt{n}}B_n\left(1+K(A_\beta+A_\sigma)\right)\left(6\sqrt{\ln 2}\left(1+\sqrt{6}+\frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)+1\right) \\
&\leq 18\sqrt{K}\frac{B_n}{\sqrt{n}}\left[m\|\boldsymbol{x}\|_{\max,n}\sqrt{K\ln(2p+1)}\,\ln n + 6\left(1+K(A_\beta+A_\sigma)\right)\right].
\end{aligned}$$

□

Part II

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

Chapter 4

Our Lasso-MLE procedure for variable selection in model-based clustering

Contents

4.1. Introduction **127**

4.2. Variable selection for clustering **131**

4.2.1. Model-based clustering 131

4.2.2. Finite Gaussian mixture models 132

4.2.3. Relevant variables for isotropic spherical clusters 133

4.3. Presentation of two competitor variable selection procedures **135**

4.3.1. Maugis and Michel's procedure for low-dimensional data 135

4.3.2. Pan and Shen's Lasso procedure for high-dimensional data 137

4.4. Some discussion on empirical centering **140**

4.4.1. Empirical centering to detect relevant variables by ℓ_1-penalization 140

4.4.2. About empirical centering during the estimation step 144

4.5. Our variable selection procedure: the Lasso-MLE procedure **149**

4.5.1. Estimation of the parameters by MLEs rather than by Lasso estimators . . 149

4.5.2. An alternative to empirical centering for the estimation step: selection of the active variables 153

4.5.3. A non-asymptotic model selection criterion 156

4.5.4. Description of our Lasso-MLE procedure 157

4.6. A major application: functional data clustering **160**

4.6.1. Variable selection for functional data clustering 162

4.6.2. Our procedure for functional data clustering using wavelets 163

4.A. The EM algorithms **166**

4.A.1. The EM algorithm for model-based clustering 166
4.A.2. An EM algorithm for ℓ_1-penalized model-based clustering 168
4.A.3. The EM algorithm for Maugis and Michel's procedure 170
4.A.4. The EM algorithms for our Lasso-MLE procedure 171
4.B. Some details about our algorithms 173
4.B.1. Construction of a grid of regularization parameters 173
4.B.2. Initialization and stopping rules 175
4.C. Proofs 176
4.C.1. Proof of Claim 4.4.2 176
4.C.2. Proof of Proposition 4.4.1 177

ABSTRACT

This chapter focuses on variable selection for clustering based on finite Gaussian mixture models with common spherical covariance matrix. We propose a global model selection procedure to simultaneously choose a number of clusters and a set of relevant variables for the clustering. It is especially suited to deal with high dimension, low sample size settings.

Following Pan and Shen (2007), we consider an ℓ_1-penalized likelihood approach to perform automatic variable selection and construct sets of potentially relevant variables. This results in an efficient construction of a data-driven model collection with reasonable cardinality even for high-dimensional datasets.

Our procedure differs from Pan and Shen's approach as regards the estimation of the mixture parameters in each model. First, we advocate for an estimation by the maximum likelihood estimator rather than by the ℓ_1-penalized maximum likelihood estimator (the Lasso) to avoid estimation problems due to ℓ_1-penalization shrinkage. Secondly, we do not perform empirical centering of the data during the estimation step because we prove that empirical centering can highly deteriorate estimation in high-dimensional contexts.

Finally, a model is selected by considering a non-asymptotic data-driven model selection criterion based on the slope heuristics introduced by Birgé and Massart (2006), and a data clustering is derived from the MAP principle.

4.1 Introduction

The goal of clustering methods is to discover structures (clusters) among individuals described by several variables. Specifically, given n observations $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ described by p variables ($\boldsymbol{Y_i} \in \mathbb{R}^p$), one aims at grouping the data into a few clusters such that the observations in the same cluster are more similar to each other than those from the other clusters. Many clustering methods exist and roughly fall into two categories. The first strategy is distance-based clustering. It includes hierarchical clusterings which build trees and K-means algorithms which classify data by minimizing the within-cluster sum of squares over all possible cluster centers. The second category is model-based clustering: each cluster is represented by a parametric distribution, the entire dataset is modeled by a mixture of these distributions, and a criterion is used to optimize the fit between the data and the model. An advantage of model-based clustering is to provide a rigorous statistical framework to assess the number of clusters and the role of each variable in the clustering process.

In principle, the more information one has about each individual, the better a clustering method

is expected to perform. However, the structure of interest may often be contained into a subset of the available variables and many attributes Y_{ij} of $\boldsymbol{Y_i}$ may be useless or even harmful to detect a reasonable clustering structure. Thus, it is important to select the relevant variables from the cluster analysis viewpoint. In addition, removing the irrelevant variables enables to get simpler modeling and may largely enhance interpretability. This interest in variable selection for clustering is motivated by the increasing use of high-dimensional datasets generated by recent advances in high-throughput biotechnologies. For instance, biologists try to extract groups of co-expressed genes according to transcriptome datasets in order to characterize more precisely their biological functions (Sharan et al., 2002; Jiang et al., 2004). But among all the transcriptome experiments considered, only a few of them are actually sufficient to reveal important biological phenomena. Selecting those experiments is expected to improve the clustering and the understanding of the underlying biological phenomena.

In different fields of applications, the observations are not vectors described by a great amount of variables but rather continuous functions, such as curves, spectra, time series or more generally signals. Ramsay (2006) gives the name "functional data analysis" to the analysis of such data. Functional data often arise from measurements on fine time grids, and if the sampling grid is sufficiently dense, the resulting data may be viewed as a sample of curves. These curves may vary in shape, amplitude and/or in phase. Examples involving functional data are studies on electricity consumption forecasting (Antoniadis et al., 2011), oil production estimation (Michel, 2008), nuclear reactor life span (Auder and Fischer, 2011)... For such applications, the main interest is to make accurate forecasts of a functional time series when the most recent curve is observed. Given a sample of curves, an important task is to search for homogeneous subgroups of curves using clustering, in order to determine accurate representative curve patterns for each cluster. For instance, as regards the French electricity consumption forecasting, daily profiles of the electricity power demand depend on factors such as seasons, holidays, working-days or weekends. These factors lead to different curve profiles. Performing a partition of these curves enables to construct an estimation of each profile by taking into account only the curves belonging to the corresponding cluster. This clustering process results in curve estimation improvement and thus in forecasting improvement, which helps EDF to predict more efficiently the French electricity power consumption at any given time of the year (Antoniadis et al., 2011). For such infinite dimensional data, curves are usually projected onto a functional basis such as a B-splines, Fourier or wavelet bases (Ramsay, 2006). By this process, each infinite dimensional curve is transformed into a high-dimensional vector constituted of the basis coefficients of the curve. So, curve clustering is recast into finite high-dimensional data clustering.

To perform variable selection, one traditional method is to conduct exhaustive best subset selection. For instance, in a finite Gaussian mixture model context, this method is tested by Maugis and Michel (2011a). However, Maugis and Michel (2011a) must restrict to very low-dimensional datasets ($p \approx 10$ for complete variable selection and $p \approx 30$ for ordered variable selection) since complete

variable selection, or even ordered variable selection, is unfeasible for moderate and high-dimensional data. For instance, with $p = 1000$, more than 10^{300} possible models are to be considered, which is prohibitive given the current standard computing power. In practice, two types of variable selection approaches are envisaged. On the one hand, the "filter" approaches select the variables before the clustering analysis (Dash et al., 2002; Jouve and Nicoloyannis, 2005; Misiti et al., 2007a; Auder and Fischer, 2011). Their main weakness is the independence between the selection step and the clustering step. In contrast, the "wrapper" approaches combine variable selection and clustering. For distance-based clustering, one can cite Fowlkes et al. (1988) for a forward selection approach with complete linkage hierarchical clustering, Devaney and Ram (1997) who propose a stepwise algorithm where the quality of the feature subsets is measured with the COBWEB algorithm, or the method of Brusco and Cradit (2001) based on the adjusted Rand index for K-means clustering. As regards model-based clustering, most wrapper methods have been developed with a Bayesian approach. For instance, Raftery and Dean (2006) and then Maugis et al. (2009) propose an approach similar to stepwise variable selection in regression which consists in sequentially comparing two nested models to determine whether an attribute must be included in or excluded from the current model based on a greedy search using Bayes factor. Rather than considering a Bayesian approach, Pan and Shen (2007) look at variable selection in model-based clustering from a frequentist point of view. In light of the success of variable selection via ℓ_1-penalization and the Lasso estimator in regression, Pan and Shen (2007) conjecture that ℓ_1-penalization may also be viable to variable selection for clustering. Therefore, they propose an ℓ_1-penalized model-based clustering approach.

From our point of view, taking advantage of the sparsity property of ℓ_1-penalization to efficiently construct sets of potentially relevant variables for clustering for high-dimensional data is an idea which is worth exploring. In particular, the simulations presented by Pan and Shen (2007), and then extended by Xie et al. (2008) and Zhou et al. (2009), are promising. Nonetheless, by analyzing Pan and Shen's procedure, we have pointed three drawbacks and we shall propose one remedy for each of this drawback:

- First, Pan and Shen (2007) use the Lasso not only to construct sets of relevant variables, but also to estimate the mixture parameters. Yet, ℓ_1-penalization induces shrinkage of the coefficients and thus biased estimators with high estimation risk. But for some problems such as curve clustering, it is crucial to get good estimations. To cope with this problem, we rather propose to use ℓ_1-penalization only to construct a collection of sets of relevant variables by varying the regularization parameter. Then, we rather advocate for an estimation by the Maximum Likelihood Estimator (MLE) in each model preselected by the Lasso algorithm. These unbiased estimators are expected to lead to much better estimation.

- Second, Pan and Shen (2007) perform empirical centering of the data before estimating the mixture parameters. Thus, they get an estimation of the mean parameters of the empirically

centered dataset. To derive an estimation of the mean parameters of the original (non-empirically centered) dataset, one must add the empirical mean to each mean coefficient to compensate for empirical centering. This process requires the estimation of the empirical mean for each variable. For high-dimensional data where the number of variables largely exceeds the number of observations, this may cause estimation problems. We rather advocate for estimation on the original data by performing preliminary dimensional reduction thanks to an additional ℓ_1-penalization.

- Third, Pan and Shen (2007) choose a modified BIC criterion to select their final Lasso estimator. We do not think that this asymptotic criterion is suited to high-dimensional data where the number of observations is small compared with the number of variables. We rather suggest to use a data-driven penalized criterion based on the non-asymptotic slope heuristics introduced by Birgé and Massart (2006). Since our estimators are MLEs (not Lasso estimators), we use an ℓ_0-penalization.

These three modifications brought to Pan and Shen's procedure lead to a new variable selection procedure in model-based clustering especially suited to high-dimensional settings. We call it the Lasso-MLE procedure: "Lasso" to indicate that we construct a model collection by the Lasso algorithm, and "MLE" to indicate that the parameters are estimated by the MLE in each model and that a final MLE estimator is selected by an ℓ_0-penalization.

The chapter is organized as follows. In Section 4.2, we introduce our framework, which is clustering based on isotropic spherical finite Gaussian mixture models, and we specify the meaning of a relevant variable for the clustering in this context. In Section 4.3, we present two variable selection procedures for model-based clustering. The first one, proposed by Maugis and Michel (2011b,a), deals with complete or ordered variable selection and is thus only feasible for very low-dimensional settings, while the second one is Pan and Shen's Lasso procedure introduced to cope with high-dimensional settings. Both procedures assume prior empirical centering of the data. Thus, in Section 4.4, we carry out some discussion on empirical centering, which is common practice in statistics but which may be to avoid in high-dimensional contexts when focusing on density estimation. We detail our variable selection procedure for model-based clustering in Section 4.5. This Lasso-MLE procedure is a mixture of both Maugis and Michel's procedure and Pan and Shen's procedure, with the difference that we avoid empirical centering for the parameter estimation step. Our procedure is particularly suited to functional data clustering. In Section 4.6, we present it in this specific context. The description of the EM algorithms, the initialization and stopping rules, and the construction of our grid of regularization parameters for the Lasso are postponed until Appendices.

4.2 Variable selection for clustering

4.2.1 Model-based clustering

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with unknown density s. Assume that the data come from several subpopulations, each of them having its own (biological, physical, sociological...) characteristics, and that the overall population is a mixture of these subpopulations. We aim at finding the number K of subpopulations and we want to determine from which subpopulation each data $\boldsymbol{Y_i}$ arises. In other words, we want to get a partition of the n data into some finite number K of clusters. We can see this problem as a missing data problem: the complete data are $((\boldsymbol{Y_1}, \boldsymbol{Z_1}), \ldots, (\boldsymbol{Y_n}, \boldsymbol{Z_n}))$ where the latent variables are $\boldsymbol{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_n})$ with $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iK})$ such that

$$Z_{ik} = \begin{cases} 1 & \text{if } \boldsymbol{Y_i} \text{ arises from subpopulation } k, \\ 0 & \text{otherwise.} \end{cases}$$

Model-based clustering consists in modeling each subpopulation by one component whose parameters are linked to its characteristics. The distribution of the whole population is then the mixture of those components, weighted by the proportion of individuals in each subpopulation. In this thesis, we restrict to Gaussian mixture models. In this case, each mixture component is modeled by a Gaussian density. The principle of model-based clustering is that identifying the correct mixing proportions and the parameters of each Gaussian density may enable to determine the latent variables $\boldsymbol{Z}$ and thus to derive from which cluster arises each data $\boldsymbol{Y_i}$. Let us formalize this method.

Let $K \in \mathbb{N}^\star$. The distribution of a random vector of $\mathbb{R}^p$ is a finite Gaussian mixture with K components if its density function with respect to Lebesgue measure is a convex combination of K Gaussian densities on $\mathbb{R}^p$:

$$\boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \Sigma_k).$$

The π_ks are the mixing proportions and $\Phi(\cdot \mid \boldsymbol{\mu_k}, \Sigma_k)$ is the p-dimensional Gaussian density with mean $\boldsymbol{\mu_k}$ and covariance matrix Σ_k:

$$\forall\, \boldsymbol{y} \in \mathbb{R}^p, \quad \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \Sigma_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu_k})^T \Sigma_k^{-1} (\boldsymbol{y} - \boldsymbol{\mu_k})\right).$$

The overall parameter vector is denoted $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \Sigma_1, \ldots, \Sigma_K)$. It belongs to the set $\Pi_K \times (\mathbb{R}^p)^K \times (\mathbb{S}_+^p)^K$ where $\mathbb{S}_+^p$ is the set of symmetric positive definite matrices on $\mathbb{R}^p$ and $\Pi_K = \left\{(\pi_1, \ldots, \pi_K);\ \pi_k > 0 \text{ for all } k \in \{1, \ldots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$.

If the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ comes from a probability distribution with density s which is a finite Gaussian mixture density on $\mathbb{R}^p$, there exists $\boldsymbol{\theta} \in \Pi_K \times (\mathbb{R}^p)^K \times (\mathbb{S}_+^p)^K$ such that $s = s_{\boldsymbol{\theta}} = \sum_{k=1}^K \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \Sigma_k)$. From the model-based clustering viewpoint, the parameter $\boldsymbol{\theta}$ can be used to determine a data clustering. Specifically, for all $i \in \{1, \ldots, n\}$, for all $k \in \{1, \ldots, K\}$, consider

$$\tau_{ik}(\boldsymbol{\theta}) = \frac{\pi_k \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}, \Sigma_k\right)}{\sum_{l=1}^K \pi_l \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_l}, \Sigma_l\right)} \tag{4.1}$$

the posterior probability of $\boldsymbol{Y_i}$ coming from component k. Then, the data are partitioned by applying the following rule, called the Maximum A Posteriori (MAP) principle:

$$Z_{ik}(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \tau_{ik}(\boldsymbol{\theta}) > \tau_{il}(\boldsymbol{\theta}) \; \forall \, l \neq k, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

4.2.2 Finite Gaussian mixture models

In practice, the density s of the data is unknown and is to be estimated by a finite Gaussian mixture density $\hat{s} = s_{\hat{\boldsymbol{\theta}}}$ for some $\hat{\boldsymbol{\theta}} \in \Pi_K \times (\mathbb{R}^p)^K \times (\mathbb{S}_+^p)^K$. If $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K)$, then one can estimate the posterior probability $\hat{\tau}_{ik} := \tau_{ik}(\hat{\boldsymbol{\theta}})$ of observation $\boldsymbol{Y_i}$ arising from component k, and obtain a data clustering by applying the MAP principle: $\widehat{Z}_{ik} := Z_{ik}(\hat{\boldsymbol{\theta}})$. To determine an estimator $s_{\hat{\boldsymbol{\theta}}}$, one can first construct a collection of estimators of s as finite Gaussian mixture densities and then choose an estimator among this collection. This can be formalized by introducing the notion of finite Gaussian mixture models.

A finite Gaussian mixture model is defined as a subset of finite Gaussian mixture densities,

$$\mathcal{S}_K = \left\{ \begin{array}{l} s_{\boldsymbol{\theta}} = \sum_{k=1}^K \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \Sigma_k); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \Sigma_1, \ldots, \Sigma_K) \in \Theta_K \end{array} \right\}$$

where $\Theta_K \subset \Pi_K \times (\mathbb{R}^p)^K \times (\mathbb{S}_+^p)^K$ depends on the constraints imposed on the parameter vectors $\boldsymbol{\theta}$. These constraints may be introduced either for technical reasons or to model geometrical assumptions on the clusters. For instance, one can impose the volumes, shapes and/or orientations to be equal for all components. One can also restrict to spherical or diagonal components. These geometric constraints induce constraints on the form of the covariance matrices. One can also add constraints on the proportions of the mixture. This way, Banfield and Raftery (1993) and then Celeux and Govaert (1995) define and study 28 different finite Gaussian mixture models which are more or less parsimonious.

In this thesis, we focus only on mean parameters, not variance parameters. Thus, we consider models where all clusters share the same form and the same volume. Besides, we assume that the clusters are spherical. This is equivalent to restricting to isotropic covariance matrices proportional to

the identity matrix: there exists $\sigma^2 > 0$ such that $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 \boldsymbol{I}$. Then, the clusters distinguish themselves only by the position of their center, which is characterized by the mean parameters. We allow free mixing parameters. This leads to finite Gaussian mixture models of the form

$$\mathcal{S}_K = \left\{ \begin{array}{l} s_{\boldsymbol{\theta}} = \sum_{k=1}^K \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+ \end{array} \right\} \tag{4.3}$$

with

$$\boldsymbol{y} = (y_1, \ldots, y_p) \in \mathbb{R}^p \mapsto \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^p (y_j - \mu_{kj})^2 \right).$$

4.2.3 Relevant variables for the clustering with isotropic spherical clusters

Currently, statistics deals with problems where data are explained by many variables. In principle, the more information one has about each individual, the better a clustering method is expected to perform. Nevertheless, some variables can be useless or even harmful to obtain a good data clustering. Thus, it is important to take into account the variable role in the clustering process. In this variable selection viewpoint, one is interested in identifying the set of irrelevant variables for the clustering, that is to say the set of variables that are not necessary to realize the clustering. A major point is to characterize such variables.

Let $\mathcal{J}$ be the collection of the non-empty subsets of $\{1, \ldots, p\}$. For all $\boldsymbol{J_r} \in \mathcal{J}$, denote by $\boldsymbol{J_r^c} := \{1, \ldots, p\} \setminus \boldsymbol{J_r}$ the complementary set of $\boldsymbol{J_r}$. For all $\boldsymbol{y} = (y_1, \ldots, y_p) \in \mathbb{R}^p$, denote the restriction of $\boldsymbol{y}$ on $\boldsymbol{J_r}$ by

$$\boldsymbol{y}_{[\boldsymbol{J_r}]} = (y_j)_{j \in \boldsymbol{J_r}} \, .$$

Consider $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ a sample of n observations described by p variables ($\boldsymbol{Y_i} \in \mathbb{R}^p$) with unknown density s. Assume that these observations come from K subpopulations (clusters), each of these subpopulations being distributed with a Gaussian density so that the density $s = \sum_{k=1}^K \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)$ is a finite Gaussian mixture density. Moreover, suppose that there exists $\boldsymbol{J_r} \subset \mathcal{J}$ such that for all $j \in \boldsymbol{J_r^c}$, the mean coefficients μ_{kj} are the same for all $k \in \{1, \ldots, K\}$: for $j \in \boldsymbol{J_r^c}$, there exists $\mu_{.j} \in \mathbb{R}$ such that

$$\mu_{1j} = \cdots = \mu_{Kj} = \mu_{.j}. \tag{4.4}$$

Put $\boldsymbol{\mu} := (\mu_{.j})_{j \in \boldsymbol{J_r^c}}$. Then, we get from (4.4) that for all $k \in \{1, \ldots, K\}$, $\boldsymbol{\mu_{k[J_r^c]}} = \boldsymbol{\mu}$ and for

all $i \in \{1, \ldots, n\}$,

$$\sum_{k=1}^{K} \pi_k \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right) = \Phi\left(\boldsymbol{Y_{i[J_r^c]}} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \sum_{k=1}^{K} \pi_k \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_{k\,[J_r]}}, \sigma^2 \boldsymbol{I}\right). \tag{4.5}$$

In (4.5), since $\boldsymbol{\mu}$ does not depend on the clusters, the variables indexed by $j \in \boldsymbol{J_r^c}$ provide no information to determine from which cluster the observation $\boldsymbol{Y_i}$ arises. This is highlighted by the calculation of the posterior probability of $\boldsymbol{Y_i}$ arising from component k:

$$\tau_{ik} = \frac{\pi_k \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \pi_l \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_l}, \sigma^2 \boldsymbol{I}\right)} \tag{4.6}$$

$$= \frac{\pi_k \, \Phi\left(\boldsymbol{Y_{i[J_r^c]}} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_{k\,[J_r]}}, \sigma^2 \boldsymbol{I}\right)}{\Phi\left(\boldsymbol{Y_{i[J_r^c]}} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \sum_{l=1}^{K} \pi_l \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_{l\,[J_r]}}, \sigma^2 \boldsymbol{I}\right)} \tag{4.7}$$

$$= \frac{\pi_k \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_{k\,[J_r]}}, \sigma^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \pi_l \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_{l\,[J_r]}}, \sigma^2 \boldsymbol{I}\right)}. \tag{4.8}$$

Since the term depending on $\boldsymbol{J_r^c}$ in the mixture density function cancels out from the numerator and the denominator of τ_{ik} in (4.7), the calculation of τ_{ik} is unchanged if we consider the set $\boldsymbol{J_r}$ (such as in (4.8)) rather than the whole set $\{1, \ldots, p\}$ (such as in (4.6)). Then, the data clustering obtained by the MAP principle described at (4.2) is also unchanged. Thus, the variables indexed by $j \in \boldsymbol{J_r^c}$ are useless for the clustering while the variables indexed by $j \in \boldsymbol{J_r}$ are sufficient to perform the clustering. We shall say that the variables indexed by $j \in \boldsymbol{J_r^c}$ are *irrelevant* for the clustering.

In practice, the number $K \in \mathbb{N}^\star$ of clusters and the index set $\boldsymbol{J_r} \in \mathcal{J}$ representing the relevant variables are to be determined. By varying K in a finite subset $\mathcal{K} = \{K_{\min}, \ldots, K_{\max}\}$ and by varying $\boldsymbol{J_r}$ in a subset $\mathcal{J}' \subset \mathcal{J}$, one gets a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{K} \times \mathcal{J}}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{4.9}$$

Each model $\mathcal{S}_{(K,\boldsymbol{J_r})}$ involves a data clustering with K clusters and $\boldsymbol{J_r}$ as representative set of relevant variables for the clustering. Typically, a variable selection procedure for clustering proceeds in four steps:

1. construction of a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{K} \times \mathcal{J}'}$;
2. for each $(K, \boldsymbol{J_r}) \in \mathcal{K} \times \mathcal{J}'$, determination of an estimator $\hat{s}_{(K,\boldsymbol{J_r})}$ of s in $\mathcal{S}_{(K,\boldsymbol{J_r})}$;
3. selection of a model $\mathcal{S}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ using a model selection criterion;

4. determination of a data clustering by the MAP principle using the estimated mixture parameters of $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$.

In Section 4.5.4, we shall propose a variable selection procedure both identifying the set of relevant variables for the clustering and providing a data clustering. Before introducing it, we present two competitor variable selection procedures (Maugis and Michel, 2011a; Pan and Shen, 2007) from which we have drawn our inspiration to construct our new procedure as a mixture of these two procedures.

4.3 Presentation of two competitor variable selection procedures

For each procedure, we detail the four steps mentioned above. The model collection is obtained by considering a range of number of clusters $\mathcal{K} = \{K_{\min}, \ldots, K_{\max}\}$ and a collection of index sets $\mathcal{J}'$ representing sets of relevant variables. The collection $\mathcal{K}$ does not depend on the procedure whereas the collection $\mathcal{J}'$ depends on the procedure. In this thesis, we only consider isotropic spherical finite Gaussian mixture models defined by (4.3). Thus, we present each procedure in this context although Maugis and Michel (2011a) and Pan and Shen (2007) studied less parsimonious models.

In the sequel, we consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from probability distribution with density s. We assume that the data come from different subpopulations. In a clustering purpose, s is to be estimated by a finite Gaussian mixture density.

4.3.1 Maugis and Michel's procedure for low-dimensional data

Maugis and Michel (2011a) propose a clustering procedure selecting the relevant variables in a low-dimensional context with $p \ll n$. The novelty of their method is more about their model selection criterion than about their model collection. Indeed, their model collection involves all the subsets of $\{1, \ldots, p\}$ as index set of potentially relevant variables. In this case,

$$\mathcal{J}' = \{\{j_1, \ldots, j_d\}; 1 \leq d \leq p,\ j_l \in \{1, \ldots, p\} \text{ for all } l \in \{1, \ldots, d\},\ j_l \neq j_{l'}\}. \tag{4.10}$$

This corresponds to complete variable selection and this results in a model collection of size $2^p \times \text{card}(\mathcal{K})$. Calculating an estimator in each model is feasible only when p is sufficiently small, in practice $p \leq 10$. When $p \geq 10$, Maugis and Michel (2011a) must restrict to ordered variable selection and rather consider the collection of all nested subsets of $\{1, \ldots, p\}$ in order to get a practicable method. In this case,

$$\mathcal{J}' = \{\{1, \ldots, d\}; 1 \leq d \leq p\}. \tag{4.11}$$

Nonetheless, even ordered variable selection becomes unfeasible as soon as $p \approx 30$. Thus, Maugis and Michel's procedure is only valid for very low-dimensional data. One aim of this thesis is to extend

their procedure to high-dimensional data by introducing a collection with fewer index sets in order to reduce the number of models and the number of estimators to calculate.

Step 0. Empirical centering

Maugis and Michel (2011a) center the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ by subtracting $\sum_{l=1}^{n} Y_{lj}/n$ to Y_{ij} for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. This leads to a new dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$ [1]. The density $\overline{s}$ of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ is to be estimated to get a clustering of the dataset $\overline{\boldsymbol{Y}}$.

Step 1. Model collection

Since the data are empirically centered, Maugis and Michel (2011a) consider that the irrelevant variables for the clustering of $\overline{\boldsymbol{Y}}$ have a homogeneous behavior around a null mean[2]. Put $\mathcal{M}_r = \mathcal{K} \times \mathcal{J}'$. Then, Maugis and Michel's model collection is $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{M}_r}$ with

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{4.12}$$

Each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ is the set of finite Gaussian mixture densities with K components and $\boldsymbol{J_r}$ as index set representing the relevant variables.

Step 2. Calculation of an estimator of $\overline{s}$ in each model

For each $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, Maugis and Michel (2011a) compute the maximum likelihood estimator of $\overline{s}$ in the model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$:

$$\widehat{\overline{s}}_{(K,\boldsymbol{J_r})} = \underset{s_{\overline{\boldsymbol{\theta}}} \in \overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}}{\arg\min} \ \gamma_n\left(s_{\overline{\boldsymbol{\theta}}}\right), \qquad \gamma_n\left(s_{\overline{\boldsymbol{\theta}}}\right) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})\right).$$

For all $\boldsymbol{y} \in \mathbb{R}^p$,

$$\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \hat{\sigma}^2 \boldsymbol{I}\right) \sum_{k=1}^{K} \hat{\pi}_k \, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right).$$

The estimated mixture parameters $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_r})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{1}}, \ldots, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{K}}, \hat{\sigma})$ are computed by the EM algorithm for model-based clustering (Dempster et al., 1977) described in Section 4.A.3. In practice, Maugis and Michel (2011a) use MIXMOD software (Biernacki et al., 2006) to run the EM algorithm.

[1] We shall add an horizontal line to quantities modified by empirical centering, that is to say the dataset $\boldsymbol{Y}$, the density s, the mean vectors $\boldsymbol{\mu_k}$, the global parameter vector $\boldsymbol{\theta}$, the model collection as well as the estimations $\hat{s}$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}$, $\hat{\boldsymbol{\theta}}$. The proportions π_k and the variance σ^2 are not modified by empirical centering (see proposition 4.4.1).

[2] In Section 4.4.1, we justify this modeling.

Step 3. Model selection

The model selection criterion used by Maugis and Michel (2011a) is the data-driven penalized criterion derived from the slope heuristics introduced by Birgé and Massart (2006) and recalled in Section 6.3.1. First, models are grouped according to their dimension D in order to obtain a model collection $\{\overline{\mathcal{S}}_D\}_{D\in\mathcal{D}}$. The dimension of a model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ is the total number of free parameters estimated in the model: $K-1$ mixing proportions, $K|\boldsymbol{J_r}|$ mean parameters and 1 variance parameter, which gives a total dimension equal to $K(1+|\boldsymbol{J_r}|)$. For each dimension $D\in\mathcal{D}$, let $\widehat{\overline{s}}_D$ be the estimator maximizing the likelihood among the estimators associated to a model of dimension D. There exists $(K_D,\boldsymbol{J}_D)$ such that $\widehat{\overline{s}}_D=\widehat{\overline{s}}_{(K_D,\boldsymbol{J}_D)}$. Then, the function $D/n\mapsto-\gamma_n(\widehat{\overline{s}}_D)$ is plotted. One has to check that this function has a linear behavior for large dimensions, otherwise the slope heuristics can not be applied. If a linear behavior is indeed observed, the slope $\hat{c}$ of the linear part is evaluated thanks to the graphical interface CAPUSHE (Baudry et al., 2011). The calibrated penalty function is $\mathrm{pen}(D)=2\hat{c}D/n$. Finally, the minimizer $\hat{D}$ of the penalized criterion

$$\hat{D}=\underset{D\in\mathcal{D}}{\arg\min}\left\{\gamma_n(\widehat{\overline{s}}_D)+2\hat{c}\,\frac{D}{n}\right\} \tag{4.13}$$

is determined and the model $\overline{\mathcal{S}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}:=\overline{\mathcal{S}}_{(K_{\hat{D}},\boldsymbol{J}_{\hat{D}})}$ is selected.

Step 4. Data clustering

The variables declared as relevant for the clustering are indexed by $\hat{\boldsymbol{J}}_{\boldsymbol{r}}$. The density $\overline{s}$ is estimated by $\widehat{\overline{s}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$. A clustering of the dataset $\overline{\boldsymbol{Y}}$ is derived from $\hat{\overline{\boldsymbol{\theta}}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ by applying the MAP principle.

4.3.2 Pan and Shen's Lasso procedure for high-dimensional data

Complete or ordered variable selection considered by Maugis and Michel (2011a) is untractable unless p is very small. Alternative variable selection procedures are to be considered for high-dimensional data. Pan and Shen (2007) propose a penalized model-based clustering approach. In light of the success of variable selection via ℓ_1-penalization in regression, Pan and Shen (2007) conjecture that ℓ_1-penalization may also be viable to variable selection for clustering. They try an appropriate ℓ_1-penalty function to adaptively shrink the mean parameters towards the average cluster means, resulting in automatic variable selection[3]. The selection of the relevant variables and the estimation of the parameters are performed during the same process. Contrary to Maugis and Michel (2011a), the novelty of Pan and Shen's procedure is more about their model collection construction than about their model selection criterion. Unlike Maugis and Michel (2011a), Pan and Shen (2007) do not consider a deterministic model collection. They rather construct a random model collection derived from a collection of sets of potentially relevant variables determined by ℓ_1-penalization.

[3]In Section 4.4.1, we explain why this process is expected to select the relevant variables for the clustering.

Step 0. Empirical centering

Pan and Shen (2007) center the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ by subtracting $\sum_{l=1}^{n} Y_{lj}/n$ to Y_{ij} for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. This leads to a new dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$. The density $\overline{s}$ of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ is to be estimated to get a clustering of the dataset $\overline{\boldsymbol{Y}}$.

Steps 1 and 2. Construction of a model collection and calculation of an estimator of $\overline{s}$ in each model

- Fix $K \in \mathcal{K}$ and consider

$$\overline{\mathcal{S}}_K = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Theta_K \end{array} \right\}$$

  with $\Theta_K := \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+$. Since the dataset $\overline{\boldsymbol{Y}}$ is centered, irrelevant variables are expected to have a homogeneous behavior around a null mean[4]. Then, to detect such variables, Pan and Shen (2007) penalize the empirical contrast

$$\gamma_n\left(s_{\overline{\boldsymbol{\theta}}}\right) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})\right)$$

  by an ℓ_1-penalty on the mean parameters,

$$\lambda |\overline{\boldsymbol{\theta}}|_1 := \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} |\overline{\mu}_{kj}| \tag{4.14}$$

  where $\lambda > 0$ is a regularization parameter.
  Introduce G_K a grid of regularization parameters and let $\lambda \in G_K$.
  Consider the Lasso estimator defined by

$$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = \underset{\overline{\boldsymbol{\theta}} \in \Theta_K}{\arg\min} \left\{ \gamma_n(s_{\overline{\boldsymbol{\theta}}}) + \lambda |\overline{\boldsymbol{\theta}}|_1 \right\}. \tag{4.15}$$

  To compute $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = (\hat{\pi}_k, \widehat{\overline{\mu}}_{kj}, \hat{\sigma})_{1 \le k \le K, 1 \le j \le p}$, Pan and Shen (2007) construct an EM algorithm for ℓ_1-penalized model-based clustering (see Section 4.A.2). The index set

$$\boldsymbol{J}_{(K,\lambda)} = \{j \in \{1, \ldots, p\} : \exists\, k \in \{1, \ldots, K\} \text{ such that } \widehat{\overline{\mu}}_{kj} \neq 0\}$$

  represents the set of relevant variables selected by the Lasso $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)}$. The density $\overline{s}$ is estimated

[4] In Section 4.4.1, we justify this modeling.

by the Lasso solution defined for all $\boldsymbol{y} \in \mathbb{R}^p$ by

$$\widehat{\overline{s}}_{(K,\boldsymbol{J}_{(K,\lambda)})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J}^c_{(K,\lambda)}]} \mid \boldsymbol{0}, \hat{\sigma}^2\boldsymbol{I}\right) \sum_{k=1}^{K} \hat{\pi}_k \, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J}_{(K,\lambda)}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right). \tag{4.16}$$

By keeping K fixed and by varying $\lambda \in G_K$, one gets a collection $\mathcal{J}_K = \cup_{\lambda \in G_K} \boldsymbol{J}_{(K,\lambda)}$ of index sets representing a collection of potentially relevant variables.

- By varying $K \in \mathcal{K}$, one gets a collection of index sets $\mathcal{M}_r = \{(K, \boldsymbol{J_r});\ K \in \mathcal{K}, \boldsymbol{J_r} \in \mathcal{J}_K\}$. This leads to a model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{M}_r}$ with

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J}^c_{\boldsymbol{r}}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I}) \ \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2\boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{4.17}$$

Each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ is the set of finite Gaussian mixture densities with K components and $\boldsymbol{J_r}$ as index set representing the relevant variables. In each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$, the density $\overline{s}$ is estimated by the Lasso $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$ defined by (4.16).

Step 3. Model selection

Pan and Shen (2007) consider BIC as model selection criterion. First, models are grouped according to their dimension D in order to obtain a model collection $\{\overline{\mathcal{S}}_D\}_{D \in \mathcal{D}}$. Following a conjecture of Efron et al. (2004) and a result of Zou et al. (2007), Pan and Shen (2007) calculate a model dimension by taking into account the sparsity of the model: the mean parameters set to zero by the EM algorithm are not considered in the dimension calculation. For a model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$, there are $K-1$ free mixing parameters, 1 variance parameter and $K|\boldsymbol{J_r}|$ non-zero mean parameters, which gives a dimension equal to $K(1+|\boldsymbol{J_r}|)$. For each dimension $D \in \mathcal{D}$, denote by $\widehat{\overline{s}}_D$ the Lasso estimator maximizing the likelihood among the Lasso estimators associated to a model of dimension D. There exists $(K_D, \boldsymbol{J}_D)$ such that $\widehat{\overline{s}}_D = \widehat{\overline{s}}_{(K_D,\boldsymbol{J}_D)}$. Then, the minimizer $\hat{D}$ of the BIC criterion

$$\hat{D} = \underset{D \in \mathcal{D}}{\arg\min} \left\{ \gamma_n(\widehat{\overline{s}}_D) + \frac{\ln n}{2} \frac{D}{n} \right\} \tag{4.18}$$

is determined and the model $\overline{\mathcal{S}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})} := \overline{\mathcal{S}}_{(K_{\hat{D}},\boldsymbol{J}_{\hat{D}})}$ is selected.

Step 4. Data clustering

The variables declared as relevant for the clustering are indexed by $\hat{\boldsymbol{J}}_{\boldsymbol{r}}$. The density $\overline{s}$ is estimated by the Lasso $\widehat{\overline{s}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$. A clustering of the dataset $\overline{\boldsymbol{Y}}$ is derived from the estimated Lasso parameter vector $\widehat{\overline{\boldsymbol{\theta}}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ by applying the MAP principle.

4.4 Some discussion on empirical centering

Before running their procedure, both Maugis and Michel (2011a) and Pan and Shen (2007) empirically center the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$. This operation is common practice in statistics. It enables to get a dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$ with small or moderate value $\overline{Y}_{ij}$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. This prevents from numerical divergence that may occur during the algorithm process if there are too many large observations Y_{ij}. These numerical problems are all the more likely to appear when considering high-dimensional datasets.

In this section, we explain the interest of empirical centering to construct a collection of sets of relevant variables. Then, we carry out some discussion on empirical centering for the estimation step.

In the sequel, we consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with unknown density s. To make the discussion simpler, we suppose that s is itself a finite Gaussian mixture density: for all $\boldsymbol{y} \in \mathbb{R}^p$, $s(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \Phi\left(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)$ with $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+$. But the arguments developed remain valid when s is rather to be estimated by such a finite Gaussian mixture density.

4.4.1 Empirical centering to construct sets of relevant variables by ℓ_1-penalization

From our point of view, Pan and Shen's approach of variable selection via ℓ_1-penalization is fruitful and we shall keep this idea in our own procedure (see Section 4.5). In particular, we shall borrow their EM algorithm for ℓ_1-penalized model-based clustering (see Section 4.A.2) in order to get a collection of sets of potentially relevant variables by computing the Lasso solution for various regularization parameters. Pan and Shen (2007) empirically center the dataset before running their EM algorithm, but they do not justify this preliminary step. Here, we are willing to prove the interest (and "necessity") of performing such a preliminary operation. More precisely, we shall come back to the definition of a relevant variable and see which suitable ℓ_1-penalty could be apply in order to detect such variables without performing empirical centering of the data. This way, we shall show that empirical centering enables to get round a too complex minimization problem.

4.4.1.1 ℓ_1-penalization to detect the irrelevant variables

From (4.4), a variable is irrelevant if it is indexed by $j \in \{1, \ldots, p\}$ such that there exists $\mu_{.j}$ such that $\mu_{kj} = \mu_{.j}$ for all $k \in \{1, \ldots, K\}$. This common value $\mu_{.j}$ can easily be computed. Since $\sum_{k=1}^{K} \pi_k = 1$, we have $\sum_{k=1}^{K} \pi_k \, \mu_{kj} = \mu_{.j} \sum_{k=1}^{K} \pi_k = \mu_{.j}$. So, a variable is irrelevant if it is indexed by j such that for all $k \in \{1, \ldots, K\}$,

$$\mu_{kj} = \sum_{l=1}^{K} \pi_l \, \mu_{lj}. \tag{4.19}$$

On the contrary, a variable is relevant if it is indexed by j such that there exists $k \in \{1, \ldots, K\}$ such that $\mu_{kj} \neq \sum_{l=1}^{K} \pi_l \, \mu_{lj}$.

Then, the degree of relevance of a variable can be measured by the number of $k \in \{1, \ldots, K\}$ such that $\mu_{kj} \neq \sum_{l=1}^{K} \pi_l \, \mu_{lj}$. The more $k \in \{1, \ldots, K\}$ such that $\mu_{kj} \neq \sum_{l=1}^{K} \pi_l \, \mu_{lj}$, the more relevant the variable indexed by j. In other words, the larger $\sum_{k=1}^{K} |\mu_{kj} - \sum_{l=1}^{K} \pi_l \, \mu_{lj}|$, the more relevant the variable indexed by j. In this viewpoint, one way to get rid of the irrelevant variables is to apply a penalty penalizing the distance between $\sum_{k=1}^{K} |\mu_{kj} - \sum_{l=1}^{K} \pi_l \, \mu_{lj}|$ and zero, for each $j \in \{1, \ldots, p\}$. A suitable penalty is a soft-thresholding ℓ_1-penalty proportional to

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \sum_{k=1}^{K} \left| \mu_{kj} - \sum_{l=1}^{K} \pi_l \, \mu_{lj} \right| . \tag{4.20}$$

For regularization parameter $\lambda > 0$, the corresponding estimator is the Lasso estimator defined by

$$\hat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta}} \left\{ \gamma_n(s_{\boldsymbol{\theta}}) + \lambda \, \mathrm{pen}(\boldsymbol{\theta}) \right\} . \tag{4.21}$$

To solve the minimization problem (4.21), one may use an EM algorithm adapted to ℓ_1-penalized model-based clustering. Similarly to the EM algorithm described in Section 4.A.2, the updating of the mean vectors $\boldsymbol{\mu_k}$ at step M of iteration r of the algorithm is given by the minimizer $\boldsymbol{\mu_k}^{(r+1)}$ of

$$\boldsymbol{\mu_k} \mapsto -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln \left(\Phi \left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2^{(r)}} \boldsymbol{I} \right) \right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \left| \mu_{kj}^{(r)} - \sum_{l=1}^{K} \pi_l^{(r)} \mu_{lj}^{(r)} \right| .$$

By differentiation, one finds that the solution is reached for mean parameters defined for all $k \in \{1, \ldots, K\}$ and for all $j \in \{1, \ldots, p\}$ by

$$\mu_{kj}^{(r+1)} = \mu_{kj}^{0\,(r+1)} - \frac{n\lambda \, \sigma^{2^{(r+1)}}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \, \mathrm{sign} \left(\mu_{kj}^{(r+1)} - \sum_{l=1}^{K} \pi_l^{(r+1)} \mu_{lj}^{(r+1)} \right) \tag{4.22}$$

with

$$\mu_{kj}^{0\,(r+1)} := \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} Y_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} .$$

The updating $\mu_{kj}^{(r+1)}$ of μ_{kj} depends on the updating of all the other mean coefficients μ_{lj}, $l \neq k$. Thus, one can not get an explicit formula for $\mu_{kj}^{(r+1)}$ and computing the solution (4.22) is difficult.

4.4.1.2 Centering of the data to compute the Lasso solution

The presence of the terms $\mu_{lj}^{(r+1)}$, $l \neq k$, in the updating formula of $\mu_{kj}^{(r+1)}$ in (4.22) comes from the penalty (4.20). So, to get an easier minimization problem than (4.22), an idea may be to eliminate the

terms μ_{lj} in the penalty (4.20) by centering the data. Let us precise this idea.

Claim 4.4.1. *Let $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ be a n-sample of p-dimensional random vectors whose probability distribution is $\sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)$. For all $i \in \{1, \ldots, n\}$, consider the random vector $\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} = (\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{ip})$ defined by*

$$\widetilde{Y}_{ij} = Y_{ij} - \sum_{l=1}^{K} \pi_l \, \mu_{lj}. \tag{4.23}$$

Then, $\widetilde{\boldsymbol{Y}} = (\widetilde{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}})$ is a n-sample of p-dimensional random vectors whose probability distribution is $\sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}\right)$ with $\tilde{\mu}_{kj} = \mu_{kj} - \sum_{l=1}^{K} \pi_l \, \mu_{lj}$ for all $k \in \{1, \ldots, K\}$ and for all $j \in \{1, \ldots, p\}$.

In particular, $\mu_{kj} = \sum_{l=1}^{K} \pi_l \, \mu_{lj}$ if and only if $\tilde{\mu}_{kj} = 0$.

From (4.19) and Claim 4.4.1, we see that performing the theoretical centering (4.23) of the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ enables to get a new dataset $\widetilde{\boldsymbol{Y}} = (\widetilde{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \widetilde{\boldsymbol{Y}}_{\boldsymbol{n}})$ for which a variable is irrelevant if it is indexed by $j \in \{1, \ldots, p\}$ such that the mean coefficients $\tilde{\mu}_{kj}$ equal zero for all $k \in \{1, \ldots, K\}$, that is to say if $\sum_{k=1}^{K} |\tilde{\mu}_{kj}| = 0$. On the contrary, a variable is relevant if it is indexed by j such that $\sum_{k=1}^{K} |\tilde{\mu}_{kj}| > 0$ and it is all the more relevant as $\sum_{k=1}^{K} |\tilde{\mu}_{kj}|$ is large. Then, a suitable penalty shape to get rid of the irrelevant variables becomes

$$|\tilde{\boldsymbol{\theta}}|_1 := \sum_{j=1}^{p} \sum_{k=1}^{K} |\tilde{\mu}_{kj}| \, . \tag{4.24}$$

For regularization parameter $\lambda > 0$, the corresponding estimator is the Lasso estimator defined by

$$\tilde{\boldsymbol{\theta}}(\lambda) := \arg\min_{\tilde{\boldsymbol{\theta}}} \left\{ \gamma_n(s_{\tilde{\boldsymbol{\theta}}}) + \lambda |\tilde{\boldsymbol{\theta}}|_1 \right\}. \tag{4.25}$$

The minimization problem (4.25) can be solved thanks to an EM algorithm (see Section 4.A.2). The updating of the mean vectors $\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}$ at step M of iteration r of the algorithm is given by the minimizer $\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}{}^{(r+1)}$ of

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}} \mapsto -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\widetilde{\boldsymbol{Y}}_{\boldsymbol{i}} \mid \tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}{}^{(r)}, \sigma^{2^{(r)}} \boldsymbol{I}\right)\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \left|\tilde{\mu}_{kj}^{(r)}\right|.$$

By differentiation, one finds that the solution is reached for mean parameters defined for all $k \in \{1, \ldots, K\}$ and for all $j \in \{1, \ldots, p\}$ by

$$\tilde{\mu}_{kj}^{(r+1)} = \operatorname{sign}\left(\tilde{\mu}_{kj}^{0(r+1)}\right) \left(\left|\tilde{\mu}_{kj}^{0(r+1)}\right| - \frac{n\lambda \, \sigma^{2^{(r+1)}}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \right)_+ \tag{4.26}$$

with

$$\tilde{\mu}_{kj}^{0(r+1)} := \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \widetilde{Y}_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \quad \text{and} \quad a_+ = \max\{a, 0\}.$$

The updating (4.26) of the mean parameters is now explicit and the Lasso solution $\tilde{\boldsymbol{\theta}}(\lambda)$ can be easily computed by the EM algorithm. Denote by $\tilde{\mu}_{kj}(\lambda)$ the mean parameters of $\tilde{\boldsymbol{\theta}}(\lambda)$. Then, the relevant variables detected by ℓ_1-penalization with regularization parameter λ are the variables indexed by j such that there exists $k \in \{1, \ldots, K\}$ such that $\tilde{\mu}_{kj}(\lambda) \neq 0$.

The advantage of performing theoretical centering of the dataset $\boldsymbol{Y}$ is to replace the implicit updating formula (4.22) by the explicit updating formula (4.26). Nonetheless, in practice, the theoretical means $\sum_{l=1}^{K} \pi_l \, \mu_{lj}$ are unknown, so we can not perform the theoretical transformation (4.23). Then, a solution may be to replace $\sum_{l=1}^{K} \pi_l \, \mu_{lj}$ by an approximating empirical quantity. Such a quantity can be derived from the following result.

Claim 4.4.2. *Let $(Y_1, \ldots, Y_p)$ be a p-dimensional random vector whose probability distribution is $\sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)$. Then, for all $j \in \{1, \ldots, p\}$, the expectation of the random variable Y_j is*

$$\mathbb{E}(Y_j) = \sum_{k=1}^{K} \pi_k \, \mu_{kj}. \tag{4.27}$$

Proof. Page 176. □

Since $\boldsymbol{Y_i} \sim \sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right)$ for all $i \in \{1, \ldots, n\}$, we deduce from (4.27) and the Law of large numbers that, for all $j \in \{1, \ldots, p\}$,

$$\frac{1}{n} \sum_{i=1}^{n} Y_{ij} \underset{n \to \infty}{\longrightarrow} \sum_{k=1}^{K} \pi_k \, \mu_{kj}. \tag{4.28}$$

Thus, we propose to replace theoretical centering of the data, which consists in subtracting the theoretical mean $\sum_{k=1}^{K} \pi_k \, \mu_{kj}$ to each Y_{ij}, by empirical centering, which consists in subtracting the empirical mean $\sum_{l=1}^{n} Y_{lj}/n$ to each Y_{ij}. From (4.28), we hope that performing empirical centering rather than theoretical centering won't affect too much the above reasoning and that it will lead to sets of relevant variables very close to the sets of relevant variables that would have been obtained by applying theoretical centering. Yet, from (4.28), this process is expected to lead to better results as the number n of observations is very large, which is not the case in practice when focusing on high-dimensional problems.

Remark 7. Let us point out that, for simulated data, one has access to the theoretical mean $\sum_{k=1}^{K} \pi_k \, \mu_{kj}$ for all $j \in \{1, \ldots, p\}$. In this specific case, one can thus apply theoretical (besides empirical) cente-

ring of the data. Then, by comparing the results for both methods, one can valid (or not) the method of empirical centering (see Section 5.3.3.1 for such a comparison).

To conclude, we think that empirical centering is essential to practically compute the Lasso solutions to get a collection of sets of potentially relevant variables for the clustering. In our own procedure, we shall thus perform empirical centering to construct such a collection (see Section 4.5.4).

4.4.2 About empirical centering during the estimation step

To provide a partition of the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ into a finite number of clusters via model-based clustering, one first estimates the unknown density s of the data $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ by a finite Gaussian mixture density. Then, the estimated mixture parameters enable to partition the data by the MAP principle. This way, model-based clustering recasts the clustering problem into a density estimation problem. Moreover, for some clustering problems, density estimation is not just an intermediate step to get a data clustering, and estimating the density can be as important as the clustering. For instance, one can think of curve clustering when one curve profile per cluster has to be estimated to get accurate forecasts of a given functional time series. Therefore, estimation is a crucial point in model-based clustering. Yet, during this thesis, we have been faced with estimation problems due to empirical centering. To solve these problems, we have thought about the consequences of performing empirical centering during the estimation step. Some repercussions are obvious while other ones are not so clear. Let us sum up our conclusions.

4.4.2.1 Consequence of empirical centering on the data structure

The first point to note is that one losses the independence of the n data $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ when performing empirical centering. The n empirically centered data $\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}}$ are no longer independent and one can no longer consider a common density $\overline{s}$. Yet, this is what Maugis and Michel (2011a) and Pan and Shen (2007) implicitly do when they model the dataset $\overline{\boldsymbol{Y}}$ with a common finite Gaussian mixture density $s_{\overline{\boldsymbol{\theta}}}$ and write the likelihood as $\prod_{i=1}^{n} s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})$ in their EM algorithm. One should be aware of this modeling uncorrectness.

Secondly, for $i \in \{1, \ldots, n\}$, it is not obvious[5] that $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ remains drawn from a probability distribution with a finite Gaussian mixture density when $\boldsymbol{Y_i}$ is drawn from a probability distribution with a finite Gaussian mixture density. If not the case, then modeling the dataset $\overline{\boldsymbol{Y}}$ by finite Gaussian mixture models – as it is done by Maugis and Michel (2011a) and Pan and Shen (2007) – is another uncorrectness.

[5] We have tried to prove that the density of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ indeed remains a finite Gaussian mixture density, but we have not managed to conclude this is true.

4.4.2.2 Consequence of empirical centering on the density estimation

Since they empirically center the data, Maugis and Michel (2011a) and Pan and Shen (2007) estimate "the" density $\overline{s}$ of the empirically centered dataset $\overline{\boldsymbol{Y}}$ rather than the density s of $\boldsymbol{Y}$. But the target is s, not $\overline{s}$. Then, an estimate of s has to be derived from the estimate of $\overline{s}$. Let us have a look at this estimate.

Performing empirical centering is subtracting $\overline{\mu}_j := \sum_{l=1}^n Y_{lj}/n$ to Y_{ij} for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. In the sequel, we set

$$\overline{\boldsymbol{\mu}} = (\overline{\mu}_1, \ldots, \overline{\mu}_p). \tag{4.29}$$

If one performs empirical centering, such as Maugis and Michel (2011a) or Pan and Shen (2007), then one considers a model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ for the dataset $\overline{\boldsymbol{Y}}$ with

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2\boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \ \mathbb{R}_+ \end{array} \right\}.$$

Then, for each $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, one computes an estimator $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$ of $\overline{s}$: for all $\boldsymbol{y} \in \mathbb{R}^p$,

$$\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \hat{\sigma}^2\boldsymbol{I}\right) \sum_{k=1}^K \hat{\pi}_k \, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right). \tag{4.30}$$

Since performing empirical centering is subtracting $\overline{\mu}_j$ to each Y_{ij}, a natural estimator of s derived from the estimation $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$ of $\overline{s}$ is $\hat{s}_{(K,\boldsymbol{J_r})}$ defined for all $\boldsymbol{y} \in \mathbb{R}^p$ by

$$\hat{s}_{(K,\boldsymbol{J_r})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2\boldsymbol{I}\right) \sum_{k=1}^K \hat{\pi}_k \, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right) \tag{4.31}$$

with $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r^c}]}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} = \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}} + \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r}]}$ where $\overline{\boldsymbol{\mu}}$ is defined by (4.29). The overall estimated mixture parameter vector is $(\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \hat{\sigma})$. So, there are $D = (K-1) + |\boldsymbol{J_r^c}| + K|\boldsymbol{J_r}| + 1 = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_r^c}|$ free parameters to estimate. Since $K \geq 1$, $D \geq |\boldsymbol{J_r^c}| + |\boldsymbol{J_r}| = p$. In particular, $D \geq n$ as soon as $p \geq n$, which is the case for high-dimensional data. Because of this large number of parameters to estimate, one may be afraid that $\hat{s}_{(K,\boldsymbol{J_r})}$ badly estimates s.

Then, since this analysis is true for all $(K, \boldsymbol{J_r})$, whatever the model selected by Maugis and Michel's procedure or by Pan and Shen's procedure (or more generally by one's procedure based on empirical centering), the density s is expected to be badly estimated. This shall be practically confirmed in our simulations in Section 5.3.

Remark 8. Remember that Maugis and Michel (2011a) only consider very low-dimensional data with $p \ll n$ since their procedure is not suited for high-dimensional data. In particular, in their context, for all K and for all $\boldsymbol{J_r}$, $D = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_r^c}| \leq K_{\max}(1 + p) + p \leq n$. Thus, Maugis and Michel (2011a) are not faced with degenerate models and they do not encounter estimation problems. On the contrary, Pan and Shen (2007) introduce their Lasso procedure in a high-dimensional context. Yet, in their article, they do not focus on density estimation but only on clustering. In Section 4.4.2.3, we shall see that, from a theoretical point of view, it is not easy to evaluate the consequence of empirical centering on the clustering. Our simulations in Section A.2 seem to indicate that one can achieve good clustering by working on empirically centered data.

To conclude, let us stress on the following point. By performing empirical centering, for each set $\boldsymbol{J_r}$ of relevant variables, one replaces the $|\boldsymbol{J_r^c}|$ a priori non-null constant mean parameters through the clusters by null mean parameters through the clusters ($\hat{\boldsymbol{\mu}}$ in (4.31) is replaced by $\mathbf{0}$ in (4.30)). Thus, performing empirical centering avoids the estimation of $|\boldsymbol{J_r^c}|$ parameters. But this dimensional reduction is just *artificial*. In fact, subtracting $\overline{\mu}_j$ to each Y_{ij} is equivalent to estimating $\overline{\mu}_j$ for each $j \in \{1, \ldots, p\}$ and this hidden estimation process of p parameters can highly deteriorate the estimations for high-dimensional data. In Section 4.5.2, we shall propose an alternative to empirical centering to perform *real* dimensional reduction for the estimation of the parameters.

4.4.2.3 Consequence of empirical centering on the clustering

Model-based clustering recasts the clustering problem into a density estimation problem. In Section 4.4.2.2, we saw that empirical centering deteriorates the density estimation. Therefore, we can question about the consequence of empirical centering on the data clustering.

If one performs empirical centering, such as Maugis and Michel (2011a) or Pan and Shen (2007), then one considers a model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ for the dataset $\overline{\boldsymbol{Y}}$ with

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \mathbf{0}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \ \mathbb{R}_+ \end{array} \right\}.$$

Then, for each $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, one computes an estimator of the density $\overline{s}$ in $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ and one can derive a clustering $\overline{\mathcal{C}}_{(K,\boldsymbol{J_r})}$ of the dataset $\overline{\boldsymbol{Y}}$ by the MAP principle. Yet, the primary goal is to get a clustering of the non-empirically centered dataset $\boldsymbol{Y}$. Here, we ask ourselves two questions:

1. To which clustering $\mathcal{C}_{(K,\boldsymbol{J_r})}$ of the dataset $\boldsymbol{Y}$ does the clustering $\overline{\mathcal{C}}_{(K,\boldsymbol{J_r})}$ of the dataset $\overline{\boldsymbol{Y}}$ correspond?

2. What about the quality of the clustering $\mathcal{C}_{(K,\boldsymbol{J_r})}$?

To answer the first question, for each $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, we find a model $\mathcal{S}_{(K,\boldsymbol{J_r})}$ and an estimator of the density s in $\mathcal{S}_{(K,\boldsymbol{J_r})}$ leading to a clustering $\mathcal{C}_{(K,\boldsymbol{J_r})}$ of the data $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ equivalent to the clustering $\overline{\mathcal{C}}_{(K,\boldsymbol{J_r})}$ of the data$\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}}$: that is to say $\boldsymbol{Y_i}$ is assigned to cluster k if and only if $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ is assigned to cluster k. This is formalized by the following proposition.

Proposition 4.4.1. *Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ and the dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$ derived from $\boldsymbol{Y}$ by empirical centering. Let $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ be a model collection for the dataset $\overline{\boldsymbol{Y}}$:*

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \ \mathbb{R}_+ \end{array} \right\}.$$

Let $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$.
Consider the maximum likelihood estimator of $\overline{s}$ in $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$:

$$\widehat{\overline{s}}_{(K,\boldsymbol{J_r})} = \underset{s_{\overline{\boldsymbol{\theta}}} \in \overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^n \ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})\right) \right\}.$$

Denote by $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_r})}$ the parameter vector of $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$. Compute

$$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_r})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{1}}, \ldots, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{K}}, \hat{\sigma}) \tag{4.32}$$

by the EM algorithm described in Section 4.A.3. Denote by $\overline{\mathcal{A}}$ this EM algorithm.

Now, for the dataset $\boldsymbol{Y}$, introduce the model

$$\mathcal{S}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu}_{\boldsymbol{1}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_r^c}|} \times (\mathbb{R}^{|\boldsymbol{J_r}|})^K \times \mathbb{R}_+ \end{array} \right\}$$

and the maximum likelihood estimator of s in $\mathcal{S}_{(K,\boldsymbol{J_r})}$:

$$\hat{s}_{(K,\boldsymbol{J_r})} = \underset{s_{\boldsymbol{\theta}} \in \mathcal{S}_{(K,\boldsymbol{J_r})}}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^n \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i})\right) \right\}. \tag{4.33}$$

Denote by $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})}$ the parameter vector of $\hat{s}_{(K,\boldsymbol{J_r})}$.

By computing $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})}$ by an EM algorithm $\mathcal{A}$ with appropriate initialization, we get that

$$\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \hat{\sigma})$$

with $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r^c}]}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} = \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}} + \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r}]}$ where $\overline{\boldsymbol{\mu}}$ is defined by (4.29) *and $\hat{\pi}_1, \ldots, \hat{\pi}_K, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{1}}, \ldots, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{K}}, \hat{\sigma}$*

are defined by (4.32).

In particular, for all $i \in \{1, \ldots, n\}$, for all $k \in \{1, \ldots, K\}$, the estimated posterior probability of observation $\boldsymbol{Y_i}$ arising from component k obtained by the EM algorithm $\mathcal{A}$ coincides with the estimated posterior probability of observation $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ arising from component k obtained by the EM algorithm $\overline{\mathcal{A}}$:

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \hat{\pi}_l \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)} = \frac{\hat{\pi}_k \, \Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \hat{\pi}_l \, \Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)}. \tag{4.34}$$

Then, by the MAP principle, $\boldsymbol{Y_i}$ is assigned to cluster k by the EM algorithm $\mathcal{A}$ if and only if $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ is assigned to cluster k by the EM algorithm $\overline{\mathcal{A}}$.

Proof. Page 177. □

Answering the second question is more delicate. From (4.34), for each $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, the quality of the clustering $\mathcal{C}_{(K,\boldsymbol{J_r})}$ of the dataset $\boldsymbol{Y}$ is equivalent to the quality of the clustering $\overline{\mathcal{C}}_{(K,\boldsymbol{J_r})}$ of the dataset $\overline{\boldsymbol{Y}}$. From (4.34),

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\sum_{l=1}^{K} \hat{\pi}_l \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)} \tag{4.35}$$

$$= \frac{\Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r^c}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2 \boldsymbol{I}\right) \; \hat{\pi}_k \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r^c}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2 \boldsymbol{I}\right) \; \sum_{l=1}^{K} \hat{\pi}_l \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{l}}, \hat{\sigma}^2 \boldsymbol{I}\right)} \tag{4.36}$$

$$= \frac{\Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r^c}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2 \boldsymbol{I}\right) \; \hat{\pi}_k \, \Phi\left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2 \boldsymbol{I}\right)}{\hat{s}_{(K,\boldsymbol{J_r})}} \tag{4.37}$$

where $\hat{s}_{(K,\boldsymbol{J_r})}$ is defined by (4.33). On the one hand, we saw in Section 4.4.2.2 that $\hat{s}_{(K,\boldsymbol{J_r})}$ is expected to be badly calculated, at least for high-dimensional data because it involves the estimation of too many parameters. So, the denominator of (4.37) is expected to be badly calculated. Moreover, for the same reasons, the numerator of (4.37) may also be badly estimated. So, at first sight, it seems difficult to trust in the estimation $\hat{\tau}_{ik}$. On the other hand, we saw in Section 4.4.2.2 that the estimation problems occur for the mean parameters in $\boldsymbol{J_r^c}$. But the density restricted on $\boldsymbol{J_r^c}$ cancels out from the numerator and the denominator of (4.36) and we may think that the ratio (4.35) is well estimated because it involves fewer parameter estimations. Furthermore, by the MAP principle, the clustering is determined by considering $\max_{1 \leq k \leq K} \hat{\tau}_{ik}$ for all $i \in \{1, \ldots, n\}$. So, even if $\hat{\tau}_{ik}$ is actually badly estimated for some $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, k\}$, $\max_{1 \leq k \leq K} \hat{\tau}_{ik}$ may be reached for the good component k, in which case the clustering is not affected by these individual bad estimations.

To conclude, although data clustering is closely linked to density estimation via model-based clustering, and although we are guaranteed that empirical centering deteriorates density estimation

(see Section 4.4.2.2), we may think that empirical centering actually not deteriorates data clustering. This is practically confirmed in our simulations (see Section A.2).

4.5 Our variable selection procedure: the Lasso-MLE procedure

In this section, we propose a new variable selection procedure for clustering suited for high-dimensional data. Similarly to Maugis and Michel's procedure and Pan and Shen's procedure, our procedure can be divided in four steps. First, a model collection is constructed; second, an estimate of the density is calculated in each model; third, a model is selected thanks to a model selection criterion; finally, a data clustering is provided by applying the MAP principle from the estimations in the selected model.

For the first step, the construction of a model collection is derived from the construction of a collection of sets of potentially relevant variables. Since we focus on high-dimensional data where complete variable selection is unfeasible, we must restrict our collection to a subcollection among the whole collection of possible sets $\{\{j_1, \ldots, j_d\}; 1 \leq d \leq p,\ j_l \in \{1, \ldots, p\},\ j_l \neq j_{l'}\}$. All the matter is to find a way to choose such a subcollection. Of course, it is preferable to choose a data-dependent subcollection rather than a deterministic subcollection. Besides, an automatic variable selection procedure is desirable. Pan and Shen (2007) suggest to take advantage of the soft-thresholding property of ℓ_1-regularization to construct a data-driven collection of sets of relevant variables by varying the regularization parameter. We find this idea fruitful and we use it to start our own procedure. Nonetheless, we think that Pan and Shen's procedure can be improved as regards the density estimation step and the model selection step. To improve estimation, we propose two modifications. First, we estimate the parameters by hard-thresholding estimators rather than by the soft-thresholding Lasso estimators in order to avoid shrinkage. Second, to get round empirical centering during the estimation step, we perform a preliminary reductional dimension step thanks to an additional ℓ_1-penalization. As regards model selection, we do not think that the asymptotic BIC criterion considered by Pan and Shen (2007) is suited to high-dimensional data. We rather suggest a non-asymptotic data-driven criterion based on the slope heuristics developed by Birgé and Massart (2006).

In this section, we first motivate each of the three modifications brought to Pan and Shen's procedure. Then, we detail the four steps of our procedure, called Lasso-MLE procedure.

4.5.1 Estimation of the parameters by MLEs rather than by Lasso estimators

Model-based clustering recasts the clustering problem into an estimation problem. In this viewpoint, one can expect that the better the estimation of the density, the better the clustering. Thus, the major point is density estimation. Let us explain the main weakness of Pan and Shen's procedure as regards estimation.

Let $\mathcal{S}$ be the set of all densities with respect to Lebesgue measure on $\mathbb{R}^p$. Consider a dataset $(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with unknown density $s \in \mathcal{S}$. In a maximum likelihood approach, the loss function considered is the Kullback-Leibler information defined for all $t \in \mathcal{S}$ by

$$\mathrm{KL}(s,t) = \int_{\mathbb{R}^p} \ln\left(\frac{s(\boldsymbol{y})}{t(\boldsymbol{y})}\right) s(\boldsymbol{y})\, d\boldsymbol{y}$$

if $sd\boldsymbol{y}$ is absolutely continuous with respect to $td\boldsymbol{y}$ and $+\infty$ otherwise. It is easy to show that $\mathrm{KL}(s,t) > 0$ for all $t \neq s$ and $\mathrm{KL}(s,s) = 0$. In particular, the density s is the unique minimizer of the Kullback-Leibler information on $\mathcal{S}$:

$$s = \arg\min_{t \in \mathcal{S}} \mathrm{KL}(s,t). \tag{4.38}$$

Moreover, by writing that

$$\mathrm{KL}(s,t) = \int_{\mathbb{R}^p} \ln\left(s(\boldsymbol{y})\right) s(\boldsymbol{y})\, d\boldsymbol{y} - \int_{\mathbb{R}^p} \ln\left(t(\boldsymbol{y})\right) s(\boldsymbol{y})\, d\boldsymbol{y},$$

we see that s is a minimizer over $\mathcal{S}$ of the risk:

$$s = \arg\min_{t \in \mathcal{S}} \mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \ln\left(t(\boldsymbol{Y_i})\right)\right]. \tag{4.39}$$

Now, let $\mathcal{K} = \{K_{\min}, \ldots, K_{\max}\}$ be a collection of numbers of clusters and let $\mathcal{J}$ be the collection of all non-empty subsets of $\{1, \ldots, p\}$. Put $\mathcal{M} = \mathcal{K} \times \mathcal{J}$. Introduce a model collection $\{S_m\}_{m \in \mathcal{M}}$ with $S_m \subset \mathcal{S}$ for all $m \in \mathcal{M}$. An idea is that replacing the risk $\mathbb{E}[-\sum_{i=1}^{n} \ln(t(\boldsymbol{Y_i}))/n]$ by the empirical contrast $-\sum_{i=1}^{n} \ln\left(t(\boldsymbol{Y_i})\right)/n$ in (4.39) and minimizing on S_m rather than on $\mathcal{S}$ must lead to a sensible estimator of s in S_m. Thus, for all $m \in \mathcal{M}$, one can consider the maximum likelihood estimator over the model S_m,

$$\hat{s}_m = \arg\min_{t \in S_m} \left\{-\frac{1}{n}\sum_{i=1}^{n} \ln\left(t(\boldsymbol{Y_i})\right)\right\}, \tag{4.40}$$

and aim at choosing the best estimator of s among the collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$. From (4.38), the ideal choice is the oracle $\hat{s}_{m_{\text{oracle}}}$ defined by

$$m_{\text{oracle}} = \arg\min_{m \in \mathcal{M}} \mathrm{KL}(s, \hat{s}_m). \tag{4.41}$$

In practice, $\hat{s}_{m_{\text{oracle}}}$ is unattainable because it depends on the unknown density s. Yet, it is a benchmark to evaluate the quality of one's estimator of s. For high-dimensional data, the number p of variables can be very large and the collection $\mathcal{M}$ is so rich that performing exhaustive best subset selection over

$\{\hat{s}_m\}_{m\in\mathcal{M}}$ is unfeasible. In this case, an idea is to consider a subset $\mathcal{M}' \subset \mathcal{M}$ so that performing best subset selection over $\{\hat{s}_m\}_{m\in\mathcal{M}'}$ becomes practicable. Then, one aims at the ideal choice

$$m_{\text{oracle}}(\mathcal{M}') = \underset{m\in\mathcal{M}'}{\arg\min}\, \mathrm{KL}(s, \hat{s}_m). \tag{4.42}$$

If $\mathcal{M}'$ is well-chosen, one expects that $\hat{s}_{m_{\text{oracle}}(\mathcal{M}')}$ is not too far from the ideal estimator $\hat{s}_{m_{\text{oracle}}}$ and that it remains a satisfactory estimator of s. A major point is the choice of a good subset $\mathcal{M}'$. On the one hand, $\mathcal{M}'$ must not be too large in order to make feasible best subset selection over the subcollection $\{\hat{s}_m\}_{m\in\mathcal{M}'}$. On the other hand, $\mathcal{M}'$ must be large enough for $\mathrm{KL}(s, \hat{s}_{m_{\text{oracle}}(\mathcal{M}')})$ to be as small as possible.

Pan and Shen (2007) propose their Lasso procedure as a method to perform variable selection during the estimation process of s. They look for an estimation of s by a finite Gaussian mixture density $s_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \pi_k\, \Phi(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I})$. Estimating s is equivalent to estimating the parameter vector $\boldsymbol{\theta}$. For all $K \in \mathcal{K}$, let G_K be a grid of regularization parameters and let $\lambda \in G_K$. Consider $\hat{\boldsymbol{\theta}}_{(K,\lambda)}$ the Lasso estimator of $\boldsymbol{\theta}$ calculated by Pan and Shen's Lasso procedure. An index set $\boldsymbol{J}_{(K,\lambda)} \subset \{1, \ldots, p\}$ is derived from the estimate $\hat{\boldsymbol{\theta}}_{(K,\lambda)}$ (see Section 4.3.2). By varying $\lambda \in G_K$ and $K \in \mathcal{K}$, one gets a subcollection of index sets

$$\mathcal{M}_{\text{Lasso}} = \left\{ \left(K, \boldsymbol{J}_{(K,\lambda)}\right); K \in \mathcal{K},\ \lambda \in G_K \right\}$$

included in the whole collection $\mathcal{M}$. Given this subset $\mathcal{M}' = \mathcal{M}_{\text{Lasso}} \subset \mathcal{M}$, according to (4.42), a natural procedure would be to consider the family of maximum likelihood estimators $\{\hat{s}_m\}_{m\in\mathcal{M}_{\text{Lasso}}}$ and to aim at choosing as final estimator among this family the estimator as close as possible to the oracle defined by

$$m_{\text{oracle}}(\mathcal{M}_{\text{Lasso}}) = \underset{m\in\mathcal{M}_{\text{Lasso}}}{\arg\min}\, \mathrm{KL}(s, \hat{s}_m). \tag{4.43}$$

Nonetheless, this is not the choice adopted by Pan and Shen (2007). Indeed, let $m \in \mathcal{M}_{\text{Lasso}}$. By definition of $\mathcal{M}_{\text{Lasso}}$, there exist $K \in \mathcal{K}$ and $\lambda > 0$ such that $m = (K, \boldsymbol{J}_{(K,\lambda)})$. Pan and Shen (2007) estimate $\boldsymbol{\theta}$ by the Lasso estimator $\hat{\boldsymbol{\theta}}_{(K,\lambda)}$ and they estimate the density s by the Lasso estimator $\hat{s}_m^L := s_{\hat{\boldsymbol{\theta}}_{(K,\lambda)}}$. Then, they consider the family of Lasso estimators $\left\{\hat{s}_m^L\right\}_{m\in\mathcal{M}_{\text{Lasso}}}$ and aim at choosing as final estimator among this family the Lasso estimator as close as possible to the Lasso oracle defined by

$$m_{\text{oracle}}^L(\mathcal{M}_{\text{Lasso}}) = \underset{m\in\mathcal{M}_{\text{Lasso}}}{\arg\min}\, \mathrm{KL}\left(s, \hat{s}_m^L\right). \tag{4.44}$$

Remark 9. Let us point out that the two families $\{\hat{s}_m\}_{m\in\mathcal{M}_{\text{Lasso}}}$ and $\left\{\hat{s}_m^L\right\}_{m\in\mathcal{M}_{\text{Lasso}}}$ are different.

Indeed, for all $m \in \mathcal{M}_{\text{Lasso}}$, we get from (4.40) that

$$\hat{s}_m = \underset{s_{\boldsymbol{\theta}} \in S_m}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}) \right) \right\} \tag{4.45}$$

whereas

$$\hat{s}_m^L = \underset{s_{\boldsymbol{\theta}} \in S_m}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}) \right) + \lambda \left| s_{\boldsymbol{\theta}} \right|_1 \right\}. \tag{4.46}$$

The minimization problems (4.45) and (4.46) being different, their solutions are different.

Now, comparing (4.41) with (4.43) and (4.44), we see that both the maximum likelihood oracle $m_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ and the Lasso oracle $m^L_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ are solutions of minimization problems over the same subset $\mathcal{M}_{\text{Lasso}}$ of the whole set $\mathcal{M}$ considered for the ideal oracle m_{oracle}. Yet, $m_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ is calculated by considering maximum likelihood estimators $\hat{s}_m$, just as the ideal oracle m_{oracle}, whereas $m^L_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ is calculated by considering Lasso estimators $\hat{s}^L_m$, that is to say ℓ_1-penalized maximum likelihood estimators, which are different from maximum likelihood estimators (see Remark 9). Therefore, the set $m_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ is expected to be closer to the ideal set m_{oracle} than the set $m^L_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$. In fact, we suspect that $m^L_{\text{oracle}}(\mathcal{M}_{\text{Lasso}})$ is much larger than m_{oracle}. Indeed, in the regression framework, it has been noticed (Zhang and Huang, 2008a; Zhao and Yu, 2007; Connault, 2011) that, when some irrelevant variables are enough correlated with relevant variables, then the Lasso tends to pick up these irrelevant variables to compensate for the under-estimation of the mean parameters of the relevant variables caused by ℓ_1-penalization shrinkage. Consequently, the Lasso minimizer of (4.46) is typically achieved for a set of variables containing some or even many irrelevant variables. On the opposite, if one considers the hard-thresholding estimators $\{\hat{s}_m\}_{m \in \mathcal{M}_{\text{Lasso}}}$ rather than the soft-thresholding Lasso estimators $\{\hat{s}^L_m\}_{m \in \mathcal{M}_{\text{Lasso}}}$, then parameters are not under-estimated and one can expect that the minimizer of (4.45) is achieved for a sparser set of variables with much fewer irrelevant variables. Now, since the oracles are benchmark for estimators, it is in one's interest to choose the procedure leading to the best oracle. This is why we recommend to estimate s by the maximum likelihood estimators rather than by the Lasso estimators.

Since the procedure we advocate first considers the models generated by the Lasso estimators, and then calculate the Maximum Likelihood Estimators (MLE) in these models, we call it Lasso-MLE procedure. To our knowledge, although it is quite natural, the Lasso-MLE estimators have never been studied in model-based clustering. Yet, this idea has emerged in other frameworks. For instance, in regression, Connault (2011) introduces and studies such a procedure called projected Lasso. In particular, he compares it to the classical Lasso procedure and concludes to better performance results for the projected Lasso than for the classical Lasso. Such an estimator is also mentioned as LARS-OLS hybrid in Efron et al. (2004, p. 421). In the density estimation framework, Bertin et al. (2011) consider such an idea to estimate densities decomposed in some dictionary.

4.5.2 An alternative to empirical centering for the estimation step: selection of the active variables

Estimating the density by maximum likelihood estimators rather than by Lasso estimators is not the only dissimilarity between our estimation step and Pan and Shen's estimation step. In Section 4.4.2.3, we pointed that performing empirical centering is likely to deteriorate the density estimation for high-dimensional data. Thus, contrary to Pan and Shen (2007), we do not empirically center the data during the estimation step. Nonetheless, when no empirical centering is done, we shall see that selecting the relevant variables for the clustering is not sufficient to get sparse models. Therefore, an additional dimensional reduction step is needed before the estimation step. This step involves the selection of "active" variables. Here, we introduce the notion of active variables and we detail our dimensional reduction process.

4.5.2.1 Definition of an active variable

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with density s. In a clustering purpose, s is to be estimated by a finite Gaussian mixture density $\sum_{k=1}^{K} \pi_k \Phi(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I})$. A variable is said to be *inactive* in the modeling if it is indexed by $j \in \{1, \ldots, p\}$ such that for all $k \in \{1, \ldots, K\}$, $\mu_{kj} = 0$. On the contrary, a variable is *active* if it is indexed by j such that there exists $k \in \{1, \ldots, K\}$ such that $\mu_{kj} \neq 0$. In other words, an inactive variable is absent from the model while an active variable is present in the model. Note that an inactive variable is a particular case of an irrelevant variable: an inactive variable is an irrelevant variable with common mean parameters through the clusters equal to zero.

To sum up, our procedure shall involve three different types of variables:

- the *relevant* variables, that we shall index by $\boldsymbol{J_r}$, are present in the modeling and provide information for the clustering;
- the *active irrelevant* variables, that we shall index by $\boldsymbol{J_a}$, are present in the modeling but provide no information for the clustering;
- the *inactive* variables, that we shall index by $\boldsymbol{J_a^c}$, are absent from the modeling.

4.5.2.2 Elimination of the inactive variables

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with density s. Assume we have constructed a collection of index sets $\mathcal{M}_r$ resulting in a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{M}_r}$ where

K is a number of clusters and $\boldsymbol{J_r}$ represents a set of relevant variables for the clustering:

$$\mathcal{S}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \quad (4.47)$$

In each model $\mathcal{S}_{(K,\boldsymbol{J_r})}$, if we estimate s by the maximum likelihood estimator $\hat{s}_{(K,\boldsymbol{J_r})}$, then there are $K-1$ free mixing parameters, $|\boldsymbol{J_r^c}|$ mean parameters on $\boldsymbol{J_r^c}$, $K|\boldsymbol{J_r}|$ mean parameters on $\boldsymbol{J_r}$ and one variance parameter to estimate, which leads to a total number of $D = K(1+|\boldsymbol{J_r}|) + |\boldsymbol{J_r^c}|$ parameters to estimate. Since $K \geq 1$, $D \geq |\boldsymbol{J_r}| + |\boldsymbol{J_r^c}| = p$. So, as soon as $p \geq n$ (which is the case for high-dimensional problems), the model $\mathcal{S}_{(K,\boldsymbol{J_r})}$ is degenerate in the sense that its dimension is larger than the number of observations. Let us stress that this is true for all models $\mathcal{S}_{(K,\boldsymbol{J_r})}$, even for models with very few (in fact even zero) relevant variables. This shows that selection of the relevant variables for the clustering does not lead to sparse models, at least for high-dimensional data. Since the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ only contains degenerate models, whatever the model selection criterion used in a further step, the final model selected $\mathcal{S}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ shall be degenerate and s is expected to be badly estimated by $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$.

To cope with this dimensional problem, we must find a strategy to reduce the dimension of the models. Note that the number of mean parameters coming from $\boldsymbol{J_r^c}$ is $|\boldsymbol{J_r^c}|$, while the number of mean parameters coming from $\boldsymbol{J_r}$ is $K|\boldsymbol{J_r}|$. But the number of clusters K remains small ($K \leq 20$ in most real clustering problems) while we are looking for a small set $\boldsymbol{J_r}$ of relevant variables, that is to say a large set $\boldsymbol{J_r^c}$. Thus, $|\boldsymbol{J_r^c}|$ is expected to be much larger than $K|\boldsymbol{J_r}|$ and the problem of too high total dimension D is due to the set of the irrelevant variables, that is to say the variables we are not interested in as regards clustering. So, we suggest to apply the dimensional reduction to the set of the irrelevant variables. For each of these variables, there is only one mean parameter to estimate, which is the common mean parameter through the K clusters. Now, this common mean parameter can either have a large value, in which case the variable is active although irrelevant for the clustering, or it may have a small value close to zero, in which case the variable is not much active. But inactive variables are expected to be absent from the modeling. So, they should be eliminated from the model. This can be done by applying ℓ_1-penalization on the mean parameters on $\boldsymbol{J_r^c}$. Let us detail this additional process.

Fix $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$ and let $s_{\boldsymbol{\theta}} \in \mathcal{S}_{(K,\boldsymbol{J_r})}$. From (4.47), for all $i \in \{1, \ldots, n\}$,

$$s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}) = \Phi\left(\boldsymbol{Y_i}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \sum_{k=1}^K \pi_k \, \Phi\left(\boldsymbol{Y_i}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right).$$

From this decomposition of $s_{\boldsymbol{\theta}}$, the density of $\boldsymbol{Y_i}_{[\boldsymbol{J_r^c}]}$ can be modeled by a $|\boldsymbol{J_r^c}|$-dimensional Gaussian density of the form $s_{\boldsymbol{\beta}} = \Phi(\cdot \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{\beta} = (\boldsymbol{\mu}, \sigma) \in \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \mathbb{R}_+$. In order to perform

dimensional reduction on the mean parameters μ_j for $j \in \boldsymbol{J_r^c}$, we shall use an ℓ_1-penalty proportional to $\|\boldsymbol{\mu}\|_1$. This soft-thresholding penalty shall shrink the smallest mean parameters $|\mu_j|$, resulting in automatic elimination of the inactive variables. For a given regularization parameter $\lambda > 0$, the Lasso estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{|\boldsymbol{J_r^c}|}\times\mathbb{R}_+}{\arg\min} \{\gamma_n(s_{\boldsymbol{\beta}}) + \lambda|\boldsymbol{\beta}|_1\} \tag{4.48}$$

with

$$\gamma_n(s_{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\beta}}\left(\boldsymbol{Y_{i[J_r^c]}}\right)\right)$$

and

$$|\boldsymbol{\beta}|_1 := \|\boldsymbol{\mu}\|_1 = \sum_{j\in\boldsymbol{J_r^c}} |\mu_j|.$$

To compute this Lasso estimator, we use an algorithm described in Section 4.A.4. This algorithm is inspired from the EM algorithm for model-based clustering introduced by Dempster et al. (1977). It is yet simpler because no mixture is involved in the minimization problem (4.48). Denote $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)} = (\hat{\boldsymbol{\mu}}, \hat{\sigma})$. The set $\boldsymbol{J_r^c}$ of irrelevant variables is split in two sets $\boldsymbol{J_a} = \{j \in \boldsymbol{J_r^c};\ \hat{\mu}_j \neq 0\}$ and $\boldsymbol{J_a^c} = \{j \in \boldsymbol{J_r^c};\ \hat{\mu}_j = 0\}$ such that $\boldsymbol{J_r^c} = \boldsymbol{J_a} \sqcup \boldsymbol{J_a^c}$. The set $\boldsymbol{J_a}$ represents the active variables among the irrelevant variables selected by the Lasso for regularization parameter λ. Coming back to the estimation of the density s of the whole dataset $\boldsymbol{Y}$, we get a new model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ defined by

$$\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}); \\ s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I})\ \Phi(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \boldsymbol{\mu}, \sigma^2\boldsymbol{I})\ \sum_{k=1}^{K} \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2\boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_a}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times\ \mathbb{R}_+ \end{array} \right\}.$$

The dimension of $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is $(K-1) + |\boldsymbol{J_a}| + K|\boldsymbol{J_r}| + 1 = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_a}| \leq n$ as soon as $\boldsymbol{J_r}$ and $\boldsymbol{J_a}$ are small enough.

Then, by varying $\lambda > 0$ for a fixed set $(K, \boldsymbol{J_r})$ and by varying $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$, we get a new model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ for some finite collection of index sets $\mathcal{M}_{(r,a)}$. Contrary to the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ that only contains degenerate models, the collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ contains models with small or moderate dimensions (besides some degenerate models). In a further step, a good model selection criterion is expected to select one of the small-dimensional models among the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$.

Remark 10. To perform significant reductional dimension thanks to the elimination of the inactive variables, our procedure assumes that there exist many inactive variables. Some datasets may not fulfill this assumption, in which case our procedure is not adapted. To deal with such situations, in Chapter A, we propose an alternative to our procedure that does not require the notion of active variable. Let us mention that the notion of active variable is well-defined for the important case of

curve clustering when signals have a sparse representation in some appropriate basis such as a wavelet basis. We refer to Section 4.6 for more details on this fundamental application.

4.5.3 A non-asymptotic model selection criterion

The third modification we bring to Pan and Shen's procedure deals with model selection. Pan and Shen (2007) consider a modified BIC criterion taking into account the model sparsity to calculate the model dimension. Although BIC is widely used, there are few theoretical properties proved on this criterion. For instance, BIC consistency is usually stated under restrictive regularity assumptions and assuming that the true density belongs to the considered Gaussian mixture family (see for instance Keribin, 2000). In particular, for high-dimensional data where the number of observations is small or moderate, one can question such an asymptotic criterion.

During the last years, a non-asymptotic approach for model selection via penalization has emerged, mainly with works of Birgé and Massart (1997) and Barron et al. (1999). With this viewpoint, the number and the dimension of the models may depend on the number n of observations. Given a model collection $\{S_D\}_{D\in\mathcal{D}}$, the penalty function derived from a non-asymptotic approach is typically of the form

$$\mathrm{pen}(D) = (C_1 + C_2 L_D)\frac{D}{n} \tag{4.49}$$

where C_1 and C_2 are constants independent on n. The role of the weight coefficients L_D is to quantify the richness of the model collection by taking into account the possible large number of models with identical dimension D in the model collection. This can be justified by the theoretical arguments developed by Birgé and Massart (2006) in a Gaussian regression setting. On the one hand, when the number of models having the same dimension is moderate (such as for ordered variable selection), the weight coefficients can be taken as a small positive constant $L_D = L$, in which case the penalty in (4.49) is proportional to the dimension. On the other hand, when the number of models having the same dimension D grows much faster with D (such as for complete variable selection), then a penalty shape proportional to D/n selects too complex models with high probability and stronger weight coefficients with a logarithm term depending on the data dimension are necessary to select smaller models. This phenomenon has been practically checked by Lebarbier (2005) for the study of multiple change points detection. It has also been observed in a density framework by Castellan (1999) who compared the penalty shape for regular histogram selection and complete irregular histogram selection. In practice, theory about model selection via penalization often fails in providing explicit multiplicative constants C_1 and C_2 in (4.49), and thus no practical model selection criterion can be derived. To cope with this drawback, Birgé and Massart (2006) proposed a practical method to define efficient penalty functions from the data. This method is the so-called "slope heuristics". It is recalled in Section 6.3.1.

In our procedure, we consider a non-asymptotic point of view and we use the slope heuristics to define a model selection criterion. Since our model collection is random, it is difficult to have an idea of the richness of our model collection and to determine whether weights with a logarithm term are necessary to get a proper penalty. For this reason, in the sequel, we consider both a penalty shape proportional to the dimension and a penalty shape with an additional logarithm term. We refer to Chapter 6 for a theoretical study and practical experiments carried out to try to determine the ideal penalty shape and to justify the two penalty shapes considered below.

4.5.4 Description of our Lasso-MLE procedure

Here, we detail the four steps of our procedure. They are to be compared with the four steps of Maugis and Michel's procedure and Pan and Shen's procedure respectively described in Section 4.3.1 and Section 4.3.2. A comparison between the three procedures is summarized in Table 4.1.

Step 1. Model collection
I. Fix $K \in \mathcal{K}$.

<u>Detection of the relevant variables</u>

(a) We empirically center the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ to get an empirically centered dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$. Consider

$$\overline{\mathcal{S}}_K = \left\{ \begin{array}{l} s_{\overline{\boldsymbol{\theta}}} = \sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2 \boldsymbol{I}\right); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Theta_K \end{array} \right\}$$

with $\Theta_K = \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+$. As explained in Section 4.4.1, since the dataset $\overline{\boldsymbol{Y}}$ is centered, irrelevant variables are expected to be detected by penalizing the empirical contrast

$$\gamma_n\left(s_{\overline{\boldsymbol{\theta}}}\right) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}})\right)$$

by an ℓ_1-penalty on the mean parameters proportional to

$$|\overline{\boldsymbol{\theta}}|_1 = \sum_{j=1}^{p} \sum_{k=1}^{K} |\overline{\mu}_{kj}|.$$

Introduce G_K a grid of regularization parameters [6].

[6] See Section 4.B.1 for the explanation of the construction of such a grid.

(b) Fix $\lambda \in G_K$. Consider the Lasso estimator defined by

$$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = \underset{\overline{\boldsymbol{\theta}} \in \Theta_K}{\arg\min} \left\{ \gamma_n(s_{\overline{\boldsymbol{\theta}}}) + \lambda |\overline{\boldsymbol{\theta}}|_1 \right\}. \tag{4.50}$$

We compute $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)} = (\hat{\pi}_k, \widehat{\overline{\mu}}_{kj}, \hat{\sigma})_{1 \leq k \leq K, 1 \leq j \leq p}$ by Pan and Shen's EM algorithm described in Section 4.A.2. The index set $\boldsymbol{J}_{(K,\lambda)} = \{j \in \{1, \ldots, p\} : \exists\, k \text{ such that } \widehat{\overline{\mu}}_{kj} \neq 0\}$ represents the set of relevant variables selected by the Lasso $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)}$.

(c) By varying $\lambda \in G_K$, we get a collection $\mathcal{J}_K = \cup_{\lambda \in G_K} \boldsymbol{J}_{(K,\lambda)}$ of index sets representing a collection of sets of relevant variables. As regards the estimation of s, this collection $\mathcal{J}_K$ leads to a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{\boldsymbol{J_r} \in \mathcal{J}_K}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{4.51}$$

Each model $\mathcal{S}_{(K,\boldsymbol{J_r})}$ is the set of finite Gaussian mixture densities with K components and $\boldsymbol{J_r}$ as index set representing the relevant variables.

Detection of the active variables among the irrelevant variables

1. Fix $\boldsymbol{J_r} \in \mathcal{J}_K$.

   (a) We come back to the non-empirically centered dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$. From (4.51), the density s restricted on $\boldsymbol{J_r^c}$ can be estimated by a $|\boldsymbol{J_r^c}|$-dimensional Gaussian density of the form $s_{\boldsymbol{\beta}} = \Phi(\cdot \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{\beta} = (\boldsymbol{\mu}, \sigma) \in \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \mathbb{R}_+$. As explained in Section 4.5.2, we want to detect the inactive variables in order to perform dimensional reduction on the mean parameters μ_j for $j \in \boldsymbol{J_r^c}$. Such variables are expected to be detected by penalizing the empirical contrast

   $$\gamma_n(s_{\boldsymbol{\beta}}) = -\frac{1}{n} \sum_{i=1}^n \ln\left(s_{\boldsymbol{\beta}}\left(\boldsymbol{Y_i}_{[\boldsymbol{J_r^c}]}\right)\right)$$

   by an ℓ_1-penalty on the mean parameters proportional to

   $$|\boldsymbol{\beta}|_1 := \|\boldsymbol{\mu}\|_1 = \sum_{j \in \boldsymbol{J_r^c}} |\mu_j|.$$

   Introduce $G_{(K,\boldsymbol{J_r})}$ a grid of regularization parameters.

(b) Fix $\lambda \in G_{(K,\boldsymbol{J_r})}$. Consider the Lasso estimator of $\boldsymbol{\beta}$ defined by

$$\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{|\boldsymbol{J_r^c}|}\times\mathbb{R}_+}{\arg\min} \left\{\gamma_n(s_{\boldsymbol{\beta}}) + \lambda|\boldsymbol{\beta}|_1\right\}. \tag{4.52}$$

We compute $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)} = (\hat{\boldsymbol{\mu}}, \hat{\sigma})$ by the second algorithm described in Section 4.A.4. The set $\boldsymbol{J}_{(\boldsymbol{J_r},\lambda)} = \{j \in \boldsymbol{J_r^c};\ \hat{\mu}_j \neq 0\}$ represents the active variables among the irrelevant variables indexed by $\boldsymbol{J_r^c}$, selected by the Lasso $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)}$.

(c) By varying $\lambda \in G_{(K,\boldsymbol{J_r})}$, we get a collection $\mathcal{J}_{(K,\boldsymbol{J_r})} = \cup_{\lambda\in G_{(K,\boldsymbol{J_r})}} \boldsymbol{J}_{(\boldsymbol{J_r},\lambda)}$ from which we derive a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{\boldsymbol{J_a}\in\mathcal{J}_{(K,\boldsymbol{J_r})}}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}); \\ s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I})\ \Phi(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \boldsymbol{\mu}, \sigma^2\boldsymbol{I})\ \sum_{k=1}^K \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2\boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_a}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times\ \mathbb{R}_+ \end{array} \right\}.$$

Each model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is the set of finite Gaussian mixture densities with K components, $\boldsymbol{J_r}$ as index set representing the relevant variables and $\boldsymbol{J_a}$ as index set representing the active irrelevant variables.

2. By varying $\boldsymbol{J_r} \in \mathcal{J}_K$, we get a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{J}_K\times\mathcal{J}_{(K,\boldsymbol{J_r})}}$.

II. By varying $K \in \mathcal{K}$, we get a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{K}\times\mathcal{J}_K\times\mathcal{J}_{(K,\boldsymbol{J_r})}}$. We put

$$\mathcal{M}_{(r,a)} = \{(K, \boldsymbol{J_r}, \boldsymbol{J_a});\ K \in \mathcal{K},\ \boldsymbol{J_r} \in \mathcal{J}_K,\ \boldsymbol{J_a} \in \mathcal{J}_{(K,\boldsymbol{J_r})}\}.$$

Step 2. Calculation of an estimator of s in each model

For each $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$, we compute the maximum likelihood estimator in the model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$:

$$\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \underset{s_{\boldsymbol{\theta}}\in\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{\arg\min}\ \gamma_n\left(s_{\boldsymbol{\theta}}\right), \qquad \gamma_n\left(s_{\boldsymbol{\theta}}\right) = -\frac{1}{n}\sum_{i=1}^n \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i})\right).$$

For all $\boldsymbol{y} \in \mathbb{R}^p$,

$$\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \hat{\sigma}^2\boldsymbol{I}\right)\ \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2\boldsymbol{I}\right)\ \sum_{k=1}^K \hat{\pi}_k\, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right).$$

The estimated mixture parameters $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \hat{\sigma})$ are computed by the third EM algorithm described in Section 4.A.4.

Step 3. Model selection

We select a model thanks to the data-driven penalized criterion derived from the slope heuristics introduced by Birgé and Massart (2006) and recalled in Section 6.3.1. First, models are grouped according to their dimension D in order to obtain a model collection $\{\mathcal{S}_D\}_{D\in\mathcal{D}}$. The dimension of a model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is the total number of free parameters estimated in the model: $K-1$ mixing proportions, $|\boldsymbol{J_a}|$ mean parameters on $\boldsymbol{J_a}$, $K|\boldsymbol{J_r}|$ mean parameters on $\boldsymbol{J_r}$ and 1 variance parameter, which gives a total dimension equal to $K(1+|\boldsymbol{J_r}|)+|\boldsymbol{J_a}|$. For each dimension $D\in\mathcal{D}$, let $\hat{s}_D$ be the estimator maximizing the likelihood among the estimators associated to a model of dimension D. There exists $(K_D,\boldsymbol{J}_{\boldsymbol{r},D},\boldsymbol{J}_{\boldsymbol{a},D})$ such that $\hat{s}_D=\hat{s}_{(K_D,\boldsymbol{J}_{\boldsymbol{r},D},\boldsymbol{J}_{\boldsymbol{a},D})}$. Denote by $D_{\max}$ the maximal dimension in $\mathcal{D}$. Then, we plot the function $D/n\mapsto-\gamma_n(\hat{s}_D)$. Two situations can occur[7]. We can either observe a linear behavior,

$$-\gamma_n(\hat{s}_D)\approx\kappa\,\frac{D}{n},\quad \kappa>0,$$

in which case the slope κ is estimated by the method implemented by Baudry et al. (2011) and the calibrated penalty function is $\mathrm{pen}(D)=2\,\hat{\kappa}\,D/n$. The minimizer $\hat{D}$ of the penalized criterion $D\mapsto\gamma_n(\hat{s}_D)+\mathrm{pen}(D)$ is determined. Or we can observe a logarithmic behavior,

$$-\gamma_n(\hat{s}_D)\approx\kappa_1\,\frac{D}{n}\left(1+\kappa_2\ln\left(\frac{D_{\max}}{D}\right)\right),\quad \kappa_1>0,\kappa_2>0,$$

in which case κ_1 and κ_2 are estimated by the method described in Section 6.3.3.1 and the calibrated penalty function is $\mathrm{pen}_{\ln}(D)=2\,\hat{\kappa}_1(D/n)(1+\hat{\kappa}_2\ln(D_{\max}/D))$. The minimizer $\hat{D}$ of the penalized criterion $D\mapsto\gamma_n(\hat{s}_D)+\mathrm{pen}_{\ln}(D)$ is determined. In both cases, the model $\mathcal{S}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}:=\mathcal{S}_{(K_{\hat{D}},\boldsymbol{J}_{\boldsymbol{r},\hat{D}},\boldsymbol{J}_{\boldsymbol{a},\hat{D}})}$ is selected.

Step 4. Data clustering

The variables declared as relevant for the clustering are indexed by $\hat{\boldsymbol{J}}_{\boldsymbol{r}}$. The density s is estimated by $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}$. A clustering of the dataset $\boldsymbol{Y}$ is derived from $\hat{\boldsymbol{\theta}}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}$ by applying the MAP principle.

4.6 A major application: functional data clustering

In different fields of applications, observations are functions observed either discretely or continuously. Given a sample of curves, an important task is to search for homogeneous subgroups of curves using clustering. In a functional context, besides identifying the individuals who are involved in the same or similar processes, clustering is useful to determine one representative curve per cluster from the noisy observations. This implies a denoising and smoothing signal process so as to remove the noise and capture only the important patterns in the data. Here, we explain how our Lasso-MLE procedure can be applied in this context. Simulations shall be presented in Section 5.4.

[7] We refer to Chapter 6 for justification of the two following penalty shapes pen and $\mathrm{pen}_{\ln}$.

procedure	model collection	parameter estimation	model selection	clustering
Lasso (Pan and Shen, 2007)	ℓ_1-penalization on the mean parameters to detect the irrelevant variables $\Rightarrow$ data-driven collection of sets of relevant variables $\Rightarrow$ collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ of models with K clusters and $\boldsymbol{J_r}$ as representative set of relevant variables	Lasso	modified BIC	MAP
MLE (Maugis and Michel, 2011a)	deterministic collection of sets of relevant variables corresponding to complete or ordered variable selection $\Rightarrow$ collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ of models with K clusters and $\boldsymbol{J_r}$ as representative set of relevant variables	MLE	non-asymptotic data-driven criterion	MAP
Lasso-MLE	1. ℓ_1-penalization on the mean parameters to detect the irrelevant variables 2. ℓ_1-penalization on the mean parameters to detect the inactive variables $\Rightarrow$ data-driven collection of sets of relevant variables and active variables $\Rightarrow$ collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ of models with K clusters, $\boldsymbol{J_r}$ as representative set of relevant variables and $\boldsymbol{J_a}$ as representative set of active irrelevant variables	MLE	non-asymptotic data-driven criterion	MAP

Table 4.1: Comparison between Pan and Shen's Lasso procedure, Maugis and Michel's procedure and our Lasso-MLE procedure. The operations performed on the original dataset $\boldsymbol{Y}$ are in green, whereas the operations performed on the empirically centered dataset $\overline{\boldsymbol{Y}}$ are in red.

4.6.1 Variable selection for functional data clustering

4.6.1.1 Functional data clustering

Consider K functions $f_1, \ldots, f_K \in L_2([0,1])$. Let $\{t_1, \ldots, t_p\}$ be a fine time grid with $t_j \in [0,1]$. For all $k \in \{1, \ldots, K\}$, denote by $\boldsymbol{f_k} = (f_k(t_1), \ldots, f_k(t_p))$ the discretization of the function f_k on the grid $\{t_1, \ldots, t_p\}$. Consider n data $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$ such that each $\boldsymbol{y_i}$ is a white-noised observation of some $\boldsymbol{f_k}$, $k \in \{1, \ldots, K\}$. For all $i \in \{1, \ldots, n\}$ and for all $k \in \{1, \ldots, K\}$, if $\boldsymbol{y_i}$ belongs to cluster k, then $\boldsymbol{y_i} = \boldsymbol{f_k} + \boldsymbol{\xi_i} \in \mathbb{R}^p$ with

$$y_{ij} = f_k(t_j) + \xi_{ij}, \qquad j = 1, \ldots, p \tag{4.53}$$

where $\boldsymbol{\xi_i} = (\xi_{i1}, \ldots, \xi_{ip}) \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is a white noise. We denote $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ this functional dataset.

Recently, several functional clustering methods have been developed. Usually, the observations $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$ are preliminary projected onto a suitable basis $\mathcal{B} = \{\phi_1, \ldots, \phi_p\}$ of the functional space, such as a spline, Fourier or wavelet basis. This leads to a new dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ where each data $\boldsymbol{Y_i}$ is the coefficient decomposition of the observation $\boldsymbol{y_i}$ into the basis $\mathcal{B}$. Then, the clustering process is performed on this coefficient dataset $\boldsymbol{Y}$. For instance, as regards distance-based methods, Abraham et al. (2003) and Garcia-Escudero and Gordaliza (2005) project the curves onto a B-spline basis, and get clusters with a functional K-means algorithm applied to the coefficients (Tarpey and Kinateder, 2003). Auder and Fischer (2011) extend this method to several Hilbert projection bases and look for the best basis by minimizing some criterion. Chiou and Li (2007) propose a method which generalizes the K-means algorithm by considering covariance structures via functional principal component analysis. Another option adopted by Rossi et al. (2004) is to use a Self-Organizing Map algorithm on the coefficients obtained by projecting the functions onto a B-spline basis. As regards model-based methods, one can cite James and Sugar (2003) who use a spline decomposition specially adapted for sparsely sample functional data, or Ma et al. (2006) who use a spline smoothing. Ray and Mallick (2006) develop a method for optimization of both the basis and the coefficients via a MCMC algorithm.

4.6.1.2 Variable selection

Once the functional data have been projected onto a given basis, the coefficient data is considered to perform the clustering. Each data is described by p variables which are the functions ϕ_j of the basis $\mathcal{B}$. Variable selection for functional data clustering is a recent topic and, to our knowledge, few methods have been developed in this context.

On the one hand, distance-based methods do not offer a rigorous statistical framework to assess

variable relevance. Therefore, most procedures perform dimensional reduction rather than real variable selection. In this case, the point is to determine a level of truncation $p_0 \leq p$ to decompose the function into a truncated basis of size p_0. Then, all the variables ϕ_j with $j \leq p_0$ are kept whereas all the variables ϕ_j with $j > p_0$ are eliminated. Yet, a few procedures have been proposed to perform real variable selection during the clustering process. For instance, Antoniadis et al. (2011) present two methods based on wavelet-based similarity measures: the smooth curves are reduced to a finite number of representative variables by considering the contribution of each wavelet coefficient to the global energy of the curve.

On the other hand, as regards model-based methods, one can cite Michel (2008) who considers ordered variable selection included in the clustering process by performing preliminary projection onto Fourier and wavelet bases. Nonetheless, from a practical point of view, his procedure is only computationally feasible for very low-dimensional data, so it can only deal with curves described by very few points ($p = 64$). Besides, his procedure suffers from the same drawback as dimensional reduction since only ordered variable selection is considered: all the variables ϕ_j with $j \leq p_0$ for some $p_0 \leq p$ are kept whereas all variables ϕ_j with $j > p_0$ are eliminated.

4.6.2 Our procedure for functional data clustering using wavelets

Here, we explain how our Lasso-MLE procedure can be applied for functional data clustering based on preliminary data projection onto some basis. For the sake of simplicity, we restrict to the important case of wavelet bases, but the following description remains valid for any other basis, such as Fourier or spline bases. We do not detail wavelet theory and we refer to Mallat (1999) or Misiti et al. (2007b) for a complete overview on wavelets.

4.6.2.1 Preliminary step: creation of a coefficient dataset from the functional dataset

Let $l \geq 0$ and $h \in \{0, \ldots, 2^l - 1\}$. For some real function ψ, denote by ψ_{lh} the function defined from ψ by dyadic dilation and translation such that $\psi_{lh}(t) = 2^{l/2}\psi(2^l t - h)$. Let φ be a scaling function and ψ be a wavelet function such that $\mathcal{B} = \{\varphi, \psi_{lh}\}_{l \geq 0,\ 0 \leq h \leq 2^l - 1}$ is an orthonormal basis of $L_2([0,1])$.

Consider the dataset $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ defined by (4.53) and assume that $p = 2^L$, $L \in \mathbb{N}^\star$. This functional dataset is to be transformed into a wavelet coefficient dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ by decomposing each functional data $\boldsymbol{y_i}$ into the basis $\mathcal{B}$ in the following way.

Let $i \in \{1, \ldots, n\}$. Assume that $\boldsymbol{y_i}$ comes from cluster k. Then, $\boldsymbol{y_i}$ can be written $\boldsymbol{y_i} = \boldsymbol{f_k} + \boldsymbol{\xi_i}$.

Introduce the wavelet expansion of f_k in the basis $\mathcal{B}$: for all $t \in [0,1]$,

$$f_k(t) = c_0(f_k)\varphi(t) + \sum_{l=0}^{\infty}\sum_{h=0}^{2^l-1} d_{lh}(f_k)\psi_{lh}(t) \tag{4.54}$$

where $c_0(f_k) = \int_a^b f_k(t)\varphi(t)dt$ and $d_{lh}(f_k) = \int_a^b f_k(t)\psi_{lh}(t)dt$. The collection $\{c_0(f_k), d_{lh}(f_k)\}_{l,h}$ is the Discrete Wavelet Transform (DWT) of f_k in the basis $\mathcal{B}$. Now, since f_k is only observed through the p discrete values $\boldsymbol{f_k} = (f_k(t_1), \ldots, f_k(t_p))$, we rewrite (4.54) using the truncation imposed by the $p = 2^L$ points. For all $j \in \{1, \ldots, p\}$,

$$f_k(t_j) = c_0(\boldsymbol{f_k})\varphi(t_j) + \sum_{l=0}^{L-1}\sum_{h=0}^{2^l-1} d_{lh}(\boldsymbol{f_k})\psi_{lh}(t_j). \tag{4.55}$$

Similarly, we have

$$\xi_{ij} = c_0(\boldsymbol{\xi_i})\varphi(t_j) + \sum_{l=0}^{L-1}\sum_{h=0}^{2^l-1} d_{lh}(\boldsymbol{\xi_i})\psi_{lh}(t_j). \tag{4.56}$$

From (4.53), (4.55) and (4.56), we get that

$$\boldsymbol{y_i} = W\boldsymbol{Y_i} \tag{4.57}$$

where $\boldsymbol{Y_i} = (c_0, \boldsymbol{d_0}, \ldots, \boldsymbol{d_{L-1}})$ with $c_0 = c_0(\boldsymbol{f_k}) + c_0(\boldsymbol{\xi_i})$ and $\boldsymbol{d_l} = (d_{l,0}(\boldsymbol{f_k}) + d_{l,0}(\boldsymbol{\xi_i}), \ldots, d_{l,2^l-1}(\boldsymbol{f_k}) + d_{l,2^l-1}(\boldsymbol{\xi_i}))$ for $l \in \{0, \ldots, L-1\}$, and W is a $p \times p$ matrix defined by

$$W = \begin{bmatrix} V^{(0)} & W^{(0)} & \ldots & W^{(L-1)} \end{bmatrix} \tag{4.58}$$

where $V^{(0)} = (\varphi(t_1), \ldots, \varphi(t_p)) \in \mathbb{R}^p$ and $W^{(l)} = (\psi_{lh}(t_j))_{1 \le j \le p, 0 \le h \le 2^l-1}$ is a $p \times 2^l$ matrix for all $l \in \{0, \ldots, L-1\}$. Depending on the wavelet basis, the matrix W can be either orthogonal or nearly orthogonal. For simplicity, assume that W is orthogonal (otherwise see for instance Donoho et al., 1997, Section 4.6). Then, $W^T W = I$ and we deduce from (4.57) that $\boldsymbol{Y_i} = W^T \boldsymbol{y_i}$. In practice, the DWT is not implemented by matrix multiplication but by a sequence of special finite-length filtering steps that results in an efficient $O(n)$ transform (Mallat, 1999). This leads to a n-sample $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ of wavelet coefficient decomposition vectors.

For all $i \in \{1, \ldots, n\}$, if the discretized curve $\boldsymbol{y_i} = \boldsymbol{f_k} + \boldsymbol{\xi_i}$ belongs to cluster k, then its wavelet coefficient decomposition vector can be written

$$\boldsymbol{Y_i} = \boldsymbol{\mu_k} + \boldsymbol{\varepsilon_i} \tag{4.59}$$

where $\boldsymbol{\mu_k} = (\mu_{k1}, \ldots, \mu_{kp})$ and $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})$ are respectively the wavelet coefficient decom-

position vectors of $\boldsymbol{f_k}$ and $\boldsymbol{\xi_i}$. Since W is orthogonal, the noises $\boldsymbol{\xi_i}$ and $\boldsymbol{\varepsilon_i}$ have the same statistical properties. So, $\boldsymbol{\varepsilon_i} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is a white noise.

4.6.2.2 Running of our procedure

From (4.59), the wavelet coefficient dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ is constituted of n observations whose probability distribution is modeled by an isotropic spherical finite Gaussian mixture density $s = \sum_{k=1}^{K} \pi_k \Phi(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I})$. Each observation $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{ip})$ is a wavelet coefficient decomposition of length $p = 2^L$. The mean vectors $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}$ are the wavelet coefficient decomposition of the unknown discretized functions $\boldsymbol{f_1}, \ldots, \boldsymbol{f_K}$. The p variables are the functions φ and ψ_{lh}, $l \in \{0, \ldots, L-1\}$, $h \in \{0, \ldots, 2^l - 1\}$, of the wavelet basis $\mathcal{B}$. By running our Lasso-MLE procedure described in Section 4.5.4 on the dataset $\boldsymbol{Y}$, we get a partition of the data $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ and an estimation $\hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}$ of the mean vectors.

Our Lasso-MLE procedure is naturally suited to functional data clustering:

- For such a problem, the notion of active variable introduced in our procedure (see Section 4.5.2.1) is natural. The function φ or a function ψ_{lh} is inactive in the model if it appears in none of the wavelet coefficient decomposition of the functions $\boldsymbol{f_1}, \ldots, \boldsymbol{f_K}$.

- The Lasso may annihilate mean coefficients μ_{kj} at each level $j \in \{1, \ldots, p\}$. Consequently, contrary to the methods mentioned in Section 4.6.1.2, there is no level p_0 such that $\mu_{kj} = 0$ for all $k \in \{1, \ldots, K\}$ and $j > p_0$ and such that $\mu_{kj} \neq 0$ for at least one $k \in \{1, \ldots, p\}$ for all $j \leq p_0$. So, we are expected to detect relevant variables ϕ_j for the clustering whatever the index level $j \in \{1, \ldots, p\}$. This can be an advantage to distinguish between very similar curves that only differ locally (see Section 5.4.2.2 for one simulation).

- Unlike Michel (2008), our procedure remains efficient for high-dimensional data.

4.6.2.3 Additional step: curve reconstruction for each cluster

Consider the $p \times K$ matrix $\hat{\boldsymbol{\mu}}$ whose k^{th} column is $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}$. Put $\hat{\boldsymbol{f}} = W\hat{\boldsymbol{\mu}}$ where W is defined by (4.58). Then, $\hat{\boldsymbol{f}}$ is a $p \times K$ matrix whose k^{th} column is $\hat{\boldsymbol{f}}_{\boldsymbol{k}} = (\hat{f}_k(t_1), \ldots, \hat{f}_k(t_p))$. Plotting $(t_1, \ldots, t_p) \mapsto (\hat{f}_k(t_1), \ldots, \hat{f}_k(t_p))$ provides a curve estimation of the function f_k for each $k \in \{1, \ldots, K\}$.

Appendices

4.A The EM (Expectation-Maximization) algorithms

4.A.1 The EM algorithm for model-based clustering (Dempster et al., 1977)

Here, we recall the principle and the main steps of the EM algorithm first introduced by Dempster et al. (1977) to compute maximum likelihood estimators in a finite Gaussian mixture density estimation framework.

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with density s. Assume that the data come from several subpopulations and that the density s is to be estimated by the maximum likelihood estimator $\hat{s}$ in the finite Gaussian mixture model $\mathcal{S}$ defined by

$$\mathcal{S} = \left\{ \begin{array}{l} s_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+ \end{array} \right\}.$$

The estimator $\hat{s}$ is the minimizer of the empirical contrast over $\mathcal{S}$: $\hat{s} = s_{\hat{\boldsymbol{\theta}}}$ with

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \gamma_n(s_{\boldsymbol{\theta}}), \qquad \gamma_n(s_{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i})\right).$$

Resolution of this minimization problem is difficult and it is usually done by recasting the problem into the framework of missing data. The complete data are $((\boldsymbol{Y_1}, \boldsymbol{Z_1}), \ldots, (\boldsymbol{Y_n}, \boldsymbol{Z_n}))$ where the latent variables are $\boldsymbol{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_n})$ with $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iK})$ such that

$$Z_{ik} = \left\{ \begin{array}{ll} 1 & \text{if } \boldsymbol{Y_i} \text{ arises from subpopulation } k, \\ 0 & \text{otherwise.} \end{array} \right.$$

If the latent variables Z_{ik} could be observed, then the empirical contrast for the complete data would be

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}, \boldsymbol{Z_i})\right) &= -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Z_i} | \boldsymbol{Y_i})\right) - \frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i})\right) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Z_i} | \boldsymbol{Y_i})\right) + \gamma_n(s_{\boldsymbol{\theta}}). \end{aligned} \tag{4.60}$$

Dempster et al. (1977) propose an algorithm – called EM (Expectation-Maximization) algorithm – to compute $\hat{\boldsymbol{\theta}}$ by an iterative process based on the minimization of the expectation of the empirical contrast for the complete data conditionally to the observations $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ and the current

estimate of the parameters $\boldsymbol{\theta}^{(r)}$ at each iteration r. More precisely, consider for $r \in \mathbb{N}$,

$$\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) := \mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}, \boldsymbol{Z_i})\right) \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(r)}\right] \tag{4.61}$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i}, \boldsymbol{Z_i})\right) \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(r)}\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)} \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)\right) \tag{4.62}$$

where

$$\tau_{ik}^{(r)} := \mathbb{P}\left(Z_{ik} = 1 \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(r)}\right) = \frac{\pi_k^{(r)} \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)}{\sum_{l=1}^{K} \pi_l^{(r)} \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_l}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)} \tag{4.63}$$

denotes the posterior probability of $\boldsymbol{Y_i}$ coming from component k.

INITIALIZATION: $\left(\pi_1^0, \ldots, \pi_K^0, \boldsymbol{\mu_1}^{(0)}, \ldots, \boldsymbol{\mu_K}^{(0)}, \sigma^{(0)}\right)$
AT ITERATION $r \geq 0$,

- **E step** (Expectation step): For all i and k, updating of $\tau_{ik}^{(r)}$ defined by (4.63).
- **M step** (Maximization step): Determination of the parameter vector $\boldsymbol{\theta}^{(r+1)}$ minimizing $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$. From (4.62), it is equivalent to determining the mixing proportions maximizing

$$(\pi_1, \ldots, \pi_K) \mapsto \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\right)$$

under the condition $\sum_{k=1}^{K} \pi_k = 1$, and the vector minimizing

$$(\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \mapsto -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)\right).$$

By differentiating, one gets

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)}}{n},$$

$$\mu_{kj}{}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} Y_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}}, \tag{4.64}$$

$$\sigma^{2(r+1)} = \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{p} \tau_{ik}^{(r)} \left(Y_{ij} - \mu_{kj}^{(r+1)}\right)^2.$$

END OF ITERATION r

These iterations are repeated until convergence.

Now, from (4.60) and (4.61),

$$\gamma_n\left(s_{\theta^{(r)}}\right) = \mathbb{E}\left[\gamma_n(s_{\boldsymbol{\theta}}) \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(r)}\right] = \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) + \mathcal{H}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) \tag{4.65}$$

with $\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) := \mathbb{E}[n^{-1}\sum_{i=1}^{n} \ln(s_{\boldsymbol{\theta}}(\boldsymbol{Z_i}|\boldsymbol{Y_i}) \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(r)})]$. By definition of $\boldsymbol{\theta}^{(r+1)}$, $\mathcal{Q}(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) \leq \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ for all $\boldsymbol{\theta}$, so in particular $\mathcal{Q}(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) \leq \mathcal{Q}(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$. Besides, by Jensen's Inequality, $\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) \leq \mathcal{H}(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$ for all $\boldsymbol{\theta}$, so in particular for $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r+1)}$. Therefore, we deduce from (4.65) that the empirical contrast $\gamma_n\left(s_{\theta^{(r)}}\right)$ is decreased at each iteration r. Moreover, it can be shown that the algorithm converges to a local (yet not necessarily global) minimum of the empirical contrast under some regularity properties (Dempster et al., 1977). In practice, the convergence of the EM algorithm strongly depends on the initialization parameters $\boldsymbol{\theta}^{(0)}$. To increase the probability of reaching the global minimum, many runs of the algorithm can be done from different initialization parameters. By assuming that the global minimum is reached at convergence of the algorithm, we get the minimizer of the empirical contrast $\hat{\boldsymbol{\theta}}$ and we derive the maximum likelihood estimator $\hat{s} = s_{\hat{\boldsymbol{\theta}}}$ of the density s in the model $\mathcal{S}$.

4.A.2 An EM algorithm for ℓ_1-penalized model-based clustering (Pan and Shen, 2007)

Pan and Shen (2007) introduce an EM algorithm to compute their Lasso estimator (see Section 4.3.2). Here, we present the main steps of this algorithm. We keep the notations used in Section 4.3.2.

Pan and Shen's EM algorithm is very close to the EM algorithm for standard model-based clustering recalled in Section 4.A.1. The only difference is that $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)})$ defined by (4.61) is replaced by its ℓ_1-penalized version

$$\begin{aligned}
\mathcal{Q}\left(\overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}^{(r)}\right) &:= \mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}}, \boldsymbol{Z_i})\right) + \lambda|\overline{\boldsymbol{\theta}}|_1 \mid \overline{\boldsymbol{Y}}, \overline{\boldsymbol{\theta}}^{(r)}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ln\left(s_{\overline{\boldsymbol{\theta}}}(\overline{\boldsymbol{Y}}_{\boldsymbol{i}}, \boldsymbol{Z_i})\right) \mid \overline{\boldsymbol{Y}}, \overline{\boldsymbol{\theta}}^{(r)}\right] + \lambda\left|\overline{\boldsymbol{\theta}}^{(r)}\right|_1 \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}} \mid \overline{\boldsymbol{\mu}_{\boldsymbol{k}}}^{(r)}, \sigma^{2^{(r)}}\boldsymbol{I}\right)\right) + \lambda\sum_{j=1}^{p}\sum_{k=1}^{K}\left|\overline{\mu}_{kj}^{(r)}\right|
\end{aligned} \tag{4.66}$$

where

$$\tau_{ik}^{(r)} := \mathbb{P}\left(Z_{ik} = 1 \mid \overline{\boldsymbol{Y}}, \overline{\boldsymbol{\theta}}^{(r)}\right) = \frac{\pi_k^{(r)}\Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}} \mid \overline{\boldsymbol{\mu}_{\boldsymbol{k}}}^{(r)}, \sigma^{2^{(r)}}\boldsymbol{I}\right)}{\sum_{l=1}^{K}\pi_l^{(r)}\Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}} \mid \overline{\boldsymbol{\mu}_{\boldsymbol{l}}}^{(r)}, \sigma^{2^{(r)}}\boldsymbol{I}\right)} \tag{4.67}$$

denotes the posterior probability of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ coming from component k.

INITIALIZATION: $\left(\pi_1^0, \ldots, \pi_K^0, \overline{\boldsymbol{\mu}_{\mathbf{1}}}^{(0)}, \ldots, \overline{\boldsymbol{\mu}_{\boldsymbol{K}}}^{(0)}, \sigma^{(0)}\right)$
AT ITERATION $r \geq 0$,

- **E step**: For all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$, updating of $\tau_{ik}^{(r)}$ defined by (4.67).
- **M step**: Determination of the parameter vector $\overline{\boldsymbol{\theta}}^{(r+1)}$ minimizing $\mathcal{Q}(\overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}^{(r)})$. From (4.66), it is equivalent to determining the mixing proportions maximizing

$$(\pi_1, \ldots, \pi_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\right)$$

under the condition $\sum_{k=1}^{K} \pi_k = 1$, and the vector minimizing

$$(\overline{\boldsymbol{\mu}_{\mathbf{1}}}, \ldots, \overline{\boldsymbol{\mu}_{\boldsymbol{K}}}, \sigma) \mapsto -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\overline{\boldsymbol{Y}_{\boldsymbol{i}}} \mid \overline{\boldsymbol{\mu}_{\boldsymbol{k}}}^{(r)}, \sigma^{2(r)} \boldsymbol{I}\right)\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} |\overline{\mu}_{kj}|.$$

By differentiating, one gets for all $k \in \{1, \ldots, K\}$, for all $j \in \{1, \ldots, p\}$,

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)}}{n}, \tag{4.68}$$

$$\overline{\mu}_{kj}^{(r+1)} = \text{sign}\left(\overline{\mu}_{kj}^{0\,(r+1)}\right) \left(\left|\overline{\mu}_{kj}^{0\,(r+1)}\right| - \frac{n\lambda\, \sigma^{2(r+1)}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}}\right)_+ \tag{4.69}$$

with

$$\overline{\mu}_{kj}^{0\,(r+1)} := \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \overline{Y}_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \quad \text{and} \quad a_+ = \max\{a, 0\}, \tag{4.70}$$

and

$$\sigma^{2(r+1)} = \frac{1}{np} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \tau_{ik}^{(r)} \left(\overline{Y}_{ij} - \overline{\mu}_{kj}^{(r+1)}\right)^2. \tag{4.71}$$

END OF ITERATION r

These iterations are repeated until convergence.

Remark 11.

1. Similarly as in Section 4.A.1, it can be shown that the ℓ_1-penalized empirical contrast is decreased at each iteration of the algorithm and that the algorithm converges to a local minimum of the ℓ_1-penalized empirical contrast under some regularity properties (Zhou et al., 2009). Provided that the EM algorithm indeed converges to the global minimum solution of (4.15), one obtains the Lasso estimator at convergence of the algorithm.

2. The two updatings $\overline{\mu}_{kj}^{(r+1)}$ and $\sigma^{(r+1)}$ in (4.69) and (4.71) depend on each other. Pan and Shen (2007) update $\sigma^{(r+1)}$ with $\overline{\mu}_{kj}^{0}{}^{(r+1)}$ defined by (4.70) and $\overline{\mu}_{kj}^{(r+1)}$ with $\sigma^{(r+1)}$. In our procedure, we use Pan and Shen's EM algorithm, but we update $\overline{\mu}_{kj}^{(r+1)}$ with $\sigma^{(r)}$ and $\sigma^{(r+1)}$ with $\overline{\mu}_{kj}^{(r+1)}$.

3. By taking $\lambda = 0$ (i.e. no penalization) in (4.69), one recovers the same updating of the mean parameters as in (4.64) for the EM algorithm computing the maximum likelihood estimator.

4.A.3 The EM algorithm for Maugis and Michel's procedure

Here, we detail the EM algorithm used at step 2 in Maugis and Michel's procedure described in Section 4.3.1. We keep the notations introduced in Section 4.3.1. Let $(K, \boldsymbol{J_r}) \in \mathcal{K} \times \mathcal{J}'$. Consider the minimizer $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$ of the empirical contrast on the dataset $\overline{\boldsymbol{Y}}$ in the model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$. For all $\boldsymbol{y} \in \mathbb{R}^p$,

$$\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \hat{\sigma}^2\boldsymbol{I}\right) \sum_{k=1}^{K} \hat{\pi}_k \, \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right). \tag{4.72}$$

The estimated mixture parameters $(\hat{\pi}_1, \ldots, \hat{\pi}_K, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{1}}, \ldots, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{K}}, \hat{\sigma})$ are computed by an EM algorithm similar to the EM algorithm described in Section 4.A.1 and applied to the empirically centered dataset $\overline{\boldsymbol{Y}}$. The specific decomposition of $\widehat{\overline{s}}_{(K,\boldsymbol{J_r})}$ in (4.72) must be taken into account during the M step of the algorithm.

INITIALIZATION: $\left(\pi_1^0, \ldots, \pi_K^0, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}{}^{(0)}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}{}^{(0)}, \sigma^{(0)}\right)$
AT ITERATION $r \geq 0$,

- **E step**: For all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$, updating of

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}{}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)}{\sum_{l=1}^{K} \pi_l^{(r)} \Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{l}}{}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)}. \tag{4.73}$$

- **M step**: Determination of the mixing proportions maximizing

$$(\pi_1, \ldots, \pi_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\right)$$

  under the condition $\sum_{k=1}^{K} \pi_k = 1$, and of the vector minimizing

$$(\overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \mapsto -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^{2(r)}\boldsymbol{I}\right) \, \Phi\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}{}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)\right).$$

By differentiating, one gets for all $k \in \{1, \ldots, K\}$, for all $j \in \boldsymbol{J_r}$,

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(r)}}{n}, \tag{4.74}$$

$$\overline{\mu}_{kj}^{(r+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(r)} \overline{Y}_{ij}}{\sum_{i=1}^n \tau_{ik}^{(r)}} \tag{4.75}$$

and

$$\sigma^{2(r+1)} = \frac{1}{np} \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(r)} \left(\sum_{j \in \boldsymbol{J_r}} \left(\overline{Y}_{ij} - \overline{\mu}_{kj}^{(r+1)} \right)^2 + \sum_{j \in \boldsymbol{J_r^c}} \overline{Y}_{ij}^2 \right). \tag{4.76}$$

END OF ITERATION r

These iterations are repeated until convergence.

4.A.4 The EM algorithms for our Lasso-MLE procedure

Here, we detail the three algorithms used in our Lasso-MLE procedure described in Section 4.5.4. We focus only on the description of the iteration r of the algorithm. The initialization and the stopping rules are specified in Section 4.B.2. We keep the notations introduced in Section 4.5.4.

First algorithm: construction of a set of relevant variables

The EM algorithm we use to construct a set of relevant variables is the EM algorithm for ℓ_1-penalized model-based clustering introduced by Pan and Shen (2007) and described in Section 4.A.2. One minor difference is about the updating of the variance parameter at each iteration of the algorithm (see Remark 11).

Second algorithm: construction of a set of active variables among the irrelevant variables

Fix a number of clusters K and an index set $\boldsymbol{J_r}$ representing a set of relevant variables. Introduce $G_{(K,\boldsymbol{J_r})}$ a grid of regularization parameters and let $\lambda \in G_{(K,\boldsymbol{J_r})}$. Consider the Lasso estimator $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)} = (\hat{\boldsymbol{\mu}}, \hat{\sigma}) \in \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \mathbb{R}_+$ defined by (4.52). We compute $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J_r},\lambda)}$ by using an algorithm similar to the EM algorithm introduced by Dempster et al. (1977) and recalled in Section 4.A.1. Yet, no mixture is involved in the minimization problem (4.52), so the following algorithm is simpler: there is no E step and no mixture parameters to update at the M step.

INITIALIZATION : $\left(\boldsymbol{\mu}^{(0)}, \sigma^{(0)}\right)$

AT ITERATION $r \geq 0$, **M step**: Determination of the vector minimizing

$$(\boldsymbol{\mu}, \sigma) \mapsto -\frac{1}{n} \sum_{i=1}^n \ln \left(\Phi \left(\boldsymbol{Y}_{\boldsymbol{i}[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}^{(r)}, \sigma^{2(r)} \boldsymbol{I} \right) \right).$$

By differentiating, we get for all $j \in \boldsymbol{J_r^c}$,

$$\mu_j^{(r+1)} = \text{sign}\left(\mu_j^{0(r+1)}\right)\left(\left|\mu_j^{0(r+1)}\right| - \lambda\,\sigma^{2(r+1)}\right)_+ \tag{4.77}$$

with

$$\mu_j^{0(r+1)} := \frac{1}{n}\sum_{i=1}^{n} Y_{ij} \quad \text{and} \quad a_+ = \max\{a, 0\}$$

and

$$\sigma^{2(r+1)} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij} - \mu_j^{(r+1)}\right)^2. \tag{4.78}$$

END OF ITERATION r

These iterations are repeated until convergence.

Remark 12.

1. The two updatings $\mu_j^{(r+1)}$ and $\sigma^{(r+1)}$ in (4.77) and (4.78) depend on each other. In practice, we update $\mu_j^{(r+1)}$ with $\sigma^{(r)}$ while we update $\sigma^{(r+1)}$ with $\mu_j^{(r+1)}$.
2. By taking $\lambda = 0$ (i.e. no penalization) in (4.77), one finds the maximum likelihood estimator which is the empirical mean.

Third algorithm: estimation of the density in each model

For each $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$, we estimate the density s by the minimizer $\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ of the empirical contrast on the dataset $\boldsymbol{Y}$ in the model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. For all $\boldsymbol{y} \in \mathbb{R}^p$,

$$\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\boldsymbol{y}) = \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \hat{\sigma}^2\boldsymbol{I}\right)\ \Phi\left(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \hat{\boldsymbol{\mu}}, \hat{\sigma}^2\boldsymbol{I}\right)\ \sum_{k=1}^{K}\hat{\pi}_k\,\Phi\left(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \hat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \hat{\sigma}^2\boldsymbol{I}\right). \tag{4.79}$$

The estimated mixture parameters $(\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \hat{\sigma})$ are computed by an EM algorithm similar to the EM algorithm described in Section 4.A.1. The specific decomposition of $\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ in (4.79) must be taken into account during the M step of the algorithm.

INITIALIZATION : $\left(\pi_1^0, \ldots, \pi_K^0, \boldsymbol{\mu}^{(0)}, \boldsymbol{\mu_1}^{(0)}, \ldots, \boldsymbol{\mu_K}^{(0)}, \sigma^{(0)}\right)$
AT ITERATION $r \geq 0$,

- **E step**: For all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$, updating of

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)}\Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)}{\sum_{l=1}^{K}\pi_l^{(r)}\Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_l}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)}.$$

- **M step**: Determination of the mixing proportions maximizing

$$(\pi_1, \ldots, \pi_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\right)$$

under the condition $\sum_{k=1}^{K} \pi_k = 1$, and of the vector minimizing

$$(\boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \mapsto$$
$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\boldsymbol{Y_{i[J_a^c]}} \mid \boldsymbol{0}, \sigma^{2(r)}\boldsymbol{I}\right) \Phi\left(\boldsymbol{Y_{i[J_a]}} \mid \boldsymbol{\mu}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right) \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)\right).$$

By differentiating, we get for all $k \in \{1, \ldots, K\}$,

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)}}{n},$$

$$\mu_{kj}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} Y_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \text{ if } j \in \boldsymbol{J_r}; \qquad \mu_j^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} Y_{ij} \text{ if } j \in \boldsymbol{J_a},$$

and

$$\sigma^{2(r+1)} = \frac{1}{np} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \left(\sum_{j \in \boldsymbol{J_r}} \left(Y_{ij} - \mu_{kj}^{(r+1)}\right)^2 + \sum_{j \in \boldsymbol{J_a}} \left(Y_{ij} - \mu_j^{(r+1)}\right)^2 \sum_{j \in \boldsymbol{J_a^c}} Y_{ij}^2 \right).$$

END OF ITERATION r

These iterations are repeated until convergence.

4.B Some details about our algorithms

4.B.1 Construction of a grid of regularization parameters

To construct a collection of sets of relevant variables, we apply ℓ_1-regularization to the empirical contrast for various regularization parameters (see Section 4.5.4). In practice, for a given number K of clusters, we introduce G_K a grid of regularization parameters and compute the Lasso solution $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)}$ defined by (4.50) by the EM algorithm described in Section 4.A.2 for each regularization parameter λ of the grid G_K. A crucial point is the construction of an appropriate grid G_K. Pan and Shen (2007) consider a regular deterministic grid. On the opposite, we favor a data-driven non-regular grid. Here, we explain the construction of our grid of regularization parameters.

Fix a number K of clusters. We compute the Lasso $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\lambda)}$ for various values of λ in order to

construct a model collection. On the one hand, this model collection must be small enough to enable the estimation of the parameters in each model and the selection of one of these models within a reasonable time-consuming algorithm. On the other hand, this model collection must be rich enough to contain a range of models with different sparsities and enable the selection of a satisfactory model. Therefore, one has to construct a grid of regularization parameters providing an intermediate number of models with various sparsities. So, it is essential to understand the link between the value of the regularization parameter λ and the sparsity of the model generated by the Lasso with regularization parameter λ. From our simulations, we can conclude that increasing the value of λ generally results in an increasing number of mean parameters set to zero, that is to say a sparser model (see Figure 4.1). Moreover, there exist breakpoint regularization parameters $0 := \lambda_0, \lambda_1, \ldots, \lambda_L$ depending on the data such that, for all $l \in \{0, \ldots, L-1\}$, the sparsity of the model generated by the Lasso with $\lambda \in [\lambda_l, \lambda_{l+1}[$ is the same as the sparsity of the model generated by the Lasso with λ_l (only the estimation of the parameters are modified), and the sparsity of the model generated by the Lasso with $\lambda \geq \lambda_L$ is maximal: all the estimates of the mean parameters equal zero. From these observations, we deduce that an optimal grid of regularization parameters should contain only breakpoint regularization parameters since all other regularization parameters increase the size of the grid, and thus the computational time, without providing a richer model collection.

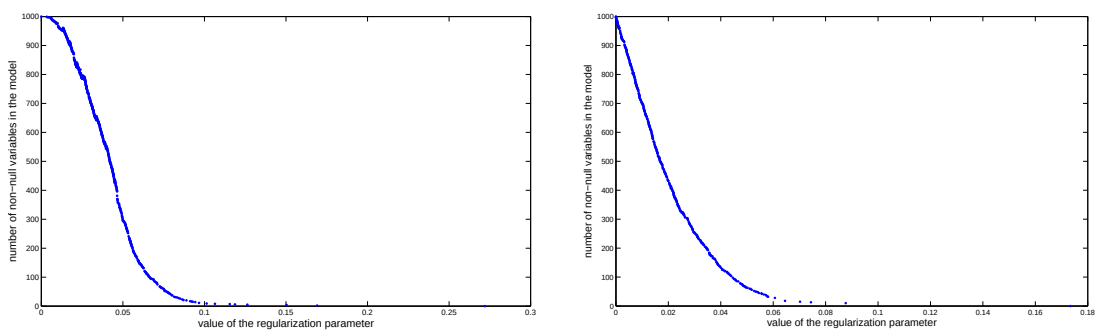

Figure 4.1: Examples of regularization paths obtained for simulations in Section 5.3.

To construct a grid of regularization parameters, an idea (adopted by Pan and Shen, 2007) is to begin at $\lambda = 0$, choose a maximum value of λ, let say λ_{max}, and consider a regular subdivision of the interval $I = [0, \lambda_{\text{max}}]$. We tried to construct such a grid, but it is sometimes difficult to choose λ_{max}. The ideal choice would be $\lambda_{\text{max}} = \lambda_L$, but it depends on the dataset and we have no concrete idea of its value. In order to guess λ_L, Pan and Shen (2007) preliminary conduct a few Lasso algorithms by increasing the value of λ until obtaining a model whose mean parameters all equal zero. But this method can reveal quite time-consuming. Besides arises the problem of the choice of the step of the subdivision of the interval I. On the one hand, if one takes a too large step, one can miss many

breakpoints and get a poor model collection. On the other hand, if one takes a too small step, one may increase the computational time without enriching one's model collection.

To construct a grid of regularization parameters, we consider another approach. The key idea of our method is to construct a data-driven grid of regularization parameters by using the updating formulas of the mixture parameters in the EM algorithm computing the Lasso solutions. Specifically, first, we run the EM algorithm described in Section 4.A.1 to calculate the estimate of the parameter vector $\hat{\boldsymbol{\theta}}(0) = (\hat{\pi}_k(0), \widehat{\overline{\mu}}_{kj}(0), \hat{\sigma}(0))_{k,j}$ for $\lambda = 0$. Then, for all $\lambda > 0$, we consider the formulas updating the values of the prior probability parameters and the mean parameters in the EM algorithm described in Section 4.A.2: at iteration r, for all $k \in \{1, \ldots, K\}$, for all $j \in \{1, \ldots, p\}$,

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(r)}}{n}, \tag{4.80}$$

$$\overline{\mu}_{kj}^{(r+1)} = \text{sign}\left(\overline{\mu}_{kj}^{0\ (r+1)}\right)\left(\left|\overline{\mu}_{kj}^{0\ (r+1)}\right| - \frac{n\lambda\,\sigma^{2(r+1)}}{\sum_{i=1}^n \tau_{ik}^{(r)}}\right)_+ . \tag{4.81}$$

From (4.80) and (4.81),

$$\overline{\mu}_{kj}^{(r+1)} = 0 \Longleftrightarrow \lambda \geq \frac{\sum_{i=1}^n \tau_{ik}^{(r)}}{n\sigma^{2(r+1)}}\left|\overline{\mu}_{kj}^{0\ (r+1)}\right| = \frac{\pi_k{}^{(r+1)}}{\sigma^{2(r+1)}}\left|\overline{\mu}_{kj}^{0\ (r+1)}\right|.$$

Now, by assuming that for all $\lambda > 0$, for all $k \in \{1, \ldots, K\}$, for all $j \in \{1, \ldots, p\}$, the values $\pi_k^{(r+1)}$, $\sigma^{(r+1)}$ and $\overline{\mu}_{kj}^{0\ (r+1)}$ are not too far from the estimates $\hat{\pi}_k(0)$, $\widehat{\overline{\mu}}_{kj}(0)$ and $\hat{\sigma}(0)$ respectively, and by replacing $\overline{\mu}_{kj}^{(r+1)}$ by its value at convergence $\widehat{\overline{\mu}}_{kj}(\lambda)$, we make the following heuristics:

$$\widehat{\overline{\mu}}_{kj}(\lambda) = 0 \Longleftrightarrow \lambda \geq \lambda_{kj} := \frac{\hat{\pi}_k(0)}{\hat{\sigma}^2(0)}\left|\widehat{\overline{\mu}}_{kj}(0)\right|. \tag{4.82}$$

From (4.82), we can expect that $G_K = \{\lambda_{11}, \ldots, \lambda_{1p}, \ldots, \lambda_{K1}, \ldots, \lambda_{Kp}\}$ is a good candidate list for the breakpoint regularization parameters. To ensure to reach λ_L, we add an extra point to the above grid, called λ_{extra}, whose value is taken to twice the maximum value of the λ_{kj}s (the factor 2 is quite arbitrary). We also add the value $\lambda = 0$ as a starting value for the grid. In practice, this method is time-efficient since it only requires to run one EM algorithm for $\lambda = 0$ and then apply Formula (4.82). Contrary to Pan and Shen's grid construction, neither preliminary runs of EM algorithms for a few values of $\lambda > 0$ nor determination of the step of the grid are needed.

4.B.2 Initialization and stopping rules

Here, we explain how we handle the initialization, the stopping-rule and the chaining of the first two EM algorithms used for our Lasso-MLE procedure (see Section 4.A.4).

Let K be a fixed number of clusters. Let $G_K = \{0 := \lambda_0, \lambda_1, \ldots, \lambda_{L-1}, \lambda_L := \lambda_{\text{extra}}\}$ be the grid of regularization parameters constructed above. Assume that it is sorted with $\lambda_0 < \lambda_1 < \cdots < \lambda_L$.

- $l = 0$: $\lambda_l = 0$.
  We run the EM algorithm for standard model-based clustering described in Section 4.A.1. We initialize the EM algorithm by running a few K-means algorithms and we take the estimation of the K-means algorithm maximizing the likelihood as starting value for the EM algorithm. Each K-means algorithm is initialized by selecting at random K observations from the whole sample as initial cluster centroid positions.
  We stop the EM algorithm as soon as we consider that the local maximum has been reached, that is to say when there is no longer significative increase in the likelihood between two successive iterations during a certain number of iterations. To prevent from a too time-consuming algorithm in case of low convergence, we also give a maximal number of iterations as an alternative stopping rule.

- $l \in \{1, \ldots, L\}$: $\lambda_l > 0$.
  We run the EM algorithm for penalized model-based clustering described in 4.A.2. We initialize the EM algorithm with the estimation of the parameter vector $\hat{\boldsymbol{\theta}}(\lambda_{l-1})$ obtained by the EM algorithm for the preceding value λ_{l-1} in the grid G_K.
  The stopping rules are the same as for $l = 0$. At convergence of the EM algorithm, we get the parameter vector estimate $\hat{\boldsymbol{\theta}}(\lambda_l)$ and the associated set of selected variables.

Remark 13. Let us point out that we perform chaining initializations: the EM algorithm with λ_l is initialized with the estimation obtained by the EM algorithm for the preceding value λ_{l-1}. Pan and Shen (2007) do not adopt this strategy. They rather initialize each EM algorithm with one common K-means initialization. In practice, we noted that the estimations and the models obtained for the different values λ_l of the grid change more regularly by performing chaining initializations rather than one common K-means initialization. Besides, the EM algorithm with λ_l tends to converge faster when initialized with the estimation obtained by the EM algorithm for the preceding value λ_{l-1} than when initialized with a K-means algorithm, especially when λ is large.

4.C Proofs

4.C.1 Proof of Claim 4.4.2

Claim 4.4.2 is an immediate consequence of the fact that, if $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a p-multivariate Gaussian random vector with density function $\Phi(\cdot \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, then, for each $j \in \{1, \ldots, p\}$, the random variable X_j is Gaussian with density function $\Phi(\cdot \mid \mu_j, \sigma^2)$.
Let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. Introduce the random variable Z representing the number

of the component from which $\boldsymbol{Y_i}$ arises, that is to say $Z = k$ if $\boldsymbol{Y_i}$ arises from the component k. Consider the $(p+1)$-dimensional random vector $(\boldsymbol{Y_i}, Z)$. Denote by $f_{(Y_{ij},Z)}$ the density of (Y_{ij}, Z) and by $f_{Y_{ij}|Z}$ the density of Y_{ij} conditionally to Z. Then, the marginal density function of Y_{ij} is defined for all $y \in \mathbb{R}$ by

$$s_{\theta_j}(y) = \sum_{k=1}^{K} f_{(Y_{ij},Z)}(y,k) = \sum_{k=1}^{K} \mathbb{P}(Z=k) f_{Y_{ij}|Z}(y|Z=k) = \sum_{k=1}^{K} \pi_k \, \Phi\left(y \mid \mu_{kj}, \sigma^2\right),$$

and

$$\begin{aligned}
\mathbb{E}(Y_{ij}) &= \int_{\mathbb{R}} y s_{\theta_j}(y)\, dy \\
&= \int_{\mathbb{R}} y \sum_{k=1}^{K} \pi_k \, \Phi\left(y \mid \mu_{kj}, \sigma^2\right) \, dy \\
&= \sum_{k=1}^{K} \pi_k \int_{\mathbb{R}} y \, \Phi\left(y \mid \mu_{kj}, \sigma^2\right) \, dy \\
&= \sum_{k=1}^{K} \pi_k \, \mathbb{E}(X) \qquad \text{with } X \sim \Phi\left(\cdot \mid \mu_{kj}, \sigma^2\right) \\
&= \sum_{k=1}^{K} \pi_k \, \mu_{kj}.
\end{aligned}$$

4.C.2 Proof of Proposition 4.4.1

Let $(K, \boldsymbol{J_r}) \in \mathcal{M}_r$. Consider $\hat{s}_{(K,\boldsymbol{J_r})}$ the maximum likelihood estimator in the model $\mathcal{S}_{(K,\boldsymbol{J_r})}$. Let $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})}$ be such that $\hat{s}_{(K,\boldsymbol{J_r})} = s_{\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})}}$. We compute the estimated mixture parameters $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})}$ by the following EM algorithm $\mathcal{A}$.

INITIALIZATION
Let $\overline{\boldsymbol{\theta}}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}^{(0)}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}^{(0)}, \sigma^{(0)})$ be the initialization of $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_r})}$ defined by (4.32) by the EM algorithm $\overline{\mathcal{A}}$. Put $\boldsymbol{\theta}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\mu_1}^{(0)}, \ldots, \boldsymbol{\mu_K}^{(0)}, \sigma^{(0)})$ with $\boldsymbol{\mu}^{(0)} = \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r^c}]}$ and $\boldsymbol{\mu_k}^{(0)} = \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}^{(0)} + \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r}]}$.

AT ITERATION $r \geq 0$,

- **E step**: For all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$, we update

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)} \boldsymbol{I}\right)}{\sum_{l=1}^{K} \pi_l^{(r)} \, \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_l}^{(r)}, \sigma^{2(r)} \boldsymbol{I}\right)}. \tag{4.83}$$

- **M step**: We determine the mixing proportions maximizing

$$(\pi_1, \ldots, \pi_K) \mapsto \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\pi_k^{(r)}\right)$$

under the condition $\sum_{k=1}^{K} \pi_k = 1$, and the vector minimizing

$$(\boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \mapsto -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \ln\left(\Phi\left(\boldsymbol{Y_{i[J_r^c]}} \mid \boldsymbol{\mu}, \sigma^{2(r)}\boldsymbol{I}\right) \ \Phi\left(\boldsymbol{Y_{i[J_r]}} \mid \boldsymbol{\mu_k}^{(r)}, \sigma^{2(r)}\boldsymbol{I}\right)\right).$$

By differentiating, we get for all $k \in \{1, \ldots, K\}$,

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)}}{n}, \tag{4.84}$$

$$\mu_{kj}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} Y_{ij}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \ \text{ if } j \in \boldsymbol{J_r}, \qquad \mu_j^{(r+1)} = \overline{\mu}_j \ \text{ if } j \in \boldsymbol{J_r^c} \tag{4.85}$$

and

$$\sigma^{2(r+1)} = \frac{1}{np} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(r)} \left(\sum_{j \in \boldsymbol{J_r}} \left(Y_{ij} - \mu_{kj}^{(r+1)}\right)^2 + \sum_{j \in \boldsymbol{J_r^c}} \left(Y_{ij} - \mu_j^{(r+1)}\right)^2 \right). \tag{4.86}$$

END OF ITERATION r

Now, let us compare this EM algorithm with the EM algorithm $\overline{\mathcal{A}}$ described in Section 4.A.3. By comparing Formulas (4.83)–(4.86) with Formulas (4.73)–(4.76) and by recalling that $\overline{Y}_{ij} = Y_{ij} - \overline{\mu}_j$, we see that the updatings of $\tau_{ik}^{(r)}$, $\pi_k^{(r+1)}$ and $\sigma^{2(r+1)}$ are the same for both EM algorithms $\overline{\mathcal{A}}$ and $\mathcal{A}$, while the updatings of the mean parameters are linked by the relation $\mu_{kj}^{(r+1)} = \overline{\mu}_{kj}^{(r+1)} + \overline{\mu}_j$. These relations remain checked at convergence of the algorithms. Thus, if $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_r})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{1}}, \ldots, \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{K}}, \hat{\sigma})$ are the parameters of the maximum likelihood estimator in the model $\overline{\mathcal{S}}_{(K,J)}$ computed by the EM algorithm $\overline{\mathcal{A}}$, then $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J_r})} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\mu}}_{\boldsymbol{K}}, \hat{\sigma})$ with $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r^c}]}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} = \widehat{\overline{\boldsymbol{\mu}}}_{\boldsymbol{k}} + \overline{\boldsymbol{\mu}}_{[\boldsymbol{J_r}]}$.

Chapter 5

Simulations

Contents

5.1. Introduction **181**

5.2. Definitions **182**

5.2.1. The oracle model 182

5.2.2. Model selection criteria 183

5.2.3. Selection of the relevant variables for the clustering 184

5.3. Validation of the Lasso-MLE procedure on simulated data **184**

5.3.1. First simulated dataset 185

5.3.2. Second simulated dataset 193

5.3.3. Some remarks on our procedure 195

5.4. Functional data clustering using wavelets **198**

5.4.1. Examples of curve reconstruction from wavelet coefficient clustering . . . 199

5.4.2. Examples of functional data clustering 206

ABSTRACT

This chapter is devoted to the application of our Lasso-MLE procedure to simulated data. Both low-dimensional and high-dimensional data are considered in order to study the penalty shape used for our data-driven model selection criterion according to the data dimension. The simulations highlight that our Lasso-MLE procedure is very competitive compared with other variable selection procedures for finite Gaussian mixture model-based clustering. The main advantages of our procedure are a fast automatic variable selection by the Lasso, a good parameter estimation by the MLE and an efficient non-asymptotic data-driven model selection criterion adaptive to the data dimension. Thanks to the combination of these advantages, our procedure is particularly suited to functional data clustering involving curve reconstruction with wavelets.

5.1 Introduction

In this chapter, we test our Lasso-MLE procedure introduced in Section 4.5 on simulated data. The aim of these simulations is manifold.

First, we want to evaluate the performance of our non-asymptotic model selection criterion based on the slope heuristics. One major point is to check whether the ideal penalty shape depends on the data dimension and to determine whether a logarithm term becomes necessary to define a proper model selection criterion as the dimension increases. To find the best penalty shape, we compare the model selected by our criterion with the oracle model. Furthermore, we compare our non-asymptotic model selection criterion with the classical asymptotic criteria AIC and BIC.

Second, we want to compare our Lasso-MLE procedure to other model selection procedures for clustering based on finite Gaussian mixture models. We restrict our comparison to two procedures sharing similarities with our own procedure, which are Pan and Shen's Lasso procedure and Maugis and Michel's procedure, both presented in Section 4.3. On the one hand, we use the Lasso to select a limited model collection such as Pan and Shen (2007). On the other hand, we estimate the parameters by the MLE and we apply a non-asymptotic criterion such as Maugis and Michel (2011a). Comparison with Maugis and Michel (2011a) is possible only for low-dimensional data since their procedure is not feasible for high-dimensional data. On the contrary, comparison with Pan and Shen (2007) is carried out whatever the data dimension. Both variable selection and clustering performance are compared.

Third, we simulate some functional data to evaluate the performance of our Lasso-MLE procedure in this fundamental application domain.

The chapter is organized as follows. In Section 5.2, we introduce a few notations and definitions used to analyze the further results. In Section 5.3, we compare the variable selection and the clustering performance of our procedure with Pan and Shen's procedure and Maugis and Michel's procedure on two simulated datasets. In Section 5.4, we apply our procedure to functional data clustering involving curve reconstruction with wavelets.

5.2 Definitions

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with (known) density s. Let $\{\mathcal{S}_D\}_{D\in\mathcal{D}}$ be a model collection. For all $D \in \mathcal{D}$, denote by $\hat{s}_D$ an estimator of s in the model $\mathcal{S}_{\mathcal{D}}$.

5.2.1 The oracle model

The oracle[1] model is $\mathcal{S}_{D_{\text{oracle}}}$ with $D_{\text{oracle}} = \arg\min_{D\in\mathcal{D}} \mathrm{KL}\left(s, \hat{s}_D\right)$. For simulated datasets, the density s is known and we have access to the oracle model. Indeed, for all $D \in \mathcal{D}$,

$$\begin{aligned}\mathrm{KL}\left(s, \hat{s}_D\right) &= \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\frac{s(\boldsymbol{x})}{\hat{s}_D(\boldsymbol{x})}\right) s(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(s(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x} - \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\hat{s}_D(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x}\,. \end{aligned} \tag{5.1}$$

Since the left-hand side integral in (5.1) does not depend on D, we get that

$$D_{\text{oracle}} = \arg\min_{D\in\mathcal{D}} \left\{ -\int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\hat{s}_D(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x} \right\}. \tag{5.2}$$

The integral in (5.2) can be approximated by a Monte Carlo procedure. A large dataset $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_N})$ is simulated with density s and $N \gg n$. Then, from the Law of large numbers,

$$\int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\hat{s}_D(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x} \approx \frac{1}{N}\sum_{i=1}^{N} \ln\left(\hat{s}_D(\boldsymbol{X_i})\right)$$

and

$$D_{\text{oracle}} \approx \arg\min_{D\in\mathcal{D}} \left\{ -\frac{1}{N}\sum_{i=1}^{N} \ln\left(\hat{s}_D(\boldsymbol{X_i})\right) \right\}. \tag{5.3}$$

Then, we can compare the model selected by some criterion with this benchmark to judge the quality of this criterion.

[1] In the literature, the oracle model is rather defined as the model $\mathcal{S}_{D^*}$ minimizing the expected risk: $D^* = \arg\min_{D\in\mathcal{D}} \mathbb{E}_{\boldsymbol{Y}}\left[\mathrm{KL}\left(s, \hat{s}_D\right)\right]$ where $\mathbb{E}_{\boldsymbol{Y}}$ is the expectation taken with respect to the sample $\boldsymbol{Y}$. Nonetheless, in our Lasso-MLE procedure, the collection $\mathcal{D}$ depends on the sample $\boldsymbol{Y}$, so D^* is not defined, and we define an oracle D_{oracle} for each dataset.

5.2.2 Model selection criteria

A model selection procedure is a method choosing some data-dependent $\hat{D} \in \mathcal{D}$ nearly as good as the ideal choice D_{oracle} . Then, the resulting estimator of s is $\hat{s}_{\hat{D}}$. The idea of model selection via penalization is to introduce a penalty function $\mathrm{pen} : \mathcal{D} \mapsto \mathbb{R}^+$ and to select $\hat{D}$ as the minimizer of the penalized criterion over $\mathcal{D}$,

$$\hat{D} = \arg\min_{D \in \mathcal{D}} \left\{ \gamma_n(\hat{s}_D) + \mathrm{pen}(D) \right\}. \tag{5.4}$$

In the density estimation framework, model selection via penalization was first introduced in the early seventies by Akaike (1973) who proposed the Akaike Information Criterion (AIC) defined from the penalty

$$\mathrm{pen}_{\text{AIC}}(D) = \frac{D}{n}.$$

Its heuristics relies on an asymptotic approximation of the log-likelihood based on Wilks Theorem. This criterion asymptotically aims at minimizing the Kullback-Leibler divergence to the true distribution, but it is not adapted to identify the true model (Yang, 2005).
Then, Schwarz (1978) suggested the Bayesian Information Criterion (BIC) defined from the penalty

$$\mathrm{pen}_{\text{BIC}}(D) = \frac{\ln n}{2} \frac{D}{n}.$$

This criterion is derived from Bayesian considerations. It relies on an asymptotic approximation of the integrated likelihood which is based on Laplace approximation (Lebarbier and Mary-Huard, 2006). Schwarz (1978) obtained this criterion in the particular case of exponential families, and under assumptions related to the introduced Bayesian framework. Contrary to AIC, BIC is expected to be a good identification criterion. In their Lasso procedure, Pan and Shen (2007) consider BIC as model selection criterion to select an estimator among their collection of Lasso estimators.

Both AIC and BIC heavily rely on asymptotic approximations. In particular, their heuristics are justified only when the dimensions and the number of models are bounded with respect to the number of observations n and n tends to infinity, which is not the case in our high-dimensional setting. Moreover, they are both deterministic criteria in the sense that the penalty associated to these criteria is the same whatever the dataset. We shall compare these two widely-used criteria to our non-asymptotic data-driven model selection criterion.

As regards our criterion, we shall consider two penalties: one proportional to the dimension and another one involving a logarithm term:

$$\mathrm{pen}(D) = 2\,\hat{c}\,\frac{D}{n}, \qquad \mathrm{pen}_{\ln}(D) = 2\left(\hat{c}_1\,\frac{D}{n} + \hat{c}_2\,\frac{D}{n}\,\ln\left(\frac{D_{\max}}{D}\right)\right). \tag{5.5}$$

The constants $\hat{c}$, $\hat{c}_1$ and $\hat{c}_2$ are estimated from the data by the data-driven slope estimation method

recalled in Section 6.3.1: $\hat{c}$ is computed by performing a simple regression on the couples of points $\{(D/n, -\gamma_n(\hat{s}_D)\}_{D \leq n \wedge p}$, while $\hat{c}_1$ and $\hat{c}_2$ are computed by performing a double regression on the triplets of points $\{(D/n, (D/n)\ln(D_{\max}/D), -\gamma_n(\hat{s}_D)\}_{D \leq n \wedge p}$.

We shall call the penalized estimator solution of (5.4) with $\mathrm{pen}_{\mathrm{AIC}}$, $\mathrm{pen}_{\mathrm{BIC}}$, pen and $\mathrm{pen}_{\ln}$ respectively the AIC estimator, the BIC estimator, the slope estimator and the $\ln$-slope estimator.

5.2.3 Selection of the relevant variables for the clustering

When considering a simulated dataset from a mixture population described by p variables, we know which variables are actually relevant for the clustering. Then, we can compare them with the set of relevant variables selected by one's procedure to judge the selection variable quality of this procedure. We shall say that a variable selected by one's procedure is a *true relevant* variable if it is actually a relevant variable, while we shall say that a variable selected by one's procedure is a *false relevant* variable if it is actually not a relevant variable. In our procedure, besides looking for the relevant variables, we also aim at eliminating every inactive variable so as to reduce the number of mean coefficients to estimate. We shall say that our procedure selects a *false active* variable if it declares a variable active whereas this variable is actually inactive.

For a simulated dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ coming from a mixture of populations, we know from which cluster arises each observation $\boldsymbol{Y_i}$. Then, we can compare the true data partition with the partition obtained by one's procedure to judge the clustering quality of this procedure. In our simulations, we shall use the Adjusted Rand Index (ARI) which is the corrected-for-chance version of the Rand index introduced by Rand (1971) to measure the similarity between two data clusterings. The ARI is a number between 0 and 1. The closer to 1 the ARI, the more similar the two clusterings.

5.3 Validation of the Lasso-MLE procedure on simulated data

In this section, we compare our Lasso-MLE procedure with Pan and Shen's procedure and Maugis and Michel's procedure on two simulated datasets. For Pan and Shen's Lasso procedure, we consider the modified BIC criterion defined by (4.18). For Maugis and Michel's procedure, we consider the data-driven penalized criterion derived from the slope heuristics with a penalty proportional to the dimension, which is defined by (4.13). For our procedure, we test both a penalty proportional to the dimension and a penalty with an additional logarithm term, as defined by (5.5). The corresponding estimators are respectively the slope estimator and $\ln$-slope estimator. They are compared with BIC and AIC. For each procedure, we also provide the results for the oracle defined by (5.3).

5.3.1 First simulated dataset

This first simulated dataset is in the spirit of the example in Maugis and Michel (2011a). The dataset consists of $n = 200$ observations described by $p \in \{30, 200, 1000\}$ variables. The data are simulated according to a mixture of four Gaussian distributions $\sum_{k=1}^{4} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ with

$$\begin{aligned}
&\boldsymbol{\mu_1} = (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025, \mathbf{0_{p-10}}),\\
&\boldsymbol{\mu_2} = \mathbf{0_p}, \quad \boldsymbol{\mu_3} = -\boldsymbol{\mu_1},\\
&\boldsymbol{\mu_4} = (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025, \mathbf{0_{p-10}}).
\end{aligned}$$

The vector $\mathbf{0_l}$ denotes the null vector of length l. The mixing proportions are $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.3, 0.2, 0.2, 0.3)$. The relevant variables for the clustering are the first ten variables. Thus, the true density belongs to the model with $K^* = 4$ components and $\boldsymbol{J_r^*} = \{1, \ldots, 10\}$. Note that the four subpopulations of the mixtures are progressively gathered together into a unique Gaussian distribution, as shown at Figure 5.1. Therefore, the discriminant power of the relevant variables decreases with respect to the variable index. Also note that the active and the relevant variables coincide. For each value of $p \in \{30, 200, 1000\}$, we perform 20 simulations of the dataset. For each simulation, we consider models with $K \in \{2, \ldots, 6\}$ clusters.

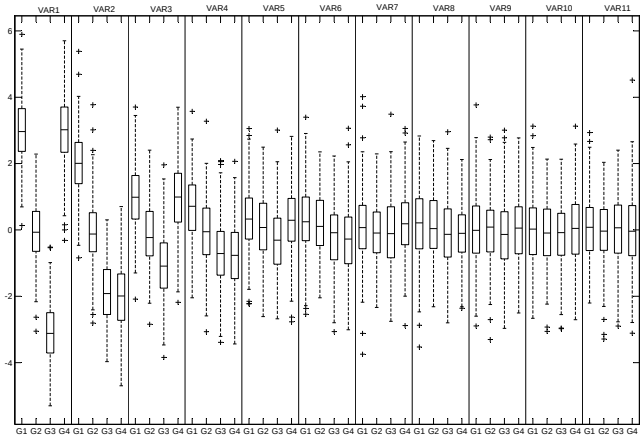

Figure 5.1: Boxplots of the first eleven variables (VAR1,...,VAR11) on the four mixture components (G1,G2,G3,G4).

5.3.1.1 Low-dimensional dataset: $p \ll n$

Here, we take $p = 30$. We compare our Lasso-MLE procedure with Pan and Shen's Lasso procedure and Maugis and Michel's MLE procedure. For Maugis and Michel's procedure, ordered variable selection is considered because of the unfeasibility of performing complete variable selection for

$p = 30$. The results are summarized in Table 5.1. The comparison between the slope graph obtained by Maugis and Michel's procedure and our procedure for one simulation is provided at Figure 5.3. Figure 5.2 gives an example of the collection of sets of relevant variables generated by the Lassos by varying the regularization parameter for the true number of clusters $K^* = 4$. This data-driven collection is to be compared with the ordered variable collection $\{\{1, \ldots, d\}; 1 \leq d \leq p\}$ considered by Maugis and Michel (2011a). Let us comment these results.

Comparison between different criteria for our procedure

First, let us comment on the data-driven slope estimation method. For this low-dimensional setting, the double regression to estimate the coefficients $\hat{c}_1$ and $\hat{c}_2$ in (5.5) often fails to lead to stable coefficients, in which case the penalty $\mathrm{pen}_{\mathrm{ln}}$ in (5.5) and the ln-slope estimator can not be defined. On the contrary, the simple regression to compute $\hat{c}$ in (5.5) leads to a stable coefficient. This suggests that no logarithm term is to be considered in the penalty shape. Thus, we present only the results obtained for the slope estimator. Let us compare them with AIC and BIC.

From Table 5.1, AIC selects too many components and many false relevant variables. The model selected by BIC is not the true model, which proves that BIC is not consistent in our context. This can be due to several reasons: we are not in an asymptotic context, the true model does not necessarily belong to our random model collection (see Figure 5.2) and the component densities are not bounded, whereas these assumptions are usually made to establish consistency results for BIC (see for instance Keribin, 2000). The data-driven slope estimation method selects a model closer to the oracle model than the model selected by BIC. Moreover, the data clustering is slightly better for the slope estimator. These results highlight the advantage of a data-driven penalty over a fixed penalty.

Comparison between our procedure and Pan and Shen's procedure

From Table 5.1, the Lasso oracle model sometimes overestimates the number of mixture components. It contains most (yet not all) true relevant variables but also many false relevant variables. This tendency to select much too many variables is not surprising because it has already been widely noted in the regression framework (Zhao and Yu, 2007; Zou, 2006; Yuan and Lin, 2007; Bach, 2008; Connault, 2011). Our simulations show that this drawback is still observed in the finite mixture Gaussian density estimation framework. As a result, the Lasso oracle model is far from the true model. On the contrary, the Lasso-MLE oracle model is close to the true model. This favors the Lasso-MLE procedure.

$p = 30$

procedure	estimator	{TR,FR,FA}		K	ARI		KL$(s, \hat{s})$	
Lasso	oracle	$\{9, 14, 0\}$	$(\{1, 1, 0\})$	$\{0, 0, 14, 6, 0\}$	0.90	(0.05)	0.18	(0.03)
	BIC	$\{6, 2, 0\}$	$(\{1, 2, 0\})$	$\{0, 0, 14, 6, 0\}$	0.87	(0.06)	0.31	(0.06)
MLE	oracle	$\{6, 0, 0\}$	$(\{2, 0, 0\})$	$\{0, 0, 20, 0, 0\}$	0.88	(0.05)	0.15	(0.03)
	AIC	$\{7, 0, 0\}$	$(\{1, 0, 0\})$	$\{0, 0, 8, 6, 6\}$	0.88	(0.04)	0.17	(0.05)
	BIC	$\{5, 0, 0\}$	$(\{1, 0, 0\})$	$\{0, 0, 20, 0, 0\}$	0.89	(0.05)	0.18	(0.04)
	slope	$\{6, 0, 0\}$	$(\{1, 0, 0\})$	$\{0, 0, 18, 2, 0\}$	0.89	(0.05)	0.16	(0.04)
Lasso-MLE	oracle	$\{6, 0, 0\}$	$(\{1, 0, 0\})$	$\{0, 0, 20, 0, 0\}$	0.89	(0.06)	0.13	(0.03)
	AIC	$\{8, 2, 7\}$	$(\{1, 1, 2\})$	$\{0, 0, 8, 8, 4\}$	0.89	(0.06)	0.19	(0.05)
	BIC	$\{5, 0, 0\}$	$(\{1, 0, 0\})$	$\{0, 0, 20, 0, 0\}$	0.89	(0.06)	0.15	(0.03)
	slope	$\{6, 1, 1\}$	$(\{1, 1, 1\})$	$\{0, 0, 18, 2, 0\}$	0.90	(0.05)	0.14	(0.04)

$p = 200$

Lasso	oracle	$\{8, 62, 0\}$	$(\{1, 9, 0\})$	$\{0, 0, 14, 4, 2\}$	0.84	(0.04)	0.70	(0.09)
	BIC	$\{6, 4, 0\}$	$(\{1, 4, 0\})$	$\{0, 0, 14, 4, 2\}$	0.79	(0.08)	1.29	(0.19)
Lasso-MLE	oracle	$\{6, 0, 1\}$	$(\{1, 0, 1\})$	$\{0, 0, 20, 0, 0\}$	0.85	(0.05)	0.21	(0.13)
	AIC	$\{7, 11, 41\}$	$(\{1, 6, 15\})$	$\{0, 0, 10, 8, 2\}$	0.82	(0.04)	0.59	(0.19)
	BIC	$\{5, 1, 1\}$	$(\{1, 1, 0\})$	$\{0, 0, 20, 0, 0\}$	0.84	(0.05)	0.25	(0.16)
	slope	$\{6, 1, 1\}$	$(\{1, 1, 0\})$	$\{0, 0, 20, 0, 0\}$	0.84	(0.05)	0.23	(0.15)
	ln-slope	$\{6, 1, 1\}$	$(\{1, 1, 0\})$	$\{0, 0, 20, 0, 0\}$	0.85	(0.05)	0.23	(0.15)

$p = 1000$

Lasso	oracle	$\{6, 100, 0\}$	$(\{1, 19, 0\})$	$\{0, 0, 8, 8, 4\}$	0.83	(0.04)	2.77	(0.12)
	BIC	$\{5, 12, 0\}$	$(\{1, 3, 0\})$	$\{0, 0, 10, 6, 4\}$	0.77	(0.07)	3.51	(0.18)
Lasso-MLE	oracle	$\{5, 1, 1\}$	$(\{1, 2, 0\})$	$\{0, 0, 20, 0, 0\}$	0.84	(0.04)	0.35	(0.12)
	AIC	$\{6, 13, 99\}$	$(\{1, 8, 28\})$	$\{0, 0, 10, 6, 4\}$	0.82	(0.09)	1.75	(0.12)
	BIC	$\{6, 5, 19\}$	$(\{1, 4, 4\})$	$\{0, 0, 20, 0, 0\}$	0.83	(0.04)	0.65	(0.13)
	slope	$\{5, 4, 10\}$	$(\{1, 6, 4\})$	$\{0, 0, 14, 2, 0\}$	0.83	(0.06)	0.54	(0.13)
	ln-slope	$\{5, 1, 1\}$	$(\{1, 2, 1\})$	$\{0, 0, 20, 0, 0\}$	0.84	(0.05)	0.36	(0.13)

Table 5.1: Mean number {TR,FR,FA} of true relevant, false relevant and false active variables, number of times $\{\nu_2, \nu_3, \nu_4, \nu_5, \nu_6\}$ a clustering with respectively $K = 2$, $K = 3$, $K = 4$, $K = 5$ and $K = 6$ components is selected, mean ARI and mean Kullback-Leibler divergence over the 20 simulations (except for the slope estimator, $p = 1000$ who fails 4 times). The standard deviations are put into brackets.

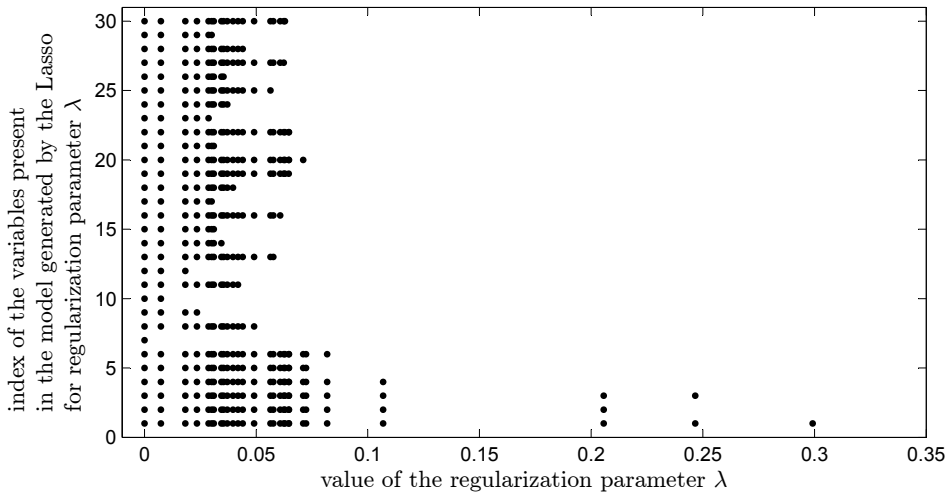

Figure 5.2: For one simulation, $p = 30$, $K = 4$, collection of the sets of relevant variables generated by the Lassos by varying the regularization parameter λ. Remember that the relevant variables are indexed by $j \in \{1, \ldots, 10\}$ and that the smaller $j \in \{1, \ldots, 10\}$, the higher the relevance of the variable indexed by j.

Comparison between our procedure and Maugis and Michel's procedure

The results obtained for both procedures are very similar. In particular, at Figure 5.3, we check that, for both procedures, the function $D/n \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior. Moreover, the values of the estimated slopes are very similar for both procedures. But two differences can be noted.

First, from Table 5.1, the models selected by Maugis and Michel's procedure never contain false relevant variables, contrary to our procedure. This can be explained by the following reason. As highlighted by Figure 5.1, the relevant variables indexed by $j \in \{7, \ldots, 10\}$ are actually not really relevant for the clustering, and they may be confused with irrelevant variables. Then, our procedure may select irrelevant variables indexed by $j \geq 11$ (possibly $j \gg 11$) instead of these relevant variables. This phenomenon is much less likely to happen for Maugis and Michel's procedure. Indeed, Maugis and Michel (2011a) only consider ordered variable selection, so selecting a model with a false relevant variable indexed by $j \gg 11$ means that this model is preferred to all the models included in this model, whereas they are all closer to Maugis and Michel's oracle model which contains only the first six variables. Let us stress that the dataset is favorable to Maugis and Michel's procedure since the variables are ordered by decreasing relevance importance, which is crucial for Maugis and Michel's ordered variable selection procedure to work. On the contrary, our procedure remains efficient whatever the order of the variables[2].

[2]We performed one simulation to check that the order of the variables has no influence on the results for our procedure,

Secondly, comparison of the Kullback-Leibler divergences in Table 5.1 shows that our procedure achieves better prediction performance than Maugis and Michel's procedure. This may be due to the fact that Maugis and Michel (2011a) estimate the parameters on the empirically centered dataset (see discussion in Section 4.4.2.2).

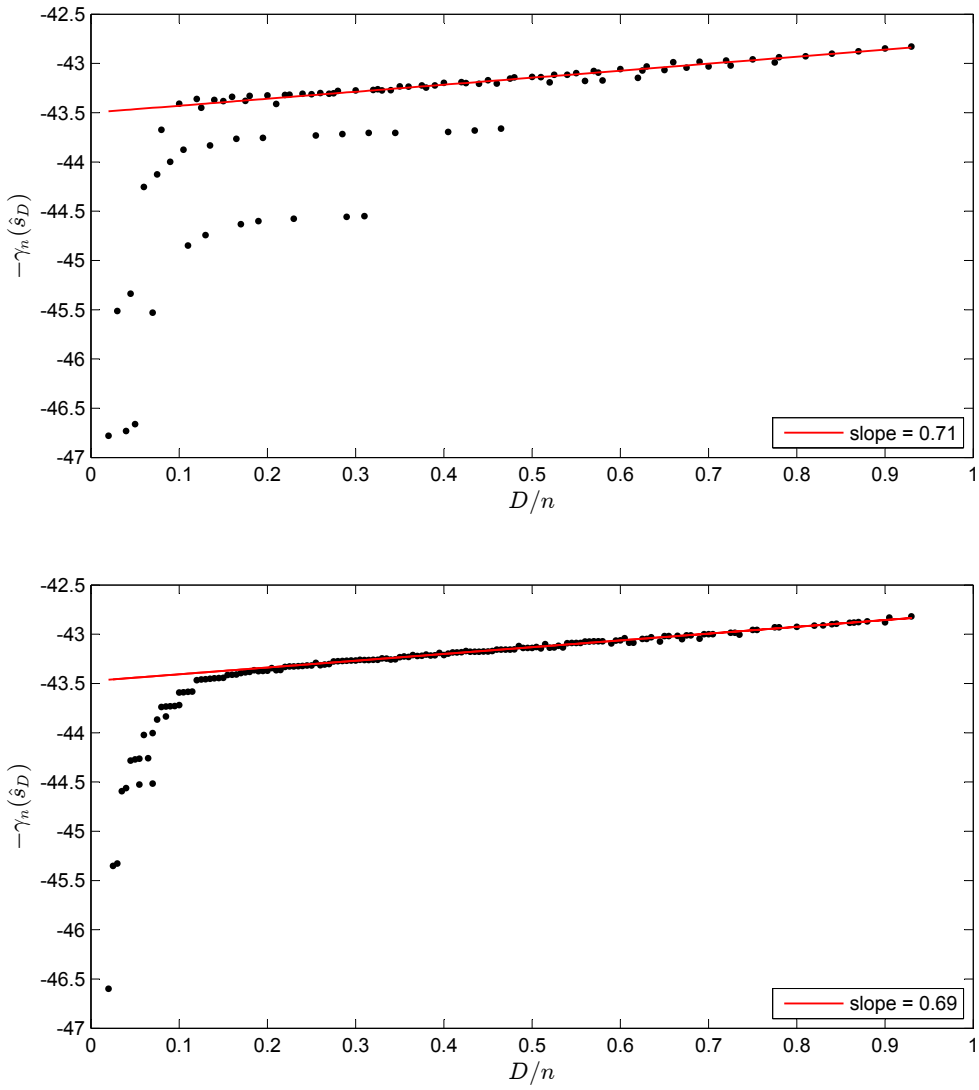

Figure 5.3: For one simulation, comparison between the slope graphs obtained by Maugis and Michel's procedure (at the top) and by our Lasso-MLE procedure (at the bottom). On both graphs, for large dimensions, we observe a linear part whose estimated slope is specified on the graphs.

which is not the case for Maugis and Michel's procedure which fails to recover the true relevant variables if they are not correctly ordered or if they are mixed up with irrelevant variables.

Note that the true model does not necessarily belong to our data-driven model collection (see Figure 5.2), while it belongs to the deterministic model collection considered by Maugis and Michel (2011a). From Table 5.1, for both procedures, the oracle model does not coincide with the true model.

5.3.1.2 Higher dimensional dataset: $p = n$ and $p \gg n$

Here, we take $p = 200$ and $p = 1000$. We compare our Lasso-MLE procedure to Pan and Shen's Lasso procedure. Maugis and Michel's procedure is unfeasible and thus ruled out for such high-dimensional datasets. The results are summarized in Table 5.1. Figure 5.5 and Figure 5.6 compare the clustering results between the Lasso procedure and the Lasso-MLE procedure. A comparison between the fixed penalty considered by BIC and the data-driven penalty derived from the slope heuristics is made at Figure 5.4. Let us focus on the results evolution when increasing p.

Slope heuristics

For $p = 30$, the double regression fails. For $p = 200$, both the simple regression and the double regression can be performed for all simulations and they lead to equivalent estimators. For $p = 1000$, the simple regression fails 4 times over the 20 simulations, and on the remaining 16 simulations, the $\ln$-slope estimator is closer to the oracle than the slope estimator. So, the ideal penalty shape seems to change when p increases: a logarithm term is to be added for large dimensions.

Variable selection

For the Lasso procedure, Table 5.1 shows that the poor variable selection performance observed for $p = 30$ for the oracle and BIC is confirmed and even gets worse when p grows: the models selected contain more and more false relevant variables while they contain fewer and fewer true relevant variables.

For the Lasso-MLE procedure, Table 5.1 shows that the models selected by AIC and, to a lesser extend by BIC, contain more (true relevant and false relevant) variables when p grows. On the opposite, the oracle model and the models selected by the slope estimator and the $\ln$-slope estimator remain stable. Figure 5.4 points out the advantage of using a data-driven penalty rather than a fixed penalty: for a penalty proportional to the dimension $\mathrm{pen}(D) = 2\hat{c}D/n$, the value of the estimated slope $\hat{c}$ globally increases with respect to p, so the associated penalty pen gets stronger as p increases, whereas the fixed BIC penalty $\mathrm{pen}_{\mathrm{BIC}}(D) = 0.5\ln(n)(D/n)$ remains the same whatever p and tends to under-penalize as p grows.

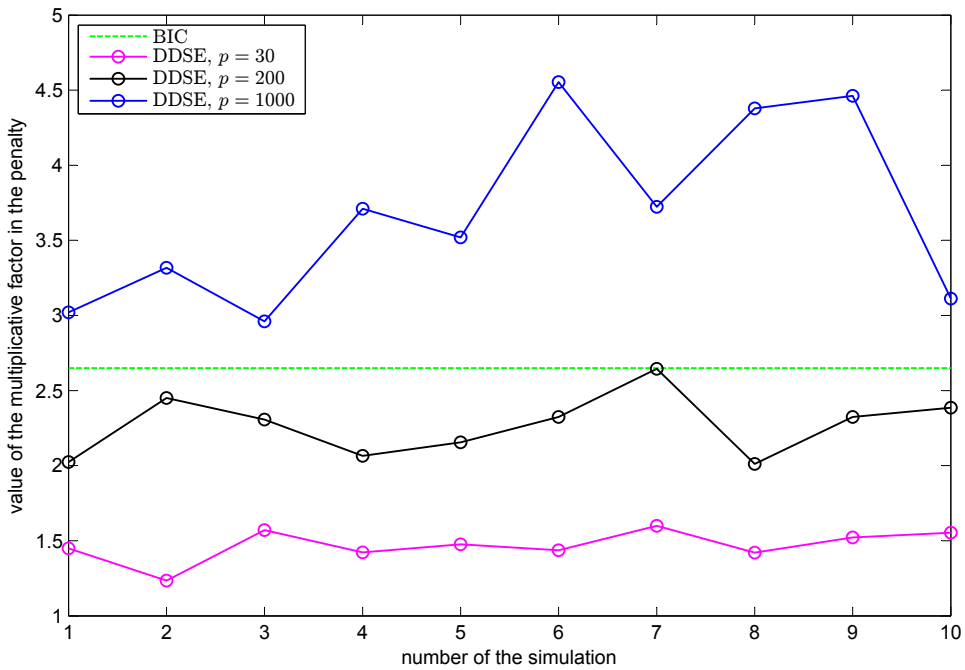

Figure 5.4: For 10 simulations, comparison between the value of the multiplicative factor C in the penalty $\mathrm{pen}(D) = CD/n$ defining the model selection criterion for BIC and the data-driven slope estimator (DDSE). For BIC, $C = 0.5 \ln n = 2.65$ is constant. For the DDSE, $C = 2\hat{c}$ is changing at each simulation according to the slope estimation $\hat{c}$.

Clustering performance

From Table 5.1, the data clustering globally lightly deteriorates as p grows. The biggest deterioration is for BIC for Pan and Shen's procedure. Figure 5.5 can explain this bad clustering performance: for the Lasso procedure, the models achieving the best clustering have moderate or high dimension ($D/n > 0.35$), while BIC selects a less complex model. On the opposite, for the Lasso-MLE procedure, Figure 5.6 shows that small models ($D/n < 0.5$) achieve good clustering. In particular, the slope estimator and the $\ln$-slope estimator, which select such small models, lead to a satisfactory data clustering.

Figure 5.5 and Figure 5.6 confirm that variable selection is useful to get a better data clustering: first, introducing the relevant variables into the models improves the clustering, but then, adding the irrelevant variables deteriorates the clustering.

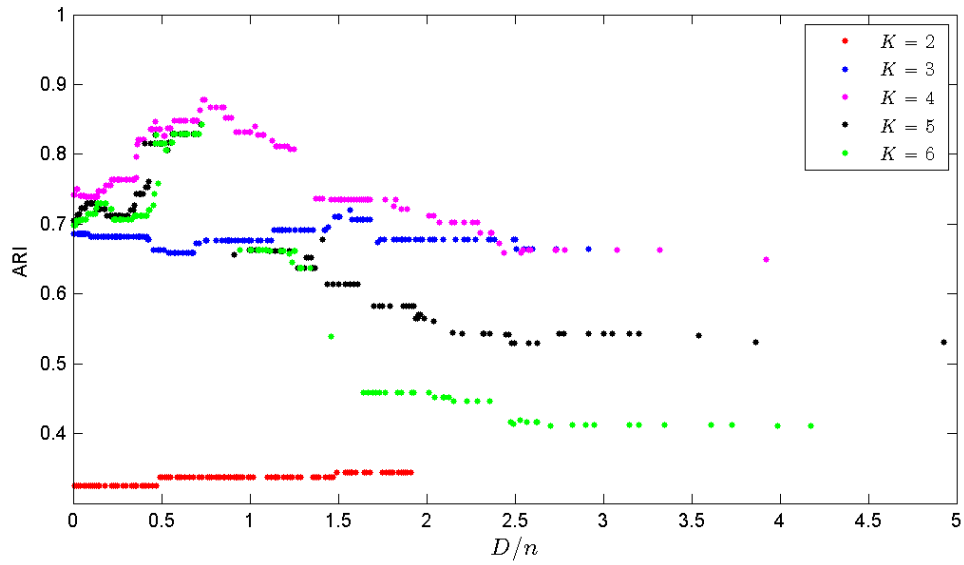

Figure 5.5: For one simulation, $p = 200$, ARI values for each model in Pan and Shen's model collection. Models with $D/n < 0.5$ have quite a low ARI. For this simulation, BIC selects a model with $D/n = 0.1$ and $K = 4$, so its ARI is only 0.74. The Lasso oracle is achieved for a more complex model with $D/n = 0.59$ and $K = 4$, so its ARI is 0.84.

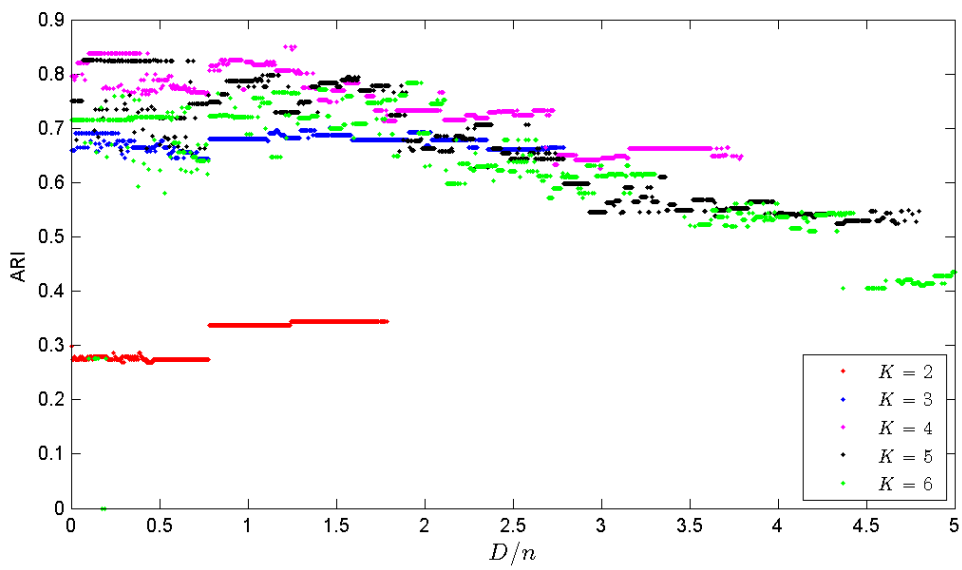

Figure 5.6: For one simulation, $p = 200$, ARI values for each model in our model collection. For this simulation, the slope estimator and the $\ln$-slope estimator select the oracle model with $D/n = 0.14$. Thus, they reach one of the highest ARI among the model collection.

5.3.2 Second simulated dataset

This second simulated dataset is taken from the example in Pan and Shen (2007). The dataset consists of $n = 200$ observations described by $p = 1000$ variables. The data are simulated according to a mixture of two Gaussian distributions $s = \sum_{k=1}^{2} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ where $\boldsymbol{\mu_1} = \mathbf{0_p}$ and $\boldsymbol{\mu_2} = (1.5, \ldots, 1.5, \mathbf{0_{950}})$. The vector $\mathbf{0_l}$ denotes the null vector of length l. The mixing proportions are $(\pi_1, \pi_2) = (0.85, 0.15)$. The relevant variables are the first fifty variables. Thus, the true density s belongs to the model with $K^* = 2$ components and $\boldsymbol{J_r^*} = \{1, \ldots, 50\}$. Note that the active variables and the relevant variables coincide. We perform 20 simulations of the dataset. For each simulation, we consider models with $K \in \{1, 2, 3\}$ clusters.

We compare our Lasso-MLE procedure with Pan and Shen's Lasso procedure. Maugis and Michel's procedure is unfeasible and thus ruled out for this second high-dimensional dataset. The results are summarized in Table 5.2. An example of a slope graph for the Lasso-MLE procedure is provided at Figure 5.7.

procedure	estimator	{TR,FR,FA}		K	ARI		$\mathrm{KL}(s, \hat{s})$	
Lasso	oracle	$\{50, 215, 0\}$	$(\{0, 79, 0\})$	$\{0, 16, 4\}$	0.90	(0.03)	2.53	(0.30)
	BIC	$\{49, 14, 0\}$	$(\{1, 3, 0\})$	$\{0, 18, 2\}$	0.86	(0.02)	3.59	(0.21)
Lasso-MLE	oracle	$\{50, 0, 1\}$	$(\{0, 0, 0\})$	$\{0, 20, 0\}$	0.95	(0.02)	0.31	(0.05)
	AIC	$\{50, 15, 68\}$	$(\{0, 4, 7\})$	$\{0, 14, 6\}$	0.90	(0.04)	1.44	(0.03)
	BIC	$\{50, 4, 22\}$	$(\{0, 2, 2\})$	$\{0, 20, 0\}$	0.92	(0.02)	0.74	(0.03)
	slope	$\{50, 1, 4\}$	$(\{0, 1, 2\})$	$\{0, 20, 0\}$	0.94	(0.02)	0.38	(0.06)
	ln-slope	$\{49, 0, 1\}$	$(\{1, 0, 1\})$	$\{0, 20, 0\}$	0.95	(0.02)	0.35	(0.12)

Table 5.2: Mean number {TR,FR,FA} of true relevant, false relevant and false active variables, number of times $\{\nu_1, \nu_2, \nu_3\}$ a clustering with respectively $K = 1$, $K = 2$ and $K = 3$ components is selected, mean ARI and mean Kullback-Leibler divergence over the 20 simulations. The standard deviations are put into brackets.

Variable selection and clustering

Table 5.2 shows that the Lasso oracle model, and to a lesser extend the model selected by BIC, contain many false relevant variables and may overestimate the number of mixture components. This confirms that Pan and Shen's Lasso procedure is not suited to recover the true model and the true relevant variables. Moreover, BIC data clustering is disappointing.

In contrast, the Lasso-MLE oracle model always coincides with the true model and leads to a very good data clustering. As expected for this high-dimensional dataset, the $\ln$-slope estimator is slightly closer to the Lasso-MLE oracle than the slope estimator. The penalty with a logarithm term is stronger than the penalty with no logarithm term (see Figure 5.7), so the $\ln$-slope estimator leads

to sparser models than the slope estimator. Note that the $\ln$-slope estimator does not always select all the 50 relevant variables, yet it leads to an excellent data clustering. Both the $\ln$-slope estimator and the slope estimator achieve better performance than BIC and AIC.

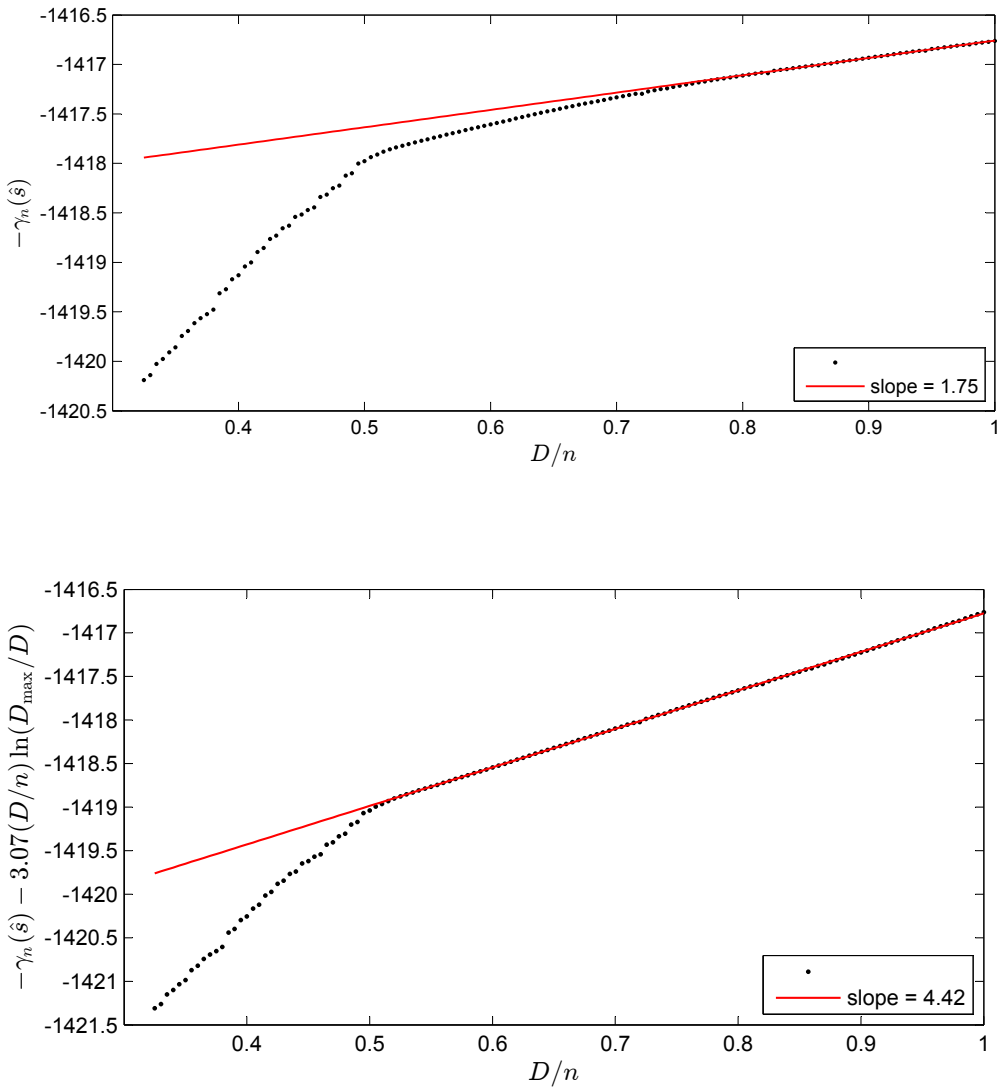

Figure 5.7: For one simulation, slope graphs obtained for the slope estimator (at the top) and the $\ln$-slope estimator (at the bottom). The estimations of the slope coefficients are calculated by restricting to the models with dimension $D \leq n$. The associated penalties are $\mathrm{pen}(D) = 2 \times 1.75 D/n$ and $\mathrm{pen}_{\ln}(D) = 2(4.42 D/n + 3.07(D/n)\ln(D_{\max}/D))$. The penalty $\mathrm{pen}_{\ln}$ is stronger and leads to a sparser model.

Density estimation

Table 5.2 shows that the risk of the Lasso oracle is much higher than the risk of the Lasso-MLE oracle. This proves that the density estimation obtained by the Lasso procedure is not reliable. This may be explained by three main reasons: a poor variable selection performance, a poor parameter estimation due to ℓ_1-regularization shrinkage and an excessive number of empirical means to be estimated due to empirical centering.

5.3.3 Some remarks on our procedure

5.3.3.1 Empirical centering versus theoretical centering for the model collection construction

In our Lasso-MLE procedure, the datasets are preliminary empirically centered to run the Lasso algorithm used to construct a collection of sets of relevant variables (see step 1 in Section 4.5.4). In Section 4.4.1, we justified this empirical centering as a practical surrogate for the unfeasible theoretical centering we would like to do to detect the relevant variables by ℓ_1-penalization. Yet, for simulated datasets, the true density is known and theoretical centering of the data is feasible. Thus, we can compare the collection of sets of relevant variables obtained by performing either empirical centering or theoretical centering of the data.

Figure 5.8 provides such a comparison for one simulation of the dataset studied in Section 5.3.1. To make easier the comparison, the low-dimensional case $p = 30$ is considered. For a fixed number of clusters, the Lasso algorithm generates a collection of sets of relevant variables by varying the Lasso regularization parameter λ in a data-driven grid (see Section 4.B.1). Figure 5.8 represents the model collection obtained for the true number $K = 4$ of clusters when either empirical centering or theoretical centering is preliminary performed. On the one hand, the two model collections share some similarities: for both model collections, the number of relevant variables in the models decreases as λ increases, the models are not necessarily nested and the true model is not in the model collection. On the other hand, the two model collections present dissimilarities: the sets of relevant variables are different, they are not obtained for the same regularization parameter values, and the variables do not enter in the same order as λ decreases. Thus, we can conclude that empirical centering and theoretical centering do not lead to the same model collection.

Now, since the final model selected belongs to the model collection, we can wonder whether these dissimilarities affect the choice of the model selected by the Lasso-MLE procedure. To answer this question, we performed theoretical centering for a few simulations for the datasets studied in Section 5.3.1 and Section 5.3.2. We compared the models selected with those in Table 5.1 and Table 5.2: the models selected by the slope estimator and the $\ln$-slope estimator, as well as the oracle model, were globally unchanged. This suggests that the model collection generated from empirical cente-

ring is close enough to the model collection generated from theoretical centering not to alter the final model selection, which is quite reassuring.

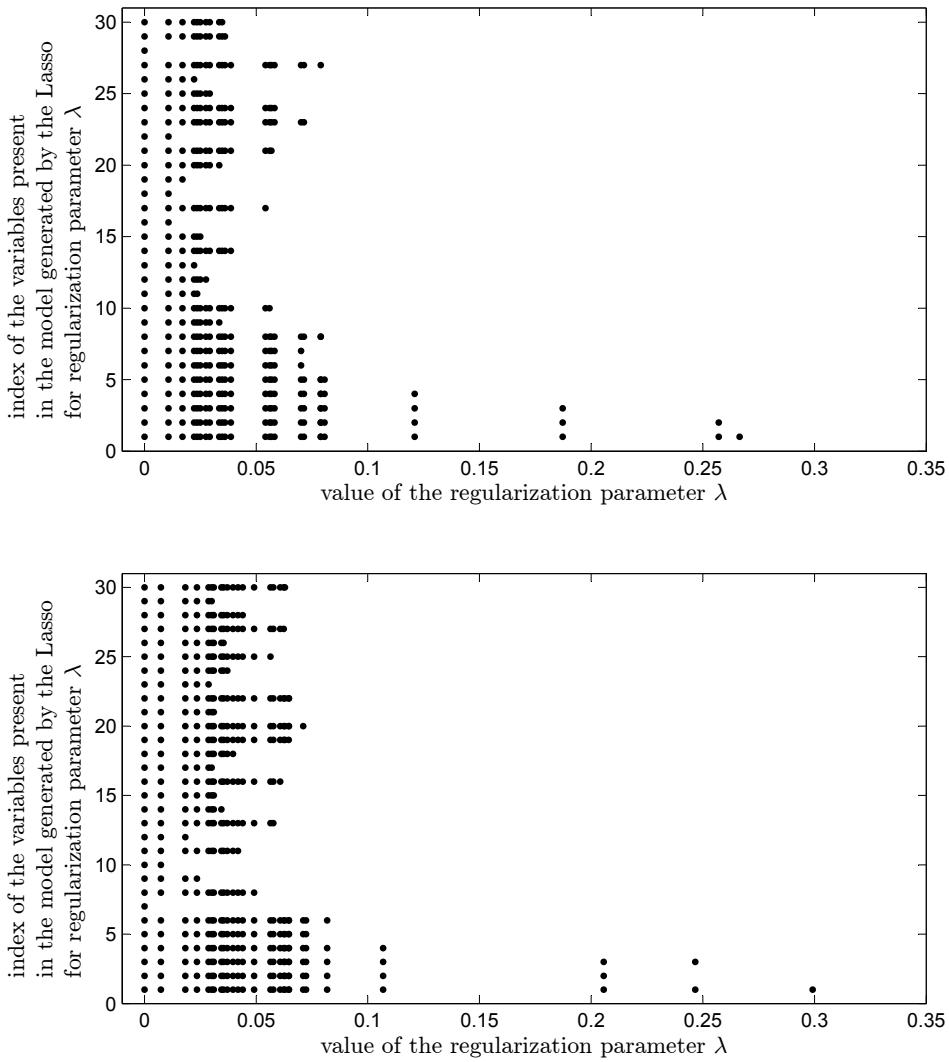

Figure 5.8: For one simulation of the dataset studied in Section 5.3.1, $p = 30$, collection of the sets of relevant variables generated by the Lassos by varying the regularization parameter λ, when either preliminary theoretical centering (at the top) or empirical centering (at the bottom) of the dataset is performed.

5.3.3.2 Slope estimator versus $\ln$-slope estimator

In Section 5.3.1 and Section 5.3.2, we noted that the $\ln$-slope estimator is to be preferred to the slope estimator as the dimension of the dataset increases since it becomes closer to the oracle whereas the

slope estimator tends to under-penalize as p increases. This comparison was possible because we worked with simulated data and we had access to the oracle. Yet, for real data, this is not possible. So, we can wonder how to choose between the two estimators in such a concrete case. We do not think it is possible or even desirable to fix a deterministic rule to decide which of the two estimators is the best one according to the number p of variables. It may depend on the number n of observations, on the number of true relevant variables, on the value of the mean coefficients, or on some other particularities of the data at hand. Yet, we can give some general indications to choose among the two estimators.

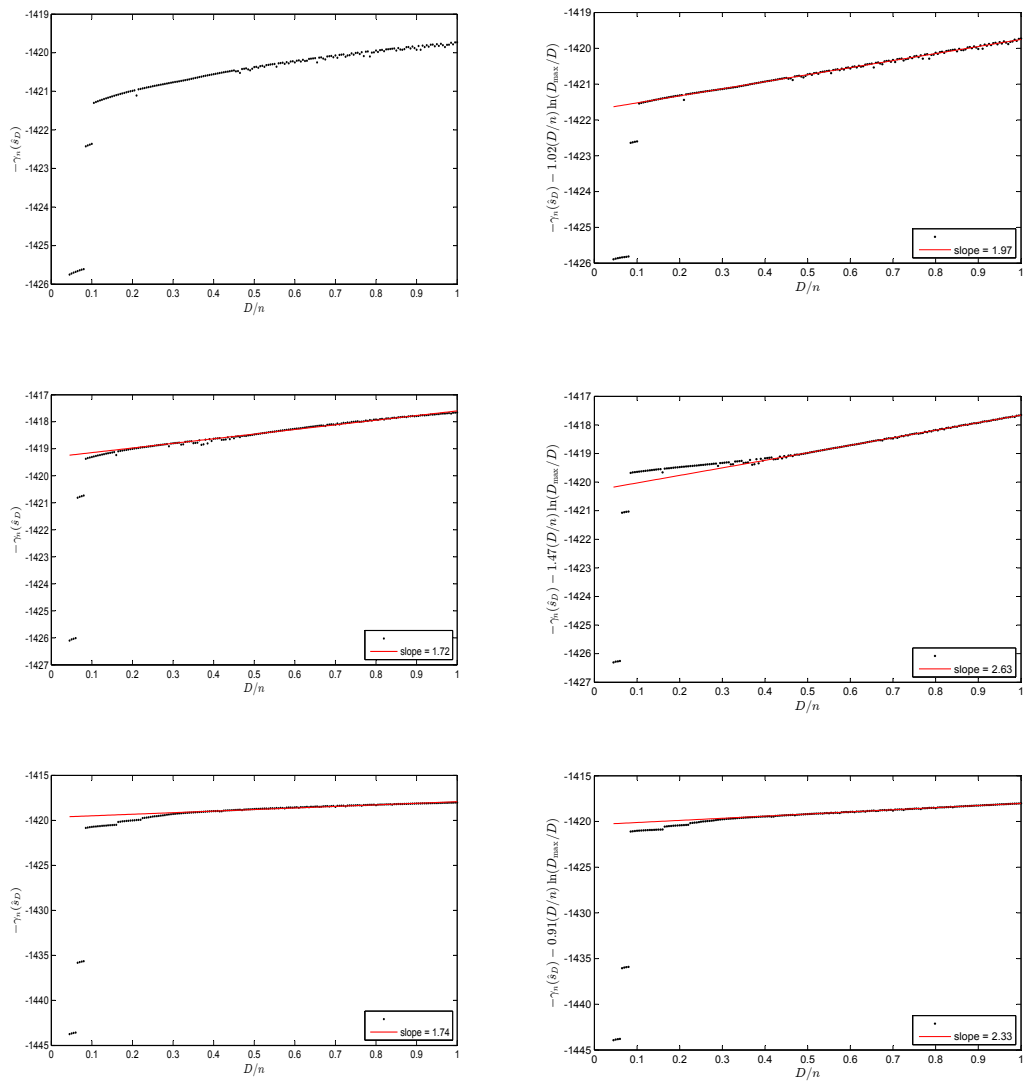

Figure 5.9: Different typical slope graphs for the slope estimator (on the left) and the $\ln$-slope estimator (on the right). At the top, the $\ln$-slope estimator is to be chosen. In the middle, the slope estimator is to be chosen. At the bottom, both estimators can be equally chosen.

Given a dataset, we advocate to try to compute both estimators. Then, several situations can occur.

If one of the two regressions fails (no stable estimation of the slope coefficient(s) can be obtained), this suggests that the associated estimator is not the good one. For instance, in Section 5.3.1, performing a double regression failed for $p = 30$, which suggested that the $\ln$-slope estimator was not the good one, whereas performing a simple regression failed 4 times over the 20 simulations for $p = 1000$, which indicated that the slope estimator was to rule out for those simulations. Figure 5.9 illustrates a situation where the simple regression fails (top left of Figure 5.9) while the double regression works well (top right of Figure 5.9).

If both regressions are feasible, then we advise the user to look at the slope graphs. It is often sufficient to guess which estimator is to be preferred. For instance, at Figure 5.7, the graph obtained for the simple regression is not so linear, while the graph obtained by subtracting a logarithm term calculated by the double regression leads to a beautiful linear graph. This suggests that the $\ln$-slope estimator is to be favored. On the contrary, Figure 5.9 presents a situation where the red slope calculated from the double regression goes under the black graph on the extreme left part of the plot (middle right of Figure 5.9). When such a graph is observed, the associated $\ln$-slope estimator is often too sparse. In this case, the slope estimator is to be preferred (middle left of Figure 5.9). Finally, if both graphs look like each other (Figure 5.9, at the bottom), then the slope estimator and the $\ln$-slope estimator are expected to lead to similar results.

5.4 Functional data clustering using wavelets

A major application of our Lasso-MLE procedure may be functional data clustering using wavelets. Here, we present some simulated data in this context to evaluate our procedure performance as regards both clustering and sparse curve reconstruction. We consider two approaches.

First, we work with datasets such that each data is assumed to be the wavelet coefficient decomposition of a noised function in some given wavelet basis. We apply our Lasso-MLE procedure to obtain both a clustering of the coefficient decompositions as well as an estimation of the coefficient decompositions for each cluster. Then, the coefficient clustering provides a clustering of the underlying functions associated to the coefficient decompositions. Moreover, by using the estimation of the coefficient decompositions for each cluster and by performing an inverse wavelet transform, we derive a sparse curve estimation for each cluster. The inverse wavelet transform is done using the function `waverec` available in the Wavelet Toolbox of MATLAB.

Secondly, we work with datasets such that each data is the noisy observation of a function. In this case, we introduce a wavelet basis and decompose each data into this basis. This preliminary step is done using the function `mdwtdec` available in the Wavelet Toolbox of MATLAB. This operation results in a new dataset such that each data is a noisy wavelet coefficient decomposition, such as in the first case. Then, we can apply our Lasso-MLE procedure on this dataset and end as in the first case.

5.4.1 Examples of curve reconstruction from wavelet coefficient clustering

We consider two simulated datasets in the spirit of the two datasets studied in Section 5.3. We assume that each data is the wavelet coefficient decomposition of a noised function in some wavelet basis. We aim at identifying the different clusters and at recovering the different denoised functions by providing a curve estimation for each cluster.

5.4.1.1 First simulated dataset

This first dataset is similar to the dataset studied by Pan and Shen (2007), with the difference that we add some active irrelevant variables in order to get less flat curve shapes. The dataset consists of $n = 200$ data described by $p = 1086$ variables. The data are simulated according to a mixture of two Gaussian distributions $\sum_{k=1}^{2} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ with mixing proportions $(\pi_1, \pi_2) = (0.85, 0.15)$ and mean vectors $\boldsymbol{\mu_1} = (\mathbf{0_{25}}, \mathbf{1.5_{25}}, \mathbf{0_{p-50}})$ and $\boldsymbol{\mu_2} = (\mathbf{1.5_{50}}, \mathbf{0_{p-50}})$, where $\boldsymbol{a_l}$ denotes the vector of length l whose coordinates equal a. Thus, the first 25 variables are relevant for the clustering while the variables 26 to 50 are active irrelevant.

We assume that each data is the wavelet coefficient decomposition of the noisy observation of a function f in the symmlet-4 basis at level 10. The scaling and the wavelet functions defining this basis are represented at Figure 5.10. By performing an inverse wavelet transform of the wavelet coefficient dataset in this basis, we get a functional dataset which is simulated according to a mixture of two Gaussian distributions $\sum_{k=1}^{2} \pi_k \, \Phi(\cdot \mid \boldsymbol{f_k}, \boldsymbol{I})$ with mixing proportions $(\pi_1, \pi_2) = (0.85, 0.15)$ and means $\boldsymbol{f_1} \in \mathbb{R}^{1024}$ and $\boldsymbol{f_2} \in \mathbb{R}^{1024}$. The means $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ are the discretization of two functions f_1 and f_2 on a grid containing 1024 points. Since $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are known, we have access to the discretized functions $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ by performing an inverse wavelet transform of $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ in the symmlet-4 basis at level 10. The functions f_1 and f_2 obtained by this process are displayed at Figure 5.11.

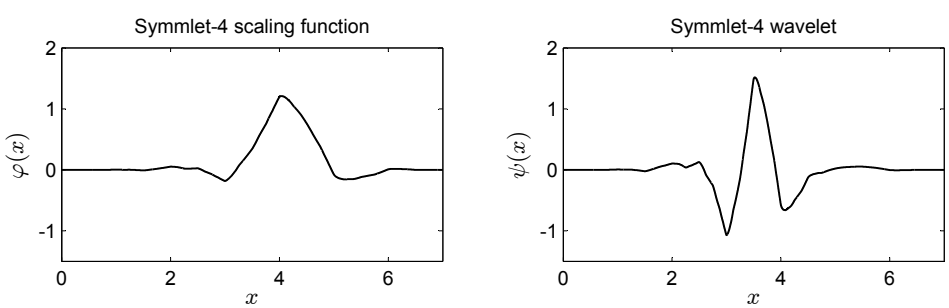

Figure 5.10: Scaling (father) function φ and wavelet (mother) function ψ for the symmlet-4 basis.

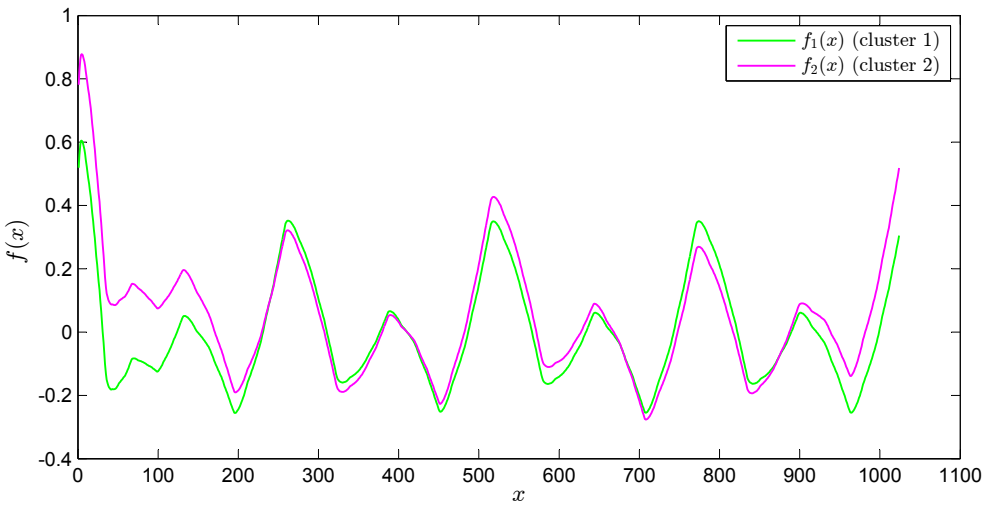

Figure 5.11: Plot of the functions f_1 for cluster 1 and f_2 for cluster 2.

The aim is to identify the two clusters and to get a good estimation of f_1 and f_2. This can be achieved by providing a clustering of the wavelet coefficient dataset, by estimating the mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, and by performing an inverse wavelet transform of $\hat{\boldsymbol{\mu}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{2}}$ to get an estimation of $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$.

For one simulation of this dataset, we compare the clustering and the curve estimation obtained by our Lasso-MLE procedure and by Pan and Shen's Lasso procedure. We consider models with $K \in \{1, 2, 3\}$ clusters. Both procedures choose a mixture with two components. Figure 5.12 and Figure 5.13 show the reconstruction of the functions f_1 and f_2 obtained by Pan and Shen's Lasso estimator (at the bottom), our Lasso-MLE procedure with the slope estimator (in the middle) and the $\ln$-slope estimator (at the top).

The $\ln$-slope estimator performs the best curve estimations. The estimations obtained by the slope estimator are quite good but we note a few extra peaks due to non-null estimations $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$ of a few truly null mean coefficients μ_{1j} and μ_{2j}. This is due to the fact that the penalty proportional to the dimension lightly under-penalizes. Figure 5.15 shows the slope graphs obtained for both estimators.

As regards Pan and Shen's Lasso estimation, it is very noisy. This is not surprising. Indeed, Pan and Shen's procedure runs on the empirically centered dataset, so one must add the $p = 1086$ empirical means to Pan and Shen's Lasso estimations to get an estimation of the mean coefficients μ_{kj} of the original (non-empirically centered) dataset. Consequently, the resulting estimation is not sparse. Pan and Shen's procedure is not adapted to curve reconstruction and this would be the case for any procedure running on empirically centered datasets. Figure 5.14 highlights another problem encountered by Pan and Shen's procedure, which is not due to empirical centering but rather to the shrinkage induced by ℓ_1-regularization. This shrinkage worsens the curve estimations.

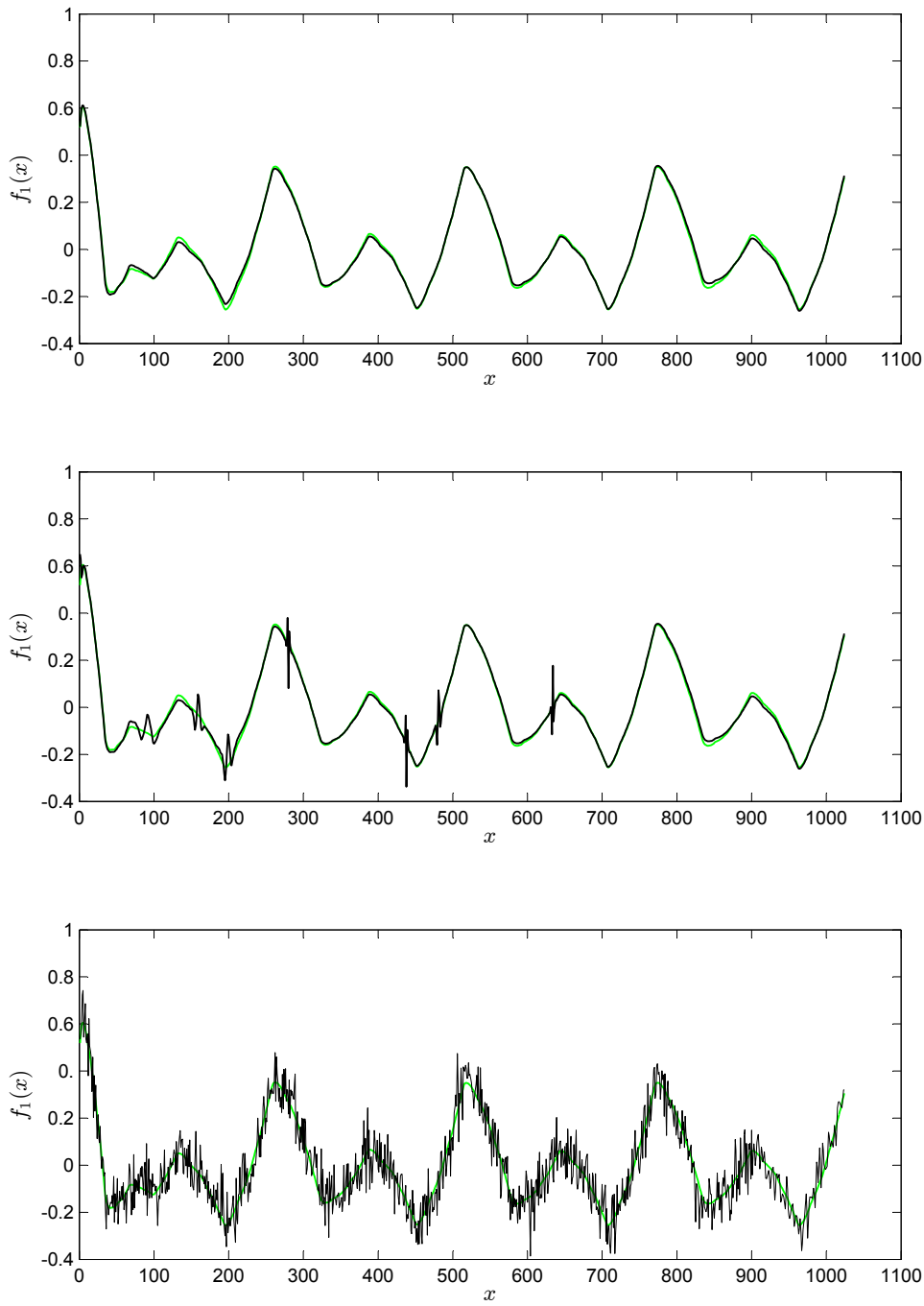

Figure 5.12: Estimation of f_1 by the $\ln$-slope estimator (at the top), the slope estimator (in the middle) and Pan and Shen's Lasso estimator (at the bottom). The true function is plotted in green while the estimations are black.

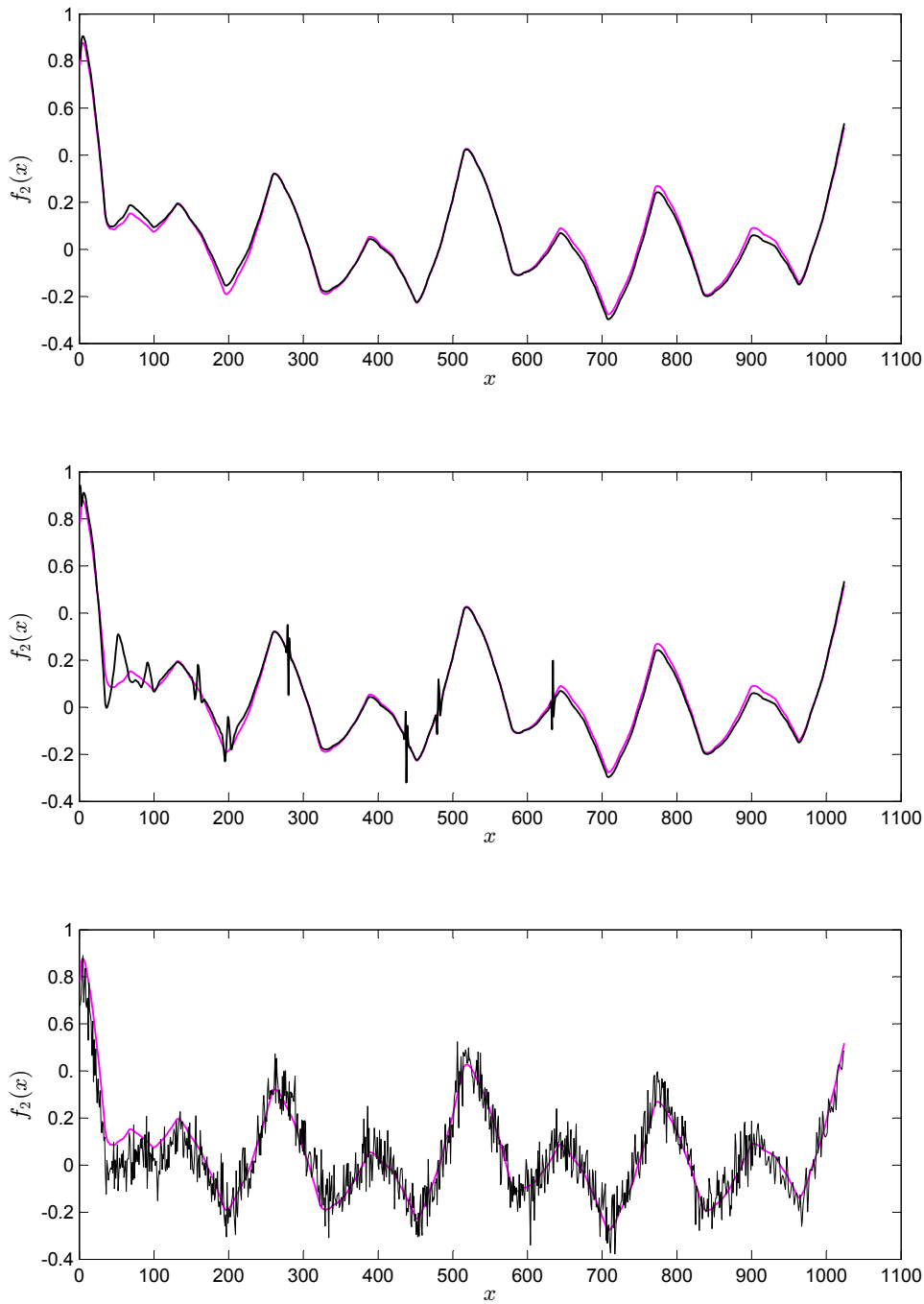

Figure 5.13: Estimation of f_2 by the $\ln$-slope estimator (at the top), the slope estimator (in the middle) and Pan and Shen's Lasso estimator (at the bottom). The true function is plotted in pink while the estimations are black. At the bottom, note that the estimated curve is below the true curve between $x = 50$ and $x = 150$. This is due to the Lasso shrinkage (see Figure 5.14).

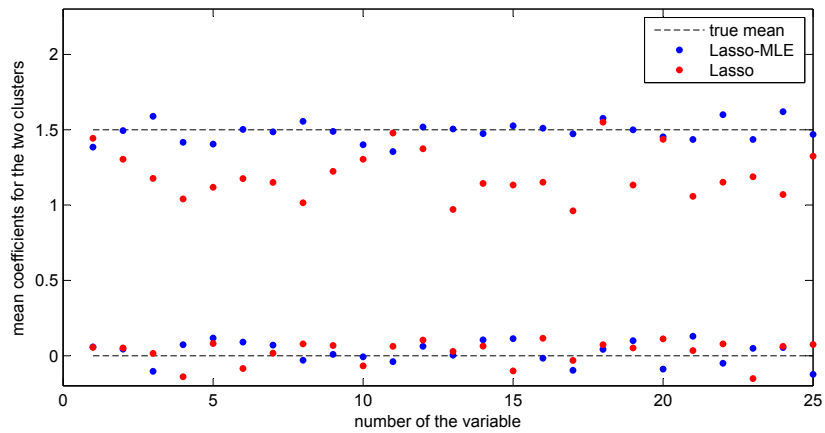

Figure 5.14: Estimations of the mean coefficients μ_{kj} for the 25 relevant variables by Pan and Shen's Lasso estimator and our $\ln$-slope estimator. The true mean coefficients all equal 0 for the first cluster and 1.5 for the second cluster. The Lasso under-estimates the mean coefficients of the second cluster.

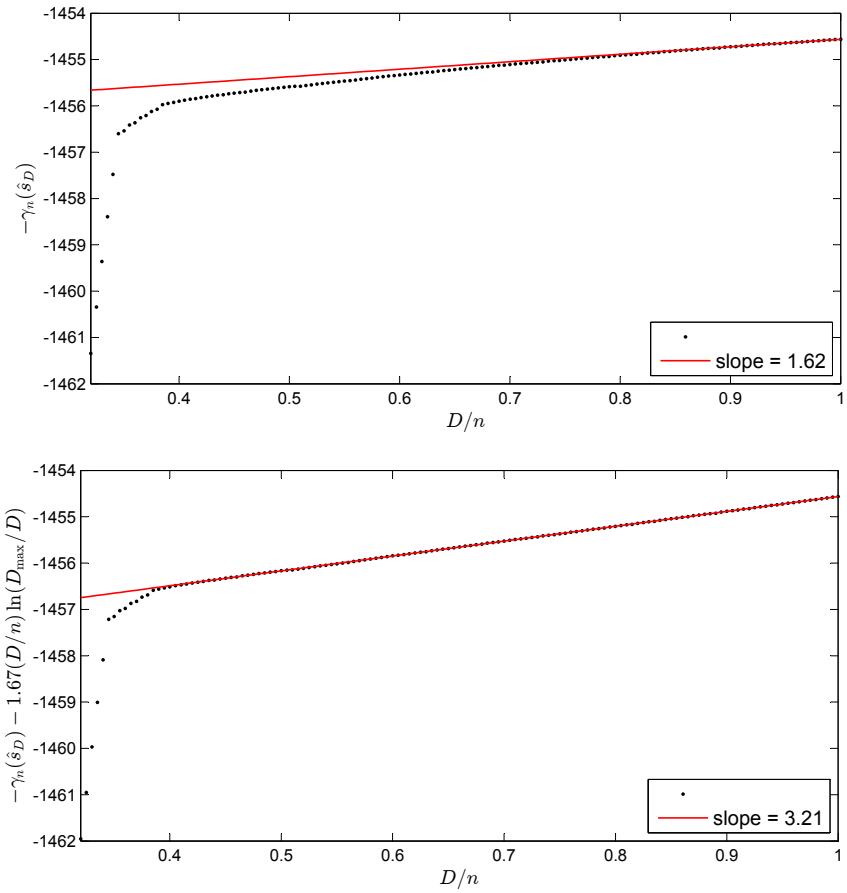

Figure 5.15: Slope graphs for the slope estimator (at the top) and the $\ln$-slope estimator (at the bottom).

5.4.1.2 Second simulated dataset

This second dataset is similar to the dataset studied in Section 5.3.1, with the difference that we add some active irrelevant variables in order to get less flat curve shapes. The dataset consists of $n = 200$ data described by $p = 1086$ variables. Denote by $\boldsymbol{a_l}$ the vector of length l whose coordinates equal a and put $\boldsymbol{\mu} = (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025)$. The data are simulated according to a mixture of four Gaussian distributions $\sum_{k=1}^{4} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ with mixing proportions $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.3, 0.2, 0.2, 0.3)$ and mean vectors constructed from positive, negative or alternate coefficients of $\boldsymbol{\mu}$:

$$\begin{aligned}
\boldsymbol{\mu_1} &= (\boldsymbol{\mu}, \boldsymbol{\mu}, \boldsymbol{0_{p-20}}),\\
\boldsymbol{\mu_2} &= (\boldsymbol{0_{10}}, \boldsymbol{\mu}, \boldsymbol{0_{p-20}}),\\
\boldsymbol{\mu_3} &= (-\boldsymbol{\mu}, \boldsymbol{\mu}, \boldsymbol{0_{p-20}}),\\
\boldsymbol{\mu_4} &= (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025, \boldsymbol{\mu}, \boldsymbol{0_{p-20}}).
\end{aligned}$$

We assume that each data is the wavelet coefficient decomposition of the noisy observation of a function f in the symmlet-4 basis at level 10. By performing an inverse wavelet transform of the wavelet coefficient dataset in this basis, we get a functional dataset which is simulated according to a mixture of four Gaussian distributions $\sum_{k=1}^{4} \pi_k \, \Phi(\cdot \mid \boldsymbol{f_k}, \boldsymbol{I})$ with mixing proportions $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.3, 0.2, 0.2, 0.3)$ and means $\boldsymbol{f_1}$, $\boldsymbol{f_2}$, $\boldsymbol{f_3}$ and $\boldsymbol{f_4} \in \mathbb{R}^{1024}$ that are the discretization of four functions f_1, f_2, f_3 and f_4 on a grid containing 1024 points. Since $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\boldsymbol{\mu_3}$ and $\boldsymbol{\mu_4}$ are known, we have access to the discretized functions $\boldsymbol{f_1}$, $\boldsymbol{f_2}$, $\boldsymbol{f_3}$ and $\boldsymbol{f_4}$ by performing an inverse wavelet transform of $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\boldsymbol{\mu_3}$ and $\boldsymbol{\mu_4}$ in the symmlet-4 basis at level 10. The functions f_1, f_2, f_3 and f_4 obtained by this process are displayed at the top of Figure 5.16.

The aim is to identify the four clusters and to get a good estimation of f_1, f_2, f_3 and f_4. This can be achieved by providing a clustering of the wavelet coefficient dataset, by estimating the mean vectors $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\boldsymbol{\mu_3}$ and $\boldsymbol{\mu_4}$ and by performing an inverse wavelet transform of $\boldsymbol{\hat{\mu}_1}$, $\boldsymbol{\hat{\mu}_2}$, $\boldsymbol{\hat{\mu}_3}$ and $\boldsymbol{\hat{\mu}_4}$ to get an estimation of $\boldsymbol{f_1}$, $\boldsymbol{f_2}$, $\boldsymbol{f_3}$ and $\boldsymbol{f_4}$.

We consider models with $K \in \{2, \ldots, 6\}$ clusters. Figure 5.16 presents the estimations obtained by the $\ln$-slope estimator of our Lasso-MLE procedure for one simulation of the dataset. The four clusters are detected and the curves are well estimated. Let us point out that the slope estimator fails to be computed, which suggests that a logarithm term in the penalty is necessary. Figure 5.17 shows the slope graph obtained for the $\ln$-slope estimator.

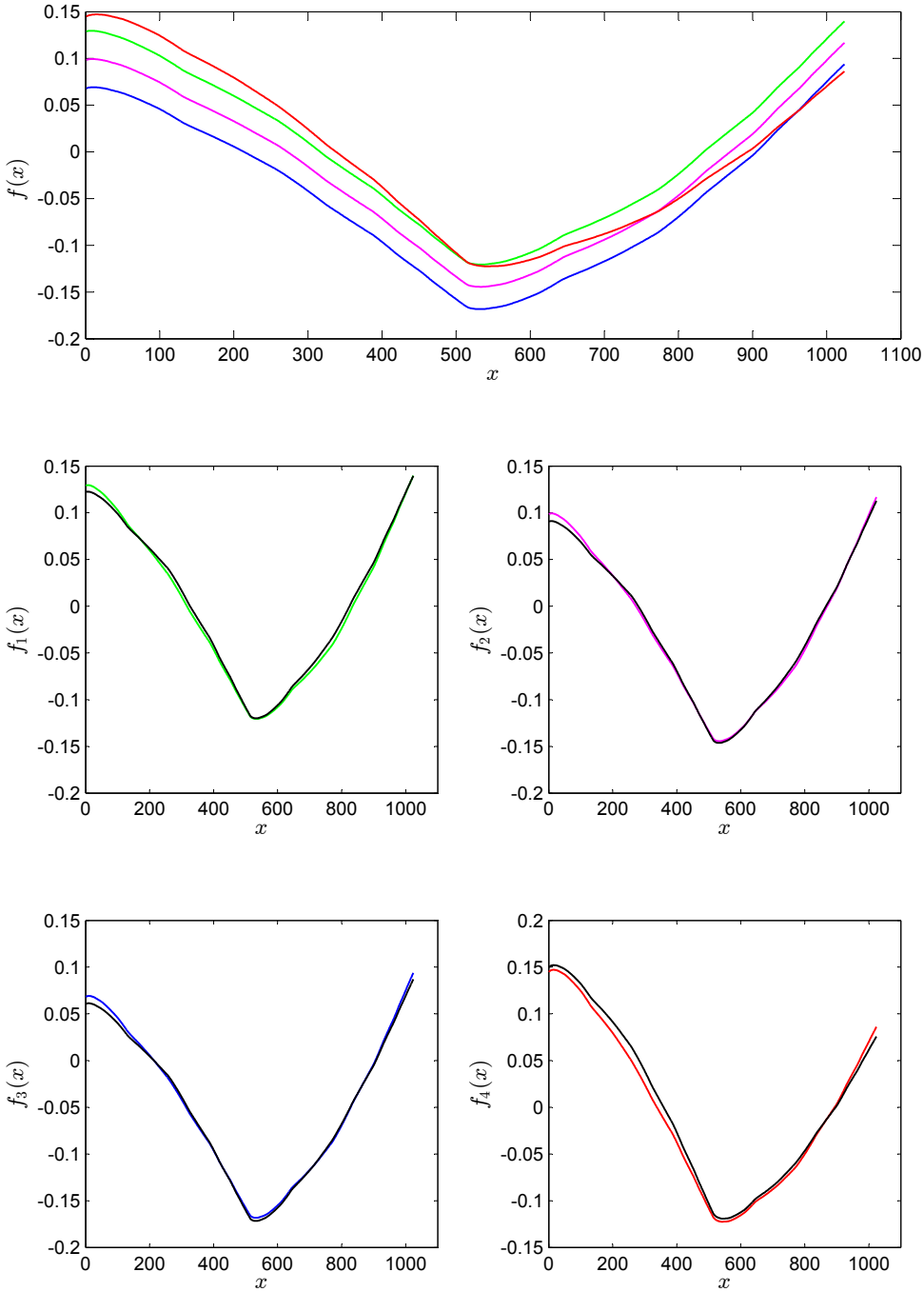

Figure 5.16: At the top: true functions f_1, f_2, f_3 and f_4 corresponding to cluster 1 (green), cluster 2 (pink), cluster 3 (blue) and cluster 4 (red) respectively. Below, estimation of each function by the $\ln$-slope estimator. The true functions are colored while the estimations are black.

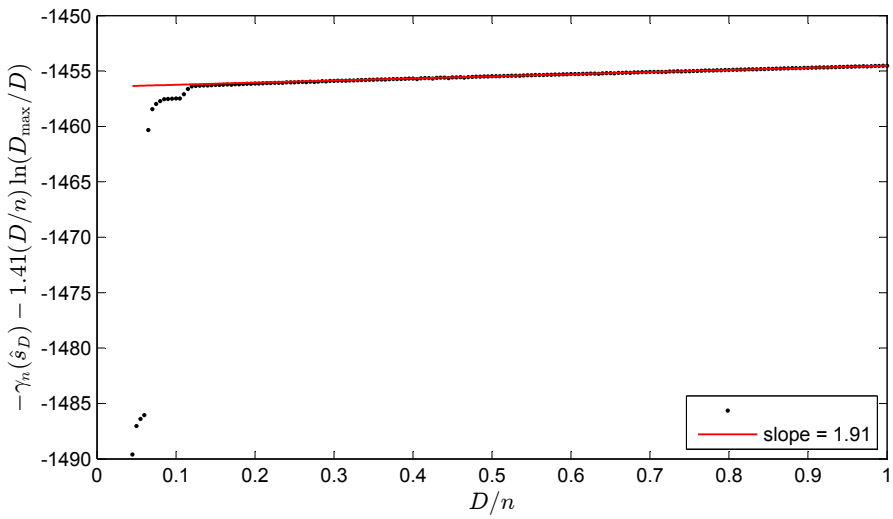

Figure 5.17: Slope graph obtained for the $\ln$-slope estimator.

5.4.2 Examples of functional data clustering

Now, we present two simulated datasets where the data are no longer wavelet coefficient decompositions of noised functions but rather discretized noised functions measured on a fine time grid. In this case, we do not apply our Lasso-MLE procedure directly on the dataset. We preliminary perform a discrete wavelet transform of each functional data in a common wavelet basis. This leads to a new dataset where the data are the wavelet coefficient decompositions of the noised functions of the original dataset.

5.4.2.1 First simulated dataset

This first dataset is a clustering problem proposed by Misiti et al. (2007a). The data $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ are simulated according to a mixture of five Gaussian distributions $\sum_{k=1}^{5} \pi_k \, \Phi(\cdot \mid \boldsymbol{f_k}, \boldsymbol{I})$ with equal mixing proportions $\pi_k = 1/5$ and mean functions f_k defined on $[0, 1]$ by

$$
\begin{aligned}
f_1(t) &= \sin(4\pi t) - (\text{sign}(t - 0.3) + \text{sign}(0.72 - t))\,/4, \\
f_2(t) &= \sin(4\pi t), \\
f_3(t) &= \sin(2\pi t), \\
f_4(t) &= 4\left(t\mathbb{1}_{\{t \leq 1/4\}} + (1/2 - t)\,\mathbb{1}_{\{1/4 < t \leq 3/4\}} + (t-1)\mathbb{1}_{\{t > 3/4\}}\right), \\
f_5(t) &= 2\left(t\mathbb{1}_{\{t \leq 1/2\}} + (1 - t)\,\mathbb{1}_{\{t > 1/2\}}\right).
\end{aligned}
$$

Figure 5.18 represents these functions. The function f_1 is the Heavisine function, which is derived from the sinusoidal function f_2 by adding two breaking points. The functions f_2 and f_3 are sinusoidal functions with different phases. The function f_4 is a piecewise linear version of the sinusoidal function f_3. The function f_5 is quite distinguishable from the other functions; in particular, it takes only positive values. The dataset consists of $n = 400$ observations described by $p = 2^{10} = 1024$ variables. Each observation $\boldsymbol{y_i}$ in cluster $k \in \{1, \ldots, 5\}$ can be written

$$y_{ij} = f_k(t_j) + \xi_{ij}, \qquad j = 1, \ldots, p,$$

where $t_j = (j-1)/p$ and ξ_{ij} are i.i.d. $\sim \mathcal{N}(0, 1)$.

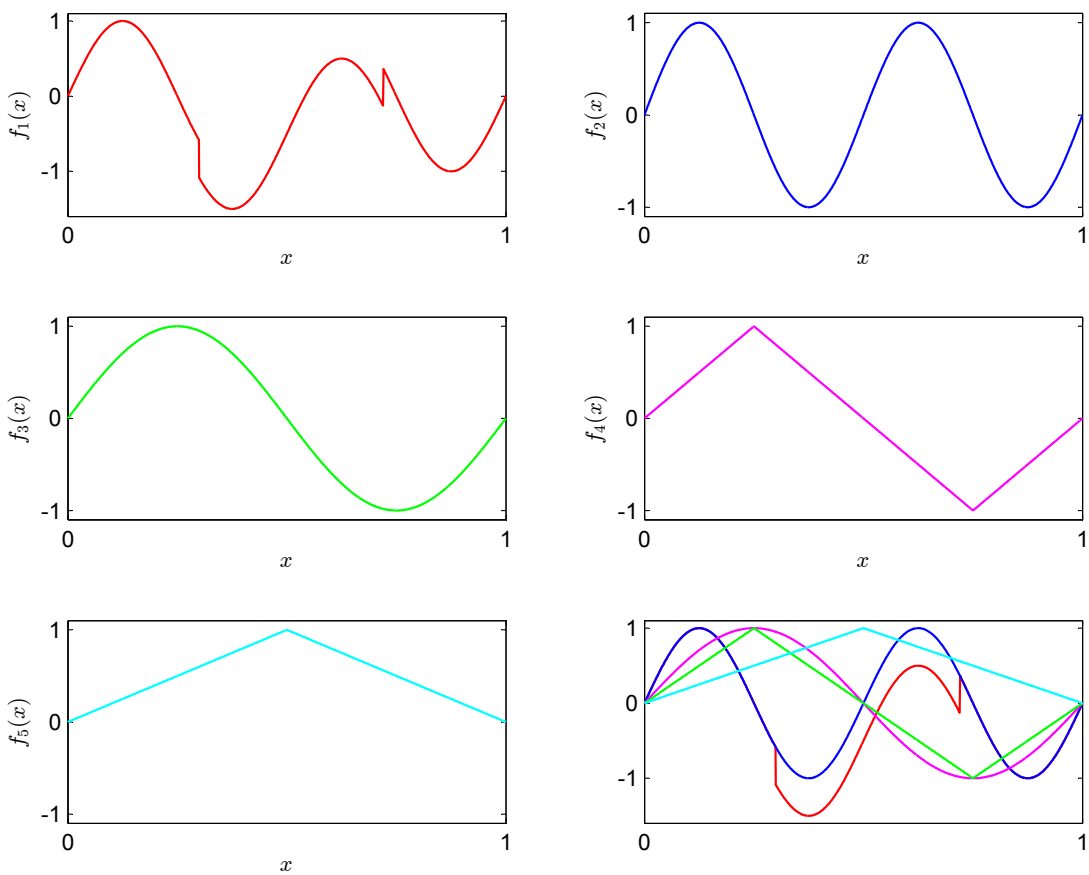

Figure 5.18: Plot of each function f_1, f_2, f_3, f_4 and f_5. On the right, at the bottom, comparison between the five functions.

First, we decompose each data $\boldsymbol{y_i}$ into the Haar basis at level 10 by performing a discrete wavelet transform. The scaling and the wavelet functions of this basis are represented at Figure 5.19. This leads to a new dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ which consists of $n = 400$ data described by $p = 1024$ variables. Each data $\boldsymbol{Y_i} \in \mathbb{R}^p$ is the wavelet coefficient decomposition of the noisy observation $\boldsymbol{y_i}$. Since the discrete wavelet transform is orthogonal, the dataset $\boldsymbol{Y}$ has the same statistical properties as the dataset $\boldsymbol{y}$. It is simulated according to a mixture of five Gaussian distributions $\sum_{k=1}^{5} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ with equal mixing proportions $\pi_k = 1/5$. The mean vectors $\boldsymbol{\mu_k} \in \mathbb{R}^p$ are represented at Figure 5.20. Since $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are quite close and $\boldsymbol{\mu_3}$ and $\boldsymbol{\mu_4}$ are quite close, we may think that one's procedure can mix up those means and detect only three clusters instead of the five true clusters. Thus, we let vary the number of clusters $K \in \{2, \ldots, 6\}$.

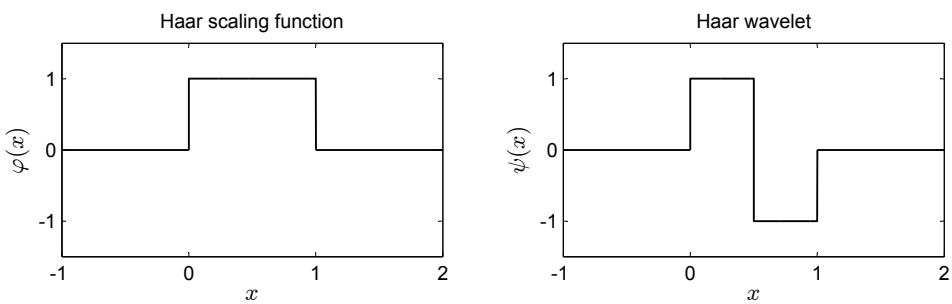

Figure 5.19: Scaling (father) function φ and wavelet (mother) function ψ for the Haar basis.

We run our Lasso-MLE procedure on the dataset $\boldsymbol{Y}$. The slope estimator fails to be computed while the $\ln$-slope estimator can be computed. Our procedure selects a model with four clusters, mixing up cluster 3 and cluster 4 whose mean vectors $\boldsymbol{\mu_3}$ and $\boldsymbol{\mu_4}$ are very similar (see Figure 5.20). Despite the similarities between the mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, our procedure distinguishes between cluster 1 and cluster 2. By performing an inverse wavelet transform of the estimated mean vectors $\hat{\boldsymbol{\mu}}_{\boldsymbol{1}}$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{2}}$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{5}}$ and mixed up mean vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{3/4}}$, we derive a curve estimation respectively for function f_1, f_2, f_5 and mixed up functions f_3 and f_4. These estimations are represented at Figure 5.21. They are globally accurate, although the second breaking point of f_1 is not detected and the peak of f_5 is not well marked. Note that the first breaking point of f_1 is detected and that the mixed up functions f_3 and f_4 are both quite well approximated by the their common estimation.

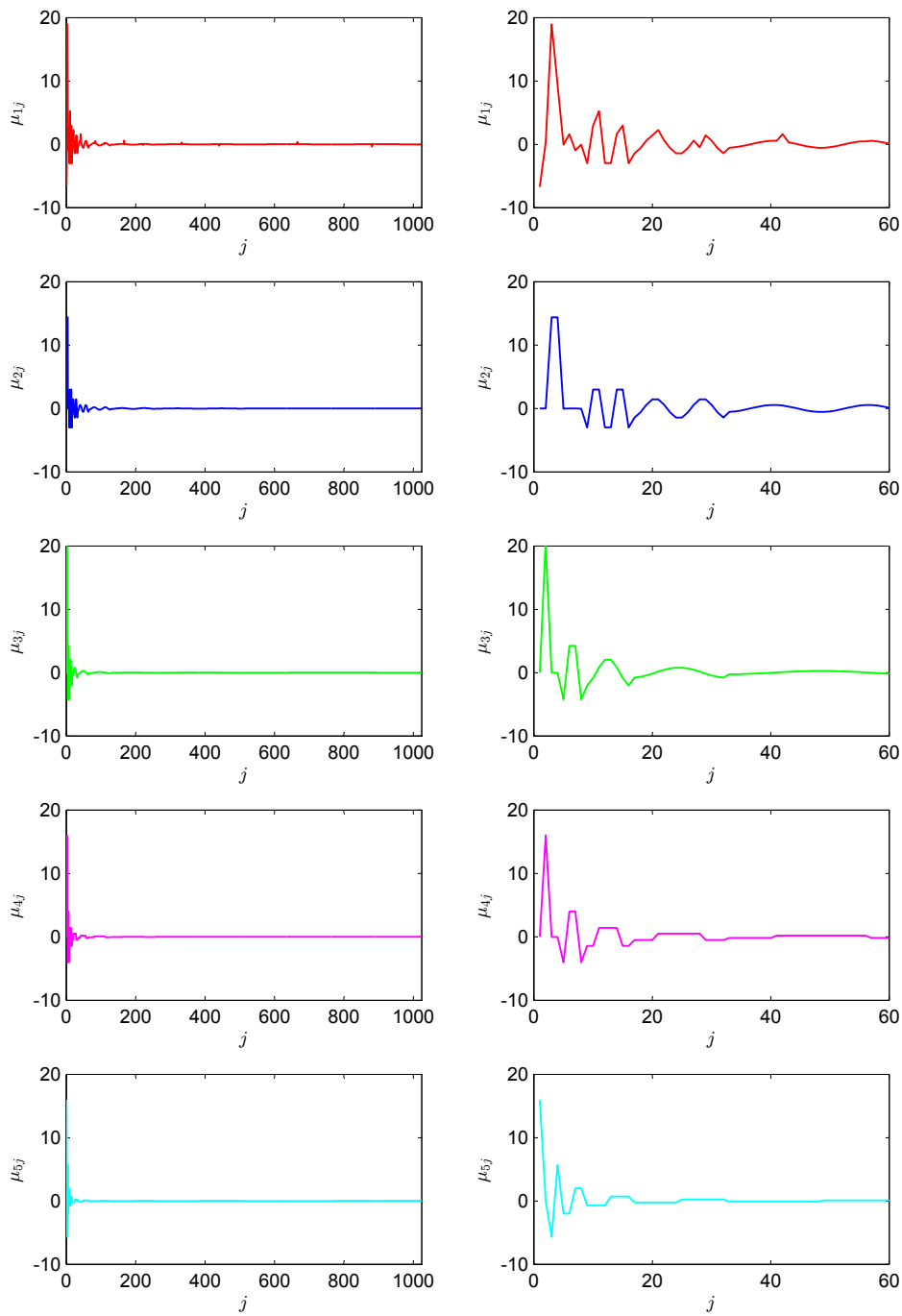

Figure 5.20: On the left, from top to bottom, plot of $j \in \{1, \ldots, p\} \mapsto \mu_{kj}$ for $k = 1, \ldots, 5$. Only around the first 40 to 100 variables seem to be active and relevant for the clustering. On the right, same plots restricted to $j \leq 60$ to make easier the comparison between the relevant part of the graphs.

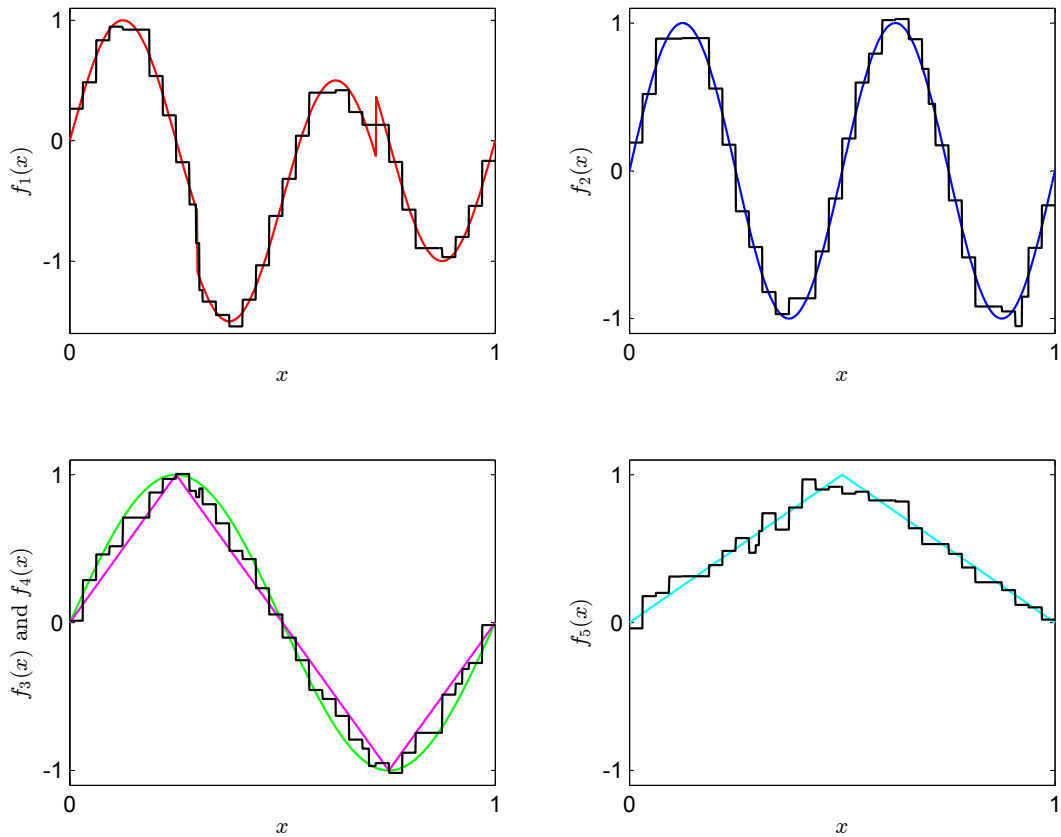

Figure 5.21: Curve estimation of f_1, f_2, mixed up f_3 and f_4, and f_5 by the $\ln$-slope estimator. The true functions are colored while the estimations are black.

5.4.2.2 Second simulated dataset

Finally, we propose a dataset chosen to evaluate our procedure capability to distinguish between two functions that differ only locally. The data are simulated according to a mixture of two Gaussian distributions $\sum_{k=1}^{2} \pi_k \, \Phi(\cdot \mid \boldsymbol{f_k}, \boldsymbol{I})$ with equal mixing proportions $\pi_k = 1/2$ and mean functions f_k defined on $\mathbb{R}$ by

$$f_1(t) = 0.23\, t \, \exp\left(-t/2\right),$$
$$f_2(t) = \frac{1}{2.5\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-2}{2.5}\right)^2\right) + 0.1\, \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(t-8\right)^2\right).$$

Figure 5.22 represents these functions. The function f_1 is proportional to the $\chi^2(4)$ probability distribution while f_2 is a mixture of two real Gaussian distributions. Their shapes are globally similar but

f_2 presents a slight distortion around $t = 7$. The dataset consists of $n = 200$ observations described by $p = 2^{10} = 1024$ variables. Each observation $\boldsymbol{y_i}$ in cluster $k \in \{1, 2\}$ can be written

$$y_{ij} = f_k(t_j) + \xi_{ij}, \qquad j = 1, \ldots, p,$$

where $t_j = 20(j-1)/p$ and ξ_{ij} are i.i.d. $\sim \mathcal{N}(0, 1)$.

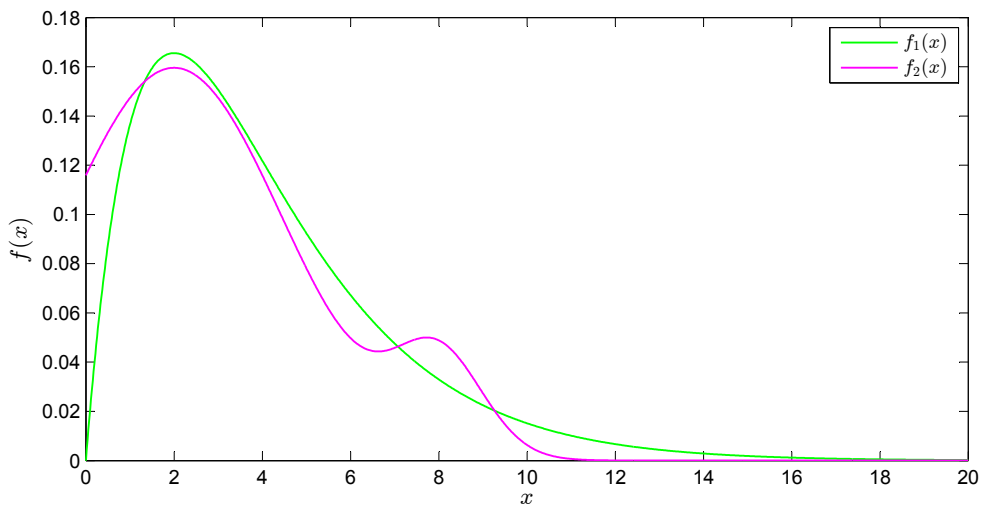

Figure 5.22: Plot of the functions f_1 and f_2.

First, we decompose each data $\boldsymbol{y_i}$ in the symmlet-4 basis at level 10 by performing a discrete wavelet transform. This leads to a new dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ which consists of $n = 200$ wavelet coefficient decompositions $\boldsymbol{Y_i}$ of the noisy observations $\boldsymbol{y_i}$. We model the dataset $\boldsymbol{Y}$ as a mixture of two Gaussian distributions $\sum_{k=1}^{2} \pi_k \, \Phi(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{I})$ with equal mixing proportions $\pi_k = 1/2$. We let vary the number of clusters $K \in \{1, 2, 3\}$.

We run our Lasso-MLE procedure on the dataset $\boldsymbol{Y}$. The slope estimator fails to be computed while the $\ln$-slope estimator can be computed. Figure 5.24 shows the slope graph obtained for the $\ln$-slope estimator. Our procedure detects the two clusters. By performing an inverse wavelet transform of the estimated mean vectors $\hat{\boldsymbol{\mu}}_{\boldsymbol{1}}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{2}}$, we derive a curve estimation for the functions f_1 and f_2 respectively. These estimations are represented at Figure 5.23.

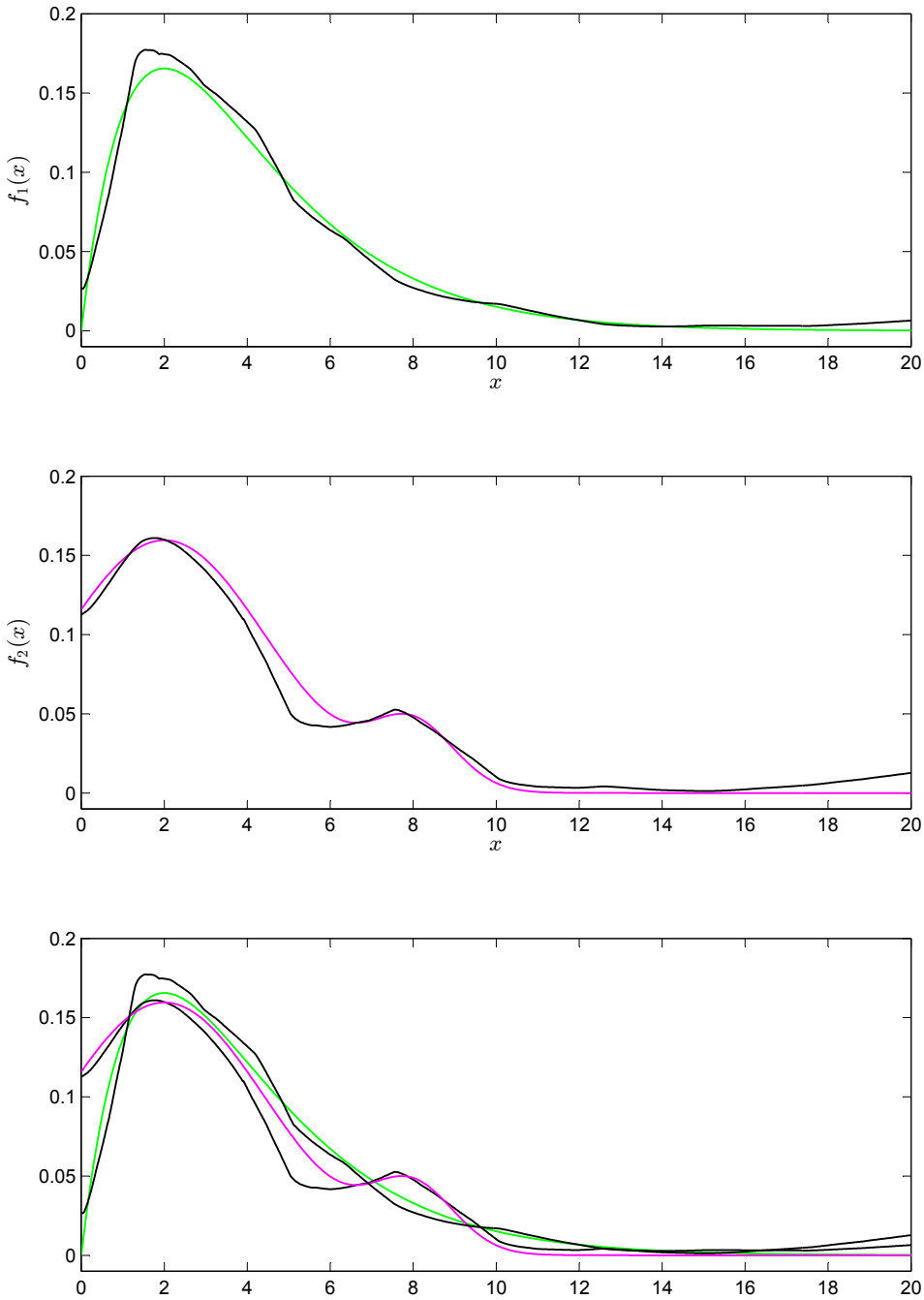

Figure 5.23: Curve estimation of f_1 and f_2 by the $\ln$-slope estimator. The true functions are colored while the estimations are black.

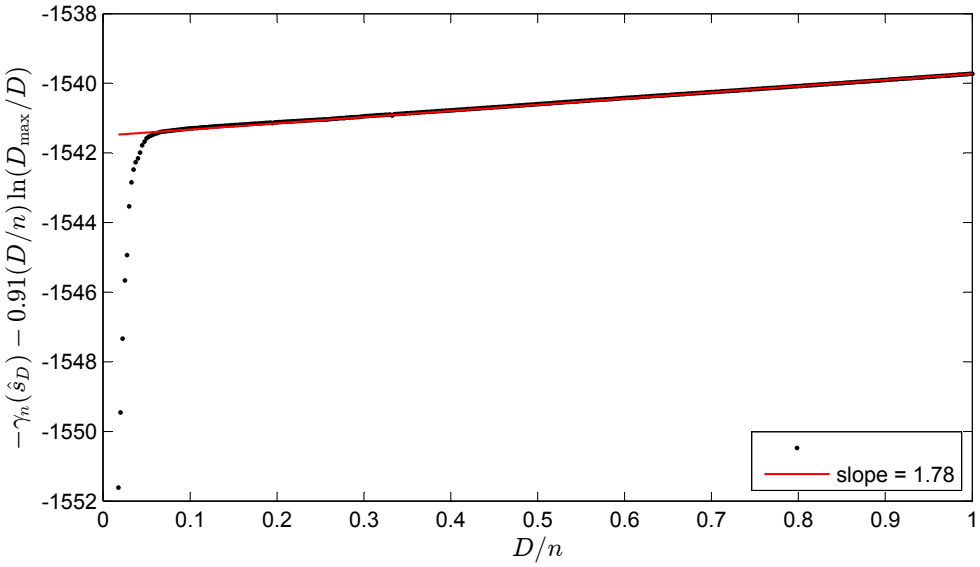

Figure 5.24: Slope graph obtained for the $\ln$-slope estimator.

Chapter 6

A non-asymptotic data-based model selection criterion

Contents

6.1. Introduction **217**

6.2. An oracle inequality for the Lasso-MLE estimator **219**

6.2.1. A model selection theorem for MLEs in random models 219

6.2.2. Application to our Lasso-MLE estimator 221

6.3. The slope heuristics under the null model principle **225**

6.3.1. The slope heuristics principle 225

6.3.2. A method to validate the slope heuristics 228

6.3.3. Which penalty shape? 232

6.A. Proofs **239**

6.A.1. Proof of Theorem 6.2.1 239

6.A.2. Sketch of the proof of Theorem 6.2.2 245

ABSTRACT

In our Lasso-MLE procedure for variable selection in clustering, our modeling recasts the variable selection and the clustering into a global model selection problem. In Section 4.5.4, we constructed a random model collection and proposed a non-asymptotic data-driven model selection criterion defined from two possible penalty shapes. In this chapter, we focus on studying and justifying these penalty shapes.

On the one hand, we establish a general model selection theorem for maximum likelihood estimators in a random model collection, and we apply it to the model collection constructed in our procedure to deduce a theoretical convenient penalty. This penalty involves a logarithm term to take into account the possible great richness of our model collection. Yet, we might doubt the optimality of this penalty shape for low-dimensional datasets. Besides, this penalty depends on unknown quantities, so it can not be used to get a practical model selection criterion.

On the other hand, we apply a practical method to determine a convenient penalty from the data. This method is based on the "slope heuristics" introduced by Birgé and Massart (2006). This heuristics has been proved only in restricted frameworks. Here, we propose a practical method to check the validity of the slope heuristics in a "null" context. This method can be applied in any framework. We apply it to assess the validity of the slope heuristics in our framework. We also carry out simulations under the "null" model to see whether our theoretical penalty is sharp: a logarithm term is indeed practically detected for high-dimensional datasets, but a penalty proportional to the dimension seems sufficient to define a proper penalty for low-dimensional datasets.

6.1 Introduction

Assume we observe a sample which is a mixture of several subpopulations, each observation being described by a large number of variables. In the last chapter, we proposed a procedure to provide a data clustering taking into account the variable role in the clustering process. This procedure is based on a modeling that recasts variable selection and clustering problems into a model selection problem in a density estimation framework. Specifically, we construct a collection of finite Gaussian mixture models with various numbers of clusters and sets of relevant and active irrelevant variables. Then, we estimate the density of the sample by the maximum likelihood estimator in each model. This leads to a collection of estimators for the density. A final estimator has to be selected among this collection, which is equivalent to selecting a model among the model collection. A global model selection criterion choosing simultaneously the best number of clusters, the best set of relevant variables and the best set of active irrelevant variables according to the observations is required.

This chapter focuses on the model selection criterion used in our procedure. We use a penalized criterion to select a model from a non-asymptotic point of view. In the density estimation framework, the principle of selecting a model by penalizing the empirical contrast emerged during the seventies. Akaike (1973) proposed the Akaike's Information Criterion (AIC) and Schwarz (1978) suggested the Bayesian Information Criterion (BIC). Both of these criteria are based on asymptotic heuristics that are valid only when the number of observations is large enough. In contrast, a non-asymptotic approach for model selection via penalization has emerged during the last ten years, mainly with works of Birgé and Massart (1997) and Barron et al. (1999). The aim of this approach is to define penalized data-driven criteria which lead to oracle inequalities. The penalty function depends on the number of parameters in each model, but also on the complexity of the whole model collection. In our context of density estimation, Barron et al. (1999) and Massart (2007) proposed a general model selection theorem for maximum likelihood estimation. But we can not apply it directly because it is stated for a deterministic model collection whereas our data-driven model collection is random. By extending the proof of Theorem 7.11 in Massart (2007) to cope with the randomness of our model collection, we establish a general model selection theorem for maximum likelihood estimators in a random model collection (see Theorem 6.2.1). Then, by applying this general theorem to the finite Gaussian mixture random model collection constructed in our procedure, we derive a convenient theoretical penalty as well as an associated non-asymptotic penalized criterion and an oracle inequality fulfilled by our Lasso-MLE estimator (see Theorem 6.2.2).

Unfortunately, our theorem does not provide a practical model selection criterion because it provides a penalty shape depending on unknown constants. This drawback is usual when deriving penalties from a general model selection theorem. To deal with such situations, Birgé and Massart (2006) proposed their so-called "slope heuristics" which leads to a data-driven method to calibrate penalties. This heuristics has been proved in some restricted frameworks. Birgé and Massart (2006) proved it for Gaussian regression with homoscedastic fixed design. Then, Arlot and Massart (2008) extended those results to heteroscedastic regression with random design, without assuming that the data are Gaussian. They had to restrict to histograms, but they suppose that this is only due to technical reasons and that the heuristics remains at least valid for the general least squares regression framework. The conjecture that the slope heuristics may be valid in a wider range of frameworks is supported by many encouraging practical studies. Indeed, the slope heuristics has been successfully applied in various model selection situations: for multiple change points detection (Lebarbier, 2005), for estimation of oil reserves (Michel, 2008), in Gaussian Markov random fields (Verzelen, 2008), for the estimation of the number of interior knots in a B-spline regression model (Denis and Molinari, 2009), for the choice of a simplicial complex in the computational geometry field (Caillerie and Michel, 2009), for the determination of the number of mixture components (Baudry, 2009) or for variable selection (Maugis and Michel, 2011a) in a finite Gaussian mixture setting...

Since the slope heuristics is widely used in frameworks where it is not theoretically proved, we propose a general method to assess the validity of the slope heuristics from a practical point of view. This general method can be used in any framework. It consists in simulating a "null model" (in a sense to be specified) and plotting some graph to check whether the slope heuristics seems valid in this specific null context. Although not ensuring that the slope heuristics remains valid when dealing with non-null models, this method can at least detect situations where the slope heuristics might not be used. The idea is to simulate the target in a model included in each model of the collection so that the target estimators are unbiased. In this specific context, the calculations involved in the slope heuristics become simpler and they only depend on quantities that can be computed from the data.

Besides using the null model principle to check the validity of the slope heuristics in our framework, we use it to guess a convenient penalty shape from the data. In fact, our general model selection theorem for a random model collection (Theorem 6.2.1) heavily relies on the fact that this random model collection is included in a larger deterministic model collection. When applying it to the finite Gaussian mixture random model collection constructed by our Lasso-MLE procedure, we take the model collection for complete variable selection as deterministic collection. Therefore, we obtain a penalty shape with a logarithm term to take into account the high richness of this latter model collection. Yet, our model collection is actually much poorer than the model collection for complete variable selection. So, we can wonder whether the logarithm term is actually necessary to define proper penalties. By performing simulations under the null model, we show that the penalty shape derived from our Theorem 6.2.2 is sharp for high-dimensional data whereas it is too pessimistic for low-dimensional data.

The chapter is organized as follows. In Section 6.2, first, we present our general model selection theorem for maximum likelihood estimators in a random model collection. Then, we apply it to the model collection constructed in our procedure to deduce a theoretical convenient penalty shape and derive a non-asymptotic penalized criterion. In Section 6.3, first, we recall the slope heuristics proposed by Birgé and Massart (2006). Then, we introduce the "null model principle" for assessing the validity of this heuristics in any framework and we apply it to our framework. Finally, we carry out simulations under the null model to check whether our theoretical penalty shape is sharp. All the proofs are postponed until the Appendices.

6.2 An oracle inequality for the Lasso-MLE estimator

6.2.1 A model selection theorem for MLEs in random models

Before stating our general MLE model selection theorem, let us recall the definition of the Hellinger distance and specify some notations. The norm $\|\sqrt{t} - \sqrt{u}\|$ between two non-negative integrable

functions t and u is denoted $\mathrm{d_H}(t,u)$. If t and u are two densities with respect to Lebesgue measure on $\mathbb{R}^p$, $\mathrm{d_H}(t,u)$ is the Hellinger distance between t and u. Consider $\mathcal{S}$ the set of all densities on $\mathbb{R}^p$. An ε-bracketing for a subset S of $\mathcal{S}$ with respect to $\mathrm{d_H}$ is a set of integrable function pairs $(l_1,u_1),\ldots,(l_N,u_N)$ such that for each $t\in S$, there exists $j\in\{1,\ldots,N\}$ such that $l_j\le t\le u_j$ and $\mathrm{d_H}(l_j,u_j)\le\varepsilon$. The bracketing number $\mathcal{N}_{[.]}(\varepsilon,S,\mathrm{d_H})$ is the smallest number of ε-brackets necessary to cover S and the bracketing entropy is defined by $\mathcal{H}_{[.]}(\varepsilon,S,\mathrm{d_H})=\ln(\mathcal{N}_{[.]}(\varepsilon,S,\mathrm{d_H}))$.

Let $\{S_m\}_{m\in\mathcal{M}}$ be some at most countable model collection such that $S_m\subset\mathcal{S}$ for all $m\in\mathcal{M}$. We shall say that $\{S_m\}_{m\in\mathcal{M}}$ fulfills Property $(\boldsymbol{\mathcal{P}})$ if, for all $m\in\mathcal{M}$, $\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m,\mathrm{d_H})}$ is integrable at 0 and if there exists a function Ψ_m on $\mathbb{R}_+$ such that Ψ_m is non-decreasing, $\xi\to\Psi_m(\xi)/\xi$ is non-increasing on $]0,+\infty[$, and for $\xi\in\mathbb{R}_+$ and $u\in S_m$, denoting $S_m(u,\xi)=\{t\in S_m;\mathrm{d_H}(t,u)\le\xi\}$,

$$\int_0^\xi\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m(u,\xi),\mathrm{d_H})}\,d\varepsilon\le\Psi_m(\xi). \tag{6.1}$$

Theorem 6.2.1. *Let $s\in\mathcal{S}$ be an unknown density to be estimated from a n-sample $(Y_1,\ldots,Y_n)$. Consider $\{S_m\}_{m\in\mathcal{M}}$ some at most countable deterministic model collection fulfilling Property $(\boldsymbol{\mathcal{P}})$. Let $\{x_m\}_{m\in\mathcal{M}}$ be some family of non-negative numbers such that*

$$\sum_{m\in\mathcal{M}}\mathrm{e}^{-x_m}=\Sigma<\infty. \tag{6.2}$$

For every $m\in\mathcal{M}$, consider Ψ_m defined by $(\boldsymbol{\mathcal{P}})$ and ξ_m such that $\Psi_m(\xi_m)=\sqrt{n}\xi_m^2$.

Let $\tau>0$ such that

$$s_m\ge\mathrm{e}^{-\tau}s \tag{6.3}$$

for all $m\in\mathcal{M}$ and $s_m\in S_m$ such that $\mathrm{KL}(s,s_m)\le 2\inf_{t\in S_m}\mathrm{KL}(s,t)$.

Introduce $\{S_m\}_{m\in\widehat{\mathcal{M}}}$ some random subcollection of $\{S_m\}_{m\in\mathcal{M}}$. Let $\rho\ge0$ and consider the collection of ρ-MLEs $\{\hat{s}_m\}_{m\in\widehat{\mathcal{M}}}$:

$$\gamma_n(\hat{s}_m)\le\inf_{t\in S_m}\gamma_n(t)+\rho.$$

Let $\mathrm{pen}:\mathcal{M}\mapsto\mathbb{R}_+$. Suppose that there exists an absolute constant $\kappa>0$ such that, for all $m\in\mathcal{M}$,

$$\mathrm{pen}(m)\ge\kappa\left(\xi_m^2+(1\vee\tau)\frac{x_m}{n}\right). \tag{6.4}$$

Let $\rho'\ge0$. Then, any penalized likelihood estimator $\hat{s}_{\hat{m}}$ with $\hat{m}\in\widehat{\mathcal{M}}$ such that

$$\gamma_n(\hat{s}_{\hat{m}})+\mathrm{pen}(\hat{m})\le\inf_{m\in\widehat{\mathcal{M}}}\{\gamma_n(\hat{s}_m)+\mathrm{pen}(m)\}+\rho' \tag{6.5}$$

satisfies

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{H}}^2\left(s,\hat{s}_{\hat{m}}\right)\right] \leq C\left(\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}}\left\{\inf_{t\in S_m}\mathrm{KL}(s,t)+\mathrm{pen}(m)\right\}\right]+(1\vee\tau)\frac{\Sigma^2}{n}+\rho+\rho'\right) \tag{6.6}$$

for some absolute positive constant C.

Proof. Page 239. □

Remark 14.

1. Inequality (6.6) is not exactly an oracle inequality since the Hellinger risk is upper bounded by the Kullback bias $\mathrm{KL}(s,s_m)$. Nevertheless, this last term is of the same order as $\mathrm{d}_{\mathrm{H}}^2(s,s_m)$ if $\ln(\|s/t\|_\infty)$ is uniformly bounded on $\cup_{m\in\mathcal{M}}S_m$ (Massart, 2007, Lemma 7.23). In our context, this condition can be achieved if all densities are assumed to be bounded and defined on a compact support, the Gaussian mixtures being truncated on this compact support. In the sequel, we shall consider such an assumption for technical reasons. So, $\mathrm{KL}(s,s_m)$ and $\mathrm{d}_{\mathrm{H}}^2(s,s_m)$ will be equivalent.

2. Condition (6.3) is useful to control the second moment of $\log$-likelihood ratios in order to apply Bernstein's Inequality to bound the empirical process of $\ln(s/s_m)$ (see Lemma 6.A.1). Note that the larger the value of the parameter τ, the larger the minimal penalty (6.4) and the less accurate Inequality (6.6). The minimizers of the Kullback-Leibler divergence s_m are densities, so they are positive and there always exists some $\tau>0$ fulfilling Condition (6.3): at worst, $\tau=+\infty$ is convenient. It seems difficult to have an idea of the minimal convenient value of τ since it depends on the unknown true density s. Nonetheless, we may think that Condition (6.3) is satisfied for reasonable values of τ because the minimizers of the Kullback-Leibler divergence are expected to be close to the true density s.

6.2.2 Application to our Lasso-MLE estimator

Here, we establish an oracle inequality for the estimator of our Lasso-MLE procedure described in Section 4.5.4. We are particularly interested in finding the shape of the minimal penalty leading to an oracle inequality.

Let $\mathcal{J}$ be the collection of all non-empty subsets of $\{1,\ldots,p\}$. Let $\mathcal{M}=\{(K,\boldsymbol{J_r},\boldsymbol{J_a});K\in\mathbb{N}^\star,$ $\boldsymbol{J_r}\in\mathcal{J},\boldsymbol{J_a}\subset\boldsymbol{J_r}\}$. For all $(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}$, consider the model

$$\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}=\left\{\begin{array}{l}\boldsymbol{y}\in\mathbb{R}^p\mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y});\\ s_{\boldsymbol{\theta}}(\boldsymbol{y})=\Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]}\mid\boldsymbol{0},\sigma^2\boldsymbol{I})\,\Phi(\boldsymbol{y}_{[\boldsymbol{J_a}]}\mid\boldsymbol{\mu},\sigma^2\boldsymbol{I})\,\sum_{k=1}^K\pi_k\,\Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]}\mid\boldsymbol{\mu_k},\sigma^2\boldsymbol{I});\\ \boldsymbol{\theta}=(\pi_1,\ldots,\pi_K,\boldsymbol{\mu},\boldsymbol{\mu_1},\ldots,\boldsymbol{\mu_K},\sigma)\in\Pi_K\times\mathbb{R}^{|\boldsymbol{J_a}|}\times\left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K\times\mathbb{R}_+\end{array}\right\}.$$

Since the Lasso-MLE procedure is based on a random model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\widehat{\mathcal{M}}}$ (where $\widehat{\mathcal{M}}$ is a random collection of index sets selected by the Lasso) included in the whole collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}}$, we shall apply Theorem 6.2.1 to provide an oracle inequality for our Lasso-MLE estimator. To apply Theorem 6.2.1, we need an upper bound of the entropy number of the models $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. To construct brackets over $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$, we shall assume that the parameter vectors are bounded. Thus, we shall restrict to bounded models included in $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$:

$$\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{ s_{\boldsymbol{\theta}} \in \mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})};\ \ \boldsymbol{\theta} \in \Pi_K \times [-A_\mu, A_\mu]^{|\boldsymbol{J_a}|} \times \left([-A_\mu, A_\mu]^{|\boldsymbol{J_r}|}\right)^K \times\ [a_\sigma, A_\sigma] \right\} \tag{6.7}$$

where A_μ, a_σ and A_σ are absolute positive constants.

Theorem 6.2.2. *Let $\widehat{\mathcal{M}}$ be a random subcollection of index sets (selected by the Lasso) included in the whole collection $\mathcal{M}$. Consider the collection of bounded models $\{\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\widehat{\mathcal{M}}}$ defined by* (6.7) *and the maximum likelihood estimators*

$$\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \underset{s_{\boldsymbol{\theta}}\in\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{\arg\min}\ \ \gamma_n(s_{\boldsymbol{\theta}}).$$

Denote by $D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = K(1+|\boldsymbol{J_r}|)+|\boldsymbol{J_a}|$ the dimension of the model $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$.
Define

$$B(A_\mu, A_\sigma, a_\sigma, p) := 1 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1+\frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p}.$$

Let $\tau > 0$ such that $s_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \geq \mathrm{e}^{-\tau}\, s$ for all $(K,\boldsymbol{J_r},\boldsymbol{J_a}) \in \mathcal{M}$ and $s_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \in \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ such that $\mathrm{KL}(s, s_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}) \leq 2\inf_{s_{\boldsymbol{\theta}}\in\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}} \mathrm{KL}(s, s_{\boldsymbol{\theta}})$.

Let $\mathrm{pen} : \mathcal{M} \mapsto \mathbb{R}_+$. Suppose that there exists an absolute constant $\kappa > 0$ such that, for all $(K,\boldsymbol{J_r},\boldsymbol{J_a}) \in \mathcal{M}$,

$$\mathrm{pen}\,(K,\boldsymbol{J_r},\boldsymbol{J_a}) \geq \kappa\frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}\left[B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1\wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}}\right)\right.$$
$$\left. +(1\vee\tau)\ln\left(\frac{p}{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\wedge p}\right)\right]. \tag{6.8}$$

Then, the estimator $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}$ with

$$(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}}) = \underset{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\widehat{\mathcal{M}}}{\arg\min}\left\{\gamma_n(\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}) + \mathrm{pen}(K,\boldsymbol{J_r},\boldsymbol{J_a})\right\}$$

satisfies

$$\begin{aligned}&\mathbb{E}\left[\mathrm{d}_{\mathrm{H}}^2\left(s, \hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}\right)\right]\\ &\leq C\left(\mathbb{E}\left[\inf_{(K,\boldsymbol{J}_{\boldsymbol{r}},\boldsymbol{J}_{\boldsymbol{a}})\in\widehat{\mathcal{M}}}\left\{\inf_{s_{\boldsymbol{\theta}}\in\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J}_{\boldsymbol{r}},\boldsymbol{J}_{\boldsymbol{a}})}} \mathrm{KL}(s, s_{\boldsymbol{\theta}}) + \mathrm{pen}(K, \boldsymbol{J}_{\boldsymbol{r}}, \boldsymbol{J}_{\boldsymbol{a}})\right\}\right] + \frac{1\vee\tau}{n}\right) \qquad (6.9)\end{aligned}$$

for some absolute positive constant C.

Proof. Page 245. □

Let us make a few comments on this result.

Contrary to classical asymptotic criteria for which p is fixed and n tends to infinity, our result is non-asymptotic and allows to study cases for which p increases with n. Since the ratio $\ln(p)/n$ appears in the right hand-side of Inequality (6.9) through the term $\mathrm{pen}(K, \boldsymbol{J}_{\boldsymbol{r}}, \boldsymbol{J}_{\boldsymbol{a}})$, our result remains meaningful for a number of variables p not exceeding e^n, which allows to consider many situations with $p \gg n$.

Theorem 6.2.2 is to be compared with Theorem 7.3.2 presented in Maugis and Michel (2011b) where complete variable selection is considered. Both the penalty shape (6.8) and the associated oracle inequality (6.9) are similar to the penalty shape and the oracle inequality established by Maugis and Michel (2011b). This is not surprising when analyzing the proof of Theorem 6.2.2, which follows the proof of Theorem 7.3.2 of Maugis and Michel (2011b), except that we take into consideration the specific form of the models $\mathcal{S}_{(K,\boldsymbol{J}_{\boldsymbol{r}},\boldsymbol{J}_{\boldsymbol{a}})}$ that differs from the form of Maugis and Michel's models. Let us point out a major difference between Theorem 6.2.2 and Maugis and Michel's Theorem 7.3.2. Maugis and Michel's procedure consists in an exhaustive research of the best model and it is untractable as soon as the number of variables becomes too large. In practice, Maugis and Michel (2011b) are limited to studying at most around ten variables unless they restrict to ordered variables. Consequently, the penalized maximum likelihood estimator satisfying their oracle inequality can not be calculated in practice. Their result has only a theoretical – yet not practical – interest. On the contrary, our Lasso-MLE procedure runs on a small random subcollection of models preselected by the Lassos and it remains feasible even for large p, possibly $p \gg n$. Thus, the estimator $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}},\hat{\boldsymbol{J}}_{\boldsymbol{a}})}$ considered in Theorem 6.2.2 is calculable in practice and Theorem 6.2.2 ensures that this estimator achieves good performance compared with the oracle as long as $p < \mathrm{e}^n$.

As expected, the penalty shape (6.8) is proportional to the model dimension D and thus penalizes models with high complexities. It also involves two additional logarithm terms. We do not trust equally in the significance of these two terms.

On the one hand, the first logarithm term into brackets is probably not necessary to define proper penalties. Its presence is certainly only due to the lack of accuracy in the proof of the result. Specifically, in order to apply Theorem 6.2.1, the local bracketing entropy $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(u,\xi), \mathrm{d_H})$ has to be controlled. Yet, it is difficult to characterize the subset $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(u,\xi)$ in function of the parameters of its mixtures, so we have only studied the global entropy bracketing $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})$. This global study is sufficient since the local entropy is upper bounded by the global entropy. Yet, it is not optimal and yields extra logarithm terms.

On the other hand, the second logarithm term may actually be necessary to define proper penalties. It quantifies the complexity of the model collection by taking into account the possible large number of models with identical dimension. Its influence depends on the data dimension. The higher this dimension, the larger the number of models having the same dimension and the higher the influence of this logarithm term. For instance, in regression, Birgé and Massart (2006) prove that, for complete variable selection, a penalty shape proportional to the dimension selects too complex models with high probability and that a logarithm term is necessary to select smaller models. This has been practically checked in some situations. One can cite Lebarbier (2005) for multiple change points detection in a regression framework or Castellan (1999) for histogram selection in a density estimation framework. Yet, this logarithm term becomes unnecessary if the number of models with the same dimension is small enough. For instance, for finite Gaussian mixture models in a very low-dimensional setting, Maugis and Michel (2011a) observe that a penalty proportional to the dimension – with no logarithm term – is sufficient to select models close to the oracle. Unlike Maugis and Michel (2011a), we focus on high-dimensional data and the number of models with the same dimension is expected to grow. Nonetheless, we do not perform complete variable selection. Thanks to preselection of sets of relevant and active variables by ℓ_1-penalization, we obtain a random model collection and we can wonder how rich is this model collection. It seems difficult to answer such a question since our model collection depends on the data. We have been faced with this randomness problem to establish Theorem 6.2.2. Since our model collection is data-dependent, we know nothing about it except that it is included in the whole model collection considered for complete variable selection. Therefore, in the proof of Theorem 6.2.2, we do not take advantage of the fact that our random model collection is just a subcollection of the whole deterministic model collection considered for complete variable selection. For this reason, we obtain a penalty shape (6.8) similar to the penalty shape obtained for complete variable selection, involving a logarithm factor. Yet, our model subcollection may actually be much poorer than the whole model collection and it may contain just a few models with the same dimension, in which case the penalty shape (6.8) may be too pessimistic. If our model collection is poor enough, then a penalty proportional to the dimension might be sufficient to select models of proper dimension. In Section 6.3.3, we shall perform simulations to determine practically a suitable penalty shape.

Theorem 6.2.2 provides a convenient penalty shape (6.8) to get an efficient penalized estimator, but it does not lead to an explicit model selection criterion since (6.8) depends on two unknown constants κ and τ. Besides, mixture parameters are not bounded in practice. That is why we applied the practical slope heuristics introduced by Birgé and Massart (2006) to calibrate the penalty function in our simulations in Chapter 5.

6.3 The slope heuristics under the null model principle

Theory about model selection via penalization often fails in providing explicit penalties and thus practical model selection criteria. Last years, to fill in the gap between theory on penalization and practical applications, Birgé and Massart (2006) proposed an heuristics based on a mixture of theoretical and heuristic ideas to define proper penalties from the data. This heuristics is called "the slope heuristics". Two data-driven methods have been developed to use this heuristics in practice: the dimension jump method and the data-driven slope estimation method (Arlot and Massart, 2008). Both methods enable to calibrate penalties once a penalty shape for the ideal penalty is known up to a multiplicative factor. Most often, the penalty shape is derived from theoretical results such as Theorem 6.2.2. The advantage of the data-driven slope estimation method over the dimension jump method is that, besides calibrating the penalty, it also provides a graphical way to validate the preliminary choice of the penalty shape for the ideal penalty. Since we are not sure that the penalty shape provided by Theorem 6.2.2 is optimal (see the discussion above), we shall use the data-driven slope estimation method rather than the dimension jump method to practically determine an efficient penalty in our context.

The slope heuristics has been proved only in some restricted frameworks (Birgé and Massart, 2006; Arlot and Massart, 2008). In particular, it has not been proved in our framework of Gaussian mixture models. Before applying the slope heuristics, we shall propose a data-driven method to decide whether the slope heuristics may be validated or not. This method is not specific to our framework and it can be applied in any framework. We encourage the user to carry out this checking when applying the slope estimation method. We call it the "null model principle".

In this section, first, we recall the slope heuristics introduced by Birgé and Massart (2006). Then, we introduce the null model principle to practically assess the validity of the slope heuristics in any framework, and we test it in our specific framework. Finally, we exploit the graphical advantage of the data-driven slope estimation method to check whether the penalty shape (6.8) provided by Theorem 6.2.2 is sharp.

6.3.1 The slope heuristics principle (Birgé and Massart, 2006)

Here, we summarize the main ideas leading to the data-driven penalized model selection criterion proposed by Birgé and Massart (2006).

Assume we observe some data $\xi_1, \ldots, \xi_n$ independent with common probability distribution P that depends on a unknown function s belonging to a set $\mathcal{S}$. We focus on estimating s.

Suppose there exists a contrast function $\gamma : \mathcal{S} \times \mathbb{R}^p \mapsto \mathbb{R}$ such that $s = \arg\min_{t \in \mathcal{S}} P\gamma(t)$. Then, each element of $\mathcal{S}$ can be evaluated thanks to the loss function defined for all $t \in \mathcal{S}$ by $l(s,t) = P\gamma(t) - P\gamma(s) \geq 0$. Since P is unknown, one can consider the empirical contrast defined for all $t \in \mathcal{S}$ by

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, \xi_i)$$

such that $P\gamma_n(t) = P\gamma(t)$.

Let $\{S_m\}_{m \in \mathcal{M}}$ be some at most countable model collection with $S_m \subset \mathcal{S}$. For $m \in \mathcal{M}$, consider $\hat{s}_m$ a minimizer of the empirical contrast over the model S_m:

$$\hat{s}_m = \arg\min_{t \in S_m} \gamma_n(t).$$

The model selection problem is to choose the best estimator of s among the collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$. Ideally, one would like to choose the so-called oracle

$$m_{\text{oracle}} = \arg\min_{m \in \mathcal{M}} l(s, \hat{s}_m) \tag{6.10}$$

and estimate s by $\hat{s}_{m_{\text{oracle}}}$. But m_{oracle} is unattainable since it depends on s which is unknown. A model selection procedure is a method choosing some data-dependent $\hat{m} \in \mathcal{M}$ nearly as good as the ideal unattainable choice m_{oracle}. Then, the resulting estimator of s is $\hat{s}_{\hat{m}}$. The basic idea of model selection via penalization is to introduce a penalty function $\text{pen} : \mathcal{M} \mapsto \mathbb{R}^+$ and to select $\hat{m}$ as the minimizer of the penalized criterion over $\mathcal{M}$,

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \text{crit}(m), \qquad \text{crit}(m) = \gamma_n(\hat{s}_m) + \text{pen}(m). \tag{6.11}$$

From (6.10) and (6.11), the ideal penalty is defined by

$$\text{pen}_{\text{id}}(m) = l(s, \hat{s}_m) - \gamma_n(\hat{s}_m) = P\gamma(\hat{s}_m) - P\gamma(s) - \gamma_n(\hat{s}_m).$$

But $P\gamma(s)$ is independent of m, so another ideal penalty is simply $\text{pen}_{\text{id}}(m) = P\gamma(\hat{s}_m) - \gamma_n(\hat{s}_m)$. Just as the oracle m_{oracle}, this ideal penalty is unknown since it depends on the unknown probability distribution P. Yet, Birgé and Massart (2006) suggest that a good approximation of this ideal penalty can be derived from the data. Their heuristics is the following.

Consider for all $m \in \mathcal{M}$,

$$s_m = \arg\min_{t \in S_m} P\gamma(t). \tag{6.12}$$

Birgé and Massart (2006) carry out the following decomposition of the ideal penalty:

$$\begin{aligned} \text{pen}_{\text{id}}(m) &= P\gamma(\hat{s}_m) - \gamma_n(\hat{s}_m) \\ &= \underbrace{P\gamma(\hat{s}_m) - P\gamma(s_m)}_{v_m} + \underbrace{\gamma_n(s_m) - \gamma_n(\hat{s}_m)}_{\hat{v}_m} + \underbrace{P\gamma(s_m) - \gamma_n(s_m)}_{\delta_n(m)}. \end{aligned} \tag{6.13}$$

But $\gamma_n(s_m)$ concentrates around its expectation $P\gamma_n(s_m) = P\gamma(s_m)$, so $\delta_n(m) \approx 0$ and the ideal penalty is thus approximately given by $\text{pen}_{\text{id}}(m) \approx v_m + \hat{v}_m$. At this stage, Birgé and Massart (2006) introduce the main hypothesis of their heuristics, which is to assume that $v_m \approx \hat{v}_m$. The reason for this assumption is that $\hat{v}_m$ is the empirical version of v_m since the empirical measure $n^{-1}\sum_{i=1}^n \delta_{\xi_i}$ and the probability measure P play a similar role in the expressions of $\hat{v}_m$ and v_m. If one permutes these measures in the definitions of $\hat{v}_m$ and v_m and in the definitions of $\hat{s}_m$ and s_m, then $\hat{v}_m$ and v_m are permuted. This heuristics leads to

$$\text{pen}_{\text{id}}(m) \approx 2\hat{v}_m. \tag{6.14}$$

The major point is that $\hat{v}_m$ can be estimated from the data provided that the penalty is known up to a multiplicative constant: assume that there exists some unknown $\kappa_{\text{id}} > 0$ and some known $\text{pen}_{\text{shape}} : \mathcal{M} \mapsto \mathbb{R}_+$ such that

$$\text{pen}_{\text{id}} = \kappa_{\text{id}}\, \text{pen}_{\text{shape}}. \tag{6.15}$$

Then, Arlot and Massart (2008) propose two ways to estimate κ_{id}: the dimension jump method and the data-driven slope estimation method. In this thesis, we shall use the latter. The idea of the data-driven slope estimation method is the following. From the slope heuristics (6.14),

$$\text{pen}_{\text{id}}(m) \approx 2\hat{v}_m = 2\left(\gamma_n(s_m) - \gamma_n(\hat{s}_m)\right) = 2\left[(\gamma_n(s_m) - \gamma_n(s)) + \gamma_n(s) - \gamma_n(\hat{s}_m)\right]. \tag{6.16}$$

The term $\gamma_n(s)$ does not depend on m. Moreover, for the most complex models, the approximation of the model cannot be appreciably improved, so the empirical bias term $\gamma_n(s_m) - \gamma_n(s)$ stabilizes itself and the behavior of pen_{id} becomes similar to the behavior of $-2\gamma_n(\hat{s}_m)$. So, from (6.15) and (6.16), one can expect that, for the most complex models,

$$-\gamma_n(\hat{s}_m) \approx \frac{\kappa_{\text{id}}}{2}\, \text{pen}_{\text{shape}}(m). \tag{6.17}$$

But the goal of a penalty function is to penalize the model complexity, so $\text{pen}_{\text{shape}}(m)$ is assumed to increase with the complexity and thus (6.17) is expected to be checked for large enough $\text{pen}_{\text{shape}}(m)$.

In other words, $-\gamma_n(\hat{s}_m)$ is expected to behave linearly with respect to $\text{pen}_{\text{shape}}(m)$ with a slope around $\kappa_{\text{id}}/2$ for large enough $\text{pen}_{\text{shape}}(m)$. Thus, if $\hat{c}$ is an estimation of the slope of the linear part of $\text{pen}_{\text{shape}}(m) \mapsto -\gamma_n(\hat{s}_m)$, one can estimate κ_{id} by $\hat{\kappa}_{\text{id}} = 2\hat{c}$. Then, from (6.15), one can choose as penalty $\text{pen}(m) = 2\,\hat{c}\,\text{pen}_{\text{shape}}(m)$ and select

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \left\{ \gamma_n(\hat{s}_m) + 2\,\hat{c}\,\text{pen}_{\text{shape}}(m) \right\}. \tag{6.18}$$

The factor 2 in the penalty comes from the slope heuristics $v_m \approx \hat{v}_m$. So, if this heuristics is not valid, the penalized criterion (6.18) will not be optimal. Yet, the slope heuristics has been proved theoretically only in some restricted frameworks. Below, we propose a general data-driven method to decide whether the slope heuristics may be validated or not in any given framework. We call it the null model principle.

6.3.2 A method to validate the slope heuristics

6.3.2.1 The null model principle

The null model principle is just the rewriting of the slope heuristics described above under the following assumption.

Assumption $\mathcal{A}_0$: There exists a model S_0 in the model collection $\{S_m\}_{m \in \mathcal{M}}$ such that $S_0 \subset S_m$ for all $m \in \mathcal{M}$.

Note that if Assumption $\mathcal{A}_0$ is fulfilled, then there is only one model S_0 fulfilling $\mathcal{A}_0$. We call S_0 the null model. In the sequel, we assume that $\mathcal{A}_0$ is fulfilled and we consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from a probability distribution with density $s \in S_0$. The major point is that, for all $m \in \mathcal{M}$, $s \in S_0 \subset S_m$, so $s_m = s$ where s_m is defined by (6.12). This enables a simpler rewriting of the slope heuristics. Following (6.13), the ideal penalty can been decomposed as:

$$\text{pen}_{\text{id}}(m) = P\gamma(\hat{s}_m) - \gamma_n(\hat{s}_m) = P\gamma(\hat{s}_m) + (\gamma_n(s) - \gamma_n(\hat{s}_m)) - \gamma_n(s).$$

Since $\gamma_n(s)$ does not depend on m, another ideal penalty is simply $\text{pen}_{\text{id}}(m) = v_m^0 + \hat{v}_m^0$ where $v_m^0 := P\gamma(\hat{s}_m)$ and $\hat{v}_m^0 := \gamma_n(s) - \gamma_n(\hat{s}_m)$. The main hypothesis of the slope heuristics is to assume that $v_m^0 \approx \hat{v}_m^0$, so that

$$\text{pen}_{\text{id}}(m) \approx 2\hat{v}_m^0. \tag{6.19}$$

The validity of this assumption can be checked by the following data-driven method.

1. Simulation of a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from probability distribution with density $s \in S_0$.

2. Calculation of $\hat{v}_m^0$ for all $m \in \mathcal{M}$.

3. Calculation of v_m^0 for all $m \in \mathcal{M}$ by a Monte Carlo procedure.

4. Plot and comparison of the graphs $m \in \mathcal{M} \mapsto \hat{v}_m^0$ and $m \in \mathcal{M} \mapsto v_m^0$.

The advantage of simulating a dataset with density $s \in S_0$ is to decompose the ideal penalty into two quantities v_m^0 and $\hat{v}_m^0$ that can be calculated from the data. On the contrary, the decomposition (6.13) of the ideal penalty for a non-specific target s involves two quantities v_m and $\hat{v}_m$ that can not be calculated from the data since they depend on s_m which is unknown. Of course, the null model principle can only validate the slope heuristics for a specific target $s \in S_0$. For datasets with density $s \notin S_0$, the null model principle can not be applied. Nonetheless, this method provides a confidence level. On the one hand, if the slope heuristics is not validated for $s \in S_0$, then the slope heuristics is to be ruled out. On the other hand, if the slope heuristics is validated for $s \in S_0$, then we can expect that the slope heuristics remains valid for more general targets.

6.3.2.2 Application of the null model principle in our framework

The slope heuristics has not been proved theoretically in our framework. Thus, we can apply the null model principle to practically assess the validity of this heuristics in our framework. The null model principle relies on the existence of a null model in the model collection. Let us first determine this null model in our case. Let $\mathcal{M}_{(r,a)}$ be the collection of index sets obtained at step 1 of our Lasso-MLE procedure (see Section 4.5.4). Consider the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}); \\ s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I})\ \Phi(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \boldsymbol{\mu}, \sigma^2\boldsymbol{I})\ \sum_{k=1}^K \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2\boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_a}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times\ \mathbb{R}_+ \end{array} \right\}.$$

Let $\mathcal{S}_0 = \{\Phi(\cdot \mid \boldsymbol{0}, \boldsymbol{I})\}$ be the model containing only the p-dimensional Gaussian density with null mean and identity covariance matrix. Denote $s = \Phi(\cdot \mid \boldsymbol{0}, \boldsymbol{I})$. Then, for all $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$, $s = s_{\boldsymbol{\theta}} \in \mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ with $\boldsymbol{\theta}$ defined by $\pi_1 = 1$, $\pi_k = 0$ for all $k \in \{2, \ldots, K\}$, $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{\mu_k} = \boldsymbol{0}$ for all $k \in \{1, \ldots, K\}$ and $\sigma = 1$. So, $\mathcal{S}_0 \subset \mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ for all $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$ and $\mathcal{S}_0$ is the null model.

According to Section 6.3.2.1, the data-driven method to check the validity of the slope heuristics is decomposed into four steps.

Step 1. Simulation of a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from density $s \in S_0$
For all $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$, consider $\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ computed at step 2 of our Lasso-MLE procedure (see Section 4.5.4). Denote by $D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ the dimension of model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. Let $\mathcal{D} =$

$\{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})};\ (K,\boldsymbol{J_r},\boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}\}$. For all $D \in \mathcal{D}$, compute

$$\hat{s}_D = \underset{\substack{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)},\\ D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}=D}}{\arg\min} \gamma_n\left(\hat{s}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\right).$$

Step 2. Calculation of $\hat{v}_D^0$ for all $D \in \mathcal{D}$

$$\hat{v}_D^0 = \gamma_n(s) - \gamma_n(\hat{s}_D) = -\frac{1}{n}\sum_{i=1}^{n}\ln\left(s(\boldsymbol{Y_i})\right) + \frac{1}{n}\sum_{i=1}^{n}\ln\left(\hat{s}_D(\boldsymbol{Y_i})\right). \tag{6.20}$$

Step 3. Calculation of v_D^0 for all $D \in \mathcal{D}$

$$\begin{aligned} v_D^0 &= \mathrm{KL}\left(s, \hat{s}_D\right) \\ &= \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\frac{s(\boldsymbol{x})}{\hat{s}_D(\boldsymbol{x})}\right) s(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(s(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x} - \int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left(\hat{s}_D(\boldsymbol{x})\right) s(\boldsymbol{x})\, d\boldsymbol{x}\,. \end{aligned} \tag{6.21}$$

These integrals can be approximated by a Monte Carlo procedure. Simulate $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_N})$ from density s with a large size $N \gg n$, in order to approximate the integrals in (6.21) by

$$v_D^0 \approx \frac{1}{N}\sum_{i=1}^{N}\ln\left(s(\boldsymbol{X_i})\right) - \frac{1}{N}\sum_{i=1}^{N}\ln\left(\hat{s}_D(\boldsymbol{X_i})\right).$$

Step 4. Plot of the graphs $D \in \mathcal{D} \mapsto \hat{v}_D^0$ and $D \in \mathcal{D} \mapsto v_D^0$.

We apply these four steps in three situations: we fix the number of observations to $n = 200$ and we consider three different number of variables, $p \in \{30, 200, 1000\}$, so as to study respectively the case $p \ll n$ of low dimension, the intermediate case $p = n$ and the case $p \gg n$ of high dimension. For each value of $p \in \{30, 200, 1000\}$, we simulate 10 datasets from the p-dimensional Gaussian density $s = \Phi(\cdot \mid \boldsymbol{0}, \boldsymbol{I})$ in the null model $\mathcal{S}_0$. We let vary the number of clusters in $\mathcal{K} = \{1, \ldots, 10\}$ to construct our model collection. An example of the graphs obtained for one simulation with $p = 1000$ is given at Figure 6.1. For each $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$, the dimension of model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is $D = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_a}|$. The dimension of the most complex model $\mathcal{S}_{(10,\{1,\ldots,p\},\varnothing)}$ is $D_{\max} = 10010 \gg n = 200$, as it can be seen at Figure 6.1. In fact, for such a value of p, our model collection contains a large number of degenerate models, that is models with dimension $D > n$. We may have eliminated the degenerate models from our model collection since we aim at choosing a non-degenerate model. However, we focus on high-dimensional problems, so it seems to us interesting to keep all the models in our collection and to look at the behavior of the empirical contrast and the

Kullback-Leibler divergence for the most complex models. From Figure 6.1, we see that $\hat{v}_D^0 \approx v_D^0$ for the smallest dimensions D, in particular for all non-degenerate models with $D \leq n$. But for larger D, $v_D^0 \ll \hat{v}_D^0$. For all simulations, we actually observe $\hat{v}_D^0 \approx v_D^0$ for all $D \leq n$. So, we can conclude that the slope heuristics is validated for $s \in \mathcal{S}_0$ if we restrict our model collection to models with dimension $D \leq n$.

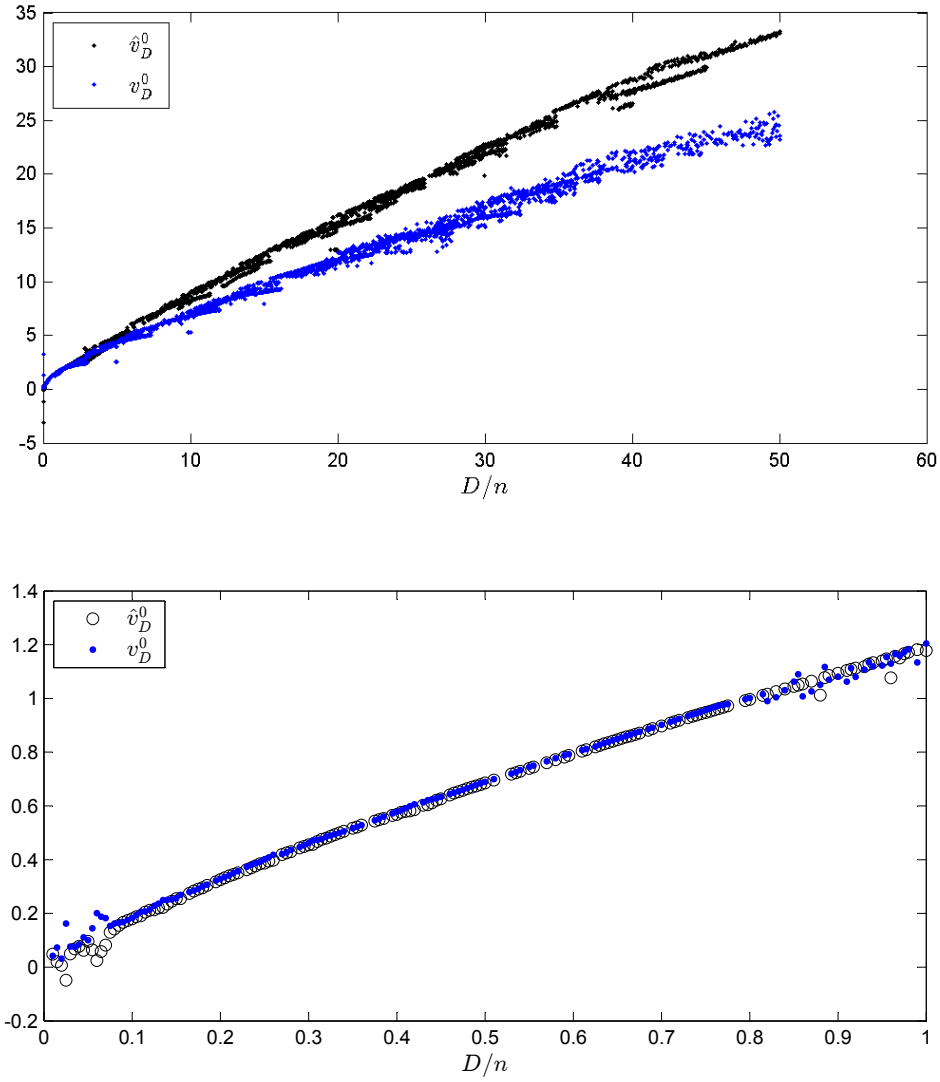

Figure 6.1: Check of the validity of the slope heuristics for a dataset simulated with density $s \in \mathcal{S}_0$, $p = 1000$. At the top, all models are considered. At the bottom, the graph is restricted to the models with dimension $D \leq n$. For $D \leq n$, $\hat{v}_D^0 \approx v_D^0$.

6.3.3 Which penalty shape?

Now that we have validated the slope heuristics, we look for a convenient penalty shape for the ideal penalty. Since we have validated the slope heuristics only for models with dimension $D \leq n$ and since we are interested in selecting such a model, we focus on determining a proper penalty shape defined for models with $D \leq n$. In our context of finite Gaussian mixture models, we know from Maugis and Michel (2011b) that a penalty proportional to the dimension $\mathrm{pen}(D) \propto D/n$ is to be considered for ordered variable selection while a penalty with a logarithm term $\mathrm{pen}_{\ln}(D) \propto (D/n)(1+\kappa \ln(p/D))$ is to be considered for complete variable selection to take into account the richness of model collection. But we have no idea of how rich is our random model collection compared with the model collection for ordered or complete variable selection. In Theorem 6.2.2, we theoretically obtained a penalty shape similar to the penalty shape for complete variable selection, involving a logarithm term. Yet, this is due to the proof of Theorem 6.2.2 (see the discussion in Section 6.2.2). If our model collection is actually much poorer than the model collection for complete variable selection, this penalty shape may be too pessimistic and a penalty proportional to the dimension may be sufficient to select models of proper dimension. To sum up, we hesitate between two penalty shapes: $\mathrm{pen}(D) \propto D/n$ and $\mathrm{pen}_{\ln}(D) \propto (D/n)(1+\kappa\, \ln(p/D))$. Here, we exploit the graphical visualization of the slope estimation method to see whether one of these two penalty shapes is more appropriate.

Given a model collection, the number of models having the same dimension is expected to increase as the number of variables p increases. Thus, the presence or the absence of a logarithm term in the penalty shape may depend on p. So, we consider datasets with $n = 200$ observations and we let vary $p \in \{30, 200, 1000\}$ to study respectively the case $p \ll n$ of low dimension, the intermediate case $p = n$ and the case $p \gg n$ of high dimension. For each value of p, we simulate 10 datasets from the p-dimensional Gaussian density $s = \Phi(\cdot \mid \mathbf{0}, \boldsymbol{I})$ in the null model $\mathcal{S}_0$. We let vary the number of clusters in $\mathcal{K} = \{K_{\min}, \ldots, K_{\max}\} = \{1, \ldots, 10\}$ to construct our model collection.

6.3.3.1 A penalty with a logarithm term?

From Theorem 6.2.2, for $D \leq p \wedge n$, a penalty

$$\mathrm{pen}_{\ln}(D) = \kappa_1 \frac{D}{n}\left(1 + \kappa_2 \ln\left(\frac{p}{D}\right)\right), \quad \kappa_1 > 0, \kappa_2 > 0, \tag{6.22}$$

guarantees to lead to a good model selection criterion. For such a penalty shape, two unknown constants are to be determined from the data to calibrate the penalty and get a data-driven criterion. This requires to perform double regression. This process is likely to more instable than the simple regression considered for a penalty proportional to the dimension. Thus, it is desirable to look for a fixed (deterministic) value of κ_2. If such a fixed value is determined, then only κ_1 remains to be estimated from the data. This can be done by applying the slope estimation method recalled in Section 6.3.1.

By comparing the simple penalty shape (6.22) with the more detailed penalty shape (6.8) in Theorem 6.2.2, we see that κ_1 and κ_2 in (6.22) are not expected to be absolute constants. Yet, in order to try to fix κ_2, it is important to consider a penalty shape for which κ_2 is likely to be an absolute constant. We think that we are more likely to find a universal constant κ_2 if we replace $\ln(p/D)$ in (6.22) by the more homogeneous quantity $\ln(D_{\max}/D)$ where $D_{\max}$ denotes the dimension of the largest model, that is $D_{\max} = K_{\max}(1+p)$. Thus, we rather consider the following penalty shape for $D \leq p \wedge n$:

$$\mathrm{pen}_{\ln}(D) = \kappa_1 \frac{D}{n}\left(1 + \kappa_2 \ln\left(\frac{D_{\max}}{D}\right)\right), \quad \kappa_1 > 0, \kappa_2 > 0. \tag{6.23}$$

For each dataset, estimations of κ_1 and κ_2 in (6.23) can not be obtained by the slope estimation method described in Section 6.3.1 because this method is valid only when the ideal penalty is known up to one multiplicative factor. Yet, this method can be easily extended to deal with the estimation of two unknown constants. Indeed, (6.23) is equivalent to

$$\mathrm{pen}_{\ln}(D) = \kappa_1' \frac{D}{n} + \kappa_2' \frac{D}{n} \ln\left(\frac{D_{\max}}{D}\right), \quad \kappa_1' > 0, \kappa_2' > 0.$$

By applying the slope heuristics recalled in Section 6.3.1, we get similarly as in (6.19)

$$\hat{v}_D^0 \approx c_1 \frac{D}{n} + c_2 \frac{D}{n} \ln\left(\frac{D_{\max}}{D}\right) \tag{6.24}$$

with $c_1 = \kappa_1'/2$ and $c_2 = \kappa_2'/2$. To estimate the coefficients c_1 and c_2, we have implemented a slope estimation method similar to the method proposed by Baudry et al. (2011) for an ideal penalty known up to one multiplicative factor. The only difference between our method and Baudry et al.'s method is that the simple robust regression at step 2 of Baudry et al.'s procedure is replaced by a double robust regression on the triplets of points $\{(D/n, (D/n)\ln(D_{\max}/D), \hat{v}_D^0)\}_{D \leq n \wedge p}$ using the `robustfit` function available in MATLAB. If we denote by $\hat{c}_1$ and $\hat{c}_2$ the estimations of c_1 and c_2 obtained by performing such a double regression, then we get from (6.24) that

$$\hat{v}_D^0 - \hat{c}_2 \frac{D}{n} \ln\left(\frac{D_{\max}}{D}\right) \approx \hat{c}_1 \frac{D}{n}.$$

Therefore, a linear behavior is expected to be observed when plotting the graph $D/n \in [0,1] \mapsto \hat{v}_D^0 - \hat{c}_2 (D/n) \ln\left(D_{\max}/D\right)$.

We perform such a process for each of the 10 datasets for each value of $p \in \{30, 200, 1000\}$.

Our aim is twofold:

1. Given a dataset, check whether we can find estimations $\hat{\kappa}_1$ and $\hat{\kappa}_2$ (both depending on the dataset) such that the penalty $\hat{\kappa}_1 (D/n)\left(1 + \hat{\kappa}_2 \ln(D_{\max}/D)\right)$ defined by (6.23) leads to a good

model selection criterion.

2. Try to find a deterministic value κ_2 such that, for all datasets, we can find an estimation $\hat{\kappa}_1$ (depending on the dataset) such that the penalty $\hat{\kappa}_1(D/n)\,(1+\kappa_2\ln(D_{\max}/D))$ defined by (6.23) leads to a good model selection criterion.

Our simulations lead to the two following conclusions:

1. The double regression fails to be computed for the 10 datasets with $p = 30$: we do not obtain stable estimations $\hat{c}_1$ and $\hat{c}_2$ defined by (6.24). This suggests that no logarithm term is actually required for such low-dimensional datasets. On the contrary, the double regression is successfully computed for the 10 datasets with $p = 200$ and $p = 1000$: we obtain stable estimations $\hat{c}_1$ and $\hat{c}_2$. This suggests that a logarithm term is detected for such higher dimensional datasets. Moreover, by applying the model selection criterion derived from (6.18),

$$\hat{D} = \underset{D\in\mathcal{D}}{\arg\min}\left\{\gamma_n(\hat{s}_D) + 2\left[\hat{c}_1\,\frac{D}{n} + \hat{c}_2\,\frac{D}{n}\,\ln\left(\frac{D_{\max}}{D}\right)\right]\right\},$$

   the null model $\mathcal{S}_0$ is actually selected.
   An example of a graph $D/n \in [0,1] \mapsto \hat{v}_D^0 - \hat{c}_2(D/n)\ln(D_{\max}/D)$ obtained for $p = 1000$ is presented at Figure 6.2.

2. By comparing the estimated ratio $\hat{c}_2/\hat{c}_1$ for the different datasets and values of p tested, no fixed value of c_2/c_1 clearly appears. So, we think that no fixed value of κ_2 in (6.23) can be envisaged.

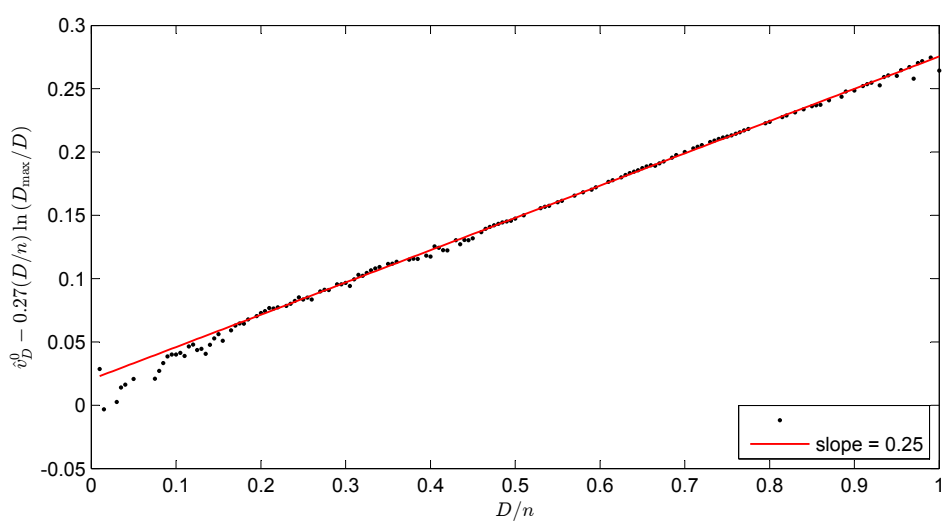

Figure 6.2: Plot of the graph $D/n \in [0,1] \mapsto \hat{v}_D^0 - \hat{c}_2(D/n)\ln(D_{\max}/D)$ for a dataset simulated with density $s \in \mathcal{S}_0$, $p = 1000$. A linear behavior is observed.

6.3.3.2 A penalty proportional to the dimension?

The double regression fails for $p = 30$. Moreover, even if it is successful for $p = 200$ and $p = 1000$, we can wonder whether a penalty proportional to the dimension can be sufficient to define a proper penalty shape. From (6.19), if the ideal penalty is proportional to the dimension, then we must observe a linear behavior of the plot $D/n \in [0,1] \mapsto \hat{v}_D^0$. A linear behavior is actually observed for all simulations: there exists $\hat{c}$ (depending on the dataset) such that $\hat{v}_D^0 \approx \hat{c}D/n$. Nonetheless, by applying the model selection criterion derived from (6.18),

$$\hat{D} = \arg\min_{D \in \mathcal{D}} \left\{ \gamma_n(\hat{s}_D) + 2\,\hat{c}\,\frac{D}{n} \right\},$$

the null model $\mathcal{S}_0$ is not always selected: it is selected for $p = 30$ and $p = 200$, but for $p = 1000$, a non-null model is selected 4 times over the 10 simulations, which suggests that a penalty proportional to the dimension may not be enough penalizing for high-dimensional datasets. An example of a graph $D/n \in [0,1] \mapsto \hat{v}_D^0$ is presented at Figure 6.3.

6.3.3.3 Some combinatorial analysis

From the above study, we can conclude that a penalty proportional to the dimension is to be preferred for low-dimensional datasets such as $p = 30$, while an additional logarithm term seems necessary to define a proper penalty for high-dimensional datasets such as $p = 1000$. For intermediate situations such as $p = 200$, both penalties seem convenient. Thus, we can not say that a penalty is always better than another one. The ideal penalty shape seems to depend on the data dimension. We think it is dangerous (maybe impossible) to try to determine from which ratio p/n a logarithm term becomes necessary, especially when we shall deal with non-null models. It may depend on the dataset, on the number of relevant variables in the true model, on the value of the true mean coefficients...

Although the above study does not enable to decide which penalty is to be considered, it highlights an interesting point, which is precisely the change of the ideal penalty shape and the apparition of a logarithm term as p increases. We know from the theory on model selection (Birgé and Massart, 2006) that this logarithm term takes into account the richness of the model collection since it stands for the number of models with the same dimension in the model collection. The richer the model collection, the stronger the penalty needed to define a proper selection criterion. Thus, the change of the penalty shape is expected to be linked to the number of models generated by the Lasso. To check that this is actually the case, for each value of $p \in \{30, 200, 1000\}$, for each simulation, we compare the number of models generated by the Lasso with the value of the data-driven estimated slope coefficient $\hat{c}$ in the penalty proportional to the dimension $\mathrm{pen}(D) = 2\hat{c}D/n$. Two interesting facts can be noted:

- The higher p, the higher the average estimated slope coefficient $\hat{c}$ (Figure 6.4, at the top).

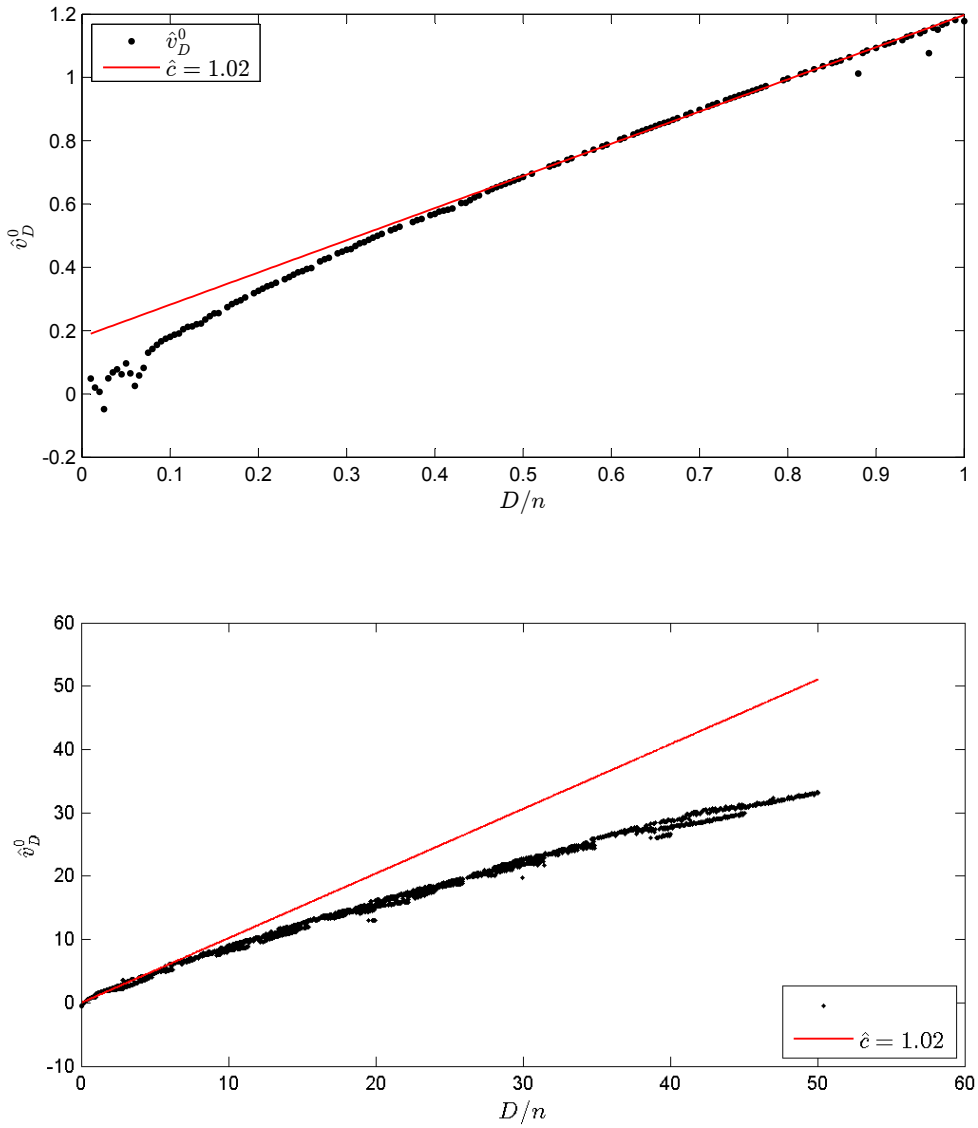

Figure 6.3: Plot of the graphs $D/n \in [0,1] \mapsto \hat{v}_D^0$ (at the top) and $D/n \mapsto \hat{v}_D^0$ (at the bottom) for a dataset simulated with density $s \in \mathcal{S}_0, p = 1000$. We observe a linear behavior of $D/n \in [0,1] \mapsto \hat{v}_D^0$. The estimated slope $\hat{c}$ is computed by considering only the models with $D \leq n$. Note that the behavior of $\hat{v}_D^0$ remains quite regular even for the most complex models with $D \gg n$.

- The higher p, the higher the variability of the estimated slope coefficient $\hat{c}$ between the different datasets with p variables (Figure 6.4, at the top). Moreover, this variability is closely related to the number of models with dimension $D \leq n$ generated by the Lasso (Figure 6.4, at the bottom).

We can wonder to which extend this phenomenon is linked to the richness of the model collections. To answer this question, let us come back to theory. Theorem 6.2.2 provides a penalty shape

$$\mathrm{pen}_{\ln}(D) \propto \frac{D}{n}\left(1 + \kappa \ln\left(\frac{p}{D}\right)\right). \tag{6.25}$$

By looking at the proof of Theorem 6.2.2, we see that the logarithm term $D\ln(p/D) = \ln[(p/D)^D]$ comes from the logarithm $\ln(\#_D)$ of the number $\#_D$ of models having the same dimension D in the model collection: $D\ln(p/D) \propto \ln(\#_D)$. Now, assume that a penalty proportional to the dimension

$$\mathrm{pen}(D) \approx 2\,\hat{c}\,\frac{D}{n} \tag{6.26}$$

is observed. By comparing (6.25) with (6.26), this suggests that

$$\frac{D}{n} + \kappa\,\frac{\ln(\#_D)}{n} \propto \hat{c}\,\frac{D}{n}\,. \tag{6.27}$$

So, $\#_D$ is expected to affect the estimation $\hat{c}$ of the slope coefficient. In particular:

- Since the number $\#_D$ of models with the same dimension D is expected to increase as p increases, (6.27) could explain why $\hat{c}$ globally increases as p increases (Figure 6.4, at the top).

- For a fixed value of p, for each simulation, our data-driven model collection changes, so the number $\#_{D\leq n} := \sum_{D\in\mathcal{D}, D\leq n} \#_D$ of models with dimension $D \leq n$ changes. This could explain the parallel observed between $\#_{D\leq n}$ (Figure 6.4, at the bottom) and the estimation $\hat{c}$ (Figure 6.4, at the top).

Remark 15. From (6.27), we could expect that $\ln(\#_D) \propto D$. We plotted $D \mapsto \ln(\#_D)$, but no clear linear behavior was observed. Although Figure 6.4 highlights that the estimation $\hat{c}$ is linked to combinatorial reasons, this global combinatorial phenomenon seems too complex to be locally analyzed.

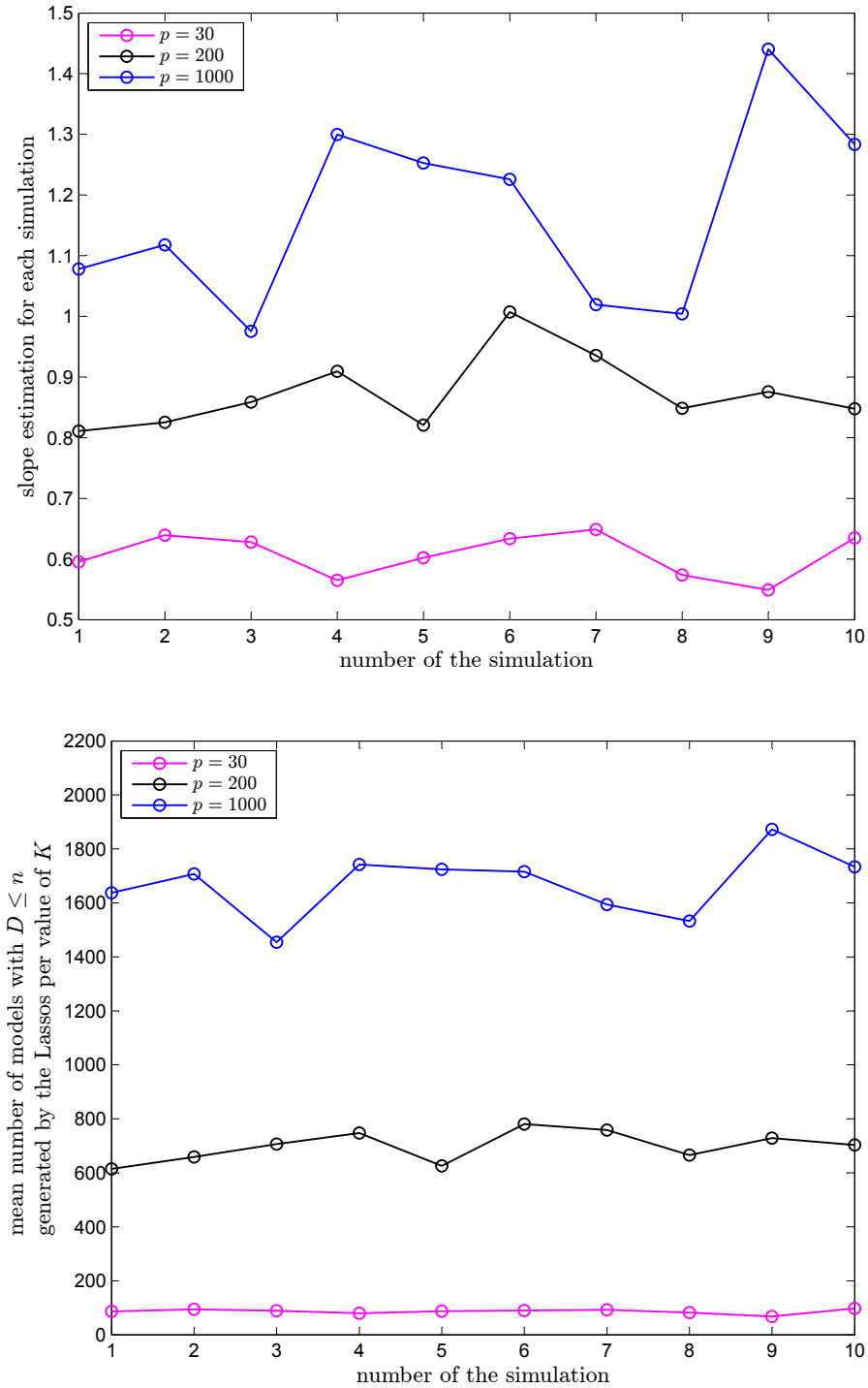

Figure 6.4: Comparison between the values of the estimated coefficients $\hat{c}$ and the average richness of model collections per value K of number of clusters, for 10 datasets simulated from probability distribution with density $s \in \mathcal{S}_0$.

Appendices

6.A Proofs

6.A.1 Proof of Theorem 6.2.1

6.A.1.1 An inequality for the second moment of $\log$-likelihood ratios.

Here, we provide an inequality for the moments of order 2 of $\log$-likelihood ratios. This inequality is based on the following Claim 6.A.1.

Claim 6.A.1. *Let $\tau > 0$. For all $x > 0$, consider*

$$f(x) = x(\ln x)^2, \tag{6.28}$$

$$h(x) = x \ln x - x + 1, \tag{6.29}$$

$$\phi(x) = \mathrm{e}^x - x - 1. \tag{6.30}$$

Then, for all $0 < x \leq \mathrm{e}^\tau$,

$$f(x) \leq \frac{\tau^2}{\phi(-\tau)}\, h(x). \tag{6.31}$$

Proof. First note that $f(1) = h(1) = 0$, so (6.31) is satisfied for $x = 1$ and we just need to prove (6.31) for $x \neq 1$. Define

$$\psi : \mathbb{R} \mapsto \mathbb{R},\ y \mapsto \begin{cases} \phi(y)/y^2 & \text{if } y \neq 0, \\ 1/2 & \text{if } y = 0 \end{cases}$$

and

$$\varphi : \mathbb{R} \mapsto \mathbb{R},\ y \mapsto \begin{cases} \phi(y)/y & \text{if } y \neq 0, \\ 0 & \text{if } y = 0. \end{cases}$$

Let us first check that ψ is non-decreasing on $\mathbb{R}$.
Since $\mathrm{e}^y = 1 + y + y^2/2 + o_{y\to 0}(y^2)$, the functions ψ and φ are continuous on $\mathbb{R}$ and, for $y \neq 0$, $\psi(y) = (\varphi(y) - \varphi(0))/(y - 0)$ is the difference quotient of φ. Thus, we just need to prove that φ is a convex function to derive that ψ is non-decreasing. By differentiating twice φ, we get that $\varphi''(y) = 2\mathrm{e}^y g(y)/y^3$ with $g(y) = 1 - y + y^2/2 - \mathrm{e}^{-y}$. The function g is non-decreasing because $g'(y) = -1 + y + \mathrm{e}^{-y} \geq 0$. But $g(0) = 0$. So, $g(y) \leq 0$ for all $y \leq 0$ and $g(y) \geq 0$ for all $y \geq 0$. It implies that $\varphi''(y) \geq 0$ for all $y \in \mathbb{R}$ and φ is convex.

Now, let $0 < x \leq \mathrm{e}^\tau$, $x \neq 1$. Put $y = -\ln x$. Then, $y \geq -\tau$ and, since ψ is non-decreasing, $\psi(y) \geq \psi(-\tau)$. Moreover, $x \neq 1$, so $y \neq 0$ and $\psi(y) = \phi(y)/y^2$. Thus, $\phi(y)/y^2 \geq \phi(-\tau)/\tau^2$.

Taking into account the definitions of $\phi(y)$ and y, it leads to

$$\ln x - 1 + \frac{1}{x} \geq \frac{\phi(-\tau)}{\tau^2}(\ln x)^2.$$

We get (6.31) by multiplying the last inequality by $x > 0$. □

Lemma 6.A.1. *Let P and Q be two probability measures with $P \ll Q$. Assume that there exists $\tau > 0$ such that $\ln\left(\|dP/dQ\|_\infty\right) \leq \tau$. Then,*

$$\int \left(\ln \frac{dP}{dQ}\right)^2 dP \leq \frac{\tau^2}{\mathrm{e}^{-\tau} + \tau - 1}\ \mathrm{KL}(P, Q). \tag{6.32}$$

Proof. Since $\ln\left(dP/dQ\right) \leq \tau$, we have $dP/dQ \leq \mathrm{e}^\tau$ and we can apply Claim 6.A.1 to $x = dP/dQ$:

$$f\left(\frac{dP}{dQ}\right) \leq \frac{\tau^2}{\phi(-\tau)}\, h\left(\frac{dP}{dQ}\right).$$

Integrating with respect to Q and taking into account (6.28), (6.29) and (6.30), we get

$$\begin{aligned}
\int \left(\ln \frac{dP}{dQ}\right)^2 dP &\leq \frac{\tau^2}{\mathrm{e}^{-\tau} + \tau - 1} \int \left[\frac{dP}{dQ} \ln\left(\frac{dP}{dQ}\right) - \frac{dP}{dQ} + 1\right] dQ \\
&\leq \frac{\tau^2}{\mathrm{e}^{-\tau} + \tau - 1} \left[\int \ln\left(\frac{dP}{dQ}\right) dP - 1 + 1\right] \\
&\leq \frac{\tau^2}{\mathrm{e}^{-\tau} + \tau - 1}\ \mathrm{KL}(P, Q).
\end{aligned}$$

□

6.A.1.2 Proof of Theorem 6.2.1

For the sake of simplicity, we assume that $\rho = \rho' = 0$. For any measurable function g, denote by ν_n the recentred process defined by

$$\nu_n(g) = \frac{1}{n} \sum_{i=1}^{n} \left(g(Y_i) - \mathbb{E}[g(Y_i)]\right). \tag{6.33}$$

For all $m \in \mathcal{M}$, consider s_m such that $\mathrm{KL}(s, s_m) \leq 2 \inf_{t \in S_m} \mathrm{KL}(s, t)$. Define the functions

$$g_m = -\frac{1}{2} \ln\left(\frac{s_m}{s}\right), \quad \hat{g}_m = -\frac{1}{2} \ln\left(\frac{\hat{s}_m}{s}\right) \tag{6.34}$$

$$\hat{f}_m = -\ln\left(\frac{s + \hat{s}_m}{2s}\right) \tag{6.35}$$

Fix $m \in \mathcal{M}$. Introduce

$$\mathcal{M}(m) = \left\{ m' \in \mathcal{M}, \gamma_n\left(\hat{s}_{m'}\right) + \operatorname{pen}\left(m'\right) \leq \gamma_n\left(\hat{s}_m\right) + \operatorname{pen}\left(m\right) \right\}.$$

Let $m' \in \mathcal{M}(m)$. We deduce from the definition of $\mathcal{M}(m)$, $\hat{s}_m$ and (6.34) that

$$\frac{2}{n}\sum_{i=1}^{n} \hat{g}_{m'}(Y_i) + \operatorname{pen}\left(m'\right) \leq \frac{2}{n}\sum_{i=1}^{n} \hat{g}_{m}(Y_i) + \operatorname{pen}\left(m\right) \leq \frac{2}{n}\sum_{i=1}^{n} g_{m}(Y_i) + \operatorname{pen}\left(m\right). \tag{6.36}$$

Besides, by concavity of the logarithm, we have $\hat{f}_{m'} \leq \hat{g}_{m'}$. So, (6.36) gives

$$\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{m'}(Y_i) \leq \frac{2}{n}\sum_{i=1}^{n} g_{m}(Y_i) + \operatorname{pen}\left(m\right) - \operatorname{pen}\left(m'\right). \tag{6.37}$$

By taking (6.33), (6.34) and (6.35) into account, this inequality implies

$$2\,\mathrm{KL}\left(s, \frac{s+\hat{s}_{m'}}{2}\right) \leq \mathrm{KL}\left(s, s_m\right) + \operatorname{pen}\left(m\right) - \operatorname{pen}\left(m'\right) + 2\left[\nu_n\left(g_m\right) - \nu_n\left(\hat{f}_{m'}\right)\right]. \tag{6.38}$$

Our purpose is now to control both $\nu_n\left(g_m\right)$ and $-\nu_n(\hat{f}_{m'})$.

To bound $-\nu_n(\hat{f}_{m'})$, we refer to the proof of Theorem 7.11 in Massart (2007). It is proved that there exists $\kappa'' > 0$ such that for all $u > 0$, for all $m' \in \mathcal{M}(m)$, for all $y > \xi_{m'}$, the following inequality holds except on a set with probability less than $2\mathrm{e}^{-u}$:

$$\frac{-\nu_n(\hat{f}_{m'})}{y^2 + \|\sqrt{s} - \sqrt{\hat{s}_{m'}}\|^2} \leq \kappa''\left(\frac{\xi_{m'} + \sqrt{u/n}}{y} + \frac{u}{ny^2}\right). \tag{6.39}$$

Let us now focus on controlling $\nu_n\left(g_m\right)$. From (6.33) and (6.34), we have

$$\nu_n(g_m) = \sum_{i=1}^{n} X_i - \mathbb{E}(X_i), \quad X_i := \frac{1}{2n}\ln\left(\frac{s(Y_i)}{s_m(Y_i)}\right). \tag{6.40}$$

To get an upper bound of $\nu_n(g_m)$, we apply Bernstein's Inequality whose statement is recalled in Lemma 6.A.2 below and whose proof can be found in Massart (2007) for instance. This inequality requires to control the moments of order k for all $k \geq 2$ of X_i defined by (6.40). Such a control is provided by Lemma 6.A.1 on condition that $\ln(\|s/s_m\|_\infty) \leq \tau$.
Assume that (6.3) is fulfilled. Then, $\ln(\|s/s_m\|_\infty) \leq \tau$ and we deduce from Lemma 6.A.1 that

$$\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \leq \frac{\tau^2}{\mathrm{e}^{-\tau} + \tau - 1}\ \mathrm{KL}(s, s_m).$$

On the one hand, $\tau^2/(\mathrm{e}^{-\tau}+\tau-1) \sim_{\tau\to\infty} \tau$, so there exists $A>0$ such that $\tau^2/(\mathrm{e}^{-\tau}+\tau-1) \le 2\tau$ for all $\tau \ge A$. On the other hand, $\tau \mapsto \tau^2/(\mathrm{e}^{-\tau}+\tau-1)$ is continuous on $]0,A]$ and $\tau^2/(\mathrm{e}^{-\tau}+\tau-1) \sim_{\tau\to 0} 2$, so there exists $B>0$ such that $\tau^2/(\mathrm{e}^{-\tau}+\tau-1) \le B$ for all $\tau \in]0,A]$. Thus, for all $\tau>0$, $\tau^2/(\mathrm{e}^{-\tau}+\tau-1) \le \delta(1\vee\tau)$ with $\delta = 2\vee B$, and

$$\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \le \delta(1\vee\tau)\,\mathrm{KL}(s,s_m). \tag{6.41}$$

From (6.40), (6.41) and the assumption $\ln(\|s/s_m\|_\infty) \le \tau$, we derive that

$$\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \le \frac{n}{(2n)^2}\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \le \frac{\delta(1\vee\tau)\,\mathrm{KL}(s,s_m)}{4n} \tag{6.42}$$

and that for all integers ≥ 3,

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\left[(X_i)_+^k\right] &\le \frac{n}{(2n)^k}\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)_+^k s(y)\,dy \\
&\le \frac{n}{(2n)^k}\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^k \mathbb{1}_{\{s(y)\ge s_m(y)\}}\, s(y)\,dy \\
&\le \frac{n}{(2n)^k}\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^{k-2}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 \mathbb{1}_{\{s(y)\ge s_m(y)\}}\, s(y)\,dy \\
&\le \frac{n}{(2n)^k}\,\tau^{k-2}\int_{\mathbb{R}^p}\left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 \mathbb{1}_{\{s(y)\ge s_m(y)\}}\, s(y)\,dy \\
&\le \frac{n}{(2n)^k}\,\tau^{k-2}\,\delta(1\vee\tau)\,\mathrm{KL}(s,s_m) \\
&\le \frac{1}{2}\left(\frac{\tau}{2n}\right)^{k-2}\frac{\delta(1\vee\tau)\,\mathrm{KL}(s,s_m)}{2n}.
\end{aligned} \tag{6.43}$$

From (6.42) and (6.43), we can apply Bernstein's Inequality with

$$v := \frac{\delta(1\vee\tau)\,\mathrm{KL}(s,s_m)}{2n}, \quad c := \frac{\tau}{2n}. \tag{6.44}$$

It gives that, for every positive u, except on a set with probability less than e^{-u},

$$\nu_n(g_m) \le \sqrt{2vu} + cu. \tag{6.45}$$

Let $z>0$ to be chosen later. Using that $z^2+\mathrm{KL}(s,s_m) \ge 2z\sqrt{\mathrm{KL}(s,s_m)}$ and $z^2+\mathrm{KL}(s,s_m) \ge z^2$, we get from (6.45) that except on a set with probability less than e^{-u},

$$\frac{\nu_n(g_m)}{z^2+\mathrm{KL}(s,s_m)} \le \frac{\sqrt{2vu}+cu}{z^2+\mathrm{KL}(s,s_m)} \le \frac{\sqrt{vu}}{\sqrt{2}z\sqrt{\mathrm{KL}(s,s_m)}} + \frac{cu}{z^2}. \tag{6.46}$$

Let us gather (6.39) and (6.46). There exists $\kappa'' > 0$ such that, for every positive u, for all $m' \in \mathcal{M}(m)$, for all $z > 0$ and for all $y \geq \xi_{m'}$, except on a set with probability less than $3\mathrm{e}^{-u}$,

$$\frac{-\nu_n\left(\hat{f}_{m'}\right)}{y^2 + \|\sqrt{s} - \sqrt{\hat{s}_{m'}}\|^2} \leq \kappa''\left(\frac{\xi_{m'}}{y} + \frac{\sqrt{u/n}}{y} + \frac{u}{ny^2}\right), \tag{6.47}$$

$$\frac{\nu_n(g_m)}{z^2 + \mathrm{KL}(s, s_m)} \leq \frac{\sqrt{vu}}{\sqrt{2}z\sqrt{\mathrm{KL}(s, s_m)}} + \frac{cu}{z^2}. \tag{6.48}$$

Now, let $x > 0$. Let x_m and $x_{m'}$ be defined by (6.2). We apply (6.47) and (6.48) to $u = x + x_m + x_{m'}$ and we choose adequately y and z by defining for some constants γ and β to be specified later,

$$y_{m,m'} := \gamma^{-1}\sqrt{\xi_{m'}^2 + \frac{x + x_m + x_{m'}}{n}}\,, \tag{6.49}$$

$$z_{m,m'} := \beta^{-1}\sqrt{\left(\frac{v}{2\,\mathrm{KL}(s, s_m)} + c\right)(x + x_m + x_{m'})}. \tag{6.50}$$

Using that $a^2 + b^2 \geq a^2$, we get that except on a set with probability less than $3\mathrm{e}^{-(x+x_m+x_{m'})}$,

$$\frac{-\nu_n\left(\hat{f}_{m'}\right)}{y_{m,m'}^2 + \|\sqrt{s} - \sqrt{\hat{s}_{m'}}\|^2} \leq \kappa''(2\gamma + \gamma^2),$$

$$\frac{\nu_n(g_m)}{z_{m,m'}^2 + \mathrm{KL}(s, s_m)} \leq \beta + \beta^2,$$

that is to say

$$-\nu_n(\hat{f}_{m'}) \leq \kappa''(2\gamma + \gamma^2)\left(y_{m,m'}^2 + \left\|\sqrt{s} - \sqrt{\hat{s}_{m'}}\right\|^2\right), \tag{6.51}$$

$$\nu_n(g_m) \leq (\beta + \beta^2)(z_{m,m'}^2 + \mathrm{KL}(s, s_m)). \tag{6.52}$$

We can now come back to Inequality (6.38). Injecting (6.51) and (6.52) into (6.38) yields

$$\begin{aligned}2\,\mathrm{KL}\left(s, \frac{s + \hat{s}_{m'}}{2}\right) &\leq \mathrm{KL}\left(s, s_m\right) + \mathrm{pen}\left(m\right) - \mathrm{pen}\left(m'\right) \\ &\quad + 2(\beta + \beta^2)(z_{m,m'}^2 + \mathrm{KL}(s, s_m)) + 2\kappa''(2\gamma + \gamma^2)\left(y_{m,m'}^2 + \left\|\sqrt{s} - \sqrt{\hat{s}_{m'}}\right\|^2\right).\end{aligned}$$

Putting $\kappa(\beta) := (1+2(\beta+\beta^2))$, using the inequality $\mathrm{KL}(s, (s+\hat{s}_{m'})/2) \geq (2\ln 2 - 1)\|\sqrt{s} - \sqrt{\hat{s}_{m'}}\|^2$ (Massart, 2007, Lemma 7.23) and choosing γ such that $2\kappa''(2\gamma + \gamma^2) = 2\ln 2 - 1 := \alpha$, we get

$$\alpha\left\|\sqrt{s} - \sqrt{\hat{s}_{m'}}\right\|^2 \leq \kappa(\beta)\,\mathrm{KL}\left(s, s_m\right) + \mathrm{pen}\left(m\right) - \mathrm{pen}\left(m'\right) + 2(\beta + \beta^2)z_{m,m'}^2 + \alpha y_{m,m'}^2.$$

From (6.50), (6.49) and (6.44), we deduce that

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 &\leq \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &\quad+(\beta+\beta^2)\beta^{-2}\left(\frac{\delta(1\vee\tau)}{2}+\tau\right)\frac{x+x_m+x_{m'}}{n}\\ &\quad+\alpha\gamma^{-2}\left(\xi_{m'}^2+\frac{x+x_m+x_{m'}}{n}\right).\end{aligned}$$

Since $\tau\leq 1\vee\tau$, if we choose β such that $(\beta+\beta^2)(\delta/2+1)=\alpha\gamma^{-2}$, we get

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 &\leq \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &\quad+\alpha\gamma^{-2}\xi_{m'}^2+\alpha\gamma^{-2}\left[\beta^{-2}(1\vee\tau)+1\right]\frac{x+x_m+x_{m'}}{n}.\end{aligned}$$

Put $\kappa=\alpha\gamma^{-2}(\beta^{-2}+1)$. Then, since $1\leq 1\vee\tau$,

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 &\leq \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &\quad+\alpha\gamma^{-2}\xi_{m'}^2+\kappa\,(1\vee\tau)\frac{x+x_m+x_{m'}}{n}\\ &\leq \kappa(\beta)\,\mathrm{KL}(s,s_m)+\left[\mathrm{pen}(m)+\kappa\,(1\vee\tau)\frac{x_m}{n}\right]\\ &\quad+\left[\alpha\gamma^{-2}\xi_{m'}^2+\kappa(1\vee\tau)\frac{x_{m'}}{n}-\mathrm{pen}(m')\right]+\kappa(1\vee\tau)\frac{x}{n}\\ &\leq \kappa(\beta)\,\mathrm{KL}(s,s_m)+\left[\mathrm{pen}(m)+\kappa\,(1\vee\tau)\frac{x_m}{n}\right]\\ &\quad+\left[\kappa\left(\xi_{m'}^2+(1\vee\tau)\frac{x_{m'}}{n}\right)-\mathrm{pen}(m')\right]+\kappa(1\vee\tau)\frac{x}{n}.\end{aligned}$$

Now, assume that Condition (6.4) on the penalty function is fulfilled for this value of κ. Then, for all $x>0$, for every $m\in\mathcal{M}$ and $m'\in\mathcal{M}(m)$, except on a set with probability less than $3\mathrm{e}^{-(x+x_m+x_{m'})}$, we have

$$\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2\leq\kappa(\beta)\,\mathrm{KL}(s,s_m)+2\,\mathrm{pen}(m)+\kappa(1\vee\tau)\frac{x}{n}. \tag{6.53}$$

It only remains to sum up the tail bounds (6.53) over all the possible values of $m\in\mathcal{M}$ and $m'\in\mathcal{M}(m)$ by taking the union of the different sets of probability less than $3\mathrm{e}^{-(x+x_m+x_{m'})}$. For all $x>0$, except on a set with probability less than

$$3\sum_{m\in\mathcal{M},m'\in\mathcal{M}(m)}\mathrm{e}^{-(x+x_m+x_{m'})}\leq 3\mathrm{e}^{-x}\sum_{(m,m')\in\mathcal{M}\times\mathcal{M}}\mathrm{e}^{-(x_m+x_{m'})}=3\mathrm{e}^{-x}\left(\sum_{m\in\mathcal{M}}\mathrm{e}^{-x_m}\right)^2=3\Sigma^2\mathrm{e}^{-x},$$

we have simultaneously for all $m \in \mathcal{M}$ and $m' \in \mathcal{M}(m)$,

$$\alpha \left\| \sqrt{s} - \sqrt{\hat{s}_{m'}} \right\|^2 \leq \kappa(\beta)\,\mathrm{KL}(s, s_m) + 2\,\mathrm{pen}(m) + \kappa(1 \vee \tau)\frac{x}{n}. \tag{6.54}$$

Inequality (6.54) is in particular satisfied for all $m \in \widehat{\mathcal{M}}$ and $m' \in \widehat{\mathcal{M}}(m)$ and, since $\hat{m}$ defined by (6.5) belongs to $\widehat{\mathcal{M}}(m)$ for all $m \in \widehat{\mathcal{M}}$, we deduce from (6.54) that for all $x > 0$, except on a set with probability less than $3\Sigma^2 \mathrm{e}^{-x}$,

$$\begin{aligned} \alpha \left\| \sqrt{s} - \sqrt{\hat{s}_{\hat{m}}} \right\|^2 &\leq \inf_{m \in \widehat{\mathcal{M}}} \{\kappa(\beta)\,\mathrm{KL}(s, s_m) + 2\,\mathrm{pen}(m)\} + \kappa(1 \vee \tau)\frac{x}{n} \\ &\leq \inf_{m \in \widehat{\mathcal{M}}} \left\{ 2\kappa(\beta) \inf_{t \in S_m} \mathrm{KL}(s, t) + 2\,\mathrm{pen}(m) \right\} + \kappa(1 \vee \tau)\frac{x}{n}\,. \end{aligned} \tag{6.55}$$

By integrating (6.55) over $x > 0$, we finally get that there exists an absolute constant $C > 0$ such that

$$\mathbb{E}\left[\left\| \sqrt{s} - \sqrt{\hat{s}_{\hat{m}}} \right\|^2 \right] \leq C \left(\mathbb{E}\left[\inf_{m \in \widehat{\mathcal{M}}} \left\{ \inf_{t \in S_m} \mathrm{KL}(s, t) + \mathrm{pen}(m) \right\} \right] + (1 \vee \tau)\frac{\Sigma^2}{n} \right).$$

□

Lemma 6.A.2. [Bernstein's Inequality] Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist some positive numbers v and c such that

$$\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] \leq v$$

and for all integers $k \geq 3$,

$$\sum_{i=1}^{n} \mathbb{E}\left[(X_i)_+^k\right] \leq \frac{k!}{2} v c^{k-2}.$$

Then, for every positive u,

$$\mathbb{P}\left[\sum_{i=1}^{n} (X_i - \mathbb{E}(X_i)) \geq \sqrt{2vu} + cu \right] \leq \mathrm{e}^{-u}.$$

6.A.2 Sketch of the proof of Theorem 6.2.2

We consider the Gaussian mixture models $\mathcal{S}^{\mathcal{B}}_{(K, \boldsymbol{J_r}, \boldsymbol{J_a})}$. We shall prove Theorem 6.2.2 by applying Theorem 6.2.1. To deduce Theorem 6.2.2 from Theorem 6.2.1, we follow the arguments developed by Maugis and Michel (2011b). The only difference between our proof and Maugis and Michel's proof is the structure of the models into consideration. Thus, we just give a sketch of the proof of Theorem 6.2.2 and we refer to Maugis and Michel (2011b) for more details.

6.A.2.1 Control of the global entropy bracketing $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})$

To apply Theorem 6.2.1, the first step is to control the bracketing entropy of the Gaussian mixture families $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. Note that Theorem 6.2.1 only requires to control the local bracketing entropy $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(u,\xi), \mathrm{d_H})$. Nevertheless, it is difficult to characterize the subset $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(u,\xi)$ in function of the parameters of its mixtures. Thus, we rather control the global entropy bracketing $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})$, which is sufficient since the local bracketing entropy is upper bounded by the global bracketing entropy.

Proposition 6.A.3. *Put $D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = K(1+|\boldsymbol{J_r}|)+|\boldsymbol{J_a}|$. For all $\varepsilon \in]0,1]$,*

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H}\right) \leq C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\left(\frac{1}{\varepsilon}\right)^{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}$$

with

$$C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p) := 4\,(2\pi \mathrm{e})^{K/2} 3^{K-1}\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right)\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|} K(3\sqrt{c}p)^{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}, \tag{6.56}$$

$c = \mathrm{sh}(1) + 49/128$ *and* $c' = 5(1-2^{-1/4})/8$.
Hence,

$$\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H}\right) \leq \ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\right) + D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \ln\left(\frac{1}{\varepsilon}\right).$$

Proof. The key idea is that the control of the bracketing entropy of $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ can be recast into the control of the bracketing entropies of the associated mixture component density families. Specifically, from (6.7), each mean vector $\boldsymbol{\mu_k}$ of a p-dimensional Gaussian mixture density in $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ can be decomposed into a $|\boldsymbol{J_a^c}|$-dimensional null mean vector, a $|\boldsymbol{J_a}|$-dimensional constant mean vector and a $|\boldsymbol{J_r}|$-dimensional free mean vector:

$$\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{\begin{array}{l} \sum_{k=1}^K \pi_k\, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right); \\ \forall\, k : \boldsymbol{\mu_k}_{[\boldsymbol{J_r}]} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_r}|}, \quad \boldsymbol{\mu_k}_{[\boldsymbol{J_a}]} = \boldsymbol{\mu} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_a}|}, \quad \boldsymbol{\mu_k}_{[\boldsymbol{J_a^c}]} = \boldsymbol{0}, \\ \forall\, k : \pi_k > 0, \sum_{k=1}^K \pi_k = 1, \quad \sigma \in [a_\sigma, A_\sigma] \end{array}\right\}. \tag{6.57}$$

Consider the $(K-1)$-dimensional simplex Π_K defined by

$$\Pi_K := \left\{(\pi_1, \ldots, \pi_K) \in (0,1)^K; \sum_{k=1}^K \pi_k = 1\right\}$$

and the family of K-tuples of p-dimensional Gaussian densities

$$\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = \left\{ \begin{array}{l} \left(\Phi\left(\cdot \mid \boldsymbol{\mu_1}, \sigma^2 \boldsymbol{I}\right), \ldots, \Phi\left(\cdot \mid \boldsymbol{\mu_K}, \sigma^2 \boldsymbol{I}\right)\right); \\ \forall\, k : \boldsymbol{\mu_{k[J_r]}} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_r}|}, \quad \boldsymbol{\mu_{k[J_a]}} = \boldsymbol{\mu} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_a}|}, \quad \boldsymbol{\mu_{k[J_a^c]}} = \boldsymbol{0}, \\ \sigma \in [a_\sigma, A_\sigma] \end{array} \right\}.$$

Following the arguments developed by Maugis (2008) (proof of Proposition 7.A.2), it is easy to show that the study of the bracketing entropy of $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ can be recast into the study of the bracketing entropy of Π_K and $\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$:

Lemma 6.A.4. *For all* $\varepsilon \in]0,1]$,

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H}\right) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \Pi_K, \mathrm{d_H}\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H}\right)$$

where

$$\mathcal{N}_{[.]}\left(\varepsilon, \Pi_K, \mathrm{d_H}\right) \leq K(2\pi \mathrm{e})^{K/2} \left(\frac{1}{\varepsilon}\right)^{K-1}.$$

From Lemma 6.A.4, all the matter is to calculate an upper bound of the bracketing entropy of $\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$.

Let $\boldsymbol{f} = (f_1, \ldots, f_K) := \left(\Phi\left(\cdot \mid \boldsymbol{\mu_1}, \sigma^2 \boldsymbol{I}\right), \ldots, \Phi\left(\cdot \mid \boldsymbol{\mu_K}, \sigma^2 \boldsymbol{I}\right)\right) \in \mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. We want to find an ε-bracket for $\boldsymbol{f}$. We shall consider shrunk and dilated Gaussian densities.

Step 1. Construction of a net for the variance

Let $\delta \in]0,1]$ to be chosen later. Let $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. We construct a regular net for the variance $\sigma^2 \in [a_\sigma^2, A_\sigma^2]$. For $l \in \{2, \ldots, r\}$, we define $\sigma_l^2 = (1+\delta)^{1-\frac{l}{2}} A_\sigma^2$ where

$$r = \left\lceil 4 \frac{\ln\left(\frac{A_\sigma}{a_\sigma}\sqrt{1+\delta}\right)}{\ln(1+\delta)} \right\rceil \tag{6.58}$$

is chosen so that $\sigma_r^2 < a_\sigma^2 < \sigma_{r-1}^2 \leq \ldots \leq \sigma_2^2 = A_\sigma^2$.

Step 2. Construction of a net for the mean vectors

Let l be the unique integer in $\{2, \ldots, r\}$ such that $\sigma_{l+1}^2 < \sigma^2 \leq \sigma_l^2$. For all $k \in \{1, \ldots, K\}$, let $\boldsymbol{\nu_k} \in \mathbb{R}^p$ to be specified later. Consider the functions defined on $\mathbb{R}^p$ by

$$\left\{ \begin{array}{l} l_k(\boldsymbol{y}) = (1+\delta)^{-p}\, \Phi\left(\boldsymbol{y} \mid \boldsymbol{\nu_k}, (1+\delta)^{-\frac{1}{4}} \sigma_{l+1}^2 \boldsymbol{I}\right) \\ u_k(\boldsymbol{y}) = (1+\delta)^{p}\, \Phi\left(\boldsymbol{y} \mid \boldsymbol{\nu_k}, (1+\delta) \sigma_l^2 \boldsymbol{I}\right). \end{array} \right.$$

Put $\boldsymbol{l} = (l_1, \ldots, l_K)$ and $\boldsymbol{u} = (u_1, \ldots, u_K)$. We now determine δ and $(\boldsymbol{\nu_1}, \ldots, \boldsymbol{\nu_K})$ so that $\boldsymbol{l}$ and

$\boldsymbol{u}$ form an ε-bracket for $\boldsymbol{f}$. On the one hand, by using the calculation of the Hellinger distance between two multivariate Gaussian densities (Maugis and Michel, 2011b, Corollary 3) and by upper bounding some usual functions, we get that, for all $k \in \{1, \ldots, K\}$, $\mathrm{d}_{\mathrm{H}}^2(l_k, u_k) \leq cp^2\delta^2$ where $c = \mathrm{sh}(1) + 49/128$. Thus, we take $\delta = \varepsilon/(\sqrt{c}p)$ so that $\mathrm{d}_{\mathrm{H}}(l_k, u_k) \leq \varepsilon$. On the other hand, by using the ratio of two multivariate Gaussian densities (Maugis and Michel, 2011b, Corollary 2), the definition of σ_l and σ_{l+1}, the inequality $\ln(1+\delta) \geq \delta/2$ for all $\delta \in]0, 1]$ and the concavity of $\delta \mapsto 1 - (1+\delta)^{-1/4}$, we get that a sufficient condition for $l_k \leq f_k \leq u_k$ for all $k \in \{1, \ldots, K\}$ is

$$\|\boldsymbol{\mu_k} - \boldsymbol{\nu_k}\|_2^2 \leq c' p \delta^2 (1+\delta)^{\frac{2-l}{2}} A_\sigma^2 \tag{6.59}$$

where $c' = 5(1 - 2^{-1/4})/8$. Put

$$U_l := \mathbb{Z} \cap \left[\left\lfloor \frac{-A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma} \right\rfloor, \left\lfloor \frac{A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma} \right\rfloor \right]. \tag{6.60}$$

For all $j \in \boldsymbol{J_a}$, choose $u_j^{(l)} = \arg\min_{v_j \in U_l} |\mu_j - \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma v_j|$, and for all $k \in \{1, \ldots, K\}$, for all $j \in \boldsymbol{J_r}$, choose $u_{kj}^{(l)} = \arg\min_{v_{kj} \in U_l} |\mu_{kj} - \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma v_{kj}|$. Define $\boldsymbol{\nu_k}^{(l)} := \left(\nu_{k1}^{(l)}, \ldots, \nu_{kp}^{(l)}\right) \in [-A_\mu, A_\mu]^p$ by

$$\begin{aligned}
&\forall\, j \in \boldsymbol{J_a^c}, \quad \nu_{kj}^{(l)} = 0, \\
&\forall\, j \in \boldsymbol{J_a}, \quad \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma u_j^{(l)}, \\
&\forall\, j \in \boldsymbol{J_r}, \quad \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma u_{kj}^{(l)}.
\end{aligned}$$

Then, $\boldsymbol{\nu_k}^{(l)}$ fulfills (6.59) and we get a net for the mean vectors.

Step 3. Upper bound of the number of ε-brackets for $\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$
From step 1 and step 2, the family

$$\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\right) = \left\{ \begin{array}{l}
[\boldsymbol{l}, \boldsymbol{u}] := \{[l_1, u_1], \ldots, [l_K, u_K]\}\,;\, \forall\, k \in \{1, \ldots, K\} : \\
l_k = (1+\delta)^{-p}\, \Phi\left(\cdot \mid \left(\nu_{k1}^{(l)}, \ldots, \nu_{kp}^{(l)}\right), (1+\delta)^{-\frac{1}{4}} \sigma_{l+1}^2 \boldsymbol{I}\right) \\
u_k = (1+\delta)^{p}\, \Phi\left(\cdot \mid \left(\nu_{k1}^{(l)}, \ldots, \nu_{kp}^{(l)}\right), (1+\delta) \sigma_l^2 \boldsymbol{I}\right) \\
\text{with} \left\{ \begin{array}{ll}
\sigma_l^2 = (1+\delta)^{1-\frac{l}{2}} A_\sigma^2, & l \in \{2, \ldots, r\}, \\
\forall\, j \in \boldsymbol{J_r},\ \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma u_{kj}^{(l)}, & u_{kj}^{(l)} \in U_l \\
\forall\, j \in \boldsymbol{J_a},\ \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}} A_\sigma u_j^{(l)}, & u_j^{(l)} \in U_l \\
\forall\, j \in \boldsymbol{J_a^c},\ \nu_{kj}^{(l)} = 0 &
\end{array} \right.
\end{array} \right\}$$

is an ε-bracket covering for $\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. Therefore, an upper bound of the number of ε-brackets necessary to cover $\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is deduced from an upper bound of the cardinal of $\mathcal{B}_\varepsilon(\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})})$.

From (6.58) and (6.60), we have

$$\begin{aligned}\left|\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\right)\right| &\leq \sum_{l=2}^{r} \prod_{(k,j)\in\{1,\ldots,K\}\times\boldsymbol{J_r}} \prod_{j\in\boldsymbol{J_a}} \left(\frac{A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma}\right) \\ &\leq \left(\frac{2A_\mu}{\sqrt{c'}\delta A_\sigma}\right)^{K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|} \sum_{l=2}^{r}(1+\delta)^{\frac{(l-2)(K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|)}{4}} \\ &\leq \left(\frac{2A_\mu}{\sqrt{c'}\delta A_\sigma}\right)^{K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|} (r-1)(1+\delta)^{\frac{(r-2)(K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|)}{4}}.\end{aligned}$$

From (6.58), $(1+\delta)^{(r-2)/4} \leq (1+\delta)^{1/4}A_\sigma/a_\sigma \leq 2^{1/4}A_\sigma/a_\sigma$ and $r-1 \leq 4(A_\sigma/a_\sigma + 1/2)/\delta$, so

$$\begin{aligned}\left|\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\right)\right| &\leq 4\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|}\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right)\delta^{-(1+K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|)} \\ &\leq 4\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|}\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right)\left(\frac{\sqrt{c}p}{\varepsilon}\right)^{1+K|\boldsymbol{J_r}|+|\boldsymbol{J_a}|}. \qquad (6.61)\end{aligned}$$

Finally, Proposition 6.A.3 is derived from Lemma 6.A.4 and (6.61). □

6.A.2.2 Determination of a function $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$

The second step is to determine a function $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ defined by Property $(\boldsymbol{\mathcal{P}})$. From Proposition 6.A.3, for all $\xi > 0$,

$$\begin{aligned}&\int_0^\xi \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})}\, d\varepsilon \\ &\leq \xi\sqrt{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\right)} + \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\int_0^{\xi\wedge 1}\sqrt{\ln\left(\frac{1}{\varepsilon}\right)}\, d\varepsilon\,.\end{aligned}$$

In order to control the last term of the right-hand side of the last inequality, we apply the following technical result taken from Maugis and Michel (2011b):

Lemma 6.A.5. (Maugis and Michel, 2011b) For all $\xi \in]0,1]$,

$$\int_0^\xi \sqrt{\ln\left(\frac{1}{\varepsilon}\right)}\, d\varepsilon \leq \xi\left[\sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\xi}\right)}\right].$$

We obtain

$$\begin{aligned}
&\int_0^{\xi} \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})}\, d\varepsilon \\
&\leq \xi \sqrt{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\right)} + \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\,(\xi \wedge 1)\left[\sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\xi \wedge 1}\right)}\right] \\
&\leq \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\,\xi\left[\sqrt{\pi} + \sqrt{\frac{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\right)}{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}} + \sqrt{\ln\left(\frac{1}{\xi \wedge 1}\right)}\right].
\end{aligned}$$

But from (6.56) and the fact that $D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_a}|$, we have

$$\begin{aligned}
&\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, \boldsymbol{J_a}, p)\right) \\
&\leq \ln 4 + \frac{K}{2}\ln(2\pi \mathrm{e}) + (K-1)\ln 3 + \ln\left(\frac{A_\sigma}{a_\sigma} + \frac{1}{2}\right) + (K|\boldsymbol{J_r}| + |\boldsymbol{J_a}|)\ln\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right) + \ln K \\
&\quad + D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\ln(3\sqrt{c}p) \\
&\leq \left[\ln 4 + \frac{\ln(2\pi\mathrm{e})}{2} + \ln 3 + \ln\left(\frac{A_\sigma}{a_\sigma} + \frac{1}{2}\right) + \ln\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right) + 1 + \ln(3\sqrt{c}p)\right] D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \\
&\leq \left[\ln\left(\frac{72\sqrt{2\pi\mathrm{e}}\,2^{5/4}\mathrm{e}\sqrt{c}}{\sqrt{c'}}\right) + \ln\left[\frac{A_\sigma}{a_\sigma}\left(1 + \frac{A_\mu}{a_\sigma}\right)\right] + \ln p\right] D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int_0^{\xi} \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}, \mathrm{d_H})}\, d\varepsilon \\
&\leq \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\,\xi\left[\sqrt{\pi} + \sqrt{\ln\left(\frac{72\sqrt{2\pi\mathrm{e}}\,2^{5/4}\mathrm{e}\sqrt{c}}{\sqrt{c'}}\right) + \ln\left[\frac{A_\sigma}{a_\sigma}\left(1 + \frac{A_\mu}{a_\sigma}\right)\right] + \ln p} + \sqrt{\ln\left(\frac{1}{\xi \wedge 1}\right)}\right] \\
&\leq \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\,\xi\left[6 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1 + \frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p} + \sqrt{\ln\left(\frac{1}{\xi \wedge 1}\right)}\right].
\end{aligned}$$

Consequently, by putting

$$B(A_\mu, A_\sigma, a_\sigma, p) := 6 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1 + \frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p},$$

we get that the function $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ defined on $\mathbb{R}_+^\star$ by

$$\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\xi) = \sqrt{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}\,\xi\left[B(A_\mu, A_\sigma, a_\sigma, p) + \sqrt{\ln\left(\frac{1}{\xi \wedge 1}\right)}\right]$$

satisfies (6.1). Besides, $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is non-decreasing and $\xi \mapsto \Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\xi)/\xi$ is non-increasing, so $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ is convenient.

6.A.2.3 Lower bound of the penalty function

Finally, according to the lower bound (6.4) of the penalty function, we need to find an upper bound of ξ_* satisfying $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\xi_*) = \sqrt{n}\,\xi_*^2$ and to calculate the weights $x_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ to take into account the richness of the family $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. This can be done along the proofs of Maugis and Michel (2011b) by replacing the dimension of the models considered by Maugis and Michel (2011b) by the dimension $D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$ of our models $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}$. This leads to the two following lemmas:

Lemma 6.A.6. *Consider ξ_* such that $\Psi_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}(\xi_*) = \sqrt{n}\,\xi_*^2$. Then,*

$$\xi_*^2 \leq \frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}\left[2B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1 \wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}}\right)\right]. \tag{6.62}$$

Lemma 6.A.7. *Consider the weight family $\left\{x_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\right\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ defined by*

$$x_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} = D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \ln\left(\frac{8\mathrm{e}p}{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \wedge p}\right). \tag{6.63}$$

Then, we have $\sum_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}} \mathrm{e}^{-x_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}} \leq 1.$

From (6.4), (6.62) and (6.63), we can apply Theorem 6.2.1 as soon as there exists $\kappa > 0$ such that $\mathrm{pen}\,(K, \boldsymbol{J_r}, \boldsymbol{J_a})$ satisfies for all $(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \in \mathcal{M}_{(r,a)}$:

$$\mathrm{pen}\,(K, \boldsymbol{J_r}, \boldsymbol{J_a}) \geq \kappa\frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}\left[2B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1 \wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}}{n}}\right)\right.$$
$$\left. +(1 \vee \tau)\ln\left(\frac{8\mathrm{e}p}{D_{(K,\boldsymbol{J_r},\boldsymbol{J_a})} \wedge p}\right)\right].$$

Applying Theorem 6.2.1 leads to Theorem 6.2.2. □

Chapter A

Alternatives to our Lasso-MLE procedure

Contents

A.1. Two alternative procedures to our Lasso-MLE procedure 255

A.1.1. The R-EC procedure 256

A.1.2. The A-R procedure 257

A.1.3. Review of the three procedures 259

A.2. Performance of the three procedures on simulated data 259

A.2.1. Comparison between the three procedures 261

A.2.2. Numerical problems for the A-R procedure 263

A.2.3. Conclusion 266

ABSTRACT

In this chapter, we present two procedures we envisaged during this thesis as possible alternatives to our Lasso-MLE procedure. For reasons we explain below, we decided to keep the Lasso-MLE procedure described in Section 4.5.4. Nonetheless, the two other procedures both present some interesting points. That is why we mention them and discuss on them here. First, we describe these alternative procedures. Then, we compare them to our Lasso-MLE procedure on two simulated datasets.

A.1 Two alternative procedures to our Lasso-MLE procedure

Our Lasso-MLE procedure described in Section 4.5.4 is not the first procedure we thought about. In fact, the key idea of our procedure is the parameter estimation by the MLE rather than by the Lasso, and we first built a simpler procedure mainly focused on this crucial point. Our initial idea was the following: first use the Lasso on the empirically centered data to construct a collection of sets of relevant variables and derive a model collection, then estimate the parameters on the empirically centered data by the MLE in each model, finally choose a model thanks to a non-asymptotic model selection criterion derived from the slope heuristics and get a data clustering according to the MAP principle. By estimating the parameters by the MLE rather than by the Lasso, this procedure is expected to lead to better estimations than Pan and Shen's procedure and thus to a better clustering by the MAP principle. In the sequel, we call this procedure the R-EC (Relevant-Empirically Centered) procedure.

We think that this procedure can provide good results in a clustering viewpoint. Besides, it does not require the notion of active variables, contrary to our Lasso-MLE procedure. Yet, the estimation of the parameters is performed on the empirically centered data, so it does not allow to deal with problems involving sparse density estimation recovery such as for curve clustering. Indeed, estimation of the density of the non-empirically centered data from the estimation of the density of the empirically centered data requires to estimate each empirical mean, which breaks sparsity. To overcome this problem, we introduced the notion of active variables that induces sparsity and enables to perform estimation of the parameters directly on the non-empirically centered data. This additional process led to our Lasso-MLE procedure.

In our Lasso-MLE procedure, the model collection construction is divided in two steps: first, we detect sets of irrelevant variables, then we detect sets of inactive variables among each set of irrelevant variables so as to reduce dimension and perform reliable estimation. But, another viewpoint can be considered by reversing these two steps, that is to say by first looking for the active variables and then searching for the relevant variables among the active variables. In the sequel, we call this alternative procedure the A-R (Active-Relevant) procedure to indicate the order of the two steps. In contrast, we refer to our Lasso-MLE procedure as the R-A (Relevant-Active) procedure.

From a theoretical viewpoint, these two procedures mainly differ in the way we look at the active variables. By first searching for the relevant variables, our R-A procedure focuses on these variables and thus on clustering, while detecting the inactive variables is above all used to reduce dimension. In particular, our procedure has no more meaning when data come from one single cluster (no mixture). On the opposite, the A-R procedure first aims at looking for the variables actually present in the model, no matter a mixture is to be considered, while identifying the relevant variables for the clustering comes in a second step. The advantage of this procedure is that it remains meaningful in the case of homogeneous data by simply dropping the second step. Thus, it may seem more natural than our R-A procedure.

Below, we detail the steps of the R-EC procedure and the A-R procedure. They are to be compared with the steps of our R-A procedure described in Section 4.5.4. We keep the notations used in Chapter 4. We assume we observe a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ from probability distribution with density s to be estimated by a finite Gaussian mixture density in a clustering purpose.

A.1.1 The R-EC procedure

Assume that we observe a dataset for which the notion of active variables introduced in Section 4.5.2.1 is not meaningful. In this case, one can not carry out dimensional reduction by eliminating the inactive variables. Then, performing parameter estimation on the non-empirically centered dataset is not sensible since it would involve much too many parameters to estimate and would conduct to degenerate models (see Section 4.5.2). To avoid such problems, one way is to perform parameter estimation on the empirically centered dataset, so that all the constant means through the clusters become null means and do not have to be estimated. This leads to the following R-EC procedure, which is a mixture of Pan and Shen's procedure and Maugis and Michel's procedure described in Section 4.3.

Steps 0-1.
As it is done at Step 0 of Pan and Shen's procedure, we empirically center the dataset $\boldsymbol{Y}$ to get an empirically centered dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{n}})$. The density $\overline{s}$ of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$ is to be estimated to get a clustering of the dataset $\overline{\boldsymbol{Y}}$.
We construct the same model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ as the model collection constructed at Step 1 of Pan and Shen's procedure (see Section 4.3.2):

$$\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I}) \ \sum_{k=1}^K \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2\boldsymbol{I}); \\ \overline{\boldsymbol{\theta}} = (\pi_1, \ldots, \pi_K, \overline{\boldsymbol{\mu}}_{\boldsymbol{1}}, \ldots, \overline{\boldsymbol{\mu}}_{\boldsymbol{K}}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{A.1}$$

Each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ is the set of finite Gaussian mixture densities with K components and $\boldsymbol{J_r}$ as index set representing the relevant variables.

Thanks to empirical centering, the mean parameters on $\boldsymbol{J_r^c}$ are taken to zero, so the dimension of a model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ is $K(1+|\boldsymbol{J_r}|) \leq n$ for small enough $\boldsymbol{J_r}$. No dimensional reduction step is necessary to perform parameter estimation[1] on the empirically centered dataset.

Steps 2-4.
Steps 2, 3 and 4 of the R-EC procedure are similar to Steps 2, 3 and 4 of Maugis and Michel's procedure described in Section 4.3.1. The only difference is that the model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ considered for Maugis and Michel's procedure is replaced by the model collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ defined by (A.1).

A.1.2 The A-R procedure

In our R-A procedure, we first construct a collection of sets of relevant variables, and then a collection of sets of active variables. Another approach consists in reversing the order of the two constructions. This leads to the following A-R procedure.
I. Fix $K \in \mathcal{K}$.

<u>Detection of the active variables</u>

(a) Consider the original (non-empirically centered) dataset $\boldsymbol{Y}$ and

$$\mathcal{S}_K = \left\{ \begin{array}{l} s_{\boldsymbol{\theta}} = \sum_{k=1}^K \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 \boldsymbol{I}\right); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Theta_K \end{array} \right\}$$

with $\Theta_K = \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+$. The inactive variables are expected to be detected by penalizing the empirical contrast

$$\gamma_n\left(s_{\boldsymbol{\theta}}\right) = -\frac{1}{n}\sum_{i=1}^n \ln\left(s_{\boldsymbol{\theta}}(\boldsymbol{Y_i})\right)$$

by an ℓ_1-penalty on the mean parameters proportional to

$$|\boldsymbol{\theta}|_1 := \sum_{j=1}^p \sum_{k=1}^K |\mu_{kj}|.$$

Introduce G_K a grid of regularization parameters.

(b) Fix $\lambda \in G_K$. Consider the Lasso estimator defined by

$$\hat{\boldsymbol{\theta}}_{(K,\lambda)} = \underset{\boldsymbol{\theta}\in\Theta_K}{\arg\min}\left\{\gamma_n(s_{\boldsymbol{\theta}}) + \lambda|\boldsymbol{\theta}|_1\right\}.$$

[1] Let us stress that we do not estimate the parameters by the Lasso estimators in each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$ as it is done at Step 1 of Pan and Shen's procedure. We shall perform estimation at Step 2 by the MLE in each model $\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}$.

We compute $\hat{\boldsymbol{\theta}}_{(K,\lambda)} = (\hat{\pi}_k, \hat{\mu}_{kj}, \hat{\sigma})_{1\leq k\leq K, 1\leq j\leq p}$ by Pan and Shen's EM algorithm described in Section 4.A.2[2]. The index set $\boldsymbol{J}_{(K,\lambda)} = \{j \in \{1,\ldots,p\} : \exists\, k \text{ such that } \hat{\mu}_{kj} \neq 0\}$ represents the active variables selected by the Lasso $\hat{\boldsymbol{\theta}}_{(K,\lambda)}$.

(c) By varying $\lambda \in G_K$, we get a collection $\mathcal{J}_K = \cup_{\lambda\in G_K} \boldsymbol{J}_{(K,\lambda)}$ of index sets representing a collection of sets of active variables. As regards the estimation of s, this collection $\mathcal{J}_K$ leads to a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_a})}\}_{\boldsymbol{J_a}\in\mathcal{J}_K}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_a})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I}) \ \sum_{k=1}^K \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_a}]} \mid \boldsymbol{\mu_k}, \sigma^2\boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1,\ldots,\pi_K,\boldsymbol{\mu_1},\ldots,\boldsymbol{\mu_K},\sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_a}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{A.2}$$

Each model $\mathcal{S}_{(K,\boldsymbol{J_a})}$ is the set of finite Gaussian mixture densities with K components and $\boldsymbol{J_a}$ as index set representing the active variables.

Detection of the relevant variables among the active variables

1. Fix $\boldsymbol{J_a} \in \mathcal{J}_K$.

   (a) We empirically center the dataset $\boldsymbol{Y} = (\boldsymbol{Y_1},\ldots,\boldsymbol{Y_n})$ to get an empirically centered dataset $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_{\boldsymbol{1}},\ldots,\overline{\boldsymbol{Y}}_{\boldsymbol{n}})$. Denote by $\overline{s}$ the density of $\overline{\boldsymbol{Y}}_{\boldsymbol{i}}$. From (A.2), the density $\overline{s}$ restricted on $\boldsymbol{J_a}$ can be modeled by a $|\boldsymbol{J_a}|$-dimensional mixture Gaussian density $s_{\overline{\boldsymbol{\theta}}} = \sum_{k=1}^K \pi_k\, \Phi(\cdot \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{k}}, \sigma^2\boldsymbol{I})$ with $\overline{\boldsymbol{\theta}} = (\pi_1,\ldots,\pi_K,\overline{\boldsymbol{\mu}}_{\boldsymbol{1}},\ldots,\overline{\boldsymbol{\mu}}_{\boldsymbol{K}},\sigma) \in \Theta_{(K,\boldsymbol{J_a})} := \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_a}|}\right)^K \times \mathbb{R}_+$. The irrelevant variables are expected to be detected by penalizing the empirical contrast

   $$\gamma_n\left(s_{\overline{\boldsymbol{\theta}}}\right) = -\frac{1}{n}\sum_{i=1}^n \ln\left(s_{\overline{\boldsymbol{\theta}}}\left(\overline{\boldsymbol{Y}}_{\boldsymbol{i}[\boldsymbol{J_a}]}\right)\right)$$

   by an ℓ_1-penalty on the mean parameters proportional to

   $$|\overline{\boldsymbol{\theta}}|_1 := \sum_{j\in\boldsymbol{J_a}}\sum_{k=1}^K |\overline{\mu}_{kj}|.$$

   Introduce $G_{(K,\boldsymbol{J_a})}$ a grid of regularization parameters.

   (b) Fix $\lambda \in G_{(K,\boldsymbol{J_a})}$. Consider the Lasso estimator of $\overline{\boldsymbol{\theta}}$ defined by

   $$\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_a},\lambda)} = \underset{\overline{\boldsymbol{\theta}}\in\Theta_{(K,\boldsymbol{J_a})}}{\arg\min}\ \left\{\gamma_n(s_{\overline{\boldsymbol{\theta}}}) + \lambda|\overline{\boldsymbol{\theta}}|_1\right\}.$$

[2]We run this algorithm on the dataset $\boldsymbol{Y}$, not on the dataset $\overline{\boldsymbol{Y}}$ considered in Section 4.A.2.

We compute $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_a},\lambda)} = (\hat{\pi}_k, \widehat{\overline{\mu}}_{kj}, \hat{\sigma})_{1\leq k\leq K, j\in \boldsymbol{J_a}}$ by Pan and Shen's EM algorithm described in Section 4.A.2 and restricted to the dataset $\overline{\boldsymbol{Y}}_{[\boldsymbol{J_a}]}$. The index set $\boldsymbol{J}_{(K,\boldsymbol{J_a},\lambda)} = \{j \in \boldsymbol{J_a} : \exists\, k \text{ such that } \widehat{\overline{\mu}}_{kj} \neq 0\}$ represents the relevant variables selected by the Lasso $\widehat{\overline{\boldsymbol{\theta}}}_{(K,\boldsymbol{J_a},\lambda)}$.

(c) By varying $\lambda \in G_{(K,\boldsymbol{J_a})}$, we get an index set collection $\mathcal{J}_{(K,\boldsymbol{J_a})} = \cup_{\lambda\in G_{(K,\boldsymbol{J_a})}} \boldsymbol{J}_{(K,\boldsymbol{J_a},\lambda)}$ representing a collection of sets of relevant variables among the active variables indexed by $\boldsymbol{J_a}$. We derive a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}\}_{\boldsymbol{J_r}\in\mathcal{J}_{(K,\boldsymbol{J_a})}}$ with

$$\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}); \\ s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_a^c}]} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I})\ \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{\mu}, \sigma^2\boldsymbol{I})\ \sum_{k=1}^K \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2\boldsymbol{I}); \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_r^c}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+ \end{array} \right\}. \tag{A.3}$$

Each model $\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}$ is the set of finite Gaussian mixture densities with K components, $\boldsymbol{J_a}$ as index set representing the active variables and $\boldsymbol{J_r}$ as index set representing the relevant variables among the active variables.

2. By varying $\boldsymbol{J_a} \in \mathcal{J}_K$, we get a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}\}_{(\boldsymbol{J_a},\boldsymbol{J_r})\in\mathcal{J}_K\times\mathcal{J}_{(K,\boldsymbol{J_a})}}$.

II. By varying $K \in \mathcal{K}$, we get a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})\in\mathcal{K}\times\mathcal{J}_K\times\mathcal{J}_{(K,\boldsymbol{J_a})}}$. We put

$$\mathcal{M}_{(a,r)} = \{(K, \boldsymbol{J_a}, \boldsymbol{J_r});\ K \in \mathcal{K},\ \boldsymbol{J_a} \in \mathcal{J}_K,\ \boldsymbol{J_r} \in \mathcal{J}_{(K,\boldsymbol{J_a})}\}.$$

Steps 2-4.
Steps 2, 3 and 4 of the A-R procedure are similar to Steps 2, 3 and 4 of the R-A procedure described in Section 4.5.4. The only difference is that the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ considered for the R-A procedure is replaced by the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})\in\mathcal{M}_{(a,r)}}$ defined by (A.3).

A.1.3 Review of the three procedures

Table A.1 summarizes the similarities and dissimilarities between the R-A procedure, the A-R procedure and the R-EC procedure.

A.2 Performance of the three procedures on simulated data

Here, we compare the R-A procedure, the A-R procedure and the R-EC procedure on the two simulated datasets studied in Section 5.4.1 in order to explain why we have decided to keep the R-A procedure rather than the two other procedures. Both of these datasets involve relevant variables as

procedure	model collection	parameter estimation	model selection	clustering
R-A	1. ℓ_1-penalization on the mean parameters to detect the irrelevant variables 2. ℓ_1-penalization on the mean parameters to detect the inactive variables $\Rightarrow$ data-driven collection on sets of relevant variables and active variables $\Rightarrow$ collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_a})\in\mathcal{M}_{(r,a)}}$ of models with K clusters, $\boldsymbol{J_r}$ as representative set of relevant variables and $\boldsymbol{J_a}$ as representative set of active irrelevant variables	MLE	data-driven slope and ln-slope estimators	MAP
A-R	1. ℓ_1-penalization on the mean parameters to detect the inactive variables 2. ℓ_1-penalization on the mean parameters to detect the irrelevant variables $\Rightarrow$ data-driven collection of sets of active variables and relevant variables $\Rightarrow$ collection $\{\mathcal{S}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_a},\boldsymbol{J_r})\in\mathcal{M}_{(a,r)}}$ of models with K clusters, $\boldsymbol{J_a}$ as representative set of active irrelevant variables and $\boldsymbol{J_r}$ as representative set of relevant variables	MLE	data-driven slope and ln-slope estimators	MAP
R-EC	ℓ_1-penalization on the mean parameters to detect the irrelevant variables $\Rightarrow$ data-driven collection of sets of relevant variables $\Rightarrow$ collection $\{\overline{\mathcal{S}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}_r}$ of models with K clusters and $\boldsymbol{J_r}$ as representative set of relevant variables	MLE	data-driven slope and ln-slope estimators	MAP

Table A.1: Comparison between the three Lasso-MLE procedures. The operations performed on the original dataset $\boldsymbol{Y}$ are in green, whereas the operations performed on the empirically centered dataset $\overline{\boldsymbol{Y}}$ are in red.

well as active irrelevant variables. So, they are interesting to compare the performance of the R-A procedure and the A-R procedure which differ in the selection order of these two kinds of variables. In the sequel, we denote by

- "TR" the true relevant variables (variables declared as relevant by the procedure and actually relevant);
- "TA" the true active irrelevant variables (variables declared as active irrelevant by the procedure and actually active irrelevant);
- "FR" the false relevant variables (variables declared as relevant by the procedure and actually irrelevant);
- "FA" the false active irrelevant variables (variables declared as active irrelevant by the procedure and actually inactive).

A.2.1 Comparison between the three procedures

First, we compare the three procedures on the simulated dataset studied in Section 5.4.1.2. For this dataset, there are 10 relevant variables, 10 active irrelevant variables and 1066 inactive variables. The results are summarized in Table A.2.

procedure	estimator	{TR,TA;FR,FA}	K	ARI	$\mathrm{KL}(s, \hat{s})$
R-A	oracle	$\{5.4, 5.0; 1.3, 0.2\}$	$\{0, 0, 19, 1, 0\}$	0.87	0.27
	slope	$\{5.4, 5.4; 1.3, 9.1\}$	$\{0, 0, 13, 2, 0\}$	0.87	0.44
	ln-slope	$\{4.9, 5.0; 0.7, 0.2\}$	$\{0, 0, 19, 1, 0\}$	0.86	0.31
A-R	oracle	$\{5.1, 5.1; 1.0, 9.8\}$	$\{0, 1, 10, 2, 0\}$	0.85	0.34
	slope	$\{5.1, 5.1; 0.9, 21.8\}$	$\{0, 1, 10, 2, 0\}$	0.85	0.50
	ln-slope	$\{4.6, 4.2; 0.0, 1.6\}$	$\{0, 1, 10, 2, 0\}$	0.84	0.48
R-EC	oracle	$\{5.4, 0.0; 1.7, 0.0\}$	$\{0, 0, 19, 1, 0\}$	0.87	2.71
	slope	$\{5.4, 0.0; 1.5, 0.0\}$	$\{0, 0, 19, 1, 0\}$	0.87	2.73

Table A.2: Mean number {TR,TA;FR,FA}, number of times $\{\nu_2, \nu_3, \nu_4, \nu_5, \nu_6\}$ a clustering with respectively $K = 2$, $K = 3$, $K = 4$, $K = 5$ and $K = 6$ components is selected, mean ARI and mean Kullback-Leibler divergence over the 20 simulations, except for the A-R procedure that fails 7 times and except for the slope estimator that fails 5 times for the R-A procedure.

Secondly, we compare the three procedures on the simulated dataset studied in Section 5.4.1.1. For this dataset, there are 25 relevant variables, 25 active irrelevant variables and 1036 inactive variables. The results are summarized in Table A.3.

procedure	estimator	{TR,TA;FR,FA}	K	ARI	$\mathrm{KL}(s, \hat{s})$
R-A	oracle	$\{25.0, 25.0; 0.2, 0.1\}$	$\{0, 18, 2\}$	0.95	0.25
	slope	$\{25.0, 25.0; 0.7, 5.5\}$	$\{0, 20, 0\}$	0.96	0.37
	ln-slope	$\{24.6, 25.0; 0.2, 0.8\}$	$\{0, 20, 0\}$	0.96	0.31
A-R	oracle	$\{24.7, 25.0; 2.2, 36.2\}$	$\{0, 9, 3\}$	0.88	0.70
	slope	$\{22.1, 25.0; 1.0, 31.0\}$	$\{0, 9, 3\}$	0.87	0.78
	ln-slope	$\{22.0, 25.0; 0.0, 16.7\}$	$\{0, 10, 2\}$	0.82	1.16
R-EC	oracle	$\{25.0, 0.0; 0.2, 0.0\}$	$\{0, 18, 2\}$	0.94	2.68
	slope	$\{25.0, 0.0; 1.0, 0.0\}$	$\{0, 19, 1\}$	0.95	2.77

Table A.3: Mean number {TR,TA;FR,FA}, number of times $\{\nu_1, \nu_2, \nu_3\}$ a clustering with respectively $K = 1$, $K = 2$ and $K = 3$ components is selected, mean ARI and mean Kullback-Leibler divergence over the 20 simulations, except for the A-R procedure that fails 8 times.

General comments

Since the R-EC procedure runs on empirically centered datasets for which the active irrelevant variables are mixed up with the inactive variables, this procedure is not adapted to detect the active irrelevant variables, hence the null values for TA and FA in Table A.2 and in Table A.3. Yet, this is not expected to alter the clustering since only the identification of the relevant variables is crucial to partition the data.

The A-R procedure is faced with numerical problems because the first Lasso algorithm runs on the non-empirically centered data. For some simulations, numerical divergence prevents from reaching models with small or moderate dimensions. These simulations are not taken into account to establish the results of Table A.2 and Table A.3. We refer to Section A.2.2 for some discussion on these numerical problems.

Theoretical performance: comparison between the oracle models

For both datasets, the best oracle model as regards variable selection, clustering and prediction is the oracle for the R-A procedure. That is why we have decided to keep this procedure.

For the R-EC procedure, the oracle achieves similar results as the oracle for the R-A procedure as regards selection of the relevant variables and clustering. Yet, the R-EC procedure is performed on the empirically centered data, so the 1086 empirical means have to be added to estimate the density of the non-empirically centered data, which results in poor prediction performance.

For the A-R procedure, the oracle model is often too complex and contains many FA variables because of the numerical problems encountered by this procedure. Thus, its clustering and prediction performances are deteriorated. This procedure is the least satisfactory.

Practical performance: comparison between the estimators

For the R-EC procedure, the $\ln$-slope estimator fails to be computed. In fact, for this procedure, only one Lasso algorithm is run, so few models are constructed and the model collection is not rich enough to justify a logarithm term in the penalty. On the opposite, two Lasso algorithms are interlocked for the R-A and the A-R procedures and many more models are constructed, often justifying a logarithm term in the penalty. For some simulations, the model collection for the R-A procedure is so rich that the slope estimator fails to be computed.

For the R-EC procedure, the slope estimator is close to the oracle model. For the R-A procedure, the $\ln$-slope estimator misses a few relevant variables. Yet, it is closer to the oracle than the slope estimator which selects many FA variables. For the A-R procedure, the slope estimator and the $\ln$-slope estimator are not so close to the oracle. They tend to miss some relevant variables and/or to catch many FA variables.

A.2.2 Numerical problems for the A-R procedure

For the A-R procedure, the first Lasso algorithm runs on the non-empirically centered data. Therefore, for high-dimensional datasets, the EM algorithm calculating the Lasso solution is faced with numerical problems, particularly to update the posterior probabilities at the E step. These problems are more likely to occur when the true number of clusters is small and the mixing proportions are very different, as for the second simulated dataset studied in Section A.2.1. The EM algorithm tends to gather the clusters, so it sets to zero all the corresponding mean coefficients together and declares some variables inactive while they may actually be active and even relevant. Thus, the first Lasso algorithm produces sets of active variables not containing all the true relevant variables. Consequently, at the end of the procedure, to compensate for this lack of relevant variables, the A-R oracle model picks up some active irrelevant variables but also many inactive variables (see Figure A.2, at the bottom), which explains the great number of FA variables selected in Table A.2 and Table A.3.

On the contrary, the R-A procedure is not faced with this numerical problem because the first Lasso algorithm runs on the empirically centered data. When the second Lasso algorithm runs on the non-empirically centered data, such numerical problems may occur for the largest sets of relevant variables selected by the first Lasso algorithm, but they do not occur for the smallest sets of relevant variables. Then, the R-A oracle chooses a model among the smallest models (see Figure A.1, at the bottom), that is to say a model for which no numerical problem occurs.

Figure A.1 and Figure A.2 highlight this difference for one simulation of the second dataset studied in Section A.2.1. For this dataset, there are 25 relevant variables and 25 active irrelevant variables. For both procedures, models with various dimensions are generated by the first Lasso algorithm by varying the regularization parameter. For the R-A procedure, this first Lasso algorithm is expected to

detect the relevant variables, so the number of TR variables in each model is expected to increase until 25 as the model dimension increases. Figure A.1 (at the top) shows that this is actually what happens.

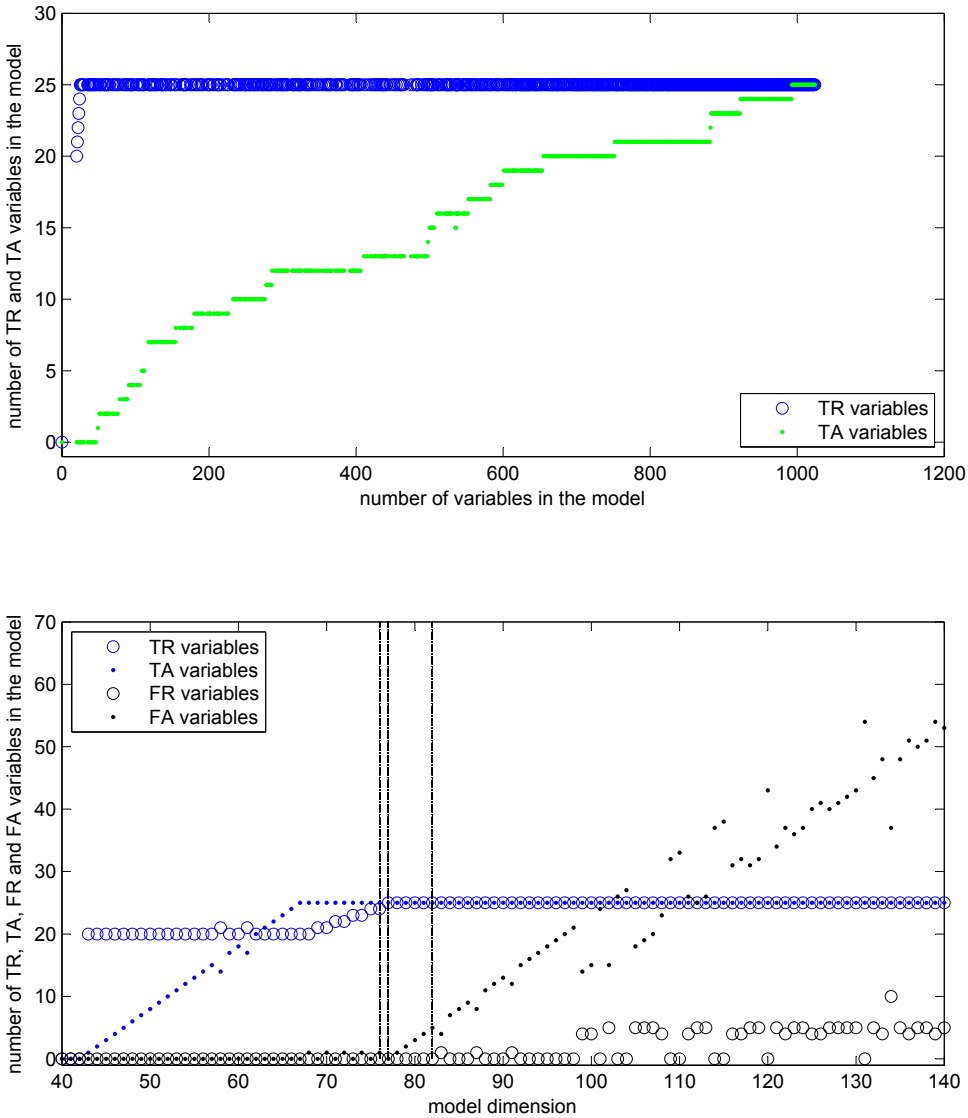

Figure A.1: For one simulation of the second dataset in Section A.2.1, model collection obtained for the R-A procedure. At the top: number of TR and TA variables in the model collection generated by the first Lasso algorithm (this algorithm is expected to detect only the 25 relevant variables). At the bottom: number of TR, TA, FR and FA variables in the final model collection obtained after the two successive Lasso algorithms. Only models with dimension between 40 and 140 are shown. The dotted lines indicate the model selected by the $\ln$-slope estimator (on the left) and by the slope estimator (on the right) as well as the oracle model (in the middle). The true model dimension is 77.

For the A-R procedure, the first Lasso algorithm is expected to detect all the active variables (the relevant variables and the active irrelevant variables), so the number of TR variables in each model is expected to increase until 25 and the number of TA variables in each model is expected to increase until 25 as the model dimension increases. Yet, Figure A.2 (at the top) shows that the Lasso does not manage to catch all the 25 relevant until (too) high model dimensions.

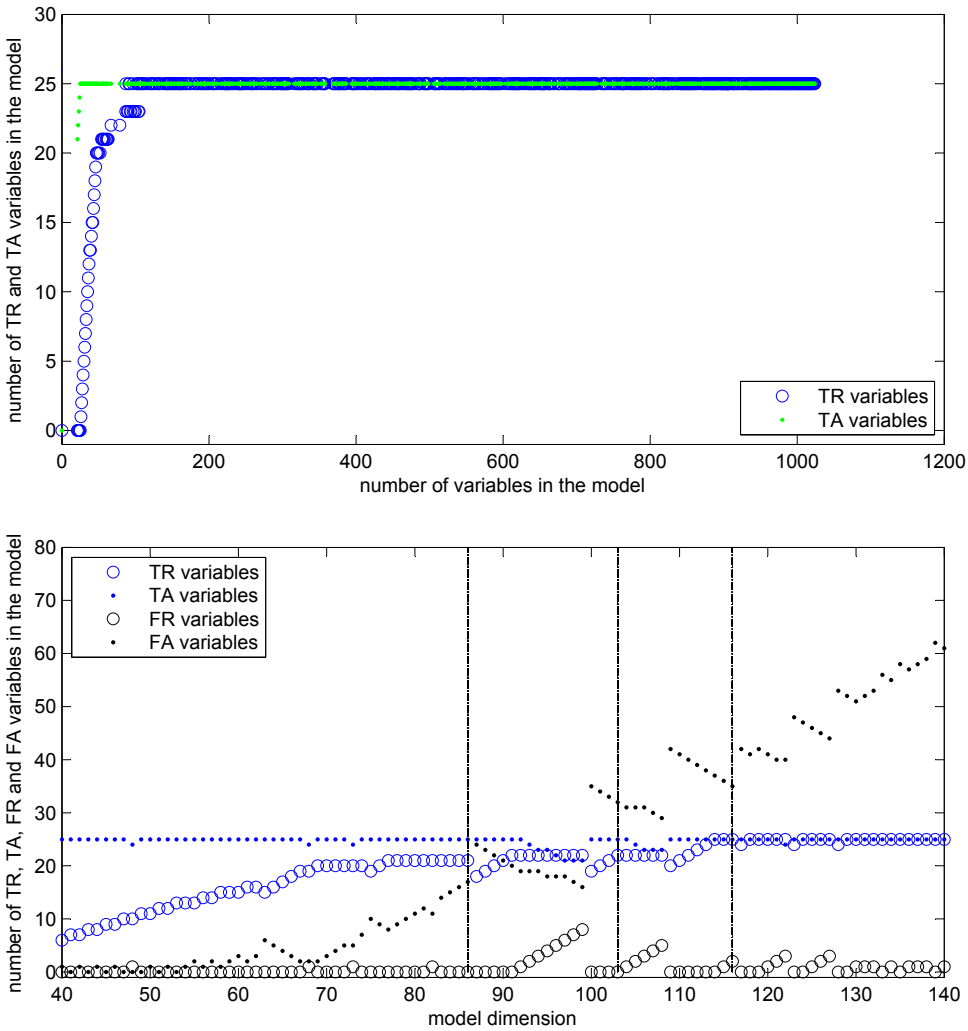

Figure A.2: For one simulation of the second dataset in Section A.2.1, model collection obtained by the A-R procedure. At the top: number of TR and TA variables in the model collection generated by the first Lasso algorithm (this algorithm is expected to detect both the 25 relevant and the 25 active irrelevant variables). At the bottom: number of TR, TA, FR and FA variables in the final model collection obtained after the two successive Lasso algorithms. Only models with dimension between 40 and 140 are shown. The dotted lines indicate the model selected by the ln-slope estimator (on the left) and by the slope estimator (in the middle) as well as the oracle model (on the right). The true model dimension is 77.

A.2.3 Conclusion

Because of the numerical problems encountered by the A-R procedure and because of the poor density estimation of the R-EC procedure, we have chosen to favor the R-A procedure as final procedure for our Lasso-MLE procedure. Yet, on the one hand, as highlighted in Table A.2 and Table A.3, the R-EC procedure seems to be a good surrogate for the R-A procedure in a clustering purpose if the notion of active variable is not meaningful. Besides, since it requires only one Lasso algorithm, it is a bit faster than the R-A procedure. On the other hand, if one manages to get rid of the numerical problems for the A-R procedure, this procedure may reveal competitive.

Bibliographie

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.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

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.

Bach, F. (2008). Bolasso : model consistent Lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3) :803–821.

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

Barron, A., Cohen, A., Dahmen, W., and DeVore, R. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*, 36(1).

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.

Beninel, F., Biernacki, C., Bouveyron, C., Jacques, J., and Lourme, A. (2012). *Parametric link models for knowledge transfer in statistical learning*. Nova Publishers.

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.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4) :1705–1732.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, 51(2) :587–600.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268.

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

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration inequalities with applications*.

Brusco, M. J. and Cradit, J. D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, 66(2) :249–270.

Bühlmann, P. and van de Geer, S. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3 :1360–1392.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. *Learning theory*, pages 379–391.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4) :1674–1697.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007b). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1 :169–194.

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

Candes, E. and Tao, T. (2007). The Dantzig selector : Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6) :2313–2351.

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical report, Université Paris-Sud 11.

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

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.

Chen, S., Donoho, D., and Saunders, M. (1999). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1) :33–61.

Chiou, J. and Li, P. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(4) :679–699.

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.

Cohen, S. and Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *arXiv :1103.2021*.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B.*, 39(1) :1–38.

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.

Devaney, M. and Ram, A. (1997). Efficient feature selection in conceptual clustering. *Machine Learning : Proceedings of the Fourteenth International Conference*, pages 92–97.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1997). Universal near minimaxity of wavelet shrinkage. *Festschrift for Lucien Le Cam*, pages 183–218.

Dudley, R. (2010). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Selected Works of RM Dudley*, pages 125–165.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2) :407–499.

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

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, 5(2) :205–228.

Garcia-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, 22(2) :185–201.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6) :971–988.

Grün, B., Leisch, F., et al. (2007). Applications of finite mixtures of regression models. *URL : http ://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf.*

Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory Series A*, 69 :217–232.

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.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462) :397–408.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11) :1370–1386.

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.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A*, 62(1) :49–66.

Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *The Annals of Statistics*, 45(1) :7–57.

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.

Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, 147(1) :39–57.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4) :1261–1269.

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

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic Press.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, 37(4) :362–372.

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.

Massart, P. and Meynet, C. (2010). An ℓ_1-oracle inequality for the Lasso. *ArXiv 1007.4791*.

Massart, P. and Meynet, C. (2011). The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5 :669–687.

Maugis, C. (2008). *Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes.* PhD thesis, Université Paris-Sud 11.

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.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience, New York.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3) :1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1) :246–270.

Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. PhD thesis, Université Paris-Sud 11.

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

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2007b). *Wavelets and their Applications*. Wiley Online Library.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8 :1145–1164.

Pisier, G. (1999). *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press.

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

Ramsay, J. (2006). *Functional data analysis*. Wiley Online Library.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(2) :305–332.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, pages 195–239.

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

Rivoirard, V. (2006). Nonlinear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli*, 12(4) :609–632.

Rossi, F., Conan-Guez, B., and El Golli, A. (2004). Clustering functional data with the SOM algorithm. In *Proceedings of ESANN*, pages 305–312.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

Sharan, R., Elkon, R., and Shamir, R. (2002). Cluster analysis and its applications to gene expression data. In *Ernst Schering Workshop on Bioinformatics and Genome Analysis*. Springer Verlag.

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

Talagrand, M. (1996). Majorizing measures : the generic chaining. *The Annals of Probability*, 24(3) :1049–1103.

Talagrand, M. (2005). *The generic chaining : upper and lower bounds of stochastic processes*. Springer Verlag.

Tarpey, T. and Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification*, 20(1) :93–114.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B.*, 58 :267–288.

van de Geer, S. (2008). High dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2) :614–645.

van de Geer, S. (2012). Generic chaining and the ℓ_1-penalty. *Arxiv preprint :1205.3703*.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5 :688–749.

van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer, Berlin.

Vapnik, V. (1982). *Estimation of dependencies based on empirical data*. Springer, New-York.

Vapnik, V. (1990). *Statistical learning theory*. J. Wiley, New-York.

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

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *Information Theory, IEEE Transactions on*, 55(5) :2183–2202.

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.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared ? A conflict between model identification and regression estimation. *Biometrika*, 92(4) :937–950.

Young, D. S. and Hunter, D. R. (2006). Random effects regression mixtures.

Yuan, M. and Lin, Y. (2007). On the non-negative garotte estimator. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(2) :143–161.

Zhang, C. and Huang, J. (2008a). Model-selection consistency of the Lasso in high-dimensional linear regression. *The Annals of Statistics*, 36 :1567–1594.

Zhang, C. and Huang, J. (2008b). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4) :1567–1594.

Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7 :2541–2567.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the ”degrees of freedom” of the Lasso. *The Annals of Statistics*, 35(5) :2173–2192.